PROCEEDINGS
OF THE
SUMMER INSTITUTE ON PARTICLE PHYSICS

August 3 – 11, 1998

GRAVITY
From the Hubble Length to the Planck Length

Program Directors:
David Burke
Lance Dixon
Charles Prescott

Edited by Lance Dixon

Sponsored by Stanford University and Stanford Linear Accelerator Center under contract with the U.S. Department of Energy, Contract DE-AC03-76SF00515

January 2001

Printed in the United States

TABLE OF CONTENTS

Part I. SUMMER SCHOOL LECTURES

R. Wagoner
"Relativistic Gravity and Some Astrophysical Applications".....1

C. Will
"The Confrontation Between General Relativity and Experiment: A 1998 Update".....15

L.R. Cominsky
"X-Ray Emission from Compact Sources".....55

T. Tyson
"Dark Matter Tomography".....89

P. Saulson
"Physics of Gravitational Wave Detection: Resonant and Interferometric Detectors".....113

A. Dekel
"Large Scale Structure" (manuscript not available).....163

S. Finn
"Gravitational Radiation Sources and Signatures ".....165

J. Polchinski
"Quantum Gravity at the Planck".....209

Part II. TOPICAL CONFERENCE

J. Kotcher
"Electroweak Physics and M_w from the Tevatron" (manuscript not available).....227

M. Martinez
"Precision Electroweak Results at LEP and LEP I".....229

R. Drucker
"Electroweak Results from NuTeV".....249

F. Taylor
"Electroweak Results from SLD".....259

W. Marciano
"Precision Electroweak, Light Higgs, etc. (Theory)".....279

G. Goldhaber
"Evidence for a Cosmological Constant".....289

N. Arkani-Hamed
"New Dimensions at a Millimeter" (manuscript not available).....299

G. Bernardi
"Deep Inelastic Scattering at HERA".....301

D. Jackson
"B Physics and QCD at SLD".....327

R. Briere
"A Selection of Recent CLEO Results: B Physics, Silicon Detector, and More...".....343

G. Apollinari
"Heavy Flavor Physics at the Tevatron" 367

U. Katz
"New HERA Results on Deep-Inelastic e⁺p Scattering at Very High Q^2".....393

S. Rosier-Lees
"Search for New Particles at LEP II".....409

L. Sorrell
"Spin Structure Measurements from E155 at SLAC".....423

W. Lorenson
"Results from HERMES".....437

G. Catanesi
"Results from *CHORUS Experiment*" 447

A. Marchionni
"Results from NOMAD" 455

G. Mills
"Results from Neutrinos LSND".....467

M. Nakahata
"Results from Super-Kamiokande".....477

R. Kessler
"Preliminary Results on CP-Violation Phenomena in the KTEV Experiment at Fermilab" 495

G. Redlinger
"Evidence and Prospects for the Rare Decays $K^+ \rightarrow \pi^+\nu\nu$.....505

J. Imazato
"T-Violation in $K^+ \rightarrow \pi^0 \mu^+ \nu_\mu$ at KEK".....513

J. Norris
"Gamma-Ray Bursts: Current Observational Picture".....524

P. Mészáros
"Gamma-Ray Bursts: The Fireball Shock Model and its Implications" 539

M. Peskin
"Unification & Gravitation" (manuscript not available).....549

APPENDIX

Author Index.....551

PREFACE

The XXVI SLAC Summer Institute on Particle Physics was held from August 3 to August 14, 1998. The topic, "GRAVITY--from the Hubble Length to the Planck Length," brought together 179 physicists from 13 countries.

The lectures in this volume cover the seven-day school portion of the Institute, which took us from the largest scales of the cosmos, to the Planck length at which gravity might be unified with the other forces of nature. Lectures by Robert Wagoner, Clifford Will, and Lynn Cominsky explored the embedding of gravity into general relativity and the confrontation of this idea with experiments in the laboratory and astrophysical settings. Avishai Deckel discussed observations and implications of the large-scale structure of the universe, and Tony Tyson presented the gravitational lensing effect and its use in the ongoing search for signatures of the unseen matter of the cosmos. The hunt for the wave nature of gravity was presented by Sam Finn and Peter Saulson, and Joe Polchinski showed us what gravity might look like in the quantum limit at the Planck scale.

The excellent lectures each morning were followed by afternoon discussion sessions, where students could further pursue questions and topics with the day's lecturers. The Institute concluded with a three-day topical conference covering recent developments in theory and experiment from around the world of elementary particle physics and cosmology; its proceedings are also presented in this volume.

We thank the lecturers for their well-prepared and stimulating lectures; the provocateurs who helped keep the afternoon discussion sessions rolling along; and most particularly the participants whose enthusiasm and intellect really make up what is most important to the success of the School. The Institute would not be possible without the support of SLAC staff and its administration. And finally, we are grateful for support from Stanford University and the U.S. Department of Energy.

Dave Burke, Lance Dixon, and Charles Prescott

RELATIVISTIC GRAVITY AND SOME ASTROPHYSICAL APPLICATIONS

Robert V. Wagoner*
Department of Physics
and
Center for Space Science and Astrophysics
Stanford University, Stanford, CA 94305-4060

ABSTRACT

These lectures are intended to provide a framework for the others at this institute. We will first discuss the foundations and structure of metric theories of gravity, developed from the consequences of various extremal action principles. Then, we restrict our consideration to scalar-tensor theories of gravity, and introduce various astrophysical applications. The weak-field applications are the index of refraction of gravity and gravitational waves. The strong-field applications are spherically symmetric compact objects, their surrounding particle orbits, and cosmology. Except for waves and cosmology, effects of the scalar field are assumed negligible, reducing the theory to general relativity.

*Supported in part by NASA ATP grant NAG 5-3102.

© 1998 by Robert V. Wagoner.

1 Key Concepts and Principles

Gravitation is the most fundamental interaction, affecting all forms of mass energy. This allows its geometrical description, at least within the classical (nonquantum) regime that we shall consider. The scope of this regime is indicated in Fig. 1. This series of three lectures is intended to provide a foundation for the others in this school, most of which deal with various astrophysical, cosmological, and quantum mechanical manifestations of gravity. Useful results rather than detailed derivations will be emphasized. The approach that we will take, and some applications that we will briefly consider are outlined in Fig. 2.

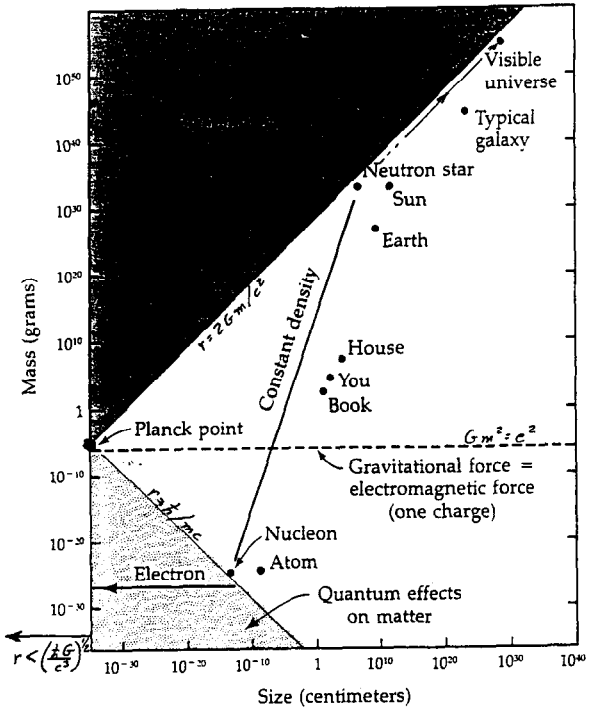


Fig. 1. The domains of our knowledge of the physical world, with the mass and size of representative objects indicated. A theory of quantum gravity is needed at distances less than the Planck length $L_P \equiv \sqrt{\hbar G/c^3}$.

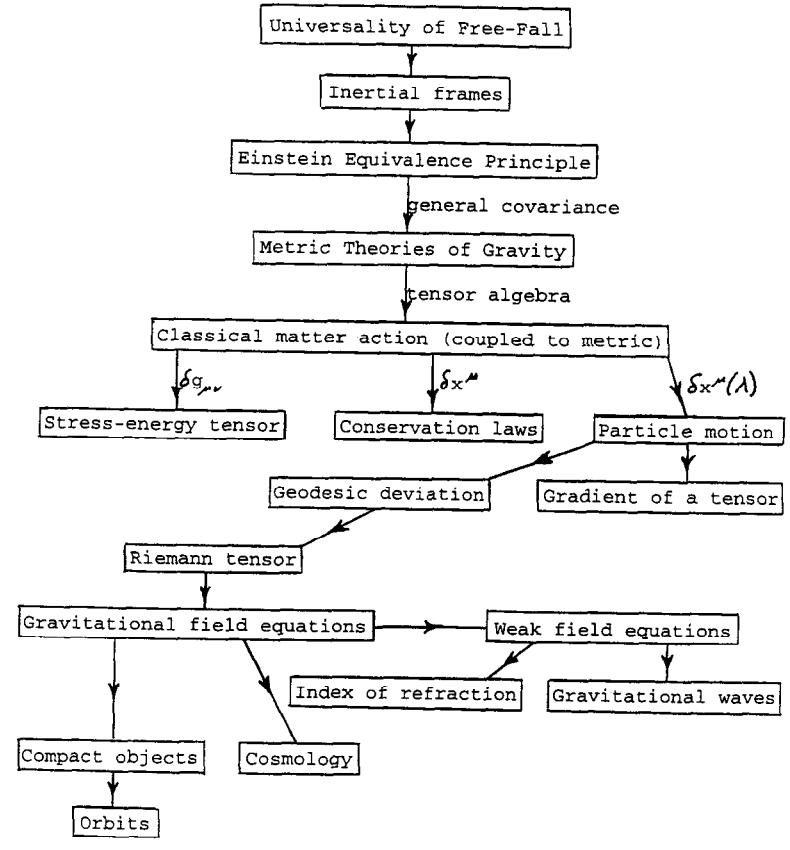


Fig. 2. The structure of these lectures.

The textbooks that are closest to our viewpoint are by Schutz[1] at the introductory level and by Misner, Thorne, and Wheeler[2] at the comprehensive level. Except for only setting the speed of light $c = 1$, but not the gravitational constant G, we adopt their notation and conventions.

We shall first consider the broad class of *metric theories of gravitation*. This class is defined by the specification of how gravity affects matter. The manner in which matter generates gravity is separately specified by each theory within this

class. Metric theories are based upon a few key concepts and principles:

1. Universality of Free-Fall (UFF)
   This principle states that if a test particle is placed at an initial event and is given an initial velocity there, its subsequent worldline through spacetime will be independent of its structure (*i.e.*, all forms of energy "fall" at the same rate). A test particle is conveniently defined as one whose charge, mass, and size are reduced until experimental results are unchanged.
2. Coordinate Frame
   A coordinate (or reference) frame can be visualized as a continuous set of spatially labeled "clocks" filling spacetime. Such generalized clocks merely provide the time label of events. Infinitesimal distances are best measured by the radar method. The proper distance is $c/2$ times the round trip travel time, as measured by an ideal clock carried by the fiducial observer. The time associated with the measurement is the average of the photon emission and reception times.
3. Inertial Reference Frame
   This is a local coordinate frame in which any free test particle is unaccelerated (to a specified accuracy) within a small specified region of spacetime. It can always be constructed at *any* point (event) in spacetime (if UFF is valid). A physical realization is a nonrotating lab in free-fall, small enough that the effects of tidal gravitational forces are negligible.
4. Einstein Equivalence Principle (EEP)
   This principle states that in all inertial frames, the *nongravitational* laws of physics are those formulated within special relativity (when local Lorentz coordinates are employed).

Employing the EEP, we will see how *gravity emerges from a local analysis.* Demanding that the laws of physics retain their form under general coordinate transformations $x^{\mu'} = x^{\mu'}(x^\alpha)$ (general covariance), will then allow us to determine how matter couples to gravity via a metric tensor. This metric tensor has components $g_{\mu\nu}(x^\alpha)$, which can be put in the Minkowski form $\eta_{\mu\nu}$, with $\partial g_{\mu\nu}/\partial x^\alpha = 0$ in every inertial frame (which is then called a local Lorentz frame).

Note that all tensor equations ($\mathbf{T} = \mathbf{0}$) are generally covariant. Such an equation will thus be true in all coordinate systems if it is known to be true in any one (such as a local Lorentz frame).

2 Tensor Algebra in Metric Spacetimes

Tensor algebra is the study of geometrical objects (scalars, vectors, tensors, ...) at any fixed point $\mathcal{P}_0$. These objects exist independent of any coordinates, and thus, form the proper description of physical reality.

2.1 Vector

In curved spacetime, we require a local definition of a vector. A familiar one is the tangent vector

$$\mathbf{v} = \left(\frac{\mathrm{d}\mathcal{P}}{d\lambda}\right)_{\mathcal{P}_0}$$

to some curve $\mathcal{P}(\lambda)$ at the point $\mathcal{P}_0$ where the vector exists, where $\mathbf{d}\mathcal{P}$ is the infinitesimal displacement vector. If the path parameter λ is chosen as the proper time τ (for a nonzero mass particle), the tangent vector is the four-velocity $\mathbf{U} = \mathrm{d}\mathcal{P}/d\tau$.

2.2 Tangent Space

The vectors at any point $\mathcal{P}_0$ form this abstract four-dimensional vector space. All geometrical objects at this point reside in this tangent space (not in spacetime).

2.3 Basis

A basis is a set of four linearly independent vectors $\mathbf{e}_\alpha$ ($\alpha = 0, 1, 2, 3$) at a point $\mathcal{P}_0$. Any vector $\mathbf{v}$ at $\mathcal{P}_0$ can be represented by its components v^α,

$$\mathbf{v} = v^\alpha \mathbf{e}_\alpha \qquad \text{(summation convention)}.$$

Consider some coordinate system: four functions $x^\alpha(\mathcal{P})$. A (global) *coordinate basis* is then

$$\mathbf{e}_\alpha = \partial\mathcal{P}/\partial x^\alpha,$$

indicated in Fig. 3. We shall only employ such bases.

Most other aspects of tensor algebra are direct generalizations from special relativity in each tangent space, as follows.

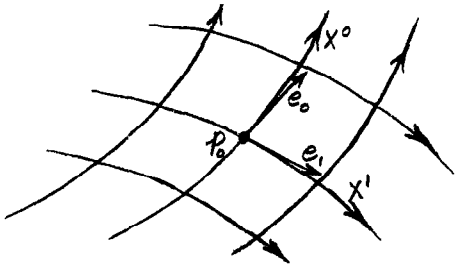


Fig. 3. Coordinate basis vectors in the tangent space at $\mathcal{P}_0$, tangent to the coordinate curves there.

2.4 Tensor

A tensor can be thought of in at least two ways:

- As a direct product: $\mathbf{T} = T^{\alpha\beta\cdots}\mathbf{e}_\alpha \otimes \mathbf{e}_\beta \otimes \cdots$.
- As a linear operator on vectors, giving a scalar (number):

$$\mathbf{T}(\mathbf{u}, \mathbf{v}, \ldots) = \mathbf{T}(\mathbf{e}_\alpha, \mathbf{e}_\beta, \ldots)u^\alpha v^\beta \cdots = T_{\alpha\beta\ldots}u^\alpha v^\beta \cdots .$$

The second equality defines the components of a tensor.

2.5 Metric Tensor

This generalization of the Minkowski metric of special relativity, with components $\eta_{\mu\nu} = \text{diag.}(-1, 1, 1, 1)$, produces the scalar product of vectors: $\mathbf{u} \cdot \mathbf{v} \equiv \mathbf{g}(\mathbf{u}, \mathbf{v})$. Its components then also represent scalar products:

$$\mathbf{g}(\mathbf{e}_\alpha, \mathbf{e}_\beta) \equiv \mathbf{e}_\alpha \cdot \mathbf{e}_\beta = g_{\alpha\beta} = \mathbf{g}(\mathbf{e}_\beta, \mathbf{e}_\alpha) = g_{\beta\alpha}.$$

Other aspects are:

- The interval $ds^2 = \mathbf{g}(\mathbf{d}\mathcal{P}, \mathbf{d}\mathcal{P}) = g_{\alpha\beta}dx^\alpha dx^\beta \quad (\mathbf{d}\mathcal{P} = dx^\sigma \mathbf{e}_\sigma)$.
- Its inverse, given by $g^{\mu\sigma}g_{\sigma\nu} = \delta^\mu{}_\nu$.
- Raising and lowering indices:

$$\begin{aligned} v_\mu &\equiv \mathbf{v} \cdot \mathbf{e}_\mu = \mathbf{g}(v^\nu \mathbf{e}_\nu, \mathbf{e}_\mu) = g_{\nu\mu}v^\nu, \\ v^\mu &= g^{\mu\sigma}g_{\sigma\nu}v^\nu = g^{\mu\sigma}v_\sigma. \end{aligned}$$

This generalizes to give

$$T_\alpha{}^{\beta\ \cdots}_{\gamma\cdots} = g_{\alpha\mu}g_{\gamma\nu}T^{\mu\beta\nu\cdots} = g^{\beta\sigma}T_{\alpha\sigma\gamma\cdots}$$

and the generalized scalar product $T^\alpha_{\beta\gamma}V_\alpha N^{\beta\gamma}$, for instance.

It is important to remember that *all measured quantities are scalars.* For instance, the energy of a photon of four-momentum $\mathbf{p}$ measured by a detector of four-velocity $\mathbf{U}$ (as shown in Fig. 4) is $E = -\mathbf{U} \cdot \mathbf{p} = -U^\sigma p_\sigma$. This reduces to $E = p^0$ in an instantaneously comoving ($U^i = 0$) local Lorentz frame.

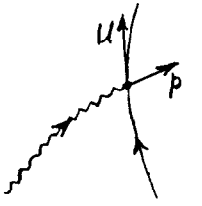


Fig. 4. The interaction of a photon of four-momentum $\mathbf{p}$ with a detector of four-velocity $\mathbf{U}$.

- Contraction produces a new tensor of rank two lower; for instance

$$Q_{\mu\nu} = M^\sigma{}_{\mu\sigma\nu} = g^{\sigma\tau}M_{\tau\mu\sigma\nu},$$

independent of choice of basis.

2.6 Change of Basis

This is represented by $\mathbf{e}_{\mu'} = L^\nu{}_{\mu'}\mathbf{e}_\nu$, generated by the tranformation matrix $L^\nu{}_{\mu'}(\mathcal{P}_0)$. With the inverse transformation matrix constructed from $L^\mu{}_{\sigma'}L^{\sigma'}{}_\nu = \delta^\mu{}_\nu$, one obtains the transformed components

$$T^{\alpha'\ \cdots}_{\beta'\cdots} = L^{\alpha'}{}_\sigma L^\tau{}_{\beta'} \cdots T^\sigma{}_{\tau\cdots}^{\ \cdots}$$

of a tensor, using the above equations. Under a tranformation of coordinate bases [generated by the four functions $x^{\mu'}(x^\alpha)$], the tranformation matrices assume the form

$$L^{\alpha'}{}_\beta = \partial x^{\alpha'}/\partial x^\beta, \quad L^\beta{}_{\alpha'} = \partial x^\beta/\partial x^{\alpha'}.$$

2.7 Four-Volume Element

The unique scalar, which generalizes all the usual properties of a volume element, is

$$dV_{(4)} = \sqrt{-g}\, dx^0 dx^1 dx^2 dx^3 \equiv \sqrt{-g}\, d^4x\ ,$$

where g is the determinant of the matrix $g_{\mu\nu}$.

2.8 Gradient of a Function

This operation on a function $f(x^\alpha)$ is represented as

$$\mathbf{d}f = f^{,\alpha}\mathbf{e}_\alpha = g^{\alpha\beta}f_{,\beta}\mathbf{e}_\alpha = g^{\alpha\beta}\frac{\partial f}{\partial x^\beta}\mathbf{e}_\alpha,$$

in terms of its partial derivatives in a coordinate basis. The *directional derivative* of a function along a curve $\mathcal{P}(\lambda)$ (at $\mathcal{P}_0$) is then

$$\frac{df}{d\lambda} = \frac{dx^\alpha}{d\lambda}\frac{\partial f}{\partial x^\alpha} = v^\alpha f_{,\alpha} = \mathbf{v}\cdot\mathbf{d}f.$$

This scalar indicates how much the function f changes in the direction of the tangent vector $\mathbf{v}$.

3 Extremal Action Principles for Matter and Their Consequences

The fundamental special-relativistic laws of physics may be obtained by extremizing the (scalar) action $\mathcal{I} = \int \mathcal{L}d^4x$, $\delta\mathcal{I} = 0$. Within metric theories, the effects of gravity *on* a classical material system may be obtained by the replacements

$$\mathcal{I} \to \int \mathcal{L}\sqrt{-g}\,d^4x\ ,\ \mathcal{L} \to \mathcal{L}_M(\eta_{\mu\nu} \to g_{\mu\nu}; A_\mu, A_{[\mu,\nu]},\ \text{"matter"}). \tag{1}$$

Variation with respect to A_μ gives Maxwell's equations in an arbitrary metric field. (Complete antisymmetrization and symmetrization of indices will be denoted by $[\alpha\beta\cdots]$ and $[\alpha\beta\cdots]$.) We will add the contribution of the gravitational field(s) to the action $\mathcal{I}$ later.

3.1 Stress-Energy Tensor

Under the variation $g_{\mu\nu} \to g_{\mu\nu} + \delta g_{\mu\nu}$, the stress-energy tensor $T^{\mu\nu}$ is defined by

$$\delta\mathcal{I}_M = \tfrac{1}{2}\int T^{\mu\nu}\delta g_{\mu\nu}\sqrt{-g}\,d^4x\,. \tag{2}$$

This definition automatically generates all of its desired properties from special relativity.

3.2 Energy-Momentum Conservation

Since the "matter" action $\mathcal{I}_M$ is a scalar, it will be unchanged under a coordinate transformation $x^\mu \to x^{\mu'} = x^\mu + \epsilon^\mu(x^\alpha)$, here taken to be infinitesimal. It will also be unchanged by a subsequent change, $x^{\mu'} \to x^\mu$, in the integration variable. This induces the net change,

$$\delta g_{\mu\nu} = -(g_{\sigma\nu}\epsilon^\sigma{}_{,\mu} + g_{\mu\sigma}\epsilon^\sigma{}_{,\nu} + g_{\mu\nu,\sigma}\epsilon^\sigma),$$

the Lie derivative of $g_{\mu\nu}$. Under this variation of the metric tensor components, one obtains $\delta\mathcal{I}_M = \int T^\mu{}_{\sigma;\mu}\epsilon^\sigma\sqrt{-g}\,d^4x$, where

$$T^\mu{}_{\sigma;\mu} \equiv (1/\sqrt{-g})[\sqrt{-g}\,T^\mu{}_\sigma]_{,\mu} - \tfrac{1}{2}g_{\mu\nu,\sigma}T^{\mu\nu}.$$

Note that the divergence, $T^{\mu\sigma}{}_{;\mu}\mathbf{e}_\sigma$ must be the components of a vector, since its contraction with the components ϵ^σ of the infinitesimal displacement vector is a scalar. Since $\delta\mathcal{I}_M = 0$ for arbitrary $\epsilon^\sigma(x^\alpha)$, we obtain the four components,

$$T^\mu{}_{\sigma;\mu} = 0 \qquad (\text{or } T^{\mu\sigma}{}_{;\mu} = 0), \tag{3}$$

representing conservation of energy and momentum. The latter are the equations of motion of the continuum.

3.3 Test-Particle Equation of Motion

The prescription (1) gives the action

$$\mathcal{I}_M = -\int\left(-g_{\mu\nu}\frac{dx^\mu}{d\lambda}\frac{dx^\nu}{d\lambda}\right)^{1/2}d\lambda$$

for a test particle, where the tangent vector to the particle's world line $x^\alpha(\lambda)$ is $\mathbf{d}\mathcal{P}/d\lambda = (dx^\mu/d\lambda)\mathbf{e}_\mu \equiv \dot{x}^\mu\mathbf{e}_\mu$. Vary the worldline, and choose λ (after the variation) so that $g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu = -m^2$. The rest mass is $m \geq 0$. The components of the particle's four-momentum are then $p^\mu = \dot{x}^\mu$. Then $\delta\mathcal{I}_M = -m^{-1}\int Q_\alpha\delta x^\alpha d\lambda = 0$, so

$$Q_\alpha \equiv dp_\alpha/d\lambda - \tfrac{1}{2}g_{\mu\nu,\alpha}p^\mu p^\nu = 0. \tag{4}$$

Thus, if the metric components $g_{\mu\nu}$ are independent of any coordinate x^α, the corresponding four-momentum component p_α is conserved.

We also find the vector components

$$Q^\alpha = \frac{d^2x^\alpha}{d\lambda^2} + \left\{ \begin{matrix} \alpha \\ \mu\nu \end{matrix} \right\} \frac{dx^\mu}{d\lambda}\frac{dx^\nu}{d\lambda} \equiv \frac{D}{D\lambda}\left(\frac{dx^\alpha}{d\lambda}\right) = \frac{Dp^\alpha}{D\lambda} = 0\,, \tag{5}$$

where the Christoffel symbol (connection coefficient in a coordinate basis, not the components of a tensor) is

$$\left\{ \begin{matrix} \alpha \\ \mu\nu \end{matrix} \right\} = \left\{ \begin{matrix} \alpha \\ \nu\mu \end{matrix} \right\} = \tfrac{1}{2}g^{\alpha\sigma}(g_{\sigma\nu,\mu} + g_{\sigma\mu,\nu} - g_{\mu\nu,\sigma})\,. \tag{6}$$

For rest masses $m > 0$, Eqs. (4) and (5) also govern the particle's four-velocity $U^\alpha = p^\alpha/m = dx^\alpha/d\tau$, where the proper time interval is $d\tau = m d\lambda$.

It can be shown that in the neighborhood of a freely falling observer, coordinates can be chosen so that $g_{\mu\nu} = \eta_{\mu\nu}$ and $g_{\mu\nu,\alpha} = 0$ along his/her worldline. It then follows from Eq. (5) that all test particles in that neighborhood are indeed unaccelerated ($dx^\alpha/d\tau = \text{constant} \Rightarrow dx^i/dx^0 = \text{constant}$), verifying that it is an inertial frame.

3.4 Gradient and Covariant Derivative of a Tensor

Since $Q^\alpha = Dp^\alpha/D\lambda$ are the components of a vector [given by Eq. (5)], it follows that for any vector field with components $V^\alpha(x^\mu)$,

$$\begin{aligned} \frac{DV^\alpha}{D\lambda} &= \frac{dV^\alpha}{d\lambda} + \left\{ \begin{matrix} \alpha \\ \mu\nu \end{matrix} \right\} V^\mu \frac{dx^\nu}{d\lambda} \\ &= \left(V^\alpha{}_{,\nu} + \left\{ \begin{matrix} \alpha \\ \mu\nu \end{matrix} \right\} V^\mu \right) \frac{dx^\nu}{d\lambda} \equiv V^\alpha{}_{;\nu}\frac{dx^\nu}{d\lambda} \end{aligned} \tag{7}$$

are also the components of a vector. It thus also follows that $V^\alpha{}_{;\nu}$ must be the components of a (rank two) tensor: the generalization of the gradient to operate on vectors. The generalization of the directional derivative is the covariant derivative, with the above components $DV^\alpha/D\lambda$.

Denoting the directional derivative operator $D/D\lambda$ along a basis vector $\mathbf{e}_\nu$ by ∇_ν, its effect on a vector $\mathbf{V}$ can also be described as

$$\nabla_\nu(V^\sigma \mathbf{e}_\sigma) = (\nabla_\nu V^\sigma)\mathbf{e}_\sigma + V^\sigma(\nabla_\nu \mathbf{e}_\sigma) \equiv V^\alpha{}_{;\nu}\mathbf{e}_\alpha.$$

Comparing with Eq. (7), we see that the connection coefficients describe how the basis vectors vary with position:

$$\nabla_\nu \mathbf{e}_\mu = \left\{ \begin{matrix} \alpha \\ \mu\nu \end{matrix} \right\} \mathbf{e}_\alpha.$$

The application to tensors of any rank follows straightforwardly to give

$$T^{\alpha\cdots}_{\beta\cdots;\mu} = T^{\alpha\cdots}_{\beta\cdots,\mu} + \left\{ \begin{matrix} \alpha \\ \sigma\mu \end{matrix} \right\} T^{\sigma\cdots}_{\beta\cdots} + \cdots - \left\{ \begin{matrix} \sigma \\ \beta\mu \end{matrix} \right\} T^{\alpha\cdots}_{\sigma\cdots} - \cdots.$$

It then follows that the gradient of the metric tensor vanishes: $g_{\alpha\beta;\mu} = 0$.

As an interesting example, consider the Maxwell field-strength tensor, whose components $F_{\mu\nu}$ are generated from the components

$$A_{\mu;\nu} = A_{\mu,\nu} - \left\{ \begin{matrix} \sigma \\ \mu\nu \end{matrix} \right\} A_\sigma$$

of the gradient of the potential via $F_{\mu\nu} = A_{\nu,\mu} - A_{\mu;\nu} \equiv 2A_{[\nu;\mu]} = 2A_{[\nu,\mu]}$. The last equality follows from the symmetry of the Christoffel symbols [Eq. (6)], and reflects the curious fact that only partial derivatives seem to appear in the matter Lagrangian [Eq. (1)].

3.5 Geodesic Deviation $\Longleftrightarrow$ Riemann Tensor

Consider two freely falling test particles with infinitesimal separation $\Delta x^\alpha \mathbf{e}_\alpha$, as indicated in Fig. 5. Their separation vector can be constrained to obey $\boldsymbol{\Delta x}\cdot\mathbf{p} = 0$. It is then purely spatial ($\Delta x_0 = 0$) in an instantaneously comoving frame ($U^i = 0$) for a nonzero mass test particle.

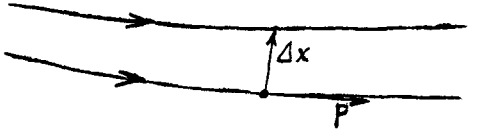


Fig. 5. The paths of two nearby test particles.

Subtracting their (geodesic) equations of motion (5) gives

$$\Delta Q^\alpha = \frac{d^2\Delta x^\alpha}{d\lambda^2} + \left\{ \begin{matrix} \alpha \\ \mu\nu \end{matrix} \right\}_{,\sigma} \frac{dx^\mu}{d\lambda}\frac{dx^\nu}{d\lambda}\Delta x^\sigma + 2\left\{ \begin{matrix} \alpha \\ \mu\nu \end{matrix} \right\} \frac{dx^\mu}{d\lambda}\frac{d\Delta x^\nu}{d\lambda} = 0.$$

Now since $D\Delta x^\alpha/D\lambda$ (but not $d\Delta x^\alpha/d\lambda$) are the components of a vector, so is the result of applying the operator $D/D\lambda$ [defined by Eq. (7)] again. Employing the above equation, this operation produces the equation of geodesic deviation

$$\frac{D^2\Delta x^\alpha}{D\lambda^2} + R^\alpha{}_{\mu\sigma\nu}\frac{dx^\mu}{d\lambda}\frac{dx^\nu}{d\lambda}\Delta x^\sigma = 0. \tag{8}$$

Since both terms in this equation are the components of vectors, the quantities,

$$R^{\alpha}{}_{\mu\sigma\nu} = \left\{ \begin{matrix} \alpha \\ \mu\nu \end{matrix} \right\}_{,\sigma} - \left\{ \begin{matrix} \alpha \\ \mu\sigma \end{matrix} \right\}_{,\nu} + \left\{ \begin{matrix} \alpha \\ \sigma\tau \end{matrix} \right\} \left\{ \begin{matrix} \tau \\ \mu\nu \end{matrix} \right\} - \left\{ \begin{matrix} \alpha \\ \tau\nu \end{matrix} \right\} \left\{ \begin{matrix} \tau \\ \mu\sigma \end{matrix} \right\}$$

must be the components of a (rank four) tensor, called the Riemann (curvature) tensor. It plays a role similar to that of the electromagnetic field tensor, $F_{\mu\nu}$ in the extension of the equation of motion Eq. (5) to charged test particles:

$$\frac{D}{D\tau}\left(\frac{dx^{\alpha}}{d\tau}\right) = \frac{q}{m} F^{\alpha}{}_{\sigma} \frac{dx^{\sigma}}{d\tau} ,$$

the generalized Lorentz force equation. We see, however, that the Riemann tensor represents the physical field *gradients* (tidal forces), and only *relative* gravitational acceleration has physical meaning.

The symmetry properties of the Riemann tensor (analogous to $F_{\mu\nu} = F_{[\mu\nu]}$) are

$$R_{\alpha\beta\mu\nu} = R_{([\alpha\beta][\mu\nu])} , \qquad R_{[\alpha\beta\mu\nu]} = 0 ;$$

the first giving $6 \cdot 7/2$ independent components and the second giving one less, for a total of 20 independent components. Its (unique) contractions are

$$R_{\mu\nu} = g^{\alpha\beta} R_{\beta\mu\alpha\nu} , \qquad R = g^{\mu\nu} R_{\mu\nu} ,$$

the components of the (symmetric) Ricci tensor and the Ricci scalar. It also obeys the Bianchi identities

$$R_{\alpha\beta[\mu\nu;\zeta]} = 0, \tag{9}$$

analogous to the homogeneous Maxwell equations $F_{[\mu\nu;\zeta]} = 0$ and their (unique) double contraction

$$G^{\mu\nu}{}_{;\nu} = 0 \qquad (G^{\mu\nu} \equiv R^{\mu\nu} - \tfrac{1}{2} g^{\mu\nu} R),$$

which involves the components of the Einstein tensor $\mathbf{G}$.

Consider the parallel transport of a vector $\mathbf{S}$ along some curve $\mathcal{P}(\lambda)$, defined by $DS^{\mu}/D\lambda = 0$. If the curve is the boundary of an infinitesimal area generated by the displacement vectors $\delta\mathbf{a}$ and $\delta\mathbf{b}$, the change in $\mathbf{S}$ after a complete circuit is

$$\Delta S^{\mu} = -R^{\mu}{}_{\alpha\beta\gamma} S^{\alpha} \delta a^{\beta} \delta b^{\gamma}. \tag{10}$$

This is another way to characterize the Riemann tensor. Now parallel transport the vector around all six faces of an infinitesimal cube, as shown in Fig. 6. Since the pairs of paths along each edge must cancel, the total change is $\Delta S^{\mu} = 0$. This has been characterized as "the boundary of a boundary equals zero."[2] Expressing this result in terms of Eq. (10) then produces the Bianchi identities (9), giving them an intriguing geometrical meaning.

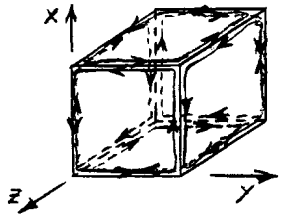


Fig. 6. The paths of vector transport which generate the Bianchi identity.

4 Gravitational Field Equations

4.1 Scalar-Tensor Theories

We now complete the implementation of the extremal action principle by adding to the matter Lagrangian density $\mathcal{L}_M$ [specified in Eq. (1)] a Lagrangian density $\mathcal{L}_G$ which depends solely on the gravitational field(s). Adopting the principle of simplicity that has worked so well in deriving the laws of physics, we are tempted to include nature's simplest (scalar, spin 0) field φ in addition to the metric tensor field $\mathbf{g}$. In analyzing the field equations, more insight is gained by employing the "spin representation," in which $\hat{\mathbf{g}}$ denotes the metric tensor which corresponds to a pure spin two field. On the other hand, $\mathbf{g} = A^2(\varphi)\hat{\mathbf{g}}$ is the metric tensor discussed above, through which gravity couples to matter.

If we require only that the field equations be of at most second differential order, the most general Lagrangian density is then[3]

$$\mathcal{L}_G = (16\pi G)^{-1}[\hat{R} - 2\hat{g}^{\mu\nu}\varphi_{,\mu}\varphi_{,\nu} - 2\Lambda(\varphi)],$$

where G is the bare gravitational constant. Thus, there are two free functions in this theory:[4,5] the matter coupling function $A(\varphi)$ and the "cosmological function" $\Lambda(\varphi)$.

Extremizing the action $\mathcal{I}_G$ with respect to variations in $\hat{g}_{\mu\nu}$ and φ then gives the field equations

$$\hat{G}_{\mu\nu} = 8\pi G\hat{T}_{\mu\nu} - \hat{g}_{\mu\nu}\Lambda(\varphi) + 2\varphi_{,\mu}\varphi_{,\nu} - \hat{g}_{\mu\nu}\hat{g}^{\alpha\beta}\varphi_{,\alpha}\varphi_{,\beta}, \tag{11}$$

$$\hat{g}^{\mu\nu}\varphi_{,\mu;\nu} = -4\pi G\alpha(\varphi)\hat{T} + \tfrac{1}{2}d\Lambda/d\varphi. \tag{12}$$

The scalar field-matter coupling function is

$$\alpha(\varphi) \equiv d\ln A/d\varphi = a_1 + a_2(\varphi - \varphi_0) + \cdots, \tag{13}$$

where the expansion is about the present cosmological value φ_0 of the scalar field. Expanding the self interaction $\Lambda(\varphi)$ in the same way shows that the effective range of the scalar field is of order $(d^2\Lambda/d\varphi^2)_0^{-1/2}$.

The stress-energy tensor $\hat{\mathbf{T}}$ is defined, as in Eq. (2), with respect to variations of $\hat{\mathbf{g}}$. It obeys the modified conservation laws [compare with Eq. (3)]

$$\hat{T}_{\mu}{}^{\nu}{}_{;\nu} = \alpha(\varphi)\hat{T}\varphi_{,\mu}\,, \tag{14}$$

indicating the separate effects of the spin 2 and spin 0 fields.

If the coupling function $A(\varphi)$ has a minimum, Damour and Nordtvedt[6] and Santiago, Kalligas, and Wagoner[7] have shown that in most cases the theory is attracted toward that minimum during the expansion of the universe, thus approaching general relativity: $\varphi =$ constant, $A(\varphi) =$ constant, and $\alpha(\varphi) = 0$. This is in accord with the small experimental limit[8] $a_1^2 < 10^{-3}$. [The Brans-Dicke theory[8] is the special case $\alpha(\varphi) =$ constant, $\Lambda(\varphi) = 0$.]

Although there is local interest in this broad class of theories, for the most part we shall concentrate on general relativity for the remainder of these lectures. The two exceptions will be gravitational waves and cosmology, where a scalar field introduces qualitatively new effects.

In addition, we will employ the fact that most matter in the universe is well approximated as a perfect fluid, described by the stress-energy tensor (obtained from the EEP)

$$T_{\mu\nu} = (\rho + p)U_\mu U_\nu + pg_{\mu\nu}, \tag{15}$$

where ρ is the mass-energy density, p is the pressure, and $\mathbf{U}$ is the four-velocity of the fluid. Such a fluid flows adiabatically (conserving specific entropy).

4.2 Weak-Field Equations

Throughout almost the entirety of all regions, much smaller than that of the observable universe, gravitational fields can be considered weak. This means that (except near black holes and neutron stars) one can choose coordinates such that the metric assumes the nearly Minkowski form

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}(x^\alpha)\,, \qquad |h_{\mu\nu}| \ll 1\,.$$

For instance, within the solar system, $|h_{\mu\nu}| \lesssim GM_\odot/R_\odot c^2 = 2.12 \times 10^{-6}$. We shall consider isolated sources $T_{\mu\nu}$, and can neglect the cosmological constant $\Lambda(\varphi_0)$ within such regions.

We work to first order in $h_{\mu\nu}$, and utilize our freedom of general infinitesimal coordinate transformations, $x^{\alpha'}(\mathcal{P}) = x^\alpha(\mathcal{P}) + \xi^\alpha(\mathcal{P})$, which produces the change

$$h_{\mu'\nu'} = h_{\mu\nu} - \xi_{\mu,\nu} - \xi_{\nu,\mu} \tag{16}$$

in the metric perturbation. This is directly analogous to the gauge transformation, $A_{\mu'} = A_\mu + \chi_{,\mu}$ in electrodynamics and leaves the Riemann tensor components $R_{\alpha\beta\mu\nu}$, like the Maxwell field-strength tensor components $F_{\alpha\beta}$, invariant to this order. We can then use our freedom in choosing the four functions $\xi^\alpha(\mathcal{P})$ to impose the coordinate condition

$$\bar{h}^{\mu\sigma}{}_{,\sigma} = 0\,, \quad \bar{h}_{\mu\nu} \equiv h_{\mu\nu} - \tfrac{1}{2}\eta_{\mu\nu}h \quad (h = \eta^{\alpha\beta}h_{\alpha\beta})\,,$$

analogous to the Lorentz gauge condition $A^\sigma{}_{,\sigma} = 0$ in electrodynamics.

The Einstein field Eq. (11) (with $\varphi = \varphi_0 =$ constant) then produces the weak field equations

$$\eta^{\alpha\beta}\bar{h}_{\mu\nu,\alpha\beta} \equiv \Box\bar{h}_{\mu\nu} = -16\pi GT_{\mu\nu}\,, \tag{17}$$

identical in structure to those in electrodynamics. (With the Lorentz gauge condition, Eq. (17) is consistent with the conservation laws $T^{\mu\nu}{}_{,\nu} = 0$; analogous to $J^\mu{}_{,\mu} = 0$ in electrodynamics.) Thus, the solution is of the same form:

$$\bar{h}_{\mu\nu}(x^\alpha) = 4G\int T_{\mu\nu}(x^0 - |x^i - x^i_*|, x^i_*)|x^i - x^i_*|^{-1}d^3x^i_*\,. \tag{18}$$

A Newtonian system is one in which all (macroscopic and microscopic) velocities are nonrelativistic (and thus the retardation in Eq. (18) is negligible), in addition to having a weak field. In such systems the dominant component of

the stress-energy tensor (15) is seen to be $T_{00} \cong \rho$. Therefore we see from Eq. (18) that the dominant component of the (trace-reversed) metric perturbation is $\bar{h}_{00} = -4\Phi$, where Φ is the Newtonian gravitational potential. (This also identifies the meaning of the coupling constant G.) Then the spacetime interval becomes

$$ds^2 = g_{\alpha\beta}dx^\alpha dx^\beta \cong -(1+2\Phi)dt^2 + (1-2\Phi)(dx^2+dy^2+dz^2). \qquad (19)$$

Incidentally, the result, $h_{00} = -2\Phi$, can be obtained more generally by comparing Newton's second law with the geodesic equation of motion (5) for slowly moving test particles. In the same limit (in which $dx^\alpha/d\tau \cong \delta^\alpha{}_0$), the spatial components of the equation of geodesic deviation (8) become

$$\frac{d^2\Delta x^i}{d\tau^2} \cong -R^i{}_{0j0}\Delta x^j \cong -\frac{\partial^2\Phi}{\partial x^i \partial x^j}\Delta x^j \ ,$$

showing how the tidal gravitational forces affect the separation of nearby particles.

5 Weak-Field Applications

5.1 Index of Refraction of Gravity

Consider a photon passing through a Newtonian gravitational potential (for which $|\partial\Phi/\partial t| \ll |\nabla\Phi|$) produced by some localized distribution of mass. This is a good approximation for all observed systems. In the geometrical optics limit, we follow a photon initially traveling in the direction $\mathbf{e}_x$ far from the masses, so subsequently $p^x = dx/d\lambda \cong p^t \equiv p$. As shown in Fig. 7, it will be deflected by a very small angle $\vec{\alpha}$, with components $\alpha^N \cong p^N/p$, where the index $N = y, z$.

The equation of motion (5) then gives

$$\frac{dp^N}{d\lambda} \cong -\left[\begin{Bmatrix} N \\ tt \end{Bmatrix} + 2\begin{Bmatrix} N \\ tx \end{Bmatrix} + \begin{Bmatrix} N \\ xx \end{Bmatrix}\right] p^2 \cong -2\Phi_{,N}p^2$$

when the weak-field metric (19) is inserted in the Christoffel symbols. With $d\ell \equiv dx \cong pd\lambda$ and $d\alpha^N/dx \cong p^{-1}dp^N/dx$, we obtain

$$\vec{\alpha} = -2\int \nabla_\perp \Phi \, d\ell = \int \nabla_\perp n \, d\ell, \qquad (20)$$

where we have identified the effective *gravitational index of refraction* $n(x^i) \equiv 1 - 2\Phi$ (≥ 1).

Fig. 7. Photon path through a localized weak field.

Using the fact that $0 = ds^2 \cong g_{tt}dt^2 + g_{xx}dx^2$ for the photon, we obtain its coordinate velocity

$$\frac{dx}{dt} \cong \left(\frac{-g_{tt}}{g_{xx}}\right)^{1/2} \cong 1 + 2\Phi \cong 1/n \ ,$$

as expected. The time delay, relative to a photon traveling between the same initial and final values of x in the absence of any mass, is then

$$\Delta t = -2\int \Phi \, d\ell = 2\int |\Phi| d\ell = \int (n-1) d\ell \ . \qquad (21)$$

However, when the emitter and receiver of the photon are very far from the mass, coordinate time t equals proper time τ at both locations. So this is the observed time delay measured by such pairs of observers.

Thus in both respects (bending and delay), empty space acts as if it had this index of refraction.

5.2 Gravitational Waves

Unfortunately, all gravitational waves will be weak when they reach our detectors. The vacuum weak-field Eq. (17) $\Box\bar{h}_{\mu\nu} = 0$ and Lorentz gauge condition $\bar{h}^\sigma_{\mu,\sigma} = 0$ allow a representation in terms of plane transverse monochromatic waves

$$\bar{h}_{\mu\nu} = A_{\mu\nu}e^{ik_\alpha x^\alpha} \ ; \qquad A_{\mu\sigma}k^\sigma = 0 \ , \quad k_\sigma k^\sigma = 0 \ ,$$

with wave vector $\mathbf{k}$. However, both the field equations and Lorentz gauge condition are preserved under another infinitesimal coordinate (gauge) transformation (16)

if the generator also satisfies the wave equation $\Box\xi_\mu = 0$. One can then use the four additional degrees of freedom to set $\bar{h}_{0i} = 0$ and $\bar{h} = 0$ (so now $\bar{h}_{\mu\nu} = h_{\mu\nu}$). In summary, we have constructed the transverse-traceless (TT) gauge, in which the eight independent conditions

$$h_{\mu 0} = h_{jk,k} = h^k{}_k = 0$$

leave two independent polarization states, again in direct analogy with electrodynamics.

We can also include the possibility of a weak scalar wave, $\varphi_1 \equiv \varphi - \varphi_0$, since the tensor field from Eq. (11) is unaffected through first order in φ_1 (except that $h_{\mu\nu} \to \hat{h}_{\mu\nu} \cong h_{\mu\nu} - 2a_1\varphi_1\eta_{\mu\nu}$ in the above equations). The scalar field equation (12) (with Λ constant) becomes $\Box\varphi_1 = 0$, giving the same plane-wave representation.

To understand the response of a gravitational-wave detector, consider slowly moving free test particles with a separation which is much less than the gravitational wavelengths involved. Now employ a local Lorentz frame (a different choice of gauge), in which physical (*e.g.*, radar) and coordinate distances are equal through first order in the particle separation $\mathbf{\Delta x}$. The Eq. (8) of geodesic deviation becomes $d^2\Delta x^i/d\tau^2 \cong -R^i{}_{0j0}\Delta x^j$ as before, involving only the matter coupling metric $\mathbf{g} = A^2(\varphi)\hat{\mathbf{g}}$. In terms of our spin representation,

$$R_{i0j0} = \hat{R}_{i0j0} + a_1(\varphi_{,ij} - \delta_{ij}\varphi_{,00}).$$

In the previous TT gauge, one obtains $\hat{R}^{TT}_{i0j0} = -\frac{1}{2}\hat{h}^{TT}_{ij,00}$. However, the gauge (coordinate) invariance of the weak-field Riemann tensor allows us to use this expression in the above equation of geodesic deviation, giving

$$d^2\Delta x^i/d\tau^2 \cong [\tfrac{1}{2}\hat{h}^{TT}_{ij,00} + a_1(\delta_{ij}\varphi_{,00} - \varphi_{,ij})]\Delta x^j. \tag{22}$$

In Fig. 8 we show the resulting positions of an initially circular ring of test particles (at phases $\pi/2$ and $3\pi/2$) for each polarization state: (a) $\hat{h}^{TT}_{xx} = -\hat{h}^{TT}_{yy}$, (b) $\hat{h}^{TT}_{xy} = \hat{h}^{TT}_{yx}$, and (c) φ_1. They remain in the plane transverse to the propagation vector $\vec{k}$ shown.

For separations $\vec{\Delta}x$ in the same direction as $\vec{k}$, Eq. (22) also shows that there is no response to any of the three wave components.

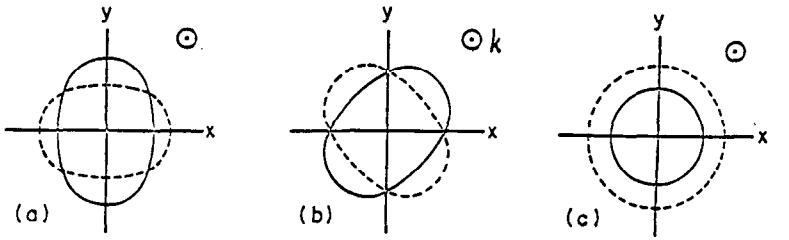


Fig. 8. Wave-induced distortions of a ring of test particles.

6 Strong-Field Applications

6.1 Compact Objects

Throughout any spherically symmetric spacetime, we can choose Schwarzschild coordinates, in which the interval assumes the form

$$ds^2 = -e^{2\Phi(r,t)}dt^2 + e^{2\lambda(r,t)}dr^2 + r^2[d\theta^2 + \sin^2\theta d\phi^2], \tag{23}$$

so that the proper area (measured by a set of observers at fixed r and t) of any spherical surface is $4\pi r^2$. We consider here isolated bodies, so the metric potentials $\Phi, \lambda \to 0$ as $r \to \infty$. In addition, we shall consider static ($U^i = 0$, $\partial/\partial t = 0$) bodies.

Then the only non-trivial momentum conservation equation, $T_r{}^\sigma{}_{;\sigma} = 0$, gives hydrostatic equilibrium:

$$\frac{dp}{dr} = -(\rho + p)\frac{d\Phi}{dr}, \tag{24}$$

also indicating that $\Phi(r)$ is the generalized Newtonian potential. The only structural difference from the Newtonian equation is the addition of the pressure to the "inertial mass-energy density."

The $\{tt\}$ component of the Einstein field equation gives, after an integration,

$$e^{2\lambda} = \left[1 - \frac{2Gm(r)}{r}\right]^{-1}, \quad m(r) = 4\pi\int_0^r \rho(r_*)r_*^2 dr_*. \tag{25}$$

Let $r = R$ be the radius of the star, defined by $p(R) = 0$. Then $m(r \geq R) = M$, the total gravitational mass (defined by applying Kepler's third law to distant orbits).

The $\{rr\}$ component of the Einstein field equation gives the generalization of the Newtonian field equation:

$$\frac{d\Phi}{dr} = G\frac{(m + 4\pi r^3 p)}{r(r - 2Gm)} . \tag{26}$$

We see that pressure also contributes to the "gravitational mass-energy density," and strong fields ($2Gm/r \to 1$) also steepen the potential gradient (and therefore, from Eq. (24), the pressure gradient). The $\{\theta\theta\}$ and $\{\phi\phi\}$ components of the field equation are redundant.

Finally, if the equation of state assumes the form

$$p = p(\rho, s(\rho)) \quad \text{(specific entropy } s \text{ determined separately)}, \tag{27}$$

we have four equations (24, 25, 26, and 27) to determine Φ, λ (and m), p, and ρ. For the major applications, white dwarfs and neutron stars, the entropy effectively vanishes since the particles providing the pressure support (electrons or neutrons, respectively) are degenerate.

Continuity of $\Phi(r)$ and $\lambda(r)$ at the stellar surface allows matching to the exterior (or black hole) Schwarzschild solution

$$e^{2\Phi} = e^{-2\lambda} = 1 - 2GM/r,$$

also obtained from the above equations. With the additional boundary condition $m \propto r^3$ as $r \to 0$, a set of stellar models is then a one-parameter [*e.g.*, $\rho_c \equiv \rho(0)$] family. Such equilibrium models are stable when $dM/d\rho_c \geq 0$.

The maximum mass of white dwarfs and neutron stars are of the same order, $M_P^3/m_n^2 = 1.86 M_\odot$, where the Planck mass $M_P \equiv \sqrt{\hbar c/G} = 2.1 \times 10^{-5}$ g and m_n is the nucleon mass. An extensive survey of this subject is provided by Shapiro and Teukolsky.[9]

6.2 Orbits

One of the most powerful probes of the strong gravitational fields near neutron stars and black holes is analysis of the orbits of test particles (or those comprising gaseous accretion disks), here taken to have rest mass $m > 0$. Within the Schwarzschild geometry, we can define the orbital plane as $\theta = \pi/2$. From Eq. (4), with $p^\mu = m dx^\mu/d\tau$, we then see that both the energy and angular momentum (per unit mass)

$$\tilde{E} = -\frac{p_t}{m} = \left(1 - \frac{2GM}{r}\right)\frac{dt}{d\tau} , \quad \tilde{L} = \frac{p_\phi}{m} = r^2\frac{d\phi}{d\tau} ,$$

are conserved along each orbit. The relation, $g_{\mu\nu}p^\mu p^\nu = -m^2$, then gives us the remaining (energy) equation

$$\left(\frac{dr}{d\tau}\right)^2 = \tilde{E}^2 - \tilde{V}^2(\tilde{L}, r) , \qquad \tilde{V}^2 = \left(1 - \frac{2GM}{r}\right)\left(1 + \frac{\tilde{L}^2}{r^2}\right) .$$

As in Newtonian theory, we can understand the orbits via plots of the effective potential $\tilde{V}$, shown in Fig. 9 for $r > 2GM$ (nonrotating black hole horizon).

For instance, radial turning points occur where $\tilde{V} = \tilde{E}$, and stable (and unstable) circular orbits correspond to the minima (and maxima) of $\tilde{V}$. However, we notice two new (relativistic) features: (a) A particle is captured by the black hole if its specific angular momentum is low enough ($\tilde{L} < 2\sqrt{3}GM$) or its specific energy is high enough ($\tilde{E} > \tilde{V}_{max}$); (b) Stable circular orbits only exist for $\tilde{L} > 2\sqrt{3}GM$, at $r > 6GM$.

For rotating (Kerr) black holes, the metric tensor is much more complicated, but effective potentials can still be obtained[2] for orbits with angular momentum parallel or anti-parallel to that of the black hole. The "dragging of inertial frames" by the rotation of the mass produces many interesting new effects, such as a region (the "ergosphere") outside the horizon where all observers must rotate in the same direction as the hole.

6.3 Cosmology

6.3.1 Kinematical Analysis

We shall first outline some general features of cosmology within the framework of any metric theory of gravity (incorporating the EEP). The analysis of the large-scale properties and evolution of the universe[10] has at its foundation one key observation, *isotropy*, and one key assumption, the *Copernican Principle* (CP). (By large scale we shall mean distances $D \gtrsim 3 \times 10^8$ light years, roughly 1/30 of the radius of the visible universe.)

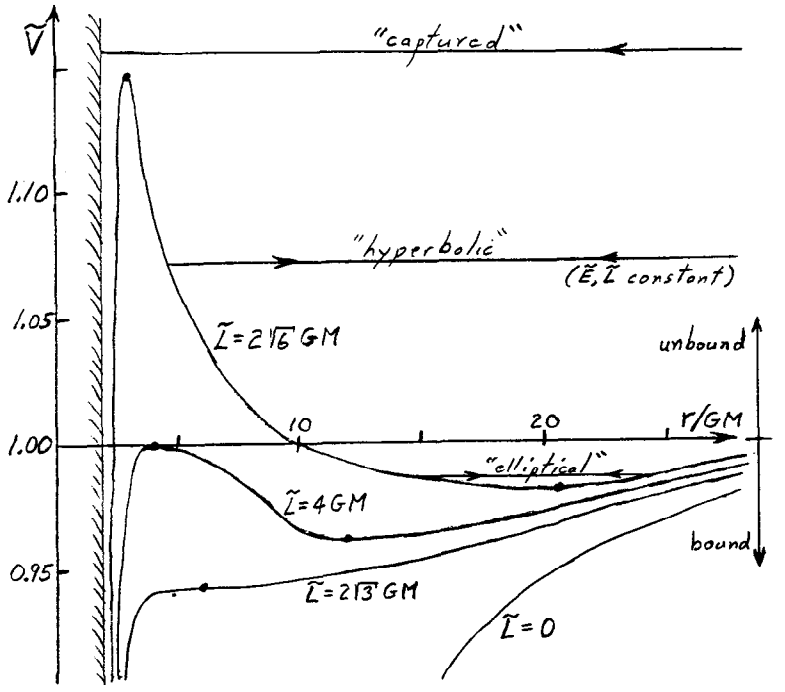


Fig. 9. The effective potential $\tilde{V}$ is plotted for various choices of $\tilde{L}$. Also shown are the three classes of orbits (constant $\tilde{E}, \tilde{L}$).

We first define a preferred set of observers: those who view the cosmic microwave background (CMB) radiation as isotropic on the corresponding angular scales ($\gtrsim 2$ degrees). The CMB is isotropic to an accuracy of 10^{-5} (after removing the effect of our motion), the size of the observed intensity variations imprinted by the primordial density fluctuations which formed galaxies, etc. Galaxies (such as ours) have random velocities with respect to this frame such that $\langle v^2/c^2\rangle^{1/2} \sim 10^{-3}$.

We shall then attach our spatial coordinates to such observers, so they become comoving with respect to the average motion of the matter. We also take the time coordinate to be their proper time, so the averaged (over volumes $\Delta V \sim D^3$) four-velocity of the matter has components $U^\alpha = \delta^\alpha_0$.

The CP asserts that our position in the universe is not special, but typical. The basis for this assumption is the similarity of our galaxy and local group of galaxies to other such systems observed throughout the universe. One then obtains the following consequence:[11]

$$\text{CP} + \text{Observed isotropy} \Rightarrow \text{Universal isotropy} \Rightarrow \text{Homogeneity}.$$

"Universal" means with respect to all comoving observers. The property of large-scale homogeneity has been dubbed the Cosmological Principle. It means that all large-scale properties of the universe depend only on the proper time t of the comoving observers. However, we see from the above that it is not an independent principle; it follows from the CP (so the same acronym is appropriate).

Another consequence of assuming the CP is the form of the metric, which corresponds to the fact that the curvature of the three-spaces $t =$ constant must also be uniform. We adopt the representation in which the interval is

$$ds^s = -dt^2 + a^2(t)[d\chi^2 + \Sigma^2(\chi)(d\theta^2 + \sin^2\theta d\phi^2)], \tag{28}$$

where χ is a dimensionless comoving radial coordinate, the curvature $K = k/a^2(t)$, and

$$\Sigma = \begin{cases} \sin\chi\,, & k = +1, \\ \chi\,, & 0, \\ \sinh\chi\,, & -1 \end{cases} .$$

Of course, a flat space ($k = 0$) is also the limit of the other two choices as the radius of curvature $a(t) \to \infty$.

Essentially all of our information about the universe comes to us via (radial) photons. So we integrate Eq. (28) for $ds = 0 = d\theta = d\phi$ from a comoving emitter (at time t and radial coordinate χ) to us (at time t_0 and $\chi = 0$) to give

$$\chi = \int_t^{t_0} a^{-1}(t')\, dt'. \tag{29}$$

Applying Eq. (29) to two photons separated by one period of oscillation, the fact that χ remains fixed gives the fundamental redshift (Z) formula

$$1 + Z \equiv \lambda_0/\lambda = a(t_0)/a(t). \tag{30}$$

Thus we see that the scale factor $a(t)$ directly stretches wavelengths, as well as determining physical (radar) distances such as the radial one $dl = a(t)d\chi$.

A final consequence of the symmetry induced by the CP plus observed isotropy is the fact that the matter stress-energy tensor $T_{\mu\nu}$ must assume the same form as that for a perfect fluid, given by equation (15). We shall now denote the corresponding effective density and pressure of the matter by ρ_m and p_m.

6.3.2 Dynamical Analysis

In order to understand what governs the dynamics of the evolution of the universe, we take the metric theory of gravity to be within the class of scalar-tensor theories presented above. We again employ the spin representation, in which the spin–2 metric $\hat{\mathbf{g}} = A^{-2}(\varphi)\mathbf{g}$, where $\mathbf{g}$ is the (physical representation) metric that couples directly to matter. However, we can retain the same form of the metric [given by the interval (28)] if we instead employ the time $d\hat{t} = A^{-1}(\varphi)dt$ and scale factor $\hat{a}(\hat{t}) = A^{-1}(\varphi)a(t)$. But we must remember that physical distances and times are given by the interval $ds^2 = A^2(\varphi)d\hat{s}^2$. Correspondingly, physical matter densities and pressures are obtained from $\rho_m = A^{-4}(\varphi)\hat{\rho}_m$ and $p_m = A^{-4}(\varphi)\hat{p}_m$.

The field Eq. (11) can then be put in the form:

$$(3/\hat{a})d^2\hat{a}/d\hat{t}^2 = -4\pi G[\hat{\rho}_m + \hat{\rho}_\varphi + 3(\hat{p}_m + \hat{p}_\varphi)], \tag{31}$$

$$(3/\hat{a}^2)[(d\hat{a}/d\hat{t})^2 + k] = 8\pi G(\hat{\rho}_m + \hat{\rho}_\varphi), \tag{32}$$

while the scalar field Eq. (12) assumes the form

$$d^2\varphi/d\hat{t}^2 + (3/\hat{a})(d\hat{a}/d\hat{t})d\varphi/d\hat{t} + \tfrac{1}{2}d\Lambda/d\varphi = -4\pi G\alpha(\varphi)[\hat{\rho}_m - 3\hat{p}_m]. \tag{33}$$

We have introduced the effective scalar field density and pressure

$$\hat{\rho}_\varphi \equiv (8\pi G)^{-1}[(d\varphi/d\hat{t})^2 + \Lambda(\varphi)], \quad \hat{p}_\varphi \equiv (8\pi G)^{-1}[(d\varphi/d\hat{t})^2 - \Lambda(\varphi)]. \tag{34}$$

In addition, we have "microscopic conservation of energy," in the form

$$d[(\hat{\rho}_m + \hat{\rho}_\varphi)\hat{a}^3]/d\hat{t} + (\hat{p}_m + \hat{p}_\varphi)d\hat{a}^3/d\hat{t} = 0. \tag{35}$$

Using Eq. (35) renders Eq. (31) redundant, while Eq. (32) can be put in the form of "macroscopic energy conservation":

$$\tfrac{1}{2}(d\hat{a}/d\hat{t})^2 + \mathcal{U}(\hat{a}) = -k/2, \tag{36}$$

where the effective gravitational potential $\mathcal{U} = -(4\pi G/3)\hat{a}^2(\hat{\rho}_m + \hat{\rho}_\varphi)$. For a small comoving sphere of radius $R_* = \chi\hat{a}(\hat{t})$ and mass $M_* = (4\pi R_*^3/3)(\hat{\rho}_m + \hat{\rho}_\varphi)$, one finds that $\chi^2\mathcal{U} = -GM_*/R_*$.

Let us consider the present epoch (when $\varphi = \varphi_0$), noting that we can set $A(\varphi_0) \equiv 1$. In addition, we use the fact that its derivative [see Eq. (13)] is small: $\alpha^2(\varphi_0) \equiv \alpha_0^2 \leq 10^{-3}$. Therefore, the spin and matter representations are approximately the same today. The measured Hubble constant is then

$$H_0 \equiv \left(\frac{1}{a}\frac{da}{dt}\right)_0 = \hat{H}_0 + \alpha_0\left(\frac{d\varphi}{dt}\right)_0.$$

The fundamental dynamical Eq. (36) then assumes the form

$$\Omega_0 + \Omega_K + \Omega_\Lambda + \Omega_{\dot{\varphi}} = [1 - \alpha_0(d\varphi/d\ln a)_0]^2, \tag{37}$$

where we adopt the standard notation

$$\Omega_0 \equiv \frac{8\pi G(\rho_m)_0}{3H_0^2}, \quad \Omega_K \equiv -\frac{k}{(a_0H_0)^2}, \quad \Omega_\Lambda \equiv \frac{\Lambda(\varphi_0)}{3H_0^2}, \quad \Omega_{\dot{\varphi}} \equiv \frac{(d\varphi/dt)_0^2}{3H_0^2}.$$

Note that Eq. (37) differs from the standard result[10] because of the contribution of a possible time dependence of the scalar field via $\Omega_{\dot{\varphi}}$ and the term which makes the right hand side no longer equal to unity. We point this out to illustrate one way in which a dynamical scalar field can impact the comparison of cosmological models with observations.

References

[1] B. F. Schutz, *A First Course in General Relativity* (Cambridge Univ. Press, Cambridge, 1986).

[2] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973).

[3] P. G. Bergmann, "Comments on the scalar-tensor theory," Int. J. Theor. Phys. **1**, 25 (1968).

[4] K. Nordtvedt, "Post-Newtonian metric for a general class of scalar-tensor gravitational theories and observational consequences," Astrophys. J. **161**, 1059 (1970).

[5] R. V. Wagoner, "Scalar-tensor theory and gravitational waves," Phys. Rev. D **1**, 3209 (1970).

[6] T. Damour and K. Nordtvedt, "Tensor-scalar cosmological models and their relaxation toward general relativity," Phys. Rev. D **48**, 3436 (1993).

[7] D. I. Santiago, D. Kalligas, and R. V. Wagoner, "Scalar-tensor cosmologies and their late time evolution," Phys. Rev. D **58**, 124005 (1998).

[8] C. M. Will, *Theory and Experiment in Gravitational Physics*, revised ed. (Cambridge Univ. Press, Cambridge, 1993).

[9] S. L. Shapiro and S. A. Teukolsky, *Black Holes, White Dwarfs, and Neutron Stars: The Physics of Compact Objects* (John Wiley & Sons, New York, 1983).

[10] P. J. E. Peebles, *Principles of Physical Cosmology* (Princeton Univ. Press, Princeton, 1993).

[11] S. Weinberg, *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity* (John Wiley & Sons, New York, 1972), pp. 378-379.

THE CONFRONTATION BETWEEN GENERAL RELATIVITY AND EXPERIMENT: A 1998 UPDATE

Clifford M. Will*
McDonnell Center for the Space Sciences
Department of Physics
Washington University, St. Louis, MO 63130

ABSTRACT

The status of experimental tests of general relativity and of theoretical frameworks for analyzing them is reviewed. Einstein's equivalence principle (EEP) is well supported by experiments such as the Eötvös experiment, tests of special relativity, and the gravitational redshift experiment. Future tests of EEP will search for new interactions arising from unification or quantum gravity. Tests of general relativity have reached high precision, including the light deflection, the Shapiro time delay, the perihelion advance of Mercury, and the Nordtvedt effect in lunar motion. Gravitational wave damping has been detected to half a percent using the binary pulsar, and new binary pulsar systems promise further improvements. When direct observation of gravitational radiation from astrophysical sources begins, new tests of general relativity will be possible.

*Support in part by NSF Grant 96-00049.

© 1998 by Clifford M. Will.

1 Introduction

At the time of the birth of general relativity (GR), experimental confirmation was almost a side issue. Einstein did calculate observable effects of general relativity, such as the deflection of light, which were tested, but compared to the inner consistency and elegance of the theory, he regarded such empirical questions as almost peripheral. But today, experimental gravitation is a major component of the field, characterized by continuing efforts to test the theory's predictions, to search for gravitational imprints of high-energy particle interactions, and to detect gravitational waves from astronomical sources.

The modern history of experimental relativity can be divided roughly into four periods: Genesis, Hibernation, a Golden Era, and the Quest for Strong Gravity. The Genesis (1887–1919) comprises the period of the two great experiments which were the foundation of relativistic physics—the Michelson-Morley experiment and the Eötvös experiment—and the two immediate confirmations of GR—the deflection of light and the perihelion advance of Mercury. Following this was a period of Hibernation (1920–1960) during which relatively few experiments were performed to test GR, and at the same time the field itself became sterile and stagnant, relegated to the backwaters of physics and astronomy.

But beginning around 1960, astronomical discoveries (quasars, pulsars, cosmic background radiation) and new experiments pushed GR to the forefront. Experimental gravitation experienced a Golden Era (1960–1980) during which a systematic, world-wide effort took place to understand the observable predictions of GR, to compare and contrast them with the predictions of alternative theories of gravity, and to perform new experiments to test them. The period began with an experiment to confirm the gravitational frequency shift of light (1960) and ended with the reported decrease in the orbital period of the binary pulsar at a rate consistent with the general relativity prediction of gravity-wave energy loss (1979). The results all supported GR, and most alternative theories of gravity fell by the wayside (for a popular review, see Ref. 1).

Since 1980, the field has entered what might be termed a Quest for Strong Gravity. Many of the remaining interesting weak-field predictions of the theory are extremely small and difficult to check, in some cases requiring further technological development to bring them into detectable range. The sense of a systematic assault on the weak-field predictions of GR has been supplanted to some extent by an opportunistic approach in which novel and unexpected (and sometimes inexpensive) tests of gravity have arisen from new theoretical ideas or experimental techniques, often from unlikely sources. Examples include the use of laser-cooled atom and ion traps to perform ultra-precise tests of special relativity, and the startling proposal of a "fifth" force, which led to a host of new tests of gravity at short ranges. Several major ongoing efforts also continue, principally the Stanford Gyroscope experiment, known as Gravity Probe-B.

Instead, much of the focus has shifted to experiments which can probe the effects of strong gravitational fields. At one extreme are the strong gravitational fields associated with Planck-scale physics. Will unification of the forces, or quantization of gravity at this scale leave observable effects accessible by experiment? Dramatically improved tests of the equivalence principle or of the "inverse square law" are being designed, to search for or bound the imprinted effects of Planck scale phenomena. At the other extreme are the strong fields associated with compact objects such as black holes or neutron stars. Astrophysical observations and gravitational-wave detectors are being planned to explore and test GR in the strong-field, highly-dynamical regime associated with the formation and dynamics of these objects.

In these lectures, we shall review theoretical frameworks for studying experimental gravitation, summarize the current status of experiments, and attempt to chart the future of the subject. We shall not provide complete references to work done in this field but instead will refer the reader to the appropriate review articles and monographs, specifically to *Theory and Experiment in Gravitational Physics*,[2] hereafter referred to as TEGP. Additional recent reviews on this subject are Refs. 3, 4, 5, and 6. Other references will be confined to reviews or monographs on specific topics, and to important recent papers that are not included in TEGP. References to TEGP will be by chapter or section, *e.g.* , "TEGP 8.9."

2 Tests of the Foundations of Gravitation Theory

2.1 The Einstein Equivalence Principle

The principle of equivalence has historically played an important role in the development of gravitation theory. Newton regarded this principle as such a cornerstone of mechanics that he devoted the opening paragraph of the *Principia* to it. In 1907, Einstein used the principle as a basic element of general relativity. We now regard the principle of equivalence as the foundation, not of Newtonian gravity or of GR, but of the broader idea that spacetime is curved.

One elementary equivalence principle is the kind Newton had in mind when he stated that the property of a body called "mass" is proportional to the "weight," and is known as the weak equivalence principle (WEP). An alternative statement of WEP is that the trajectory of a freely falling body (one not acted upon by such forces as electromagnetism and too small to be affected by tidal gravitational forces) is independent of its internal structure and composition. In the simplest case of dropping two different bodies in a gravitational field, WEP states that the bodies fall with the same acceleration.

A more powerful and far-reaching equivalence principle is known as the Einstein equivalence principle (EEP). It states that (i) WEP is valid, (ii) the outcome of any local non-gravitational experiment is independent of the velocity of the freely-falling reference frame in which it is performed, and (iii) the outcome of any local non-gravitational experiment is independent of where and when in the universe it is performed. The second piece of EEP is called local Lorentz invariance (LLI), and the third piece is called local position invariance (LPI).

For example, a measurement of the electric force between two charged bodies is a local non-gravitational experiment; a measurement of the gravitational force between two bodies (Cavendish experiment) is not.

The Einstein equivalence principle is the heart and soul of gravitational theory, for it is possible to argue convincingly that if EEP is valid, then gravitation must be a "curved spacetime" phenomenon, in other words, the effects of gravity must be equivalent to the effects of living in a curved spacetime. As a consequence of this argument, the only theories of gravity that can embody EEP are those that satisfy the postulates of "metric theories of gravity," which are (i) spacetime is endowed with a symmetric metric, (ii) the trajectories of freely falling bodies are geodesics of that metric, and (iii) in local freely falling reference frames, the non-gravitational laws of physics are those written in the language of special relativity. The argument that leads to this conclusion simply notes that, if EEP is valid, then in local freely falling frames, the laws governing experiments must be independent of the velocity of the frame (local Lorentz invariance), with constant values for the various atomic constants (in order to be independent of location). The only laws we know of that fulfill this are those that are compatible with special relativity, such as Maxwell's equations of electromagnetism. Furthermore, in local freely falling frames, test bodies appear to be unaccelerated; in other words they move on straight lines, but such "locally straight" lines simply correspond to "geodesics" in a curved spacetime (TEGP 2.3).

General relativity is a metric theory of gravity, but then so are many others, including the Brans-Dicke theory. The nonsymmetric gravitation theory (NGT) of Moffat is not a metric theory, neither is superstring theory (see Sec. 2.3). So the notion of curved spacetime is a very general and fundamental one, and therefore it is important to test the various aspects of the Einstein Equivalence Principle thoroughly.

A direct test of WEP is the comparison of the acceleration of two laboratory-sized bodies of different composition in an external gravitational field. If the principle were violated, then the accelerations of different bodies would differ. The simplest way to quantify such possible violations of WEP in a form suitable for comparison with experiment is to suppose that for a body with inertial mass m_I, the passive gravitational mass m_P is no longer equal to m_I, so that in a gravitational field g, the acceleration is given by $m_I a = m_P g$. Now the inertial mass of a typical laboratory body is made up of several types of mass-energy: rest energy, electromagnetic energy, weak-interaction energy, and so on. If one of these forms of energy contributes to m_P differently than it does to m_I, a violation of WEP would result. One could then write

$$m_P = m_I + \sum_A \eta^A E^A / c^2 , \tag{1}$$

where E^A is the internal energy of the body generated by interaction A, and η^A is a dimensionless parameter that measures the strength of the violation of WEP induced by that interaction, and c is the speed of light. A measurement or limit on the fractional difference in acceleration between two bodies then yields a quantity

called the "Eötvös ratio" given by

$$\eta \equiv \frac{2|a_1 - a_2|}{|a_1 + a_2|} = \sum_A \eta^A \left(\frac{E_1^A}{m_1 c^2} - \frac{E_2^A}{m_2 c^2} \right) , \qquad (2)$$

where we drop the subscript I from the inertial masses. Thus, experimental limits on η place limits on the WEP-violation parameters η^A.

Many high-precision Eötvös-type experiments have been performed, from the pendulum experiments of Newton, Bessel, and Potter, to the classic torsion-balance measurements of Eötvös, Dicke, Braginsky, and their collaborators. In the modern torsion-balance experiments, two objects of different composition are connected by a rod or placed on a tray and suspended in a horizontal orientation by a fine wire. If the gravitational acceleration of the bodies differs, there will be a torque induced on the suspension wire, related to the angle between the wire and the direction of the gravitational acceleration **g**. If the entire apparatus is rotated about some direction with angular velocity ω, the torque will be modulated with period $2\pi/\omega$. In the experiments of Eötvös and his collaborators, the wire and **g** were not quite parallel because of the centripetal acceleration on the apparatus due to the Earth's rotation; the apparatus was rotated about the direction of the wire. In the Princeton (Dicke *et al.*) and Moscow (Braginsky *et al.*) experiments, **g** was that of the Sun, and the rotation of the Earth provided the modulation of the torque at a period of 24 hrs [TEGP 2.4(a)]. Beginning in the late 1980s, numerous experiments were carried out primarily to search for a "fifth force" (see Sec. 2.3), but their null results also constituted tests of WEP. In the "free-fall Galileo experiment" performed at the University of Colorado, the relative free-fall acceleration of two bodies made of uranium and copper was measured using a laser interferometric technique. The "Eöt-Wash" experiment carried out at the University of Washington used a sophisticated torsion balance tray to compare the accelerations of beryllium and copper toward local topography on Earth, movable laboratory masses, the Sun and the galaxy.[7] The resulting upper limits on η are summarized in Figure 1 (TEGP 14.1; for a bibliography of experiments, see Ref. 8).

The second ingredient of EEP, local Lorentz invariance, can be said to be tested every time that special relativity is confirmed in the laboratory. However, many such experiments, especially in high-energy physics, are not "clean" tests, because in many cases it is unlikely that a violation of Lorentz invariance could be distinguished from effects due to the complicated strong and weak interactions.

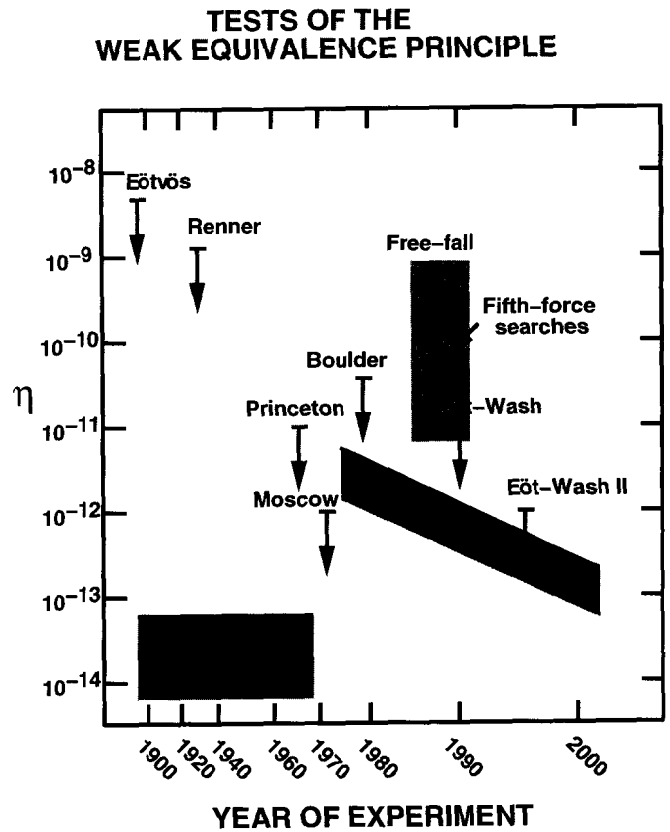


Fig. 1. Selected tests of the Weak Equivalence Principle, showing bounds on η, which measures fractional difference in acceleration of different materials or bodies. Free-fall and Eöt-Wash experiments originally performed to search for fifth force. Solid line shows current bounds on η for gravitating bodies from lunar laser ranging (LURE).

However, there is one class of experiments that can be interpreted as "clean," high-precision tests of local Lorentz invariance. These are the "mass anisotropy" experiments: the classic versions are the Hughes-Drever experiments, performed in the period 1959-60 independently by Hughes and collaborators at Yale University, and by Drever at Glasgow University [TEGP 2.4(b)]. Dramatically improved versions were carried out during the late 1980s using laser-cooled trapped atom techniques (TEGP 14.1). A simple and useful way of interpreting these experiments is to suppose that the electromagnetic interactions suffer a slight violation of Lorentz invariance, through a change in the speed of electromagnetic radiation c relative to the limiting speed of material test particles (c_0, chosen to be unity via a choice of units), in other words, $c \neq 1$ (see Sec. 2.2.3). Such a violation necessarily selects a preferred universal rest frame, presumably that of the cosmic background radiation, through which we are moving at about 300 km/s. Such a Lorentz-non-invariant electromagnetic interaction would cause shifts in the energy levels of atoms and nuclei that depend on the orientation of the quantization axis of the state relative to our universal velocity vector, and on the quantum numbers of the state. The presence or absence of such energy shifts can be examined by measuring the energy of one such state relative to another state that is either unaffected or is affected differently by the supposed violation. One way is to look for a shifting of the energy levels of states that are ordinarily equally spaced, such as the four J=3/2 ground states of the ^{7}Li nucleus in a magnetic field (Drever experiment); another is to compare the levels of a complex nucleus with the atomic hyperfine levels of a hydrogen maser clock. These experiments have all yielded extremely accurate results, quoted as limits on the parameter $\delta \equiv c^{-2} - 1$ in Figure 2. Also included for comparison is the corresponding limit obtained from Michelson-Morley type experiments.

Recent advances in atomic spectroscopy and atomic timekeeping have made it possible to test LLI by checking the isotropy of the speed of light using one-way propagation (as opposed to round-trip propagation, as in the Michelson-Morley experiment). In one experiment, for example, the relative phases of two hydrogen maser clocks at two stations of NASA's Deep Space Tracking Network were compared over five rotations of the Earth by propagating a light signal one-way along an ultrastable fiberoptic link connecting them (see Sec. 2.2.3). Although the bounds from these experiments are not as tight as those from mass-anisotropy experiments, they probe directly the fundamental postulates of special relativity,

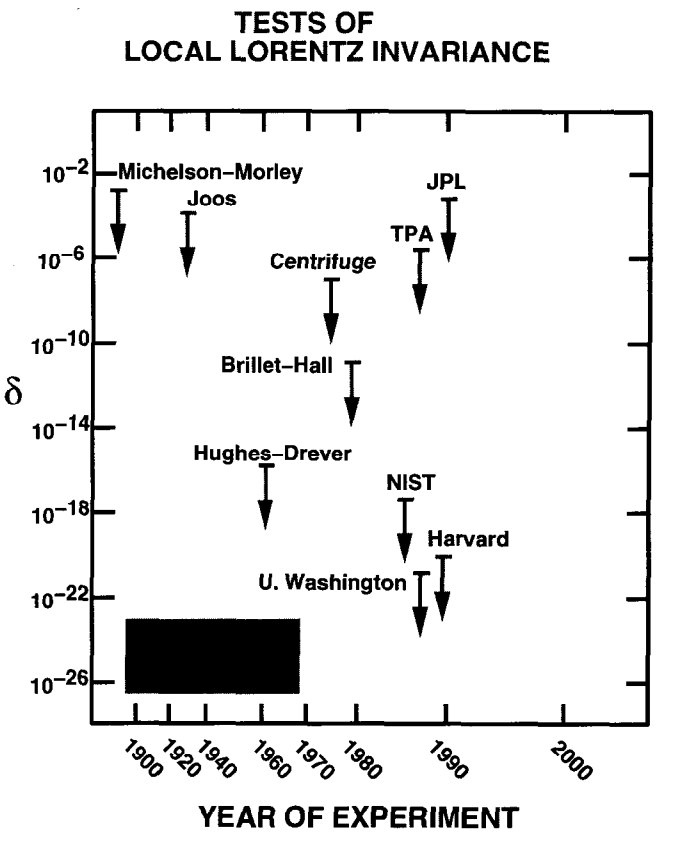


Fig. 2. Selected tests of local Lorentz invariance showing bounds on parameter δ, which measures degree of violation of Lorentz invariance in electromagnetism. Michelson-Morley, Joos, and Brillet-Hall experiments test isotropy of round-trip speed of light, the latter experiment using laser technology. Centrifuge, two-photon absorption (TPA), and JPL experiments test isotropy of light speed using one-way propagation. Remaining four experiments test isotropy of nuclear energy levels. Limits assume speed of Earth of 300 km/s relative to the mean rest frame of the universe.

been performed since 1960 [TEGP 2.4(c)].

The most precise experiment to date was the Vessot-Levine rocket redshift experiment that took place in June 1976. A hydrogen-maser clock was flown on a rocket to an altitude of about 10,000 km and its frequency compared to a similar clock on the ground. The experiment took advantage of the masers' frequency stability by monitoring the frequency shift as a function of altitude. A sophisticated data acquisition scheme accurately eliminated all effects of the first-order Doppler shift due to the rocket's motion, while tracking data were used to determine the payload's location and the velocity (to evaluate the potential difference ΔU, and the special relativistic time dilation). Analysis of the data yielded a limit $|\alpha| < 2 \times 10^{-4}$.

A "null" redshift experiment performed in 1978 tested whether the *relative* rates of two different clocks depended upon position. Two hydrogen maser clocks and an ensemble of three superconducting-cavity stabilized oscillator (SCSO) clocks were compared over a ten-day period. During this period, the solar potential U/c^2 changed sinusoidally with a 24-hour period by 3×10^{-13} because of the Earth's rotation, and changed linearly at 3×10^{-12} per day because the Earth is 90 degrees from perihelion in April. However, analysis of the data revealed no variations of either type within experimental errors, leading to a limit on the LPI violation parameter $|\alpha^{\rm H} - \alpha^{\rm SCSO}| < 2 \times 10^{-2}$. This bound is likely to be improved using more stable frequency standards.[9,10] The varying gravitational redshift of Earth-bound clocks relative to the highly stable Millisecond Pulsar, caused by the Earth's motion in the solar gravitational field around the Earth-Moon center of mass (amplitude 4000 km), has been measured to about ten percent, and the redshift of stable oscillator clocks on the Voyager spacecraft caused by Saturn's gravitational field yielded a one percent test. The solar gravitational redshift has been tested to about two percent using infrared oxygen triplet lines at the limb of the Sun, and to one percent using oscillator clocks on the Galileo spacecraft [TEGP 2.4(c) and 14.1(a)].

Modern advances in navigation using Earth-orbiting atomic clocks and accurate time-transfer must routinely take gravitational redshift and time-dilation effects into account. For example, the Global Positioning System (GPS) provides absolute accuracies of around 15 m (even better in its military mode) anywhere on Earth, which corresponds to 50 nanoseconds in time accuracy at all times. Yet the difference in rate between satellite and ground clocks as a result of special and and thereby of LLI. (TEGP 14.1).

The principle of local position invariance, the third part of EEP, can be tested by the gravitational redshift experiment, the first experimental test of gravitation proposed by Einstein. Despite the fact that Einstein regarded this as a crucial test of GR, we now realize that it does not distinguish between GR and any other metric theory of gravity; instead it is a test only of EEP. A typical gravitational redshift experiment measures the frequency or wavelength shift $Z \equiv \Delta\nu/\nu = -\Delta\lambda/\lambda$ between two identical frequency standards (clocks) placed at rest at different heights in a static gravitational field. If the frequency of a given type of atomic clock is the same when measured in a local, momentarily comoving freely falling frame (Lorentz frame), independent of the location or velocity of that frame, then the comparison of frequencies of two clocks at rest at different locations boils down to a comparison of the velocities of two local Lorentz frames, one at rest with respect to one clock at the moment of emission of its signal, the other at rest with respect to the other clock at the moment of reception of the signal. The frequency shift is then a consequence of the first-order Doppler shift between the frames. The structure of the clock plays no role whatsoever. The result is a shift

$$Z = \Delta U/c^2 , \tag{3}$$

where ΔU is the difference in the Newtonian gravitational potential between the receiver and the emitter. If LPI is not valid, then it turns out that the shift can be written

$$Z = (1+\alpha)\Delta U/c^2 , \tag{4}$$

where the parameter α may depend upon the nature of the clock whose shift is being measured [see TEGP 2.4(c) for details].

The first successful, high-precision redshift measurement was the series of Pound-Rebka-Snider experiments of 1960-1965, that measured the frequency shift of gamma-ray photons from ^{57}Fe as they ascended or descended the Jefferson Physical Laboratory tower at Harvard University. The high accuracy achieved—one percent—was obtained by making use of the Mössbauer effect to produce a narrow resonance line whose shift could be accurately determined. Other experiments since 1960 measured the shift of spectral lines in the Sun's gravitational field and the change in rate of atomic clocks transported aloft on aircraft, rockets, and satellites. Figure 3 summarizes the important redshift experiments that have

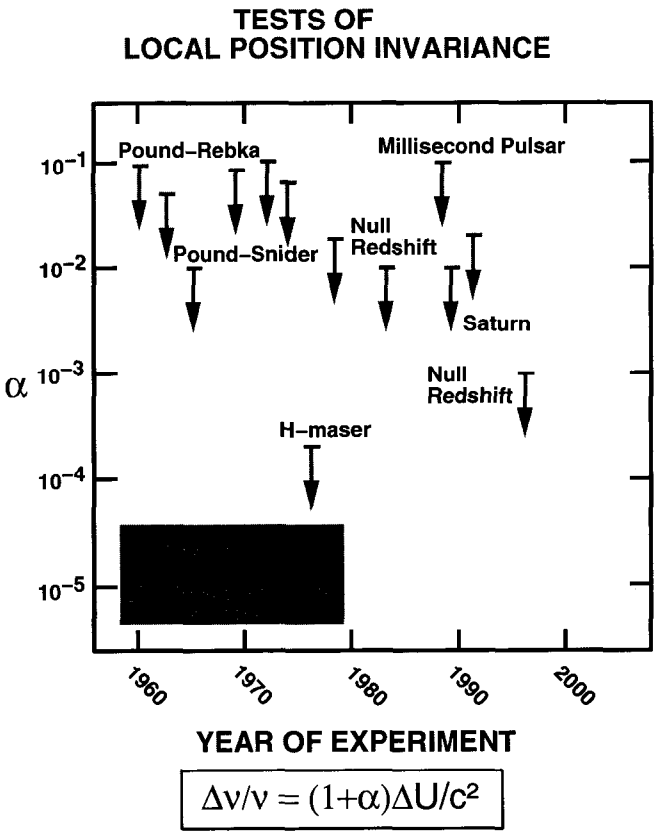


Fig. 3. Selected tests of local position invariance via gravitational redshift experiments, showing bounds on α, which measures degree of deviation of redshift from the formula $\Delta\nu/\nu = \Delta U/c^2$.

Constant k	Limit on $\dot{k}/k$ per Hubble time 1.2×10^{10} yr	Method
Fine structure constant $\alpha = e^2/\hbar c$	4×10^{-4}	H-maser vs Hg ion clock[10]
	6×10^{-7}	Oklo Natural Reactor[11]
	6×10^{-5}	21-cm vs molecular absorption at $Z = 0.7$ (Ref. 12)
Weak interaction constant $\beta = G_f m_p^2 c/\hbar^3$	1	^{187}Re, ^{40}K decay rates
	0.1	Oklo Natural Reactor[11]
	0.06	Big bang nucleosynthesis[13]
e-p mass ratio	1	Mass shift in quasar spectra at $Z \sim 2$
Proton g-factor (g_p)	10^{-5}	21-cm vs molecular absorption at $Z = 0.7$ (Ref. 12)

Table 1. Bounds on cosmological variation of fundamental constants of non-gravitational physics. For references to earlier work, see TEGP 2.4(c).

general relativistic effects is a whopping 40 *microseconds* per day (60μs from the gravitational redshift, and $-20\,\mu$s from time dilation). If these effects were not accurately accounted for, GPS would fail to function at its stated accuracy. This represents a welcome practical application of GR!

Local position invariance also refers to position in time. If LPI is satisfied, the fundamental constants of non-gravitational physics should be constants in time. Table 1 shows current bounds on cosmological variations in selected dimensionless constants. For discussion and references to early work, see TEGP 2.4(c).

2.2 Theoretical Frameworks for Analyzing EEP

2.2.1 Schiff's Conjecture

Because the three parts of the Einstein equivalence principle discussed above are so very different in their empirical consequences, it is tempting to regard them as independent theoretical principles. On the other hand, any complete and self-

consistent gravitation theory must possess sufficient mathematical machinery to make predictions for the outcomes of experiments that test each principle, and because there are limits to the number of ways that gravitation can be meshed with the special relativistic laws of physics, one might not be surprised if there were theoretical connections between the three sub-principles. For instance, the same mathematical formalism that produces equations describing the free fall of a hydrogen atom must also produce equations that determine the energy levels of hydrogen in a gravitational field, and thereby the ticking rate of a hydrogen maser clock. Hence a violation of EEP in the fundamental machinery of a theory that manifests itself as a violation of WEP might also be expected to show up as a violation of local position invariance. Around 1960, Schiff conjectured that this kind of connection was a necessary feature of any self-consistent theory of gravity. More precisely, Schiff's conjecture states that *any complete, self-consistent theory of gravity that embodies WEP necessarily embodies EEP*. In other words, the validity of WEP alone guarantees the validity of local Lorentz and position invariance, and thereby of EEP.

If Schiff's conjecture is correct, then Eötvös experiments may be seen as the direct empirical foundation for EEP, hence for the interpretation of gravity as a curved-spacetime phenomenon. Of course, a rigorous proof of such a conjecture is impossible (indeed, some special counter-examples are known) yet a number of powerful "plausibility" arguments can be formulated.

The most general and elegant of these arguments is based upon the assumption of energy conservation. This assumption allows one to perform very simple cyclic gedanken experiments in which the energy at the end of the cycle must equal that at the beginning of the cycle. This approach was pioneered by Dicke, Nordtvedt and Haugan. A system in a quantum state A decays to state B, emitting a quantum of frequency ν. The quantum falls a height H in an external gravitational field and is shifted to frequency ν', while the system in state B falls with acceleration g_B. At the bottom, state A is rebuilt out of state B, the quantum of frequency ν', and the kinetic energy $m_B g_B H$ that state B has gained during its fall. The energy left over must be exactly enough, $m_A g_A H$, to raise state A to its original location. (Here an assumption of local Lorentz invariance permits the inertial masses m_A and m_B to be identified with the total energies of the bodies.) If g_A and g_B depend on that portion of the internal energy of the states that was involved in the quantum transition from A to B according to

$$g_A = g(1+\alpha E_A/m_A c^2)\,, \qquad g_B = g(1+\alpha E_B/m_B c^2)\,, \qquad E_A - E_B \equiv h\nu \quad (5)$$

(violation of WEP), then by conservation of energy, there must be a corresponding violation of LPI in the frequency shift of the form (to lowest order in $h\nu/mc^2$)

$$Z = (\nu' - \nu)/\nu' = (1+\alpha) gH/c^2 = (1+\alpha)\Delta U/c^2\,. \quad (6)$$

Haugan generalized this approach to include violations of LLI (TEGP 2.5).

2.2.2 The $TH\epsilon\mu$ Formalism

The first successful attempt to prove Schiff's conjecture more formally was made by Lightman and Lee. They developed a framework called the $TH\epsilon\mu$ formalism that encompasses all metric theories of gravity and many non-metric theories (Box 1). It restricts attention to the behavior of charged particles (electromagnetic interactions only) in an external static spherically symmetric (SSS) gravitational field, described by a potential U. It characterizes the motion of the charged particles in the external potential by two arbitrary functions $T(U)$ and $H(U)$, and characterizes the response of electromagnetic fields to the external potential (gravitationally modified Maxwell equations) by two functions $\epsilon(U)$ and $\mu(U)$. The forms of T, H, ϵ, and μ vary from theory to theory, but every metric theory satisfies

$$\epsilon = \mu = (H/T)^{1/2}\,, \quad (7)$$

for all U. This consequence follows from the action of electrodynamics with a "minimal" or metric coupling:

$$\begin{aligned} I &= -\sum_a m_{0a} \int (g_{\mu\nu} v_a^\mu v_a^\nu)^{1/2} dt + \sum_a e_a \int A_\mu(x_a^\nu) v_a^\mu dt \\ &- \frac{1}{16\pi} \int \sqrt{-g}\, g^{\mu\alpha} g^{\nu\beta} F_{\mu\nu} F_{\alpha\beta} d^4x\,, \end{aligned} \quad (8)$$

where the variables are defined in Box 1, and where $F_{\mu\nu} \equiv A_{\nu,\mu} - A_{\mu,\nu}$. By identifying $g_{00} = T$ and $g_{ij} = H\delta_{ij}$ in a SSS field, $F_{i0} = E_i$ and $F_{ij} = \epsilon_{ijk} B_k$, one obtains Equation 7.

Conversely, every theory within this class that satisfies Equation 7 can have its electrodynamic equations cast into "metric" form. Lightman and Lee then calculated explicitly the rate of fall of a "test" body made up of interacting charged

Box 1. The $TH\epsilon\mu$ Formalism

1. **Coordinate System and Conventions:**

   $x^0 = t$ = time coordinate associated with the static nature of the static spherically symmetric (SSS) gravitational field; $\mathbf{x} = (x, y, z)$ = isotropic quasi-Cartesian spatial coordinates; spatial vector and gradient operations as in Cartesian space.

2. **Matter and Field Variables:**

   - m_{0a} = rest mass of particle a.
   - e_a = charge of particle a.
   - $x_a^\mu(t)$ = world line of particle a.
   - $v_a^\mu = dx_a^\mu/dt$ = coordinate velocity of particle a.
   - A_μ = electromagnetic vector potential; $\mathbf{E} = \nabla A_0 - \partial\mathbf{A}/\partial t$, $\mathbf{B} = \nabla \times \mathbf{A}$.

3. **Gravitational Potential:** $U(\mathbf{x})$.

4. **Arbitrary Functions:**

   $T(U)$, $H(U)$, $\epsilon(U)$, $\mu(U)$; EEP is satisfied iff $\epsilon = \mu = (H/T)^{1/2}$ for all U.

5. **Action:**

$$I = -\sum_a m_{0a}\int(T - Hv_a^2)^{1/2}dt + \sum_a e_a\int A_\mu(x_a^\nu)v_a^\mu dt + (8\pi)^{-1}\int(\epsilon E^2 - \mu^{-1}B^2)d^4x\,.$$

6. **Non-Metric Parameters:**

$$\begin{aligned}\Gamma_0 &= -c_0^2(\partial/\partial U)\ln[\epsilon(T/H)^{1/2}]_0\,,\\ \Lambda_0 &= -c_0^2(\partial/\partial U)\ln[\mu(T/H)^{1/2}]_0\,,\\ \Upsilon_0 &= 1 - (TH^{-1}\epsilon\mu)_0\,,\end{aligned}$$

   where $c_0 = (T_0/H_0)^{1/2}$ and subscript "0" refers to a chosen point in space. If EEP is satisfied, $\Gamma_0 = \Lambda_0 = \Upsilon_0 = 0$.

particles, and found that the rate was independent of the internal electromagnetic structure of the body (WEP) if and only if Equation 7 was satisfied. In other words WEP $\rightarrow$ EEP and Schiff's conjecture was verified, at least within the restrictions built into the formalism.

Certain combinations of the functions T, H, ϵ, and μ reflect different aspects of EEP. For instance, position or U-dependence of either of the combinations $\epsilon(T/H)^{1/2}$ and $\mu(T/H)^{1/2}$ signals violations of LPI, the first combination playing the role of the locally measured electric charge or fine structure constant. The "non-metric parameters" Γ_0 and Λ_0 (Box 1) are measures of such violations of EEP. Similarly, if the parameter $\Upsilon_0 \equiv 1 - (TH^{-1}\epsilon\mu)_0$ is non-zero anywhere, then violations of LLI will occur. This parameter is related to the difference between the speed of light, c, and the limiting speed of material test particles, c_o, given by

$$c = (\epsilon_0\mu_0)^{-1/2}\,, \qquad c_o = (T_0/H_0)^{1/2}\,. \tag{9}$$

In many applications, by suitable definition of units, c_0 can be set equal to unity. If EEP is valid, $\Gamma_0 \equiv \Lambda_0 \equiv \Upsilon_0 = 0$ everywhere.

The rate of fall of a composite spherical test body of electromagnetically interacting particles then has the form

$$\mathbf{a} = (m_P/m)\nabla U\,, \tag{10}$$

$$m_P/m = 1 + (E_B^{ES}/Mc_0^2)[2\Gamma_0 - \frac{8}{3}\Upsilon_0] + (E_B^{MS}/Mc_0^2)[2\Lambda_0 - \frac{4}{3}\Upsilon_0] + \ldots, \tag{11}$$

where E_B^{ES} and E_B^{MS} are the electrostatic and magnetostatic binding energies of the body, given by

$$E_B^{ES} = -\frac{1}{4}T_0^{1/2}H_0^{-1}\epsilon_0^{-1}\left\langle\sum_{ab} e_a e_b r_{ab}^{-1}\right\rangle\,, \tag{12}$$

$$E_B^{MS} = -\frac{1}{8}T_0^{1/2}H_0^{-1}\mu_0\left\langle\sum_{ab} e_a e_b r_{ab}^{-1}[\mathbf{v}_a\cdot\mathbf{v}_b + (\mathbf{v}_a\cdot\mathbf{n}_{ab})(\mathbf{v}_b\cdot\mathbf{n}_{ab})]\right\rangle\,, \tag{13}$$

where $r_{ab} = |\mathbf{x}_a - \mathbf{x}_b|$, $\mathbf{n}_{ab} = (\mathbf{x}_a - \mathbf{x}_b)/r_{ab}$, and the angle brackets denote an expectation value of the enclosed operator for the system's internal state. Eötvös experiments place limits on the WEP-violating terms in Equation 11, and ultimately place limits on the non-metric parameters $|\Gamma_0| < 2\times10^{-10}$ and $|\Lambda_0| < 3\times10^{-6}$. (We set $\Upsilon_0 = 0$ because of very tight constraints on it from tests of LLI .) These limits are sufficiently tight to rule out a number of non-metric theories of gravity thought previously to be viable [TEGP 2.6(f)].

The $TH\epsilon\mu$ formalism also yields a gravitationally modified Dirac equation that can be used to determine the gravitational redshift experienced by a variety of atomic clocks. For the redshift parameter α (Equation 4), the results are [TEGP 2.6(c)]

$$\alpha = \begin{cases} -3\Gamma_0 + \Lambda_0 & \text{hydrogen hyperfine transition, H} - \text{Maser clock} \\ -\frac{1}{2}(3\Gamma_0 + \Lambda_0) & \text{electromagnetic mode in cavity, SCSO clock} \\ -2\Gamma_0 & \text{phonon mode in solid, principal transition in hydrogen.} \end{cases} \tag{14}$$

The redshift is the standard one ($\alpha = 0$), independently of the nature of the clock if and only if $\Gamma_0 \equiv \Lambda_0 \equiv 0$. Thus the Vessot-Levine rocket redshift experiment sets a limit on the parameter combination $|3\Gamma_0 - \Lambda_0|$ (Figure 3); the null-redshift experiment comparing hydrogen-maser and SCSO clocks sets a limit on $|\alpha_H - \alpha_{SCSO}| = \frac{3}{2}|\Gamma_0 - \Lambda_0|$.

2.2.3 The c^2 Formalism

The $TH\epsilon\mu$ formalism can also be applied to tests of local Lorentz invariance, but in this context it can be simplified. Since most such tests do not concern themselves with the spatial variation of the functions T, H, ϵ, and μ, but rather with observations made in moving frames, we can treat them as spatial constants. Then by rescaling the time and space coordinates, the charges and the electromagnetic fields, we can put the action in Box 1 into the form [TEGP 2.6(a)].

$$I = -\sum_a m_{0a} \int (1-v_a^2)^{1/2} dt + \sum_a e_a \int A_\mu(x_a^\nu) v_a^\mu dt + (8\pi)^{-1} \int (E^2 - c^2 B^2) d^4x \,, \tag{15}$$

where $c^2 \equiv H_0/T_0\epsilon_0\mu_0 = (1 - \Upsilon_0)^{-1}$. This amounts to using units in which the limiting speed c_o of massive test particles is unity, and the speed of light is c. If $c \neq 1$, LLI is violated; furthermore, the form of the action above must be assumed to be valid only in some preferred universal rest frame. The natural candidate for such a frame is the rest frame of the microwave background.

The electrodynamical equations which follow from Equation 15 yield the behavior of rods and clocks, just as in the full $TH\epsilon\mu$ formalism. For example, the length of a rod moving through the rest frame in a direction parallel to its length will be observed by a rest observer to be contracted relative to an identical rod perpendicular to the motion by a factor $1 - V^2/2 + O(V^4)$. Notice that c does not appear in this expression. The energy and momentum of an electromagnetically bound body which moves with velocity $\mathbf{V}$ relative to the rest frame are given by

$$\begin{aligned} E &= M_R + \frac{1}{2} M_R V^2 + \frac{1}{2} \delta M_I^{ij} V^i V^j \,, \\ P^i &= M_R V^i + \delta M_I^{ij} V^j \,, \end{aligned} \tag{16}$$

where $M_R = M_0 - E_B^{ES}$, M_0 is the sum of the particle rest masses, E_B^{ES} is the electrostatic binding energy of the system (Equation 12 with $T_0^{1/2} H_0 \epsilon_0^{-1} = 1$), and

$$\delta M_I^{ij} = -2(\frac{1}{c^2} - 1)[\frac{4}{3} E_B^{ES} \delta^{ij} + \tilde{E}_B^{ESij}] \,, \tag{17}$$

where

$$\tilde{E}_B^{ESij} = -\frac{1}{4} \left\langle \sum_{ab} e_a e_b r_{ab}^{-1} \left((n_{ab}^i n_{ab}^j - \frac{1}{3} \delta^{ij} \right) \right\rangle \,. \tag{18}$$

Note that $(c^{-2} - 1)$ corresponds to the parameter δ plotted in Figure 2.

The electrodynamics given by Equation 15 can also be quantized, so that we may treat the interaction of photons with atoms via perturbation theory. The energy of a photon is $\hbar$ times its frequency ω, while its momentum is $\hbar\omega/c$. Using this approach, one finds that the difference in round trip travel times of light along the two arms of the interferometer in the Michelson-Morley experiment is given by $L_0(v^2/c)(c^{-2} - 1)$. The experimental null result then leads to the bound on $(c^{-2} - 1)$ shown on Figure 2. Similarly the anisotropy in energy levels is clearly illustrated by the tensorial terms in Equations 16 and 18; by evaluating $\tilde{E}_B^{ESij}$ for each nucleus in the various Hughes-Drever-type experiments and comparing with the experimental limits on energy differences, one obtains the extremely tight bounds also shown on Figure 2.

The behavior of moving atomic clocks can also be analyzed in detail, and bounds on $(c^{-2} - 1)$ can be placed using results from tests of time dilation and of the propagation of light. In some cases, it is advantageous to combine the c^2 framework with a "kinematical" viewpoint that treats a general class of boost transformations between moving frames. Such kinematical approaches have been discussed by Robertson, Mansouri and Sexl, and Will (see Ref. 14).

For example, in the "JPL" experiment, in which the phases of two hydrogen masers connected by a fiberoptic link were compared as a function of the Earth's orientation, the predicted phase difference as a function of direction is, to first order in $\mathbf{V}$, the velocity of the Earth through the cosmic background,

$$\Delta\phi/\tilde{\phi} \approx -\frac{4}{3}(1 - c^2)(\mathbf{V} \cdot \mathbf{n} - \mathbf{V} \cdot \mathbf{n}_0) \,, \tag{19}$$

where $\tilde{\phi} = 2\pi\nu L$, ν is the maser frequency, $L = 21$ km is the baseline, and where $\mathbf{n}$ and $\mathbf{n}_0$ are unit vectors along the direction of propagation of the light, at a given time, and at the initial time of the experiment, respectively. The observed limit on a diurnal variation in the relative phase resulted in the bound $|c^{-2} - 1| < 3 \times 10^{-4}$. Tighter bounds were obtained from a "two-photon absorption" (TPA) experiment, and a 1960s series of "Mössbauer-rotor" experiments, which tested the isotropy of time dilation between a gamma ray emitter on the rim of a rotating disk and an absorber placed at the center.[14]

2.3 EEP, Particle Physics, and the Search for New Interactions

In 1986, as a result of a detailed reanalysis of Eötvös' original data, Fischbach *et al.* suggested the existence of a fifth force of nature, with a strength of about a percent that of gravity, but with a range (as defined by the range λ of a Yukawa potential, $e^{-r/\lambda}/r$) of a few hundred meters. This proposal dovetailed with earlier hints of a deviation from the inverse-square law of Newtonian gravitation derived from measurements of the gravity profile down deep mines in Australia, and with ideas from particle physics suggesting the possible presence of very low-mass particles with gravitational-strength couplings. During the next four years numerous experiments looked for evidence of the fifth force by searching for composition-dependent differences in acceleration, with variants of the Eötvös experiment or with free-fall Galileo-type experiments. Although two early experiments reported positive evidence, the others all yielded null results. Over the range between one and 10^4 meters, the null experiments produced upper limits on the strength of a postulated fifth force between 10^{-3} and 10^{-6} of the strength of gravity. Interpreted as tests of WEP (corresponding to the limit of infinite-range forces), the results of the free-fall Galileo experiment, and of the Eöt-Wash III experiment are shown in Figure 1. At the same time, tests of the inverse-square law of gravity were carried out by comparing variations in gravity measurements up tall towers or down mines or boreholes with gravity variations predicted using the inverse square law together with Earth models and surface gravity data mathematically "continued" up the tower or down the hole. Despite early reports of anomalies, independent tower, borehole, and seawater measurements now show no evidence of a deviation. The consensus at present is that there is no credible experimental evidence for a fifth force of nature. For reviews and bibliographies, see Refs. 8, 15, 16, and 17.

Nevertheless, theoretical evidence continues to mount that EEP is *likely* to be violated at some level, whether by quantum gravity effects, by effects arising from string theory, or by hitherto undetected interactions, albeit at levels below those that motivated the fifth-force searches. Roughly speaking, in addition to the pure Einsteinian gravitational interaction, which respects EEP, theories such as string theory predict other interactions which do not. In string theory, for example, the existence of such EEP-violating fields is assured, but the theory is not yet mature enough to enable calculation of their strength (relative to gravity), or their range (whether they are long range, like gravity, or short range, like the nuclear and weak interactions, or too short-range to be detectable) (for further discussion see Ref. 6).

In one simple example, one can write the Lagrangian for the low-energy limit of string theory in the so-called "Einstein frame," in which the gravitational Lagrangian is purely general relativistic:

$$\tilde{\mathcal{L}} = \sqrt{-\tilde{g}}\Big(\tilde{g}^{\mu\nu}\Big[\tfrac{1}{2\kappa}\tilde{R}_{\mu\nu} - \tfrac{1}{2}\tilde{G}(\varphi)\partial_\mu\varphi\partial_\nu\varphi\Big] - U(\varphi)\tilde{g}^{\mu\nu}\tilde{g}^{\alpha\beta}F_{\mu\alpha}F_{\nu\beta} + \bar{\tilde{\psi}}\Big[i\tilde{e}^\mu_a\gamma^a(\partial_\mu + \tilde{\Omega}_\mu + qA_\mu) - \tilde{M}(\varphi)\Big]\tilde{\psi}\Big), \tag{20}$$

where $\tilde{g}_{\mu\nu}$ is the non-physical metric, $\tilde{R}_{\mu\nu}$ is the Ricci tensor derived from it, φ is a dilaton field, and $\tilde{G}$, U, and $\tilde{M}$ are functions of φ. The Lagrangian includes that for the electromagnetic field $F_{\mu\nu}$, and that for particles, written in terms of Dirac spinors $\tilde{\psi}$. This is not a metric representation because of the coupling of φ to matter via $\tilde{M}(\varphi)$ and $U(\varphi)$. A conformal transformation $\tilde{g}_{\mu\nu} = F(\varphi)g_{\mu\nu}$, $\tilde{\psi} = F(\varphi)^{-3/4}\psi$, puts the Lagrangian in the form ("Jordan" frame)

$$\mathcal{L} = \sqrt{-g}\Big(g^{\mu\nu}\Big[\tfrac{1}{2\kappa}F(\varphi)R_{\mu\nu} - \tfrac{1}{2}F(\varphi)\tilde{G}(\varphi)\partial_\mu\varphi\partial_\nu\varphi + \tfrac{3}{4\kappa F(\varphi)}\partial_\mu F\partial_\nu F\Big] - U(\varphi)g^{\mu\nu}g^{\alpha\beta}F_{\mu\alpha}F_{\nu\beta} + \bar{\psi}\Big[ie^\mu_a\gamma^a(\partial_\mu + \Omega_\mu + qA_\mu) - \tilde{M}(\varphi)F^{1/2}\Big]\psi\Big). \tag{21}$$

One may choose $F(\varphi) = \text{const}/\tilde{M}(\varphi)^2$ so that the particle Lagrangian takes the metric form (no coupling to φ), but the electromagnetic Lagrangian will still couple non-metrically to $U(\varphi)$. The gravitational Lagrangian here takes the form of a scalar-tensor theory (Sec. 3.3.2). But the non-metric electromagnetic term will, in general, produce violations of EEP.

Thus, EEP and related tests are now viewed as ways to discover or place constraints on new physical interactions, or as a branch of "non-accelerator particle physics," searching for the possible imprints of high-energy particle effects in the low-energy realm of gravity. Whether current or proposed experiments can actually probe these phenomena meaningfully is an open question at the moment, largely because of a dearth of firm theoretical predictions. Despite this uncertainty, a number of experimental possibilities are being explored.

The concept of an equivalence principle experiment in space has been developed, with the potential to test WEP to 10^{-18}. Known generically as Satellite Test of the Equivalence Principle (STEP), various versions of the project are under consideration as possible joint efforts of NASA and the European Space Agency (ESA). The gravitational redshift could be improved to the 10^{-10} level using atomic clocks on board a spacecraft which would travel to within four solar radii of the Sun. Laboratory tests of the gravitational inverse square law at sub-millimeter scales are being developed as ways to search for new short-range interactions; the challenge of these experiments is to distinguish gravitation-like interactions from electromagnetic and quantum mechanical (Casimir) effects.

3 Tests of Post-Newtonian Gravity

3.1 Metric Theories of Gravity and the Strong Equivalence Principle

3.1.1 Universal Coupling and the Metric Postulates

The overwhelming empirical evidence supporting the Einstein equivalence principle, discussed in the previous section, supports the conclusion that the only theories of gravity that have a hope of being viable are metric theories, or possibly theories that are metric apart from possible weak or short-range non-metric couplings (as in string theory). Therefore for the remainder of these lectures, we shall turn our attention exclusively to metric theories of gravity, which assume that (i) there exists a symmetric metric, (ii) test bodies follow geodesics of the metric, and (iii) in local Lorentz frames, the non-gravitational laws of physics are those of special relativity.

The property that all non-gravitational fields should couple in the same manner to a single gravitational field is sometimes called "universal coupling." Because of it, one can discuss the metric as a property of spacetime itself rather than as a field over spacetime. This is because its properties may be measured and studied using a variety of different experimental devices, composed of different non-gravitational fields and particles, and, because of universal coupling, the results will be independent of the device. Thus, for instance, the proper time between two events is a characteristic of spacetime and of the location of the events, not of the clocks used to measure it.

Consequently, if EEP is valid, the non-gravitational laws of physics may be formulated by taking their special relativistic forms in terms of the Minkowski metric $\boldsymbol{\eta}$ and simply "going over" to new forms in terms of the curved spacetime metric $\mathbf{g}$, using the mathematics of differential geometry. The details of this "going over" can be found in standard textbooks (Refs. 18 and 19; TEGP 3.2)

3.1.2 The Strong Equivalence Principle

In any metric theory of gravity, matter and non-gravitational fields respond only to the spacetime metric $\mathbf{g}$. In principle, however, there could exist other gravitational fields besides the metric, such as scalar fields, vector fields, and so on. If, by our strict definition of metric theory, matter does not couple to these fields, what can their role in gravitation theory be? Their role must be that of mediating the manner in which matter and non-gravitational fields generate gravitational fields and produce the metric; once determined, however, the metric alone acts back on the matter in the manner prescribed by EEP.

What distinguishes one metric theory from another, therefore, is the number and kind of gravitational fields it contains in addition to the metric, and the equations that determine the structure and evolution of these fields. From this viewpoint, one can divide all metric theories of gravity into two fundamental classes: "purely dynamical" and "prior-geometric."

By "purely dynamical metric theory" we mean any metric theory whose gravitational fields have their structure and evolution determined by coupled partial differential field equations. In other words, the behavior of each field is influenced to some extent by a coupling to at least one of the other fields in the theory. By "prior geometric" theory, we mean any metric theory that contains

"absolute elements," fields or equations whose structure and evolution are given *a priori*, and are independent of the structure and evolution of the other fields of the theory. These "absolute elements" typically include flat background metrics $\boldsymbol{\eta}$, cosmic time coordinates t, algebraic relationships among otherwise dynamical fields, such as $g_{\mu\nu} = h_{\mu\nu} + k_\mu k_\nu$, where $h_{\mu\nu}$ and k_μ may be dynamical fields.

General relativity is a purely dynamical theory since it contains only one gravitational field, the metric itself, and its structure and evolution are governed by partial differential equations (Einstein's equations). Brans-Dicke theory and its generalizations are purely dynamical theories; the field equation for the metric involves the scalar field (as well as the matter as source), and that for the scalar field involves the metric. Rosen's bimetric theory is a prior-geometric theory: it has a non-dynamical, Riemann-flat background metric, $\boldsymbol{\eta}$, and the field equations for the physical metric $\mathbf{g}$ involve $\boldsymbol{\eta}$.

By discussing metric theories of gravity from this broad point of view, it is possible to draw some general conclusions about the nature of gravity in different metric theories, conclusions that are reminiscent of the Einstein equivalence principle, but that are subsumed under the name "strong equivalence principle."

Consider a local, freely falling frame in any metric theory of gravity. Let this frame be small enough that inhomogeneities in the external gravitational fields can be neglected throughout its volume. On the other hand, let the frame be large enough to encompass a system of gravitating matter and its associated gravitational fields. The system could be a star, a black hole, the solar system or a Cavendish experiment. Call this frame a "quasi-local Lorentz frame." To determine the behavior of the system we must calculate the metric. The computation proceeds in two stages. First we determine the external behavior of the metric and gravitational fields, thereby establishing boundary values for the fields generated by the local system, at a boundary of the quasi-local frame "far" from the local system. Second, we solve for the fields generated by the local system. But because the metric is coupled directly or indirectly to the other fields of the theory, its structure and evolution will be influenced by those fields, and in particular by the boundary values taken on by those fields far from the local system. This will be true even if we work in a coordinate system in which the asymptotic form of $g_{\mu\nu}$ in the boundary region between the local system and the external world is that of the Minkowski metric. Thus the gravitational environment in which the local gravitating system resides can influence the metric generated by the local system via the boundary values of the auxiliary fields. Consequently, the results of local gravitational experiments may depend on the location and velocity of the frame relative to the external environment. Of course, local *non*-gravitational experiments are unaffected since the gravitational fields they generate are assumed to be negligible, and since those experiments couple only to the metric, whose form can always be made locally Minkowskian at a given spacetime event. Local gravitational experiments might include Cavendish experiments, measurement of the acceleration of massive self-gravitating bodies, studies of the structure of stars and planets, or analyses of the periods of "gravitational clocks." We can now make several statements about different kinds of metric theories.

(i) A theory which contains only the metric $\mathbf{g}$ yields local gravitational physics which is independent of the location and velocity of the local system. This follows from the fact that the only field coupling the local system to the environment is $\mathbf{g}$, and it is always possible to find a coordinate system in which $\mathbf{g}$ takes the Minkowski form at the boundary between the local system and the external environment. Thus the asymptotic values of $g_{\mu\nu}$ are constants independent of location, and are asymptotically Lorentz invariant, thus independent of velocity. General relativity is an example of such a theory.

(ii) A theory which contains the metric $\mathbf{g}$ and dynamical scalar fields φ_A yields local gravitational physics which may depend on the location of the frame but which is independent of the velocity of the frame. This follows from the asymptotic Lorentz invariance of the Minkowski metric and of the scalar fields, but now the asymptotic values of the scalar fields may depend on the location of the frame. An example is Brans-Dicke theory, where the asymptotic scalar field determines the effective value of the gravitational constant, which can thus vary as φ varies.

(iii) A theory which contains the metric $\mathbf{g}$ and additional dynamical vector or tensor fields or prior-geometric fields yields local gravitational physics which may have both location and velocity-dependent effects.

These ideas can be summarized in the strong equivalence principle (SEP), which states that (i) WEP is valid for self-gravitating bodies as well as for test bodies, (ii) the outcome of any local test experiment is independent of the velocity of the (freely falling) apparatus, and (iii) the outcome of any local test experiment is independent of where and when in the universe it is performed. The distinction between SEP and EEP is the inclusion of bodies with self-gravitational interactions (planets, stars) and of experiments involving gravitational forces (Cavendish

experiments, gravimeter measurements). Note that SEP contains EEP as the special case in which local gravitational forces are ignored.

The above discussion of the coupling of auxiliary fields to local gravitating systems indicates that if SEP is strictly valid, there must be one and only one gravitational field in the universe, the metric **g**. These arguments are only suggestive however, and no rigorous proof of this statement is available at present. Empirically it has been found that every metric theory other than GR introduces auxiliary gravitational fields, either dynamical or prior geometric, and thus predicts violations of SEP at some level (here we ignore quantum-theory inspired modifications to GR involving "R^2" terms). General relativity seems to be the only metric theory that embodies SEP completely. This lends some credence to the conjecture SEP $\rightarrow$ General Relativity. In Sec. 3.6, we shall discuss experimental evidence for the validity of SEP.

3.2 The Parametrized Post-Newtonian Formalism

Despite the possible existence of long-range gravitational fields in addition to the metric in various metric theories of gravity, the postulates of those theories demand that matter and non-gravitational fields be completely oblivious to them. The only gravitational field that enters the equations of motion is the metric **g**. The role of the other fields that a theory may contain can only be that of helping to generate the spacetime curvature associated with the metric. Matter may create these fields, and they plus the matter may generate the metric, but they cannot act back directly on the matter. Matter responds only to the metric.

Thus the metric and the equations of motion for matter become the primary entities for calculating observable effects, and all that distinguishes one metric theory from another is the particular way in which matter and possibly other gravitational fields generate the metric.

The comparison of metric theories of gravity with each other and with experiment becomes particularly simple when one takes the slow-motion, weak-field limit. This approximation, known as the post-Newtonian limit, is sufficiently accurate to encompass most solar-system tests that can be performed in the foreseeable future. It turns out that, in this limit, the spacetime metric **g** predicted by nearly every metric theory of gravity has the same structure. It can be written as an expansion about the Minkowski metric $[\eta_{\mu\nu} = \mathrm{diag}(-1,1,1,1)]$ in terms

Parameter	What it measures relative to GR	Value in GR	Value in semi-conservative theories	Value in fully-conservative theories
γ	How much space-curvature produced by unit rest mass?	1	γ	γ
β	How much "nonlinearity" in the superposition law for gravity?	1	β	β
ξ	Preferred-location effects?	0	ξ	ξ
α_1	Preferred-frame effects?	0	α_1	0
α_2		0	α_2	0
α_3		0	0	0
α_3	Violation of conservation	0	0	0
ζ_1	of total momentum?	0	0	0
ζ_2		0	0	0
ζ_3		0	0	0
ζ_4		0	0	0

Table 2. The PPN Parameters and their significance (note that α_3 has been shown twice to indicate that it is a measure of two effects).

of dimensionless gravitational potentials of varying degrees of smallness. These potentials are constructed from the matter variables (Box 2) in imitation of the Newtonian gravitational potential

$$U(\mathbf{x},t) \equiv \int \rho(\mathbf{x}',t)|\mathbf{x}-\mathbf{x}'|^{-1}d^3x' . \qquad (22)$$

The "order of smallness" is determined according to the rules $U \sim v^2 \sim \Pi \sim p/\rho \sim O(2)$, $v^i \sim |d/dt|/|d/dx| \sim 0(1)$, and so on.

A consistent post-Newtonian limit requires determination of g_{00} correctly through O(4), g_{0i} through O(3), and g_{ij} through O(2) (for details see TEGP 4.1). The only way that one metric theory differs from another is in the numerical values of the coefficients that appear in front of the metric potentials. The parametrized

post-Newtonian (PPN) formalism inserts parameters in place of these coefficients, parameters whose values depend on the theory under study. In the current version of the PPN formalism, summarized in Box 2, ten parameters are used, chosen in such a manner that they measure or indicate general properties of metric theories of gravity (Table 2). The parameters γ and β are the usual Eddington-Robertson-Schiff parameters used to describe the "classical" tests of GR; ξ is non-zero in any theory of gravity that predicts preferred-location effects such as a galaxy-induced anisotropy in the local gravitational constant G_L (also called "Whitehead" effects); α_1, α_2, α_3 measure whether or not the theory predicts post-Newtonian preferred-frame effects; α_3, ζ_1, ζ_2, ζ_3, ζ_4 measure whether or not the theory predicts violations of global conservation laws for total momentum. In Table 2 we show the values these parameters take (i) in GR, (ii) in any theory of gravity that possesses conservation laws for total momentum, called "semi-conservative" (any theory that is based on an invariant action principle is semi-conservative), and (iii) in any theory that in addition possesses six global conservation laws for angular momentum, called "fully conservative" (such theories automatically predict no post-Newtonian preferred-frame effects). Semi-conservative theories have five free PPN parameters (γ, β, ξ, α_1, α_2) while fully conservative theories have three (γ, β , ξ).

The PPN formalism was pioneered by Kenneth Nordtvedt, who studied the post-Newtonian metric of a system of gravitating point masses, extending earlier work by Eddington, Robertson, and Schiff (TEGP 4.2). A general and unified version of the PPN formalism was developed by Will and Nordtvedt. The canonical version, with conventions altered to be more in accord with standard textbooks such as MTW, is discussed in detail in TEGP, Chapter 4. Other versions of the PPN formalism have been developed to deal with point masses with charge, fluid with anisotropic stresses, bodies with strong internal gravity, and post-post-Newtonian effects (TEGP 4.2, 14.2).

3.3 Competing Theories of Gravity

One of the important applications of the PPN formalism is the comparison and classification of alternative metric theories of gravity. The population of viable theories has fluctuated over the years as new effects and tests have been discovered, largely through the use of the PPN framework, which eliminated many theories

Box 2. The Parametrized Post-Newtonian Formalism

1. **Coordinate System:** The framework uses a nearly globally Lorentz coordinate system in which the coordinates are (t, x^1, x^2, x^3). Three-dimensional, Euclidean vector notation is used throughout. All coordinate arbitrariness ("gauge freedom") has been removed by specialization of the coordinates to the standard PPN gauge (TEGP 4.2). Units are chosen so that $G = c = 1$, where G is the physically measured Newtonian constant far from the solar system.
2. **Matter Variables:**
   - ρ = density of rest mass as measured in a local freely falling frame momentarily comoving with the gravitating matter.
   - $v^i = (dx^i/dt)$ = coordinate velocity of the matter.
   - w^i = coordinate velocity of PPN coordinate system relative to the mean rest-frame of the universe.
   - p = pressure as measured in a local freely falling frame momentarily comoving with the matter.
   - Π = internal energy per unit rest mass. It includes all forms of non-rest-mass, non-gravitational energy, *e.g.*, energy of compression and thermal energy.
3. **PPN Parameters:**

   $\gamma\,,\beta\,,\xi\,,\alpha_1\,,\alpha_2\,,\alpha_3\,,\zeta_1\,,\zeta_2\,,\zeta_3\,,\zeta_4\,.$
4. **Metric:**

$$
\begin{aligned}
g_{00} &= -1 + 2U - 2\beta U^2 - 2\xi\Phi_W + (2\gamma + 2 + \alpha_3 + \zeta_1 - 2\xi)\Phi_1 \\
&\quad +2(3\gamma - 2\beta + 1 + \zeta_2 + \xi)\Phi_2 + 2(1 + \zeta_3)\Phi_3 + 2(3\gamma + 3\zeta_4 - 2\xi)\Phi_4 \\
&\quad -(\zeta_1 - 2\xi)\mathcal{A} - (\alpha_1 - \alpha_2 - \alpha_3)w^2 U - \alpha_2 w^i w^j U_{ij} + (2\alpha_3 - \alpha_1)w^i V_i \\
g_{0i} &= -\frac{1}{2}(4\gamma + 3 + \alpha_1 - \alpha_2 + \zeta_1 - 2\xi)V_i - \frac{1}{2}(1 + \alpha_2 - \zeta_1 + 2\xi)W_i \\
&\quad -\frac{1}{2}(\alpha_1 - 2\alpha_2)w^i U - \alpha_2 w^j U_{ij} \\
g_{ij} &= (1 + 2\gamma U)\delta_{ij}.
\end{aligned}
$$

Box 2. (continued)

1. **Metric Potentials:**

$$
\begin{aligned}
U &= \int \frac{\rho'}{|\mathbf{x}-\mathbf{x}'|} d^3x', \qquad U_{ij} = \int \frac{\rho'(x-x')_i(x-x')_j}{|\mathbf{x}-\mathbf{x}'|^3} d^3x' \\
\Phi_W &= \int \frac{\rho'\rho''(\mathbf{x}-\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|^3} \cdot \left(\frac{\mathbf{x}'-\mathbf{x}''}{|\mathbf{x}-\mathbf{x}''|} - \frac{\mathbf{x}-\mathbf{x}''}{|\mathbf{x}'-\mathbf{x}''|} \right) d^3x' d^3x'' \\
\mathcal{A} &= \int \frac{\rho'[\mathbf{v}'\cdot(\mathbf{x}-\mathbf{x}')]^2}{|\mathbf{x}-\mathbf{x}'|^3} d^3x', \qquad \Phi_1 = \int \frac{\rho' v'^2}{|\mathbf{x}-\mathbf{x}'|} d^3x' \\
\Phi_2 &= \int \frac{\rho' U'}{|\mathbf{x}-\mathbf{x}'|} d^3x', \quad \Phi_3 = \int \frac{\rho'\Pi'}{|\mathbf{x}-\mathbf{x}'|} d^3x', \quad \Phi_4 = \int \frac{p'}{|\mathbf{x}-\mathbf{x}'|} d^3x' \\
V_i &= \int \frac{\rho' v_i'}{|\mathbf{x}-\mathbf{x}'|} d^3x', \qquad W_i = \int \frac{\rho'[\mathbf{v}'\cdot(\mathbf{x}-\mathbf{x}')](x-x')_i}{|\mathbf{x}-\mathbf{x}'|^3} d^3x'.
\end{aligned}
$$

2. **Stress-Energy Tensor** (perfect fluid)

$$
\begin{aligned}
T^{00} &= \rho(1+\Pi+v^2+2U) \\
T^{0i} &= \rho v^i(1+\Pi+v^2+2U+p/\rho) \\
T^{ij} &= \rho v^i v^j(1+\Pi+v^2+2U+p/\rho) + p\delta^{ij}(1-2\gamma U).
\end{aligned}
$$

3. **Equations of Motion**

- Stressed Matter, $T^{\mu\nu}{}_{;\nu} = 0$
- Test Bodies, $d^2x^\mu/d\lambda^2 + \Gamma^\mu{}_{\nu\lambda}(dx^\nu/d\lambda)(dx^\lambda/d\lambda) = 0$
- Maxwell's Equations, $F^{\mu\nu}{}_{;\nu} = 4\pi J^\mu \qquad F_{\mu\nu} = A_{\nu;\mu} - A_{\mu;\nu}$.

thought previously to be viable. The theory population has also fluctuated as new, potentially viable theories have been invented.

In these lectures, we shall focus on general relativity and the general class of scalar-tensor modifications of it, of which the Jordan-Fierz-Brans-Dicke theory (Brans-Dicke, for short) is the classic example. The reasons are several-fold:

- A full compendium of alternative theories is given in TEGP, Chapter 5.
- Many alternative metric theories developed during the 1970s and 1980s could be viewed as "straw-man" theories, invented to prove that such theories exist or to illustrate particular properties. Few of these could be regarded as well-motivated theories from the point of view, say, of field theory or particle physics. Examples are the vector-tensor theories studied by Will, Nordtvedt, and Hellings.
- A number of theories fall into the class of "prior-geometric" theories, with absolute elements such as a flat background metric in addition to the physical metric. Most of these theories predict "preferred-frame" effects, that have been tightly constrained by observations (see Sec. 3.6.2). An example is Rosen's bimetric theory.
- A large number of alternative theories of gravity predict gravitational-wave emission substantially different from that of general relativity, in strong disagreement with observations of the binary pulsar (see Sec. 4).
- Scalar-tensor modifications of GR have recently become very popular in cosmological model building and in unification schemes, such as string theory.

3.3.1 General Relativity

The metric $\mathbf{g}$ is the sole dynamical field and the theory contains no arbitrary functions or parameters, apart from the value of the Newtonian coupling constant G, which is measurable in laboratory experiments. Throughout these lectures, we ignore the cosmological constant λ. Although λ has significance for quantum field theory, quantum gravity, and cosmology, on the scale of the solar-system or of stellar systems, its effects are negligible, for values of λ corresponding to a cosmological closure density.

The field equations of GR are derivable from an invariant action principle

Theory	Arbitrary Functions or Constants	Cosmic Matching Parameters	PPN Parameters				
			γ	β	ξ	α_1	α_2
General Relativity	none	none	1	1	0	0	0
Scalar-Tensor							
Brans-Dicke	ω	ϕ_0	$\frac{(1+\omega)}{(2+\omega)}$	1	0	0	0
General	$A(\varphi)$, $V(\varphi)$	φ_0	$\frac{(1+\omega)}{(2+\omega)}$	$1+\Lambda$	0	0	0
Rosen's Bimetric	none	c_0, c_1	1	1	0	0	$\frac{c_0}{c_1}-1$

Table 3. Metric Theories and Their PPN Parameter Values ($\alpha_3 = \zeta_i = 0$ for all cases).

$\delta I = 0$, where

$$I = (16\pi G)^{-1}\int R(-g)^{1/2}d^4x + I_m(\psi_m, g_{\mu\nu})\,, \tag{23}$$

where R is the Ricci scalar, and I_m is the matter action, which depends on matter fields ψ_m universally coupled to the metric $\mathbf{g}$. By varying the action with respect to $g_{\mu\nu}$, we obtain the field equations

$$G_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = 8\pi G T_{\mu\nu}\,, \tag{24}$$

where $T_{\mu\nu}$ is the matter energy-momentum tensor. General covariance of the matter action implies the equations of motion $T^{\mu\nu}{}_{;\nu} = 0$; varying I_m with respect to ψ_M yields the matter field equations. By virtue of the *absence* of prior-geometric elements, the equations of motion are also a consequence of the field equations via the Bianchi identities $G^{\mu\nu}{}_{;\nu} = 0$.

The general procedure for deriving the post-Newtonian limit is spelled out in TEGP 5.1, and is described in detail for GR in TEGP 5.2. The PPN parameter values are listed in Table 3.

3.3.2 Scalar-Tensor Theories

These theories contain the metric $\mathbf{g}$, a scalar field φ, a potential function $V(\varphi)$, and a coupling function $A(\varphi)$ (generalizations to more than one scalar field have also been carried out[20]). For some purposes, the action is conveniently written in a non-metric representation, sometimes denoted the "Einstein frame," in which the gravitational action looks exactly like that of GR:

$$\tilde{I} = (16\pi G)^{-1}\int[\tilde{R} - 2\tilde{g}^{\mu\nu}\partial_\mu\varphi\partial_\nu\varphi - V(\varphi)](-\tilde{g})^{1/2}d^4x + I_m(\psi_m, A^2(\varphi)\tilde{g}_{\mu\nu})\,, \tag{25}$$

where $\tilde{R} \equiv \tilde{g}^{\mu\nu}\tilde{R}_{\mu\nu}$ is the Ricci scalar of the "Einstein" metric $\tilde{g}_{\mu\nu}$. [Apart from the scalar potential term $V(\varphi)$, this corresponds to Equation (20) with $\tilde{G}(\varphi) \equiv (4\pi G)^{-1}$, $U(\varphi) \equiv 1$, and $\tilde{M}(\varphi) \propto A(\varphi)$.] This representation is a "non-metric" one because the matter fields ψ_m couple to a combination of φ and $\tilde{g}_{\mu\nu}$. Despite appearances, however, it is a metric theory, because it can be put into a metric representation by identifying the "physical metric"

$$g_{\mu\nu} \equiv A^2(\varphi)\tilde{g}_{\mu\nu}\,. \tag{26}$$

The action can then be rewritten in the metric form

$$I = (16\pi G)^{-1}\int[\phi R - \phi^{-1}\omega(\phi)g^{\mu\nu}\partial_\mu\phi\partial_\nu\phi - \phi^2 V](-g)^{1/2}d^4x + I_m(\psi_m, g_{\mu\nu})\,, \tag{27}$$

where

$$\begin{aligned} \phi &\equiv A(\varphi)^{-2}\,, \\ 3+2\omega(\phi) &\equiv \alpha(\varphi)^{-2}\,, \\ \alpha(\varphi) &\equiv d(\ln A(\varphi))/d\varphi\,. \end{aligned} \tag{28}$$

The Einstein frame is useful for discussing general characteristics of such theories, and for some cosmological applications, while the metric representation is most useful for calculating observable effects. The field equations, post-Newtonian limit and PPN parameters are discussed in TEGP 5.3, and the values of the PPN parameters are listed in Table 3.

The parameters that enter the post-Newtonian limit are

$$\omega \equiv \omega(\phi_0), \qquad \Lambda \equiv [(d\omega/d\phi)(3+2\omega)^{-2}(4+2\omega)^{-1}]_{\phi_0}\,, \tag{29}$$

where ϕ_0 is the value of ϕ today far from the system being studied, as determined by appropriate cosmological boundary conditions. The following formula is also useful: $1/(2+\omega) = 2\alpha_0^2/(1+\alpha_0^2)$. In Brans-Dicke theory [$\omega(\phi) = $ constant], the larger the value of ω, the smaller the effects of the scalar field, and in the

limit $\omega \to \infty$ ($\alpha_0 \to 0$), the theory becomes indistinguishable from GR in all its predictions. In more general theories, the function $\omega(\phi)$ could have the property that, at the present epoch, and in weak-field situations, the value of the scalar field ϕ_0 is such that ω is very large and Λ is very small (theory almost identical to GR today), but that for past or future values of ϕ, or in strong-field regions such as the interiors of neutron stars, ω and Λ could take on values that would lead to significant differences from GR. Indeed, Damour and Nordtvedt have shown that in such general scalar-tensor theories, GR is a natural "attractor": regardless of how different the theory may be from GR in the early universe (apart from special cases), cosmological evolution naturally drives the fields toward small values of the function α, thence to large ω. Estimates of the expected relic deviations from GR today in such theories depend on the cosmological model, but range from 10^{-5} to a few times 10^{-7} for $1-\gamma$ (Ref. 21).

Scalar fields coupled to gravity or matter are also ubiquitous in particle-physics-inspired models of unification, such as string theory. In some models, the coupling to matter may lead to violations of WEP, which are tested by Eötvös-type experiments. In many models the scalar field is massive; if the Compton wavelength is of macroscopic scale, its effects are those of a "fifth force." Only if the theory can be cast as a metric theory with a scalar field of infinite range or of range long compared to the scale of the system in question (solar system) can the PPN framework be strictly applied. If the mass of the scalar field is sufficiently large that its range is microscopic, then, on solar-system scales, the scalar field is suppressed, and the theory is essentially equivalent to general relativity. This is the case, for example, in the "oscillating-G" models of Accetta, Steinhardt, and Will (see Ref. 22), in which the potential function $V(\varphi)$ contains both quadratic (mass) and quartic (self-interaction) terms, causing the scalar field to oscillate (the initial amplitude of oscillation is provided by an inflationary epoch); high-frequency oscillations in the "effective" Newtonian constant $G_{\mathit{eff}} \equiv G/\phi = GA(\varphi)^2$ then result. The energy density in the oscillating scalar field can be enough to provide a cosmological closure density without resorting to dark matter, yet the value of ω today is so large that the theory's local predictions are experimentally indistinguishable from GR. In other models, explored by Damour and Esposito-Farèse,[23] non-linear scalar-field couplings can lead to "spontaneous scalarization" inside strong-field objects such as neutron stars, leading to large deviations from GR, even in the limit of very large ω.

3.4 Tests of the Parameter γ

With the PPN formalism in hand, we are now ready to confront gravitation theories with the results of solar-system experiments. In this section we focus on tests of the parameter γ, consisting of the deflection of light and the time delay of light.

3.4.1 The Deflection of Light

A light ray (or photon) which passes the Sun at a distance d is deflected by an angle

$$\delta\theta = \frac{1}{2}(1+\gamma)(4m_\odot/d)[(1+\cos\Phi)/2] \tag{30}$$

(TEGP 7.1), where $m_\odot$ is the mass of the Sun and Φ is the angle between the Earth-Sun line and the incoming direction of the photon (Figure 4). For a grazing ray, $d \approx d_\odot$, $\Phi \approx 0$, and

$$\delta\theta \approx \frac{1}{2}(1+\gamma)1.''75\,, \tag{31}$$

independent of the frequency of light. Another more useful expression gives the change in the relative angular separation between an observed source of light and a nearby reference source as both rays pass near the Sun:

$$\delta\theta = \frac{1}{2}(1+\gamma)\left[-\frac{4m_\odot}{d}\cos\chi + \frac{4m_\odot}{d_r}\left(\frac{1+\cos\Phi_r}{2}\right)\right], \tag{32}$$

where d and d_r are the distances of closest approach of the source and reference rays respectively, Φ_r is the angular separation between the Sun and the reference source, and χ is the angle between the Sun-source and the Sun-reference directions, projected on the plane of the sky (Figure 4). Thus, for example, the relative angular separation between the two sources may vary if the line of sight of one of them passes near the Sun ($d \sim R_\odot$, $d_r \gg d$, χ varying with time).

It is interesting to note that the classic derivations of the deflection of light that use only the principle of equivalence or the corpuscular theory of light yield only the "1/2" part of the coefficient in front of the expression in Equation 30. But the result of these calculations is the deflection of light relative to local straight lines, as defined for example by rigid rods; however, because of space curvature around the Sun, determined by the PPN parameter γ, local straight lines are bent relative to asymptotic straight lines far from the Sun by just enough to yield the remaining factor "$\gamma/2$." The first factor "1/2" holds in any metric theory, the

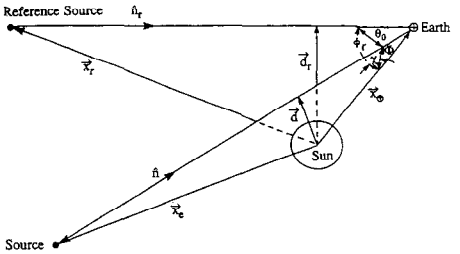


Fig. 4. Geometry of light deflection measurements.

second "$\gamma/2$" varies from theory to theory. Thus, calculations that purport to derive the full deflection using the equivalence principle alone are incorrect.

The prediction of the full bending of light by the Sun was one of the great successes of Einstein's GR. Eddington's confirmation of the bending of optical starlight observed during a solar eclipse in the first days following World War I helped make Einstein famous. However, the experiments of Eddington and his co-workers had only 30 percent accuracy, and succeeding experiments were not much better: the results were scattered between one half and twice the Einstein value (Figure 5), and the accuracies were low.

However, the development of very-long-baseline radio interferometry (VLBI) produced greatly improved determinations of the deflection of light. These techniques now have the capability of measuring angular separations and changes in angles as small as 100 microarcseconds. Early measurements took advantage of a series of heavenly coincidences: each year, groups of strong quasistellar radio sources pass very close to the Sun (as seen from the Earth), including the group 3C273, 3C279, and 3C48, and the group 0111+02, 0119+11, and 0116+08. As the Earth moves in its orbit, changing the lines of sight of the quasars relative to the Sun, the angular separation $\delta\theta$ between pairs of quasars varies (Equation 32). The time variation in the quantities d, d_r, χ, and Φ_r in Equation 32 is determined using an accurate ephemeris for the Earth and initial directions for the quasars, and the resulting prediction for $\delta\theta$ as a function of time is used as a basis for a least-squares fit of the measured $\delta\theta$, with one of the fitted parameters being the coefficient $\frac{1}{2}(1+\gamma)$. A number of measurements of this kind over the period

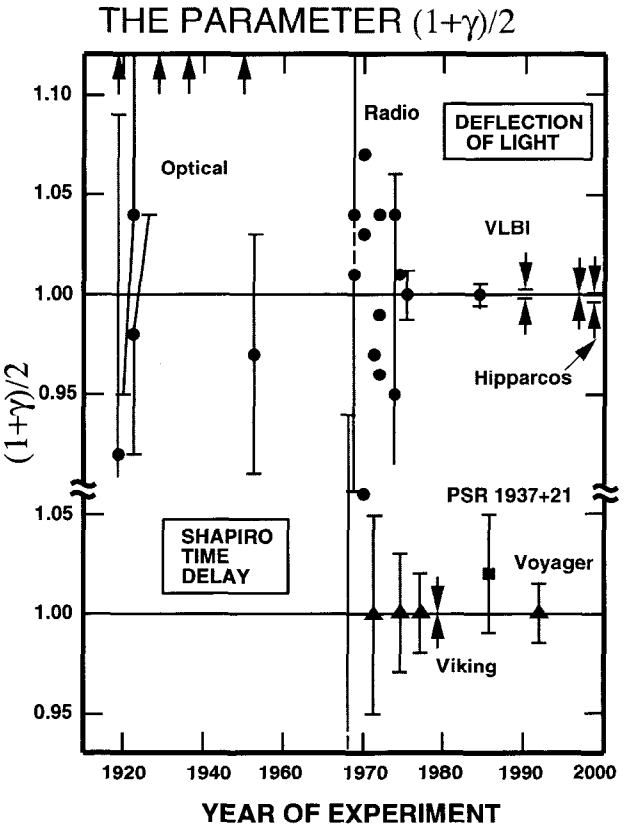


Fig. 5. Measurements of the coefficient $(1+\gamma)/2$ from light deflection and time delay measurements. General relativity value is unity. Arrows denote anomalously large values from 1929 and 1936 eclipse expeditions. Shapiro time-delay measurements using Viking spacecraft and VLBI light deflection measurements yielded agreement with GR to 0.1 percent. Hipparcos denotes the optical astrometry satellite.

1969–1975 yielded an accurate determination of the coefficient $\frac{1}{2}(1+\gamma)$ which has the value unity in GR. Their results are shown in Figure 5.

A recent series of transcontinental and intercontinental VLBI quasar and radio galaxy observations made primarily to monitor the Earth's rotation ("VLBI " in Figure 5) was sensitive to the deflection of light over almost the entire celestial sphere (at 90° from the Sun, the deflection is still 4 milliarcseconds). A recent analysis of VLBI data yielded $(1+\gamma)/2 = 0.99997 \pm 0.00016$ (Ref. 24). Analysis of observations made by the Hipparcos optical astrometry satellite yielded a test at the level of 0.3 percent.[25] A VLBI measurement of the deflection of light by Jupiter was reported; the predicted deflection of about 300 microarcseconds was seen with about 50 percent accuracy.

3.4.2 The Time Delay of Light

A radar signal sent across the solar system past the Sun to a planet or satellite and returned to the Earth suffers an additional non-Newtonian delay in its round-trip travel time, given by (see Figure 4)

$$\delta t = 2(1+\gamma)m_\odot \ln[(r_\oplus + \mathbf{x}_\oplus \cdot \mathbf{n})(r_e - \mathbf{x}_e \cdot \mathbf{n})/d^2] \qquad (33)$$

(TEGP 7.2). For a ray which passes close to the Sun,

$$\delta t \approx \frac{1}{2}(1+\gamma)[240 - 20\ln(d^2/r)]\ \mu\mathrm{s}\,, \qquad (34)$$

where d is the distance of closest approach of the ray in solar radii, and r is the distance of the planet or satellite from the Sun, in astronomical units.

In the two decades following Irwin Shapiro's 1964 discovery of this effect as a theoretical consequence of general relativity, several high-precision measurements were made using radar ranging to targets passing through superior conjunction. Since one does not have access to a "Newtonian" signal against which to compare the round-trip travel time of the observed signal, it is necessary to do a differential measurement of the variations in round-trip travel times as the target passes through superior conjunction, and to look for the logarithmic behavior of Equation 34. In order to do this accurately however, one must take into account the variations in round-trip travel time due to the orbital motion of the target relative to the Earth. This is done by using radar-ranging (and possibly other) data on the target taken when it is far from superior conjunction (*i.e.*, when the time-delay term is negligible) to determine an accurate ephemeris for the target, using the ephemeris to predict the PPN coordinate trajectory $\mathbf{x}_e(t)$ near superior conjunction, then combining that trajectory with the trajectory of the Earth $\mathbf{x}_\oplus(t)$ to determine the Newtonian round-trip time and the logarithmic term in Equation 34. The resulting predicted round-trip travel times in terms of the unknown coefficient $\frac{1}{2}(1+\gamma)$ are then fit to the measured travel times using the method of least-squares, and an estimate obtained for $\frac{1}{2}(1+\gamma)$.

The targets employed included planets, such as Mercury or Venus, used as passive reflectors of the radar signals ("passive radar"), and artificial satellites, such as Mariners 6 and 7, and Voyager 2, and the Viking Mars landers and orbiters, used as active retransmitters of the radar signals ("active radar").

The results for the coefficient $\frac{1}{2}(1+\gamma)$ of all radar time-delay measurements performed to date (including a measurement of the one-way time delay of signals from the millisecond pulsar PSR 1937+21) are shown in Figure 5 (see TEGP 7.2 for discussion and references). The Viking experiment resulted in a 0.1 percent measurement.

From the results of light-deflection experiments, we can conclude that the coefficient $\frac{1}{2}(1+\gamma)$ must be within at most 0.016 percent of unity. Scalar-tensor theories must have $\omega > 3300$ to be compatible with this constraint.

3.5 The Perihelion Shift of Mercury

The explanation of the anomalous perihelion shift of Mercury's orbit was another of the triumphs of GR. This had been an unsolved problem in celestial mechanics for over half a century, since the announcement by Le Verrier in 1859 that, after the perturbing effects of the planets on Mercury's orbit had been accounted for, and after the effect of the precession of the equinoxes on the astronomical coordinate system had been subtracted, there remained in the data an unexplained advance in the perihelion of Mercury. The modern value for this discrepancy is 43 arcseconds per century. A number of *ad hoc* proposals were made in an attempt to account for this excess, including, among others, the existence of new planet Vulcan near the Sun, a ring of planetoids, a solar quadrupole moment, and a deviation from the inverse-square law of gravitation but none was successful. General relativity accounted for the anomalous shift in a natural way without disturbing the agreement with other planetary observations.

Parameter	Effect	Limit	Remarks
$\gamma - 1$	time delay	2×10^{-3}	Viking ranging
	light deflection	3×10^{-4}	VLBI
$\beta - 1$	perihelion shift	3×10^{-3}	$J_2 = 10^{-7}$ from helioseismology
	Nordtvedt effect	6×10^{-4}	$\eta = 4\beta - \gamma - 3$ assumed
ξ	Earth tides	10^{-3}	gravimeter data
α_1	orbital polarization	4×10^{-4}	Lunar laser ranging
		2×10^{-4}	PSR J2317+1439
α_2	spin precession	4×10^{-7}	solar alignment with ecliptic
α_3	pulsar acceleration	2×10^{-20}	pulsar $\dot{P}$ statistics
η^1	Nordtvedt effect	10^{-3}	lunar laser ranging
ζ_1	–	2×10^{-2}	combined PPN bounds
ζ_2	binary acceleration	4×10^{-5}	$\ddot{P}_p$ for PSR 1913+16
ζ_3	Newton's 3rd law	10^{-8}	Lunar acceleration
ζ_4	–	–	not independent

[1]Here $\eta = 4\beta - \gamma - 3 - 10\xi/3 - \alpha_1 - 2\alpha_2/3 - 2\zeta_1/3 - \zeta_2/3$.

Table 4. Current Limits on the PPN Parameters.

The predicted advance, $\Delta\tilde{\omega}$, per orbit, including both relativistic PPN contributions and the Newtonian contribution resulting from a possible solar quadrupole moment, is given by

$$\Delta\tilde{\omega} = (6\pi m/p)[\frac{1}{3}(2+2\gamma-\beta)+\frac{1}{6}\,(2\alpha_1-\alpha_2+\alpha_3+2\zeta_2)\mu/m+J_2(R^2/2mp)]\,, \quad (35)$$

where $m \equiv m_1 + m_2$ and $\mu \equiv m_1 m_2/m$ are the total mass and reduced mass of the two-body system respectively; $p \equiv a(1-e^2)$ is the semi-latus rectum of the orbit, with a the semi-major axis and e the eccentricity; R is the mean radius of the oblate body; and J_2 is a dimensionless measure of its quadrupole moment, given by $J_2 = (C - A)/m_1 R^2$, where C and A are the moments of inertia about the body's rotation and equatorial axes, respectively (for details of the derivation see TEGP 7.3). We have ignored preferred-frame and galaxy-induced contributions to $\Delta\tilde{\omega}$; these are discussed in TEGP 8.3.

The first term in Equation 35 is the classical relativistic perihelion shift, which depends upon the PPN parameters γ and β. The second term depends upon the ratio of the masses of the two bodies; it is zero in any fully conservative theory of gravity ($\alpha_1 \equiv \alpha_2 \equiv \alpha_3 \equiv \zeta_2 \equiv 0$); it is also negligible for Mercury, since $\mu/m \approx m_{\rm Merc}/m_\odot \approx 2\times10^{-7}$. We shall drop this term henceforth. The third term depends upon the solar quadrupole moment J_2. For a Sun that rotates uniformly with its observed surface angular velocity, so that the quadrupole moment is produced by centrifugal flattening, one may estimate J_2 to be $\sim 1 \times 10^{-7}$. This actually agrees reasonably well with values inferred from rotating solar models that are in accord with observations of the normal modes of solar oscillations (helioseismology). Substituting standard orbital elements and physical constants for Mercury and the Sun, we obtain the rate of perihelion shift $\dot{\tilde{\omega}}$, in seconds of arc per century,

$$\dot{\tilde{\omega}} = 42.''98\left[\frac{1}{3}(2+2\gamma-\beta)+3\times10^{-4}(J_2/10^{-7})\right]. \quad (36)$$

Now, the measured perihelion shift of Mercury is known accurately: after the perturbing effects of the other planets have been accounted for, the excess shift is known to about 0.1 percent from radar observations of Mercury since 1966. The solar oblateness effect is smaller than the observational error, so we obtain the PPN bound $|2\gamma - \beta - 1| < 3 \times 10^{-3}$.

3.6 Tests of the Strong Equivalence Principle

The next class of solar-system experiments that test relativistic gravitational effects may be called tests of the strong equivalence principle (SEP). In Sec. 3.1.2, we pointed out that many metric theories of gravity (perhaps all except GR) can be expected to violate one or more aspects of SEP. Among the testable violations of SEP are a violation of the weak equivalence principle for gravitating bodies that leads to perturbations in the Earth-Moon orbit, preferred-location and preferred-frame effects in the locally measured gravitational constant that could produce observable geophysical effects, and possible variations in the gravitational constant over cosmological timescales.

3.6.1 The Nordtvedt Effect and the Lunar Eötvös Experiment

In a pioneering calculation using his early form of the PPN formalism, Nordtvedt showed that many metric theories of gravity predict that massive bodies violate the weak equivalence principle—that is, fall with different accelerations depending on their gravitational self-energy. For a spherically symmetric body, the acceleration from rest in an external gravitational potential U has the form

$$\begin{aligned} \mathbf{a} &= (m_p/m)\nabla U\,, \\ m_p/m &= 1-\eta(E_g/m)\,, \\ \eta &= 4\beta-\gamma-3-\frac{10}{3}\xi-\alpha_1+\frac{2}{3}\alpha_2-\frac{2}{3}\zeta_1-\frac{1}{3}\zeta_2\,, \end{aligned} \tag{37}$$

where E_g is the negative of the gravitational self-energy of the body ($E_g > 0$). This violation of the massive-body equivalence principle is known as the "Nordtvedt effect." The effect is absent in GR ($\eta = 0$) but present in scalar-tensor theory [$\eta = 1/(2+\omega) + 4\Lambda$]. The existence of the Nordtvedt effect does not violate the results of laboratory Eötvös experiments, since for laboratory-sized objects, $E_g/m \leq 10^{-27}$, far below the sensitivity of current or future experiments. However, for astronomical bodies, E_g/m may be significant (10^{-5} for the Sun, 10^{-8} for Jupiter, 4.6×10^{-10} for the Earth, 0.2×10^{-10} for the Moon). If the Nordtvedt effect is present ($\eta \neq 0$) then the Earth should fall toward the Sun with a slightly different acceleration than the Moon. This perturbation in the Earth-Moon orbit leads to a polarization of the orbit that is directed toward the Sun as it moves around the Earth-Moon system, as seen from Earth. This polarization represents a perturbation in the Earth-Moon distance of the form

$$\delta r = 13.1\eta\cos(\omega_0-\omega_s)t \quad \mathrm{m}\,, \tag{38}$$

where ω_0 and ω_s are the angular frequencies of the orbits of the Moon and Sun around the Earth (see TEGP 8.1, for detailed derivations and references; for improved calculations of the numerical coefficient, see Refs. 26 and 27).

Since August 1969, when the first successful acquisition was made of a laser signal reflected from the Apollo 11 retroreflector on the Moon, the lunar laser-ranging experiment (LURE) has made regular measurements of the round-trip travel times of laser pulses between a network of observatories and the lunar retroreflectors, with accuracies that are approaching 50 ps (1 cm). These measurements are fit using the method of least-squares to a theoretical model for the lunar motion that takes into account perturbations due to the Sun and the other planets, tidal interactions, and post-Newtonian gravitational effects. The predicted round-trip travel times between retroreflector and telescope also take into account the librations of the Moon, the orientation of the Earth, the location of the observatory, and atmospheric effects on the signal propagation. The "Nordtvedt" parameter η along with several other important parameters of the model are then estimated in the least-squares method.

Several independent analyses of the data found no evidence, within experimental uncertainty, for the Nordtvedt effect (for recent results see Ref. 28). Their results can be summarized by the bound $|\eta| < 0.001$. These results represent a limit on a possible violation of WEP for massive bodies of five parts in 10^{13} (compare Figure 1). For Brans-Dicke theory, these results force a lower limit on the coupling constant ω of 1000. Note that, at this level of precision, one cannot regard the results of lunar laser ranging as a "clean" test of SEP because the precision exceeds that of laboratory tests of WEP. Because the chemical compositions of the Earth and Moon differ in the relative fractions of iron and silicates, an extrapolation from laboratory Eötvös-type experiments to the Earth-Moon systems using various non-metric couplings to matter yields bounds on violations of WEP only of the order of 2×10^{-12}. Thus if lunar laser ranging is to test SEP at higher accuracy, tests of WEP must keep pace; to this end, a proposed satellite test of the equivalence principle (Sec. 2.3) will be an important advance.

In GR, the Nordtvedt effect vanishes; at the level of several centimeters and below, a number of non-null general relativistic effects should be present.[26]

3.6.2 Preferred-Frame and Preferred-Location Effects

Some theories of gravity violate SEP by predicting that the outcomes of local gravitational experiments may depend on the velocity of the laboratory relative to the mean rest frame of the universe (preferred-frame effects) or on the location of the laboratory relative to a nearby gravitating body (preferred-location effects). In the post-Newtonian limit, preferred-frame effects are governed by the values of the PPN parameters α_1, α_2, and α_3, and some preferred-location effects are governed by ξ (see Table 2).

The most important such effects are variations and anisotropies in the locally-measured value of the gravitational constant, which lead to anomalous Earth tides and variations in the Earth's rotation rate; anomalous contributions to the orbital dynamics of Mercury and Mars; self-accelerations of pulsars, and anomalous torques on the Sun that would cause its spin axis to be randomly oriented relative to the ecliptic [see TEGP 8.2, 8.3, 9.3, and 14.3(c)]. An improved bound on α_3 of 2×10^{-20} from the period derivatives of 20 millisecond pulsars was reported in Ref. 29; an improved bound on α_1 was achieved using observations of the circular binary orbit of the pulsar J2317+1439 (Ref. 30). Negative searches for these effects have produced strong constraints on the PPN parameters (Table 4).

3.6.3 Constancy of the Newtonian Gravitational Constant

Most theories of gravity that violate SEP predict that the locally measured Newtonian gravitational constant may vary with time as the universe evolves. For the scalar-tensor theories listed in Table 3, the predictions for $\dot{G}/G$ can be written in terms of time derivatives of the asymptotic scalar field. Where G does change with cosmic evolution, its rate of variation should be of the order of the expansion rate of the universe, *i.e.*, $\dot{G}/G \sim H_0$, where H_0 is the Hubble expansion parameter and is given by $H_0 = 100h$ km s^{-1} Mpc^{-1} $= h \times 10^{-10}$ yr^{-1}, where current observations of the expansion of the universe give $\frac{1}{2} < h < 1$.

Several observational constraints can be placed on $\dot{G}/G$ using methods that include studies of the evolution of the Sun, observations of lunar occultations (including analyses of ancient eclipse data), lunar laser-ranging measurements, planetary radar-ranging measurements, and pulsar timing data. Laboratory experiments may one day lead to interesting limits [for review and references to past work see TEGP 8.4 and 14.3(c)]. Recent results are shown in Table 5.

Method	$\dot{G}/G(10^{-12}$ yr$^{-1})$
Lunar Laser Ranging	0 ± 8
Viking Radar	2 ± 4
	-2 ± 10
Binary Pulsar[1]	11 ± 11
Pulsar PSR 0655+64[1]	< 55

[1] Bounds dependent upon theory of gravity in strong-field regime and on neutron star equation of state.

Table 5. Constancy of the Gravitational Constant.

The best limits on $\dot{G}/G$ still come from ranging measurements to the Viking landers and lunar laser ranging measurements.[28] It has been suggested that radar observations of a Mercury orbiter over a two-year mission (30 cm accuracy in range) could yield $\Delta(\dot{G}/G) \sim 10^{-14}$ yr^{-1}.

Although bounds on $\dot{G}/G$ using solar-system measurements can be obtained in a phenomenological manner through the simple expedient of replacing G by $G_0 + \dot{G}_0(t - t_0)$ in Newton's equations of motion, the same does not hold true for pulsar and binary pulsar timing measurements. The reason is that, in theories of gravity that violate SEP, such as scalar-tensor theories, the "mass" and moment of inertia of a gravitationally bound body may vary with variation in G. Because neutron stars are highly relativistic, the fractional variation in these quantities can be comparable to $\Delta G/G$, the precise variation depending both on the equation of state of neutron star matter and on the theory of gravity in the strong-field regime. The variation in the moment of inertia affects the spin rate of the pulsar, while the variation in the mass can affect the orbital period in a manner that can add to or subtract from the direct effect of a variation in G, given by $\dot{P}_b/P_b = -\frac{1}{2}\dot{G}/G$. Thus, the bounds quoted in Table 5 for the binary pulsar PSR 1913+16 and the pulsar PSR 0655+64 are theory-dependent and must be treated as merely suggestive.

3.7 Other Tests of Post-Newtonian Gravity

3.7.1 Tests of Post-Newtonian Conservation Laws

Of the five "conservation law" PPN parameters ζ_1, ζ_2, ζ_3, ζ_4, and α_3, only three, ζ_2, ζ_3, and α_3, have been constrained directly with any precision; ζ_1 is constrained indirectly through its appearance in the Nordtvedt effect parameter η, Equation 37. There is strong theoretical evidence that ζ_4, which is related to the gravity generated by fluid pressure, is not really an independent parameter—in any reasonable theory of gravity there should be a connection between the gravity produced by kinetic energy (ρv^2), internal energy ($\rho\Pi$), and pressure (p). From such considerations, there follows[31] the additional theoretical constraint

$$6\zeta_4 = 3\alpha_3 + 2\zeta_1 - 3\zeta_3 \,. \tag{39}$$

A non-zero value for any of these parameters would result in a violation of conservation of momentum, or of Newton's third law in gravitating systems. An alternative statement of Newton's third law for gravitating systems is that the "active gravitational mass," that is, the mass that determines the gravitational potential exhibited by a body, should equal the "passive gravitational mass," the mass that determines the force on a body in a gravitational field. Such an equality guarantees the equality of action and reaction and of conservation of momentum, at least in the Newtonian limit.

A classic test of Newton's third law for gravitating systems was carried out in 1968 by Kreuzer, in which the gravitational attraction of fluorine and bromine were compared to a precision of five parts in 10^5.

A remarkable planetary test was reported by Bartlett and van Buren. They noted that current understanding of the structure of the Moon involves an iron-rich, aluminum-poor mantle whose center of mass is offset about 10 km from the center of mass of an aluminum-rich, iron-poor crust. The direction of offset is toward the Earth, about 14° to the east of the Earth-Moon line. Such a model accounts for the basaltic maria which face the Earth, and the aluminum-rich highlands on the Moon's far side, and for a 2 km offset between the observed center of mass and center of figure for the Moon. Because of this asymmetry, a violation of Newton's third law for aluminum and iron would result in a momentum non-conserving self-force on the Moon, whose component along the orbital direction would contribute to the secular acceleration of the lunar orbit. Improved knowledge of the lunar orbit through lunar laser ranging, and a better understanding of tidal effects in the Earth-Moon system (which also contribute to the secular acceleration) through satellite data, severely limit any anomalous secular acceleration, with the resulting limit

$$\left|\frac{(m_A/m_P)_{\rm Al} - (m_A/m_P)_{\rm Fe}}{(m_A/m_P)_{\rm Fe}}\right| < 4 \times 10^{-12} \,. \tag{40}$$

According to the PPN formalism, in a theory of gravity that violates conservation of momentum, but that obeys the constraint of Equation 39, the electrostatic binding energy E_e of an atomic nucleus could make a contribution to the ratio of active to passive mass of the form

$$m_A = m_P + \frac{1}{2}\zeta_3 E_e/c^2 \,. \tag{41}$$

The resulting limits on ζ_3 from the lunar experiment is $\zeta_3 < 1 \times 10^{-8}$ [TEGP 9.2, 14.3(d)].

Another consequence of a violation of conservation of momentum is a self-acceleration of the center of mass of a binary stellar system, given by

$$\mathbf{a}_{\rm CM} = \frac{1}{2}(\zeta_2 + \alpha_3)\frac{m}{a^2}\frac{\mu}{a}\frac{\delta m}{m}\frac{e}{(1-e^2)^{3/2}}\mathbf{n}_P \,, \tag{42}$$

where $\delta m = m_1 - m_2$, a is the semi-major axis, and $\mathbf{n}_P$ is a unit vector directed from the center of mass to the point of periastron of m_1 (TEGP 9.3). A consequence of this acceleration would be non-vanishing values for d^2P/dt^2, where P denotes the period of any intrinsic process in the system (orbit, spectra, pulsar periods). The observed upper limit on d^2P_p/dt^2 of the binary pulsar PSR 1913+16 places a strong constraint on such an effect, resulting in the bound $|\alpha_3 + \zeta_2| < 4 \times 10^{-5}$. Since α_3 has already been constrained to be much less than this (Table 4), we obtain a strong bound on ζ_2 alone.[32]

3.7.2 Geodetic Precession

A gyroscope moving through curved spacetime suffers a precession of its axis given by

$$d\mathbf{S}/d\tau = \mathbf{\Omega}_G \times \mathbf{S} \,, \qquad \mathbf{\Omega}_G = (\gamma + \frac{1}{2})\mathbf{v} \times \nabla U \,, \tag{43}$$

where $\mathbf{v}$ is the velocity of the gyroscope, and U is the Newtonian gravitational potential of the source (TEGP 9.1). The Earth-Moon system can be considered

as a "gyroscope," with its axis perpendicular to the orbital plane. The predicted precession is about two arcseconds per century, an effect first calculated by de Sitter. This effect has been measured to about 0.7 percent using lunar laser ranging data.[28]

For a gyroscope orbiting the Earth, the precession is about eight arcseconds per year. The Stanford Gyroscope Experiment has as one of its goals the measurement of this effect to 5×10^{-5} (see below).

3.7.3 Search for Gravitomagnetism

According to GR, moving or rotating matter should produce a contribution to the gravitational field that is the analogue of the magnetic field of a moving charge or a magnetic dipole. Although gravitomagnetism plays a role in a variety of measured relativistic effects, it has not been seen to date, isolated from other post-Newtonian effects. The Relativity Gyroscope Experiment (Gravity Probe B or GP-B) at Stanford University, in collaboration with NASA and Lockheed-Martin Corporation, is in the advanced stage of developing a space mission to detect this phenomenon directly.[33] A set of four superconducting-niobium-coated, spherical quartz gyroscopes will be flown in a low polar Earth orbit, and the precession of the gyroscopes relative to the distant stars will be measured. In the PPN formalism, the predicted effect of gravitomagnetism is a precession (also known as the Lense-Thirring effect, or the dragging of inertial frames), given by

$$d\mathbf{S}/d\tau = \mathbf{\Omega}_{\rm LT} \times \mathbf{S}\,, \qquad \mathbf{\Omega}_{\rm LT} = -\frac{1}{2}(1+\gamma+\frac{1}{4}\alpha_1)[\mathbf{J} - 3\mathbf{n}(\mathbf{n}\cdot\mathbf{J})]/r^3\,, \qquad (44)$$

where $\mathbf{J}$ is the angular momentum of the Earth, $\mathbf{n}$ is a unit radial vector, and r is the distance from the center of the Earth (TEGP 9.1). For a polar orbit at about 650 km altitude, this leads to a secular angular precession at a rate $\frac{1}{2}(1+\gamma+\frac{1}{4}\alpha_1)42 \times 10^{-3}$ arcsec/yr. The accuracy goal of the experiment is about 0.5 milliarcseconds per year. The science instrument package and the spacecraft are in the final phases of construction, with launch scheduled for early 2000.

Another proposal to look for an effect of gravitomagnetism is to measure the relative precession of the line of nodes of a pair of laser-ranged geodynamics satellites (LAGEOS), ideally with supplementary inclination angles; the inclinations must be supplementary in order to cancel the dominant nodal precession caused by the Earth's Newtonian gravitational multipole moments. Unfortunately, the two existing LAGEOS satellites are not in appropriately inclined orbits, and no plans exist at present to launch a third satellite in a supplementary orbit. Nevertheless, by combing nodal precession data from LAGEOS I and II with perigee advance data from the slightly eccentric orbit of LAGEOS II, Ciufolini *et al.* reported a partial cancellation of multipole effects, and a resulting 20 percent confirmation of GR.[34]

3.7.4 Improved PPN Parameter Values

A number of advanced space missions have been proposed in which spacecraft orbiters or landers and improved tracking capabilities could lead to significant improvements in values of the PPN parameters, of J_2 of the Sun, and of $\dot{G}/G$. For example, a Mercury orbiter, in a two-year experiment, with 3 cm range capability, could yield improvements in the perihelion shift to a part in 10^4, in γ to 4×10^{-5}, in $\dot{G}/G$ to 10^{-14} yr^{-1}, and in J_2 to a few parts in 10^8. Proposals are being developed, primarily in Europe, for advanced space missions which will have tests of PPN parameters as key components, including GAIA, a high-precision astrometric telescope (successor to Hipparcos), which could measure light-deflection and γ to the 10^{-6} level, and a solar orbit relativity test, which could measure γ to 10^{-7} from time delay and light deflection measurements.

3.7.5 Gravitational-Wave Astronomy

A significant part of the field of experimental gravitation is devoted to building and designing sensitive devices to detect gravitational radiation and to use gravity waves as a new astronomical tool. The centerpieces of this effort are the US Laser Interferometric Gravitational-wave Observatory (LIGO) and the European VIRGO projects, which are currently under construction. This important topic has been reviewed thoroughly elsewhere,[35] and in this volume.

4 Stellar System Tests of Gravitational Theory

4.1 Binary Pulsars and General Relativity

The majority of tests of gravitational theory described so far have involved solar-system dynamics or laboratory experiments. Although the results confirm GR,

they test only a limited portion of the "space of predictions" of a theory. This portion corresponds to the weak-field, slow-motion, post-Newtonian limit. The 1974 discovery of the binary pulsar PSR 1913+16 by Taylor and Hulse opened up the possibility of probing new aspects of gravitational theory: the effects of strong relativistic internal gravitational fields on orbital dynamics, and the effects of gravitational radiation reaction. For reviews of the discovery and current status, see the published Nobel Prize lectures by Hulse and Taylor.[36,37]

The system consists of a pulsar of nominal period 59 ms in a close binary orbit with an as yet unseen companion. From detailed analyses of the arrival times of pulses (which amounts to an integrated version of the Doppler-shift methods used in spectroscopic binary systems), extremely accurate orbital and physical parameters for the system have been obtained (Table 6). Because the orbit is so close ($\approx 1R_\odot$) and because there is no evidence of an eclipse of the pulsar signal or of mass transfer from the companion, it is generally believed that the companion is compact: evolutionary arguments suggest that it is most likely a dead pulsar. Thus the orbital motion is very clean, free from tidal or other complicating effects. Furthermore, the data acquisition is "clean" in the sense that the observers can keep track of the pulsar phase with an accuracy of 15μs, despite gaps of up to six months between observing sessions. The pulsar has shown no evidence of "glitches" in its pulse period.

Three factors make this system an arena where relativistic celestial mechanics must be used: the relatively large size of relativistic effects [$v_{\rm orbit} \approx (m/r)^{1/2} \approx 10^{-3}$]; the short orbital period (8 hours), allowing secular effects to build up rapidly; and the cleanliness of the system, allowing accurate determinations of small effects. Just as Newtonian gravity is used as a tool for measuring astrophysical parameters of ordinary binary systems, so GR is used as a tool for measuring astrophysical parameters in the binary pulsar.

The observational parameters that are obtained from a least squares solution of the arrival time data fall into three groups: (i) non-orbital parameters, such as the pulsar period and its rate of change, and the position of the pulsar on the sky; (ii) five "Keplerian" parameters, most closely related to those appropriate for standard Newtonian systems, such as the eccentricity e and the orbital period P_b; and (iii) five "post-Keplerian" parameters. The five post-Keplerian parameters are $\langle\dot{\omega}\rangle$, the average rate of periastron advance; γ', the amplitude of delays in arrival of pulses caused by the varying effects of the gravitational redshift and

Parameter	Symbol (units)	Value[1]
(i) **"Physical" Parameters**		
Right Ascension	α	$19^h15^m28.^s00018(15)$
Declination	δ	$16°06'27.''4043(3)$
Pulsar Period	P_p (ms)	59.029997929613(7)
Derivative of Period	$\dot{P}_p$	$8.62713(8) \times 10^{-18}$
(ii) **"Keplerian" Parameters**		
Projected semimajor axis	$a_p \sin i$ (s)	2.3417592(19)
Eccentricity	e	0.6171308(4)
Orbital Period	P_b (day)	0.322997462736(7)
Longitude of periastron	ω_0 (°)	226.57528(6)
Julian ephemeris date of periastron	T_0 (MJD)	46443.99588319(3)
(iii) **"Post-Keplerian" Parameters**		
Mean rate of periastron advance	$\langle\dot{\omega}\rangle$ (° yr^{-1})	4.226621(11)
Gravitational redshift/time dilation	γ' (ms)	4.295(2)
Orbital period derivative	$\dot{P}_b\,(10^{-12})$	−2.422(6)

[1]Numbers in parentheses denote errors in last digit.

Table 6. Parameters of the Binary Pulsar PSR 1913+16.

time dilation as the pulsar moves in its elliptical orbit at varying distances from the companion and with varying speeds; $\dot{P}_b$, the rate of change of orbital period, caused predominantly by gravitational radiation damping; and r and $s = \sin i$, respectively the "range" and "shape" of the Shapiro time delay caused by the companion, where i is the angle of inclination of the orbit relative to the plane of the sky.

In GR, these post-Keplerian parameters can be related to the masses of the two bodies and to measured Keplerian parameters by the equations [TEGP 12.1, 14.6(a)]

$$\langle \dot{\omega} \rangle = 3(2\pi/P_b)^{5/3} m^{2/3} (1-e^2)^{-1} , \quad (45)$$

$$\gamma' = e(P_b/2\pi)^{1/3} m_2 m^{-1/3} (1 + m_2/m) , \quad (46)$$

$$\dot{P}_b = -(192\pi/5)(2\pi m/P_b)^{5/3} (\mu/m) \left(1 + \frac{73}{24}e^2 + \frac{37}{96}e^4\right) (1-e^2)^{-7/2} , \quad (47)$$

$$s = \sin i , \quad (48)$$

$$r = m_2 , \quad (49)$$

where m_1 and m_2 denote the pulsar and companion masses, respectively. The formula for $\langle \dot{\omega} \rangle$ ignores possible non-relativistic contributions to the periastron shift, such as tidally or rotationally induced effects caused by the companion [for discussion of these effects, see TEGP 12.1(c)]. The formula for $\dot{P}_b$ represents the effect of energy loss through the emission of gravitational radiation, and makes use of the "quadrupole formula" of GR (for a survey of the quadrupole and other approximations for gravitational radiation, see Ref. 38); it ignores other sources of energy loss, such as tidal dissipation [TEGP 12.1(f)].

The timing model that contains these parameters was developed by Damour, Deruelle, and Taylor, superseding earlier treatments by Haugan, Blandford, Teukolsky, and Epstein. The current values for Keplerian and post-Keplerian parameters are shown in Table 6.

The most convenient way to display these results is to plot the constraints they imply for the two masses m_1 and m_2, via Equations 45 and 46. These are shown in Figure 6. From $\langle \dot{\omega} \rangle$ and γ' we obtain the values $m_1 = 1.4411 \pm 0.0007 M_\odot$ and $m_2 = 1.3873 \pm 0.0007 M_\odot$. Equation 47 then predicts the value $\dot{P}_b = -2.40243 \pm 0.00005 \times 10^{-12}$. In order to compare the predicted value for $\dot{P}_b$ with the observed value, it is necessary to take into account the effect of a relative acceleration between the binary pulsar system and the solar system caused by the differential rotation of the galaxy. This effect was previously considered unimportant when $\dot{P}_b$ was known only to ten percent accuracy. Damour and Taylor carried out a careful estimate of this effect using data on the location and proper motion of the pulsar, combined with the best information available on galactic rotation, and found

$$\dot{P}_b^{\rm GAL} \simeq -(1.7 \pm 0.5) \times 10^{-14} . \quad (50)$$

Subtracting this from the observed $\dot{P}_b$ (Table 5) gives the residual

$$\dot{P}_b^{\rm OBS} = -(2.408 \pm 0.010[\rm OBS] \pm 0.005[\rm GAL]) \times 10^{-12} , \quad (51)$$

which agrees with the prediction, within the errors. In other words,

$$\frac{\dot{P}_b^{\rm GR}}{\dot{P}_b^{\rm OBS}} = 1.0023 \pm 0.0041[\rm OBS] \pm 0.0021[\rm GAL] . \quad (52)$$

The parameters r and s are not separately measurable with interesting accuracy for PSR 1913+16 because the orbit's 47° inclination does not lead to a substantial Shapiro delay.

The consistency among the measurements is also displayed in Figure 6, in which the regions allowed by the three most precise constraints have a single common overlap. This consistency provides a test of the assumption that the two bodies behave as "point" masses, without complicated tidal effects, obeying the general relativistic equations of motion including gravitational radiation. It is also a test of the strong equivalence principle, in that the highly relativistic internal structure of the neutron star does not influence its orbital motion, as predicted by GR. Recent observations[39] indicate variations in the pulse profile, which suggests that the pulsar is undergoing geodetic precession (Sec. 3.7.2) in an amount consistent with GR, assuming that the pulsar's spin is suitably misaligned with the orbital angular momentum.

4.2 A Population of Binary Pulsars?

Since 1990, several new massive binary pulsars similar to PSR 1913+16 were discovered, leading to the possibility of new or improved tests of GR.

PSR 1534+12. This is a binary pulsar system in our galaxy. Its pulses are significantly stronger and narrower than those of PSR 1913+16, so timing measurements are more precise, reaching $3\mu s$ accuracy. Its parameters are listed

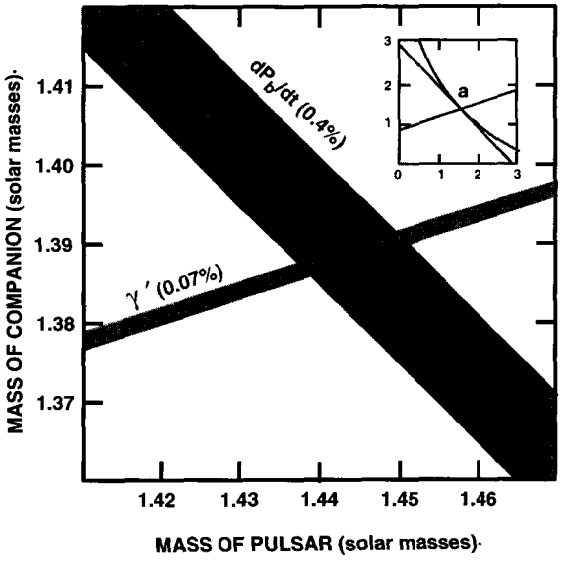


Fig. 6. Constraints on masses of pulsar and companion from data on PSR 1913+16, assuming GR to be valid. Width of each strip in the plane reflects observational accuracy, shown as a percentage. Inset shows the three constraints on the full mass plane; intersection region (a) has been magnified 400 times for the full figure.

Parameter	B1534+12	B2127+11C	B1855+09
(i) "Keplerian" Parameters			
$a_p \sin i$ (s)	3.729463(3)	2.520(3)	9.2307802(4)
e	0.2736777(5)	0.68141(2)	0.00002168(5)
P_b(day)	0.42073729929(4)	0.335282052(6)	12.3271711905(6)
(ii) "Post-Keplerian" Parameters[1]			
$\langle\dot{\omega}\rangle$ ($^\circ$ yr^{-1})	1.75576(4)	4.457(12)	
γ' (ms)	2.066(10)	4.9(1.1)	
$\dot{P}_b$ (10^{-12})	−0.129(14)		
$r(\mu s)$	6.7(1.3)		1.27(10)
$s = \sin i$	0.982(7)		0.9992(5)

[1] From Refs. 40 and 41 and `http://puppsr8.princeton.edu/psrcat.html`.

Table 7. Parameters of New Binary Pulsars.

in Table 6 (Ref. 41). The orbital plane appears to be almost edge-on relative to the line of sight ($i \simeq 80°$); as a result the Shapiro delay is substantial, and separate values of the parameters r and s have been obtained with interesting accuracy. Assuming general relativity, one infers that the two masses are virtually identical at $1.339 \pm 0.003 M_\odot$. The rate of orbit decay $\dot{P}_b$ agrees with GR to about 15 percent, the precision limited by the poorly known distance to the pulsar, which introduces a significant uncertainty into the subtraction of galactic acceleration.

PSR 2127+11C. This system appears to be a clone of the Hulse-Taylor binary pulsar, with very similar values for orbital period and eccentricity (see Table 7). The inferred total mass of the system is $2.706 \pm 0.011 M_\odot$. Because the system is in the globular cluster M15 (NGC 7078), it will suffer Doppler shifts resulting from local accelerations, either by the mean cluster gravitational field or by nearby stars, that are more difficult to estimate than was the case with the galactic system PSR 1913+16. This may make a separate, precision measurement of the relativistic contribution to $\dot{P}_b$ impossible.

PSR 1855+09. This binary pulsar system is not particularly relativistic, with a long period (12 days) and highly circular orbit. However, because we observe the orbit nearly edge on, the Shapiro delay is large and measurable, as reflected in the post-Keplerian parameters r and s.

4.3 Binary Pulsars and Alternative Theories

Soon after the discovery of the binary pulsar, it was widely hailed as a new testing ground for relativistic gravitational effects. As we have seen in the case of GR, in most respects, the system has lived up to, indeed exceeded, the early expectations.

In another respect, however, the system has only partially lived up to its promise, namely as a direct testing ground for alternative theories of gravity. The origin of this promise was the discovery that alternative theories of gravity generically predict the emission of dipole gravitational radiation from binary star systems. This additional form of gravitational radiation damping could, at least in principle, be significantly stronger than the usual quadrupole damping, because it depends on fewer powers of the parameter v/c, where v is the orbital velocity and c is the speed of light, and it depends on the gravitational binding energy per unit mass of the bodies, which, for neutron stars, could be as large as 40 per cent. As one fulfillment of this promise, Will and Eardley worked out in detail the effects of dipole gravitational radiation in the bimetric theory of Rosen, and, when the first observation of the decrease of the orbital period was announced in 1978, the Rosen theory suffered a terminal blow (TEGP 12.3). A wide class of alternative theories also fail the binary pulsar test because of dipole gravitational radiation.

On the other hand, the early observations already indicated that, in GR, the masses of the two bodies were nearly equal, so that, in theories of gravity that are in some sense "close" to GR, dipole gravitational radiation would not be a strong effect, because of the apparent symmetry of the system (technically, the amount of dipole radiation depends on the difference between the gravitational binding energy per unit mass for the two bodies). The Rosen theory, and others like it, are not "close" to general relativity, except in their predictions for the weak-field, slow-motion regime of the solar system. When relativistic neutron stars are present, theories like these can predict strong effects on the motion of the bodies resulting from their internal highly relativistic gravitational structure (violations of the Strong Equivalence Principle). As a consequence, the masses inferred from observations such as the periastron shift may be significantly different from those inferred using general relativity, and may be different from each other, leading to strong dipole gravitational radiation damping. By contrast, the Brans-Dicke theory, which was the basis for Eardley's discovery of the dipole radiation phenomenon, is "close" to GR, roughly speaking within $1/\omega_{\rm BD}$ of the predictions of the latter, for large values of the coupling constant $\omega_{\rm BD}$ (henceforth this notation for the coupling constant is adopted to avoid confusion with the periastron angle). Thus, despite the presence of dipole gravitational radiation, the binary pulsar provides at present only a weak test of Brans-Dicke theory, not yet competitive with solar-system tests.

4.4 Binary Pulsars and Scalar-Tensor Theories

Making the usual assumption that both members of the system are neutron stars, and using the methods summarized in TEGP Chapters 10–12, one can obtain formulas for the periastron shift, the gravitational redshift/second-order Doppler shift parameter, and the rate of change of orbital period, analogous to Equations 45–47. These formulas depend on the masses of the two neutron stars, on their self-gravitational binding energy, represented by "sensitivities" s and κ^* and on the Brans-Dicke coupling constant $\omega_{\rm BD}$. First, there is a modification of Kepler's third law, given by

$$P_b/2\pi = (a^3/\mathcal{G}m)^{1/2}\,. \tag{53}$$

Then, the predictions for $\langle\dot\omega\rangle$, γ' and $\dot P_b$ are

$$\langle\dot\omega\rangle = 3(2\pi/P_b)^{5/3}m^{2/3}(1-e^2)^{-1}\mathcal{P}\mathcal{G}^{-4/3}\,, \tag{54}$$

$$\gamma' = e(P_b/2\pi)^{1/3}m_2m^{-1/3}\mathcal{G}^{-1/3}(\alpha_2^* + \mathcal{G}m_2/m + \kappa_1^*\eta_2^*)\,, \tag{55}$$

$$\dot P_b = -(192\pi/5)(2\pi m/P_b)^{5/3}(\mu/m)\mathcal{G}^{-4/3}F(e) - 4\pi(2\pi m/P_b)(\mu/m)\xi\mathcal{S}^2G(e)\,, \tag{56}$$

where, to first order in $\xi \equiv (2+\omega_{\rm BD})^{-1}$, we have

$$F(e) = \frac{1}{12}(1-e^2)^{-7/2}[\kappa_1(1+\frac{7}{2}e^2+\frac{1}{2}e^4) - \kappa_2(\frac{1}{2}e^2+\frac{1}{8}e^4)]\,, \tag{57}$$

$$G(e) = (1-e^2)^{-5/2}(1+\frac{1}{2}e^2)\,, \tag{58}$$

$$\mathcal{S} = s_1 - s_2\,, \tag{59}$$

$$\mathcal{G} = 1 - \xi(s_1+s_2-2s_1s_2)\,, \tag{60}$$

$$\mathcal{P} = \mathcal{G}[1 - \frac{2}{3}\xi + \frac{1}{3}\xi(s_1+s_2-2s_1s_2)]\,, \tag{61}$$

$$\alpha_2^* = 1 - \xi s_2\,, \tag{62}$$

$$\eta_2^* = (1-2s_2)\xi\,, \tag{63}$$

$$\kappa_1 = \mathcal{G}^2[12(1-\frac{1}{2}\xi)+\xi\Gamma^2], \quad (64)$$

$$\kappa_2 = \mathcal{G}^2[11(1-\frac{1}{2}\xi)+\frac{1}{2}\xi(\Gamma^2-5\Gamma\Gamma'-\frac{15}{2}\Gamma'^2)], \quad (65)$$

$$\Gamma = 1-2(m_1s_2+m_2s_1)/m, \quad (66)$$

$$\Gamma' = 1-s_1-s_2. \quad (67)$$

The quantities s_a and κ_a^* are defined by

$$s_a = -\left(\frac{\partial(\ln m_a)}{\partial(\ln G)}\right)_N, \qquad \kappa_a^* = -\left(\frac{\partial(\ln I_a)}{\partial(\ln G)}\right)_N, \quad (68)$$

and measure the "sensitivity" of the mass m_a and moment of inertia I_a of each body to changes in the scalar field (reflected in changes in G) for a fixed baryon number N [see TEGP 11, 12, and 14.6(c) for further details].

The first term in $\dot{P}_b$ (Equation 56) is the effect of quadrupole and monopole gravitational radiation, while the second term is the effect of dipole radiation. Equations 53–59 are quite general, applying to a class of "conservative" theories of gravity whose equations of motion for compact objects can be described by a Lagrangian (TEGP 11.3). The forms of the variables in the remaining equations are specific to Brans-Dicke theory.

In order to estimate the sensitivities s_1, s_2, $\mathcal{S}$, and κ^*, we must adopt an equation of state for the neutron stars. We restrict attention to relatively stiff neutron star equations of state in order to guarantee neutron stars of sufficient mass, approximately $1.4M_\odot$; the M and O equations of state of Arnett and Bowers give similar results. The lower limit on $\omega_{\rm BD}$ required to give consistency among the constraints on $\langle\dot{\omega}\rangle$, γ, and $\dot{P}_b$ as in Figure 6 is 105. The combination of $\langle\dot{\omega}\rangle$ and γ give a constraint on the masses that is relatively weakly dependent on ξ; thus the constraint on ξ is dominated by $\dot{P}_b$ and is directly proportional to the measurement error in $\dot{P}_b$. In order to achieve a constraint comparable to the solar system value of 3×10^{-4}, the error in $\dot{P}_b^{\rm OBS}$ would have to be reduced by more than a factor of ten.

Damour and Esposito-Farèse[20] have devised a scalar-tensor theory in which two scalar fields are tuned so that their effects in the weak-field slow-motion regime of the solar system are suppressed, so that the theory is identical to general relativity in the post-Newtonian approximation. Yet in the regime appropriate to binary pulsars, it predicts strong-field SEP-violating effects and radiative effects that distinguish it from GR. It gives formulae for the post-Keplerian parameters of Equations 54–56 as well as for the parameters r and s that have corrections dependent upon the sensitivities of the relativistic neutron stars. The theory depends upon two arbitrary parameters β' and β''; GR corresponds to the values β'=β''=0. It turns out that the binary pulsar PSR 1913+16 alone constrains the two parameters to a narrow but long strip in the (β',β'')-plane that includes the origin (GR) but that could include some highly non-general relativistic theories. The sensitivity of PSR 1534+12 to r and s provides an orthogonal constraint that cuts the strip. In this class of theories, then, *both* binary pulsars are needed to provide a strong test. Interestingly, it is the strong-field aspects of PSR 1534+12, not gravitational radiation, that provide the non-trivial new constraint. In scalar-tensor theories subject to "spontaneous-scalarization" in neutron-star interiors, the bounds from binary pulsars can be significantly stronger than those from the solar-system.[23]

5 Gravitational-Radiation Tests of Gravitational Theory

5.1 Laser-Interferometric Gravitational Observatories

Some time in the next decade, a new opportunity for testing relativistic gravity may be realized, with the completion and operation of kilometer-scale, laser interferometric gravitational-wave observatories in the U.S. (LIGO project), Europe (VIRGO and GEO-600 projects), and Japan (TAMA 300 project). Gravitational-wave searches at these observatories are scheduled to commence around 2002. The LIGO broad-band antennae will have the capability of detecting and measuring the gravitational waveforms from astronomical sources in a frequency band between about 10 Hz (the seismic noise cutoff) and 500 Hz (the photon counting noise cutoff), with a maximum sensitivity to strain at around 100 Hz of $\Delta l/l \sim 10^{-22}$ (rms). The most promising source for detection and study of the gravitational-wave signal is the "inspiralling compact binary"—a binary system of neutron stars or black holes (or one of each) in the final minutes of a death dance leading to a violent merger. Such is the fate, for example, of the Hulse-Taylor binary pulsar PSR 1913+16 in about 300 M years. Given the expected sensitivity of the "advanced LIGO" (around 2007), which could see such sources out to hundreds of

megaparsecs, it has been estimated that from 3 to 100 annual inspiral events could be detectable. Other sources, such as supernova core collapse events, instabilities in rapidly rotating nascent neutron stars, signals from non-axisymmetric pulsars, and a stochastic background of waves, may be detectable (for reviews, see Ref. 35 and articles in this volume).

Detailed observation of gravitational waves from such sources may provide the means to test general relativistic predictions for the polarization and speed of the waves, and for gravitational radiation damping.

Several of these tests can be realized using gravitational-waves from inspiralling compact binaries. The analysis of gravitational-wave data from such sources will involve some form of matched filtering of the noisy detector output against an ensemble of theoretical "template" waveforms which depend on the intrinsic parameters of the inspiralling binary, such as the component masses, spins, and so on, and on its inspiral evolution. How accurate must a template be in order to "match" the waveform from a given source (where by a match we mean maximizing the cross-correlation or the signal-to-noise ratio)? In the total accumulated phase of the wave detected in the sensitive bandwidth, the template must match the signal to a fraction of a cycle. For two inspiralling neutron stars, around 16,000 cycles should be detected; this implies a phasing accuracy of 10^{-5} or better. Since $v/c \sim 1/10$ during the late inspiral, this means that correction terms in the phasing at the level of $(v/c)^5$ or higher are needed. More formal analyses confirm this intuition.[42]

Because it is a slow-motion system ($v/c \sim 10^{-3}$), the binary pulsar is sensitive only to the lowest-order effects of gravitational radiation as predicted by the quadrupole formula. Nevertheless, the first correction terms of order v/c and $(v/c)^2$ to the quadrupole formula were calculated as early as 1976 (TEGP 10.3). These are now conventionally called "post-Newtonian" (PN) corrections, with each power of v/c corresponding to half a post-Newtonian order (1/2PN), in analogy with post-Newtonian corrections to the Newtonian equations of motion.[43]

But for laser-interferometric observations of gravitational waves, the bottom line is that, in order to measure the astrophysical parameters of the source and to test the properties of the gravitational waves, it is necessary to derive the gravitational waveform and the resulting radiation back-reaction on the orbit phasing at least to 2PN, or second post-Newtonian order, $O[(v/c)^4]$, beyond the quadrupole approximation, and probably to 3PN order.

5.2 Post-Newtonian Generation of Gravitational Waves

The generation of gravitational radiation is a long-standing problem that dates back to the first years following the publication of GR, when Einstein calculated the gravitational radiation emitted by a laboratory-scale object using the linearized version of GR. Shortly after the discovery of the binary pulsar PSR 1913+16 in 1974, questions were raised about the foundations of the "quadrupole formula" for gravitational radiation damping (and in some quarters, even about its quantitative validity). These questions were answered in part by theoretical work designed to shore up the foundations of the quadrupole approximation,[44] and in part (perhaps mostly) by the agreement between the predictions of the quadrupole formula and the *observed* rate of damping of the pulsar's orbit. In 1976, the post-Newtonian corrections to the quadrupole formula were of purely academic, rather than observational, interest.

The challenge of providing accurate templates for LIGO-VIRGO data analysis has led to major efforts to calculate gravitational waves to high PN order. Three approaches have been developed.

The BDI approach of Blanchet, Damour, and Iyer is based on a mixed post-Newtonian and "post-Minkowskian" framework for solving Einstein's equations approximately, developed in a long series of papers by Damour and colleagues.[45] The idea is to solve the vacuum Einstein equations in the exterior of the material sources extending out to the radiation zone in an expansion ("post-Minkowskian") in "nonlinearity" (effectively an expansion in powers of Newton's constant G), and to express the asymptotic solutions in terms of a set of formal, time-dependent, symmetric and trace-free (STF) multipole moments.[46] Then, in a near zone within one characteristic wavelength of the radiation, the equations including the material source are solved in a slow-motion approximation (expansion in powers of $1/c$) that yields a set of STF source multipole moments expressed as integrals over the "effective" source, including both matter and gravitational field contributions. The solutions involving the two sets of moments are then matched in an intermediate zone, resulting in a connection between the formal radiative moments and the source moments. The matching also provides a natural way, using analytic continuation, to regularize integrals involving the non-compact contributions of gravitational stress-energy, that might otherwise be divergent.

An approach called DIRE is based on a framework developed by Epstein and

Wagoner (EW),[47] and extended by Will, Wiseman, and Pati. We shall describe DIRE briefly below.

A third approach, valid only in the limit in which one mass is much smaller than the other, is that of black-hole perturbation theory. This method provides numerical results that are exact in v/c, as well as analytical results expressed as series in powers of v/c, both for non-rotating and for rotating black holes. For non-rotating holes, the analytical expansions have been carried to *5.5* PN order.[48] In all cases of suitable overlap, the results of all three methods agree precisely.

5.2.1 Direct Integration of the Relaxed Einstein Equations (DIRE)

Like the BDI approach, DIRE involves rewriting the Einstein equations in their "relaxed" form, namely as an inhomogeneous, flat-spacetime wave equation for a field $h^{\alpha\beta}$, whose formal solution can be written

$$h^{\alpha\beta}(t,\mathbf{x}) = \quad 4\int_C \frac{\tau^{\alpha\beta}(t-|\mathbf{x}-\mathbf{x}'|,\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}d^3x' \,, \tag{69}$$

where the source $\tau^{\alpha\beta}$ consists of both the material stress-energy, and a "gravitational stress-energy" made up of all the terms non-linear in $h^{\alpha\beta}$, and the integration is over the past flat-spacetime null-cone $\mathcal{C}$ of the field point $(t,\mathbf{x})$ (see Fig. 7). The wave equation is accompanied by a harmonic or deDonder gauge condition $h^{\alpha\beta}{}_{,\beta}=0$, which serves to specify a coordinate system, and also imposes equations of motion on the sources. Unlike the BDI approach, a *single* formal solution is written down, valid everywhere in spacetime. This formal solution is then iterated in a slow-motion ($v/c<1$), weak-field ($||h^{\alpha\beta}||<1$) approximation, that is very similar to the corresponding procedure in electromagnetism. However, because the integrand of this retarded integral is not compact by virture of the non-linear field contributions, the original EW formalism quickly runs up against integrals that are not well-defined, or worse, are divergent. Although at the lowest quadrupole and first few PN orders, various arguments can be given to justify sweeping such problems under the rug,[49] they are not very rigorous, and provide no guarantee that the divergences do not become insurmountable at higher orders. As a consequence, despite efforts to cure the problem, the EW formalism fell into some disfavor as a route to higher orders, although an extension to 3/2PN order was accomplished.[50]

The resolution of this problem involves taking literally the statement that the

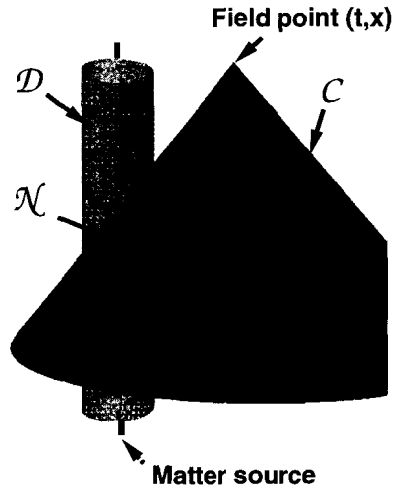


Fig. 7. Past harmonic null cone $\mathcal{C}$ of the field point $(t,\mathbf{x})$ intersects the near zone $\mathcal{D}$ in the hypersurface $\mathcal{N}$.

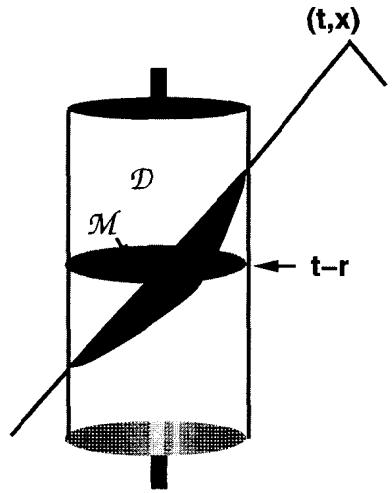


Fig. 8. Taylor expansion of retarded time dependence on $\mathcal{N}$ results in multipole moments integrated over the spatial hypersurface $\mathcal{M}$, corresponding to fixed retarded time $t-r$.

solution is a *retarded* integral, *i.e.,* an integral over the *entire* past null cone of the field point.[51] To be sure, that part of the integral that extends over the intersection $\mathcal{N}$ between the past null cone and the material source and the near zone is still approximated as usual by a slow-motion expansion involving spatial integrals over $\mathcal{M}$ of moments of the source, including the non-compact gravitational contributions, just as in the BDI framework (Fig. 8). But instead of cavalierly extending the spatial integrals to infinity as was implicit in the original EW framework, and risking undefined or divergent integrals, we terminate the integrals at the boundary of the near zone, chosen to be at a radius $\mathcal{R}$ given roughly by one wavelength of the gravitational radiation. For the integral over the rest of the past null cone exterior to the near zone ("radiation zone"), we neither make a slow-motion expansion nor continue to integrate over a spatial hypersurface; instead we use a coordinate transformation in the integral from the spatial coordinates d^3x' to quasi-null coordinates du', $d\theta'$, $d\phi'$, where

$$t - u' = r' + |\mathbf{x} - \mathbf{x}'|\,, \tag{70}$$

to convert the integral into a convenient, easy-to-calculate form, that is manifestly convergent, subject only to reasonable assumptions about the past behavior of the source:

$$h^{\alpha\beta}_{\mathcal{C}-\mathcal{N}}(t,x) = 4\int_{-\infty}^{u} du' \int \frac{\tau^{\alpha\beta}(u'+r',\mathbf{x}')}{t-u'-\mathbf{n}'\cdot\mathbf{x}}[r'(u',\Omega')]^2 d^2\Omega'\,. \tag{71}$$

This transformation was suggested by earlier work on a non-linear gravitational-wave phenomenon called the Christodoulou memory.[52] Not only are all integrations now explicitly finite and convergent, one can show that all contributions from the finite, near-zone spatial integrals that depend upon $\mathcal{R}$ are actually *cancelled* by corresponding terms from the radiation-zone integrals, valid for both positive and negative powers of $\mathcal{R}$ and for terms logarithmic in $\mathcal{R}$.[53] Thus the procedure, as expected, has no dependence on the artificially chosen boundary radius $\mathcal{R}$ of the near-zone. In addition, the method can be carried to higher orders in a straightforward manner. The result is a manifestly finite, well-defined procedure for calculating gravitational radiation to high orders.

The result of the DIRE or BDI calculations is an explicit formula for two-body systems in general orbits, for the transverse-traceless (TT) part of the radiation-zone field, denoted h^{ij}, and representing the deviation of the metric from flat spacetime. In terms of an expansion beyond the quadrupole formula, it has the schematic form,

$$h^{ij} = \frac{2G\mu}{R}\left\{\tilde{Q}^{ij}[1 + O(\epsilon^{1/2}) + O(\epsilon) + O(\epsilon^{3/2}) + O(\epsilon^2)\ldots]\right\}_{TT}\,, \tag{72}$$

where μ is the reduced mass, R is the distance to the source, and $\tilde{Q}^{ij}$ represents two time derivatives of the mass quadrupole moment tensor (the series actually contains multipole orders beyond quadrupole). The expansion parameter ϵ is related to the orbital variables by $\epsilon \sim m/r \sim v^2$, where r is the distance between the bodies, v is the relative velocity, and $m = m_1 + m_2$ is the total mass ($G = c = 1$). The 1/2PN and 1PN terms were derived by Wagoner and Will,[49] the 3/2PN terms by Wiseman.[50] The contribution of gravitational-wave "tails," caused by backscatter of the outgoing radiation off the background spacetime curvature, at $O(\epsilon^{3/2})$, were derived and studied by several authors. The 2PN terms including 2PN tail contributions were derived by two independent groups and are in complete agreement (see Refs. 51 and 54 for details and references to earlier work).

There are also contributions to the waveform due to intrinsic spin of the bodies, which occur at $O(\epsilon^{3/2})$ (spin-orbit) and $O(\epsilon^2)$ (spin-spin); these have been calculated elsewhere.[55]

Given the gravitational waveform, one can compute the rate at which energy is carried off by the radiation (schematically $\int \dot{h}\dot{h}d\Omega$, the gravitational analog of the Poynting flux). For the special case of non-spinning bodies moving on quasi-circular orbits (*i.e.,* circular apart from a slow inspiral), the energy flux through 2PN order has the form

$$\begin{aligned}\frac{dE}{dt} &= \frac{32}{5}\eta^2\left(\frac{m}{r}\right)^5\left[1 - \frac{m}{r}\left(\frac{2927}{336} + \frac{5}{4}\eta\right)\right. \\ &\quad \left. +4\pi\left(\frac{m}{r}\right)^{3/2} + \left(\frac{m}{r}\right)^2\left(\frac{293383}{9072} + \frac{380}{9}\eta\right)\right]\,,\end{aligned} \tag{73}$$

where $\eta = m_1 m_2/m^2$. The first term is the quadrupole contribution, the second term is the 1PN contribution, the third term, with the coefficient 4π, is the "tail" contribution, and the fourth term is the 2PN contribution first reported jointly by Blanchet *et al.*[56] Calculation of the 3PN contributions is nearing completion.

Similar expressions can be derived for the loss of angular momentum and linear momentum. These losses react back on the orbit to circularize it and cause it to inspiral. The result is that the orbital phase (and consequently the gravitational-wave phase) evolves non-linearly with time. It is the sensitivity of

the broad-band LIGO and VIRGO-type detectors to phase that makes the higher-order contributions to dE/dt so observationally relevant. A ready-to-use set of formulae for the 2PN gravitational waveform template, including the non-linear evolution of the gravitational-wave frequency (not including spin effects) have been published[57] and incorporated into the Gravitational Radiation Analysis and Simulation Package (GRASP), a publically available software toolkit.[58]

5.3 Polarization of Gravitational Waves

A laser-interferometric or resonant bar gravitational-wave detector measures the local components of a symmetric 3×3 tensor which is composed of the "electric" components of the Riemann tensor, R_{0i0j}. These six independent components can be expressed in terms of polarizations (modes with specific transformation properties under null rotations). Three are transverse to the direction of propagation, with two representing quadrupolar deformations and one representing a monopole "breathing" deformation. Three modes are longitudinal, with one an axially symmetric stretching mode in the propagation direction, and one quadrupolar mode in each of the two orthogonal planes containing the propagation direction. General relativity predicts only the first two transverse quadrupolar modes, independently of the source, while scalar-tensor gravitational waves can in addition contain the transverse breathing mode. More general metric theories predict up to the full complement of six modes (TEGP 10.2). A suitable array of gravitational antennas could delineate or limit the number of modes present in a given wave. The strategy depends on whether or not the source direction is known. In general there are eight unknowns (six polarizations and two direction cosines), but only six measurables (R_{0i0j}). If the direction can be established by either association of the waves with optical or other observations, or by time-of-flight measurements between separated detectors, then six suitably oriented detectors suffice to determine all six components. If the direction cannot be established, then the system is underdetermined, and no unique solution can be found. However, if one assumes that only transverse waves are present, then there are only three unknowns if the source direction is known, or five unknowns otherwise. Then the corresponding number (three or five) of detectors can determine the polarization. If distinct evidence were found of any mode other than the two transverse quadrupolar modes of GR, the result would be disastrous for GR. On the other hand, the absence of a breathing mode would not necessarily rule out scalar-tensor gravity, because the strength of that mode depends on the nature of the source.

Although some of the details of implementing such polarization observations have been worked out for arrays of resonant cylindrical, disk-shaped, and spherical detectors (TEGP 10.2, Ref. 59) only initial work has been done to assess whether the ground-based laser-interferometers (LIGO, VIRGO, GEO600, TAMA300) could perform interesting polarization measurements. The results depend sensitively on the relative orientation of the detectors' arms, which are now cast (literally) in concrete.

5.4 Speed of Gravitational Waves

According to GR, in the limit in which the wavelength of gravitational waves is small compared to the radius of curvature of the background spacetime, the waves propagate along null geodesics of the background spacetime, *i.e.,* they have the same speed, c, as light (henceforth, we do not set $c = 1$). In other theories, the speed could differ from c because of coupling of gravitation to "background" gravitational fields. For example, in the Rosen bimetric theory with a flat background metric $\boldsymbol{\eta}$, gravitational waves follow null geodesics of $\boldsymbol{\eta}$, while light follows null geodesics of $\mathbf{g}$ (TEGP 10.1).

Another way in which the speed of gravitational waves could differ from c is if gravitation were propagated by a massive field (a massive graviton), in which case, v_g would be given by, in a local inertial frame,

$$\frac{v_g^2}{c^2} = 1 - \frac{m_g^2 c^4}{E^2}, \tag{74}$$

where m_g and E are the graviton rest mass and energy, respectively. For a recent review of the idea of a massive graviton, see Ref. 60.

The most obvious way to test this is to compare the arrival times of a gravitational wave and an electromagnetic wave from the same event, *e.g.,* a supernova. For a source at a distance D, the resulting value of the difference $1 - v_g/c$ is

$$1 - \frac{v_g}{c} = 5 \times 10^{-17} \left(\frac{200\text{Mpc}}{D}\right)\left(\frac{\Delta t}{1\text{s}}\right), \tag{75}$$

where $\Delta t \equiv \Delta t_a - (1+Z)\Delta t_e$ is the "time difference," where Δt_a and Δt_e are the differences in arrival time and emission time, respectively, of the two signals, and Z is the redshift of the source. In many cases, Δt_e is unknown, so that the best

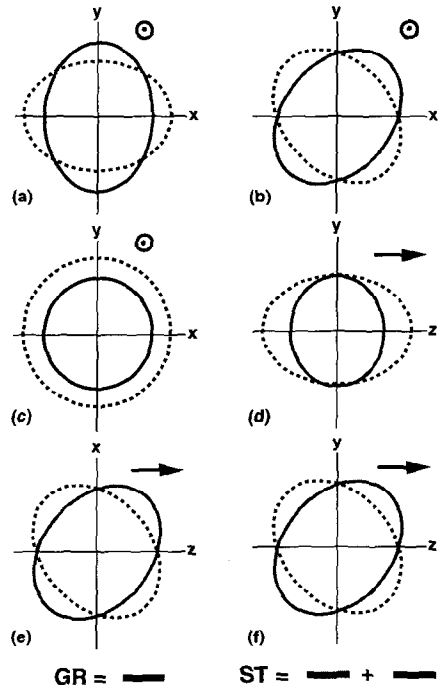


Fig. 9. Six polarization modes for gravitational waves permitted in any metric theory of gravity. Shown is the displacement that each mode induces on a ring of test particles. The wave propagates in the $+z$ direction. There is no displacement out of the plane of the picture. In (a), (b), and (c), the wave propagates out of the plane; in (d), (e), and (f), the wave propagates in the plane. In general relativity, only (a) and (b) are present; in scalar-tensor gravity, (c) may also be present.

one can do is employ an upper bound on Δt_e based on observation or modelling. The result will then be a bound on $1 - v_g/c$.

For a massive graviton, if the frequency of the gravitational waves is such that $hf \gg m_g c^2$, where h is Planck's constant, then $v_g/c \approx 1 - \frac{1}{2}(c/\lambda_g f)^2$, where $\lambda_g = h/m_g c$ is the graviton Compton wavelength, and the bound on $1 - v_g/c$ can be converted to a bound on λ_g, given by

$$\lambda_g > 3 \times 10^{12}\ \mathrm{km} \left(\frac{D}{200\ \mathrm{Mpc}} \frac{100\ \mathrm{Hz}}{f}\right)^{1/2} \left(\frac{1}{f\Delta t}\right)^{1/2} . \qquad (76)$$

The foregoing discussion assumes that the source emits *both* gravitational and electromagnetic radiation in detectable amounts, and that the relative time of emission can be established to sufficient accuracy, or can be shown to be sufficiently small.

However, there is a situation in which a bound on the graviton mass can be set using gravitational radiation alone.[61] That is the case of the inspiralling compact binary. Because the frequency of the gravitational radiation sweeps from low frequency at the initial moment of observation to higher frequency at the final moment, the speed of the gravitons emitted will vary, from lower speeds initially to higher speeds (closer to c) at the end. This will cause a distortion of the observed phasing of the waves and result in a shorter than expected overall time Δt_a of passage of a given number of cycles. Furthermore, through the technique of matched filtering, the parameters of the compact binary can be measured accurately, (assuming that GR is a good approximation to the orbital evolution, even in the presence of a massive graviton), and thereby the emission time Δt_e can be determined accurately. Roughly speaking, the "phase interval" $f\Delta t$ in Eq. (76) can be measured to an accuracy $1/\rho$, where ρ is the signal-to-noise ratio.

Thus one can estimate the bounds on λ_g achievable for various compact inspiral systems, and for various detectors. For stellar-mass inspirals (neutron stars or black holes) observed by the LIGO/VIRGO class of ground-based interferometers, $D \approx 200$Mpc, $f \approx 100$Hz, and $f\Delta t \sim \rho^{-1} \approx 1/10$. The result is $\lambda_g > 10^{13}$km. For supermassive binary black holes (10^4 to $10^7 M_\odot$) observed by the proposed laser-interferometer space antenna (LISA), $D \approx 3$Gpc, $f \approx 10^{-3}$Hz, and $f\Delta t \sim \rho^{-1} \approx 1/1000$. The result is $\lambda_g > 10^{17}$ km.

A full noise analysis using proposed noise curves for the advanced LIGO and for LISA weakens these crude bounds by factors between two and 10. These

potential bounds can be compared with the solid bound $\lambda_g > 2.8\times10^{12}$ km, derived from solar system dynamics, which limit the presence of a Yukawa modification of Newtonian gravity of the form $V(r) = (GM/r)\exp(-r/\lambda_g)$,[62] and with the model-dependent bound $\lambda_g > 6 \times 10^{19}$ km from consideration of galactic and cluster dynamics.[60]

5.5 Gravitational Radiation Back-Reaction

This plays an important role only in the inspiral of compact objects. The equations of motion of inspirals include the non-radiative, non-linear post-Newtonian corrections of Newtonian motion, as well as radiation back-reaction and its non-linear post-Newtonian corrections. The evolution of the orbit is imprinted on the phasing of the inspiral waveform, to which broad-band laser interferometers are especially sensitive through the use of matched filtering of the data against theoretical templates derived from GR.[63] A number of tests of GR using matched filtering of binary inspirals have been discussed, including putting a bound on scalar-tensor gravity,[64] measuring the non-linear "tail term" in gravitational radiation damping[65] [the "4π" term in Eq. (73)], and testing the GR "no hair" theorems by mapping spacetime outside black holes.[66]

For example, the dipole gravitational radiation predicted by scalar-tensor theories results in modifications in gravitational-radiation back-reaction. The effects are strongest for systems involving a neutron star and a black hole. Double neutron star systems are less promising because the small range of masses available near 1.4 $M_\odot$ results in suppression of dipole radiation by symmetry. Double black-hole systems turn out to be observationally identical in the two theories.

For example, for a 1.4 $M_\odot$ neutron star and a 10 $M_\odot$ (3 $M_\odot$) black hole at 200 Mpc, the bound on $\omega_{\rm BD}$ could be 600 (1800) (using advanced LIGO noise curves). The bound increases linearly with signal-to-noise ratio. If one demands that this test be performed annually, thus requiring observation of frequent, and therefore more distant, weaker sources, the bounds on $\omega_{\rm BD}$ will be too weak to compete with existing bounds. However, if one is prepared to wait ten years for the lucky observation of a nearby, strong source, the resulting bound could exceed the current bound. The bounds are illustrated in Figure 10 by the curves marked $N = 1$ and $N = 1/10$. Figure 10 assumes a double-neutron-star inspiral rate of 10^{-6} per year per galaxy, and a black-hole-neutron-star rate β times that, where

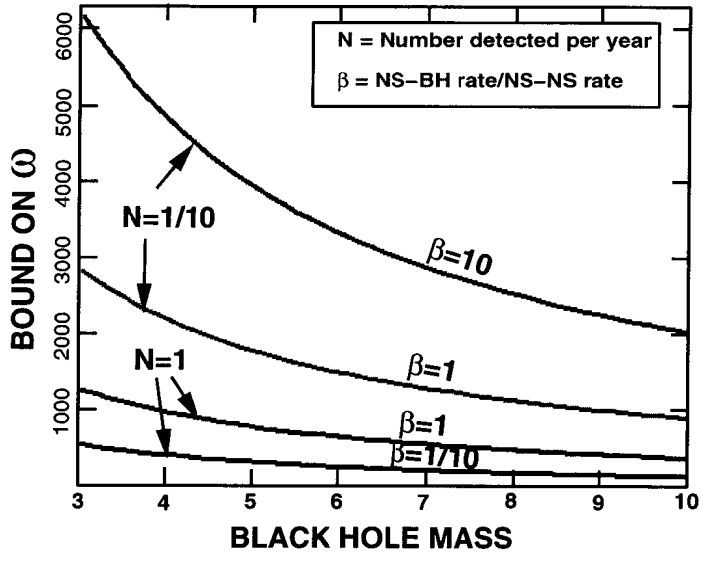


Fig. 10. Bounds on the scalar-tensor coupling constant $\omega_{\rm BD}$ from gravitational-wave observations of inspiralling black-hole neutron-star systems. The solar system bound is around $\omega_{\rm BD} = 3000$.

β is highly uncertain.

6 Conclusions

In 1998 we find that general relativity has held up under extensive experimental scrutiny. The question then arises, why bother to continue to test it? One reason is that gravity is a fundamental interaction of nature, and as such requires the most solid empirical underpinning we can provide. Another is that all attempts to quantize gravity and to unify it with the other forces suggest that the standard general relativity of Einstein is not likely to be the last word. Furthermore, the predictions of general relativity are fixed; the theory contains no adjustable constants so nothing can be changed. Thus every test of the theory is either a

potentially deadly test or a possible probe for new physics. Although it is remarkable that this theory, born 80 years ago out of almost pure thought, has managed to survive every test, the possibility of finding a discrepancy will continue to drive experiments for years to come.

Acknowledgments

This work was supported in part by the National Science Foundation, Grant Number PHY 96-00049.

References

[1] C. M. Will, *Was Einstein Right?* 2nd Ed. (Basic Books, New York, 1993).

[2] C. M. Will, *Theory and Experiment in Gravitational Physics*, 2nd Ed. (Cambridge University Press, Cambridge, 1993).

[3] C. M. Will, in *300 Years of Gravitation*, edited by S. W. Hawking and W. Israel, 80–127 (Cambridge University Press, Cambridge, 1987), p. 80.

[4] C. M. Will, Int. J. Mod. Phys. D **1**, 13 (1992).

[5] C. M. Will, in *General Relativity: Proceedings of the 46th Scottish Universities Summer School in Physics*, ed. G. S. Hall and J. R. Pulham (Institute of Physics Publishing, Bristol, 1996), p. 239.

[6] T. Damour, in *Proceedings of the 5th Hellenic School on Elementary Particle Physics*, in press (gr-qc/9606079).

[7] Y. Su, B. R. Heckel, E. G. Adelberger, J. H. Gundlach, M. Harris, G. L. Smith, and H. E. Swanson, Phys. Rev. D **50**, 3614 (1994).

[8] E. Fischbach, G. T. Gillies, D. E. Krause, J. G. Schwan, and C. Talmadge, Metrologia **29**, 213 (1992).

[9] A. Godone, C. Novero, and P. Tavella, Phys. Rev. D **51**, 319 (1995).

[10] J. D. Prestage, R. L. Tjoelker, and L. Maleki, Phys. Rev. Lett. **74**, 3511 (1995).

[11] F. Dyson and T. Damour, Nucl. Phys. **B480**, 37 (1996) (hep-ph/9606486).

[12] M. J. Drinkwater, J. K. Webb, J. D. Barrow, and V. V. Flambaum, Mon. Not. R. Astron. Soc. **295**, 457 (1998) (astro-ph/9711290).

[13] R. A. Malaney and G. Mathews, Phys. Rep. **229**, 147 (1993); H. Reeves, Rev. Mod. Phys. **66**, 193 (1994).

[14] C. M. Will, Phys. Rev. D **45** 403 (1992).

[15] E. Fischbach and C. Talmadge, Nature **356**, 207 (1992).

[16] C. M. Will, Sky and Telescope **80**, 472 (1990).

[17] E. G. Adelberger, B. R. Heckel, C. W. Stubbs, and W. F. Rogers, Ann. Rev. Nucl. Particle Sci. **41**, 269 (1991).

[18] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman Publishing Co., San Francisco, 1973).

[19] S. W. Weinberg, *Gravitation and Cosmology* (Wiley, New York, 1992).

[20] T. Damour and G. Esposito-Farèse, Class. Quantum Grav. **9**, 2093 (1992).

[21] T. Damour and K. Nordtvedt, Phys. Rev. Lett. **70**, 2217 (1993); Phys. Rev. D **48**, 3436 (1993).

[22] P. J. Steinhardt and C. M. Will, Phys. Rev. D **52**, 628 (1995) (astro-ph/9409041).

[23] T. Damour and G. Esposito-Farèse, Phys.Rev. D **54**, 1474 (1996) (gr-qc/9602056).

[24] T. M. Eubanks, D. N. Matsakis, J. O. Martin, B. A. Archinal, D. D. McCarthy, S. A. Klioner, S. Shapiro, and I. I. Shapiro, Bull. Am. Phys. Soc., Abstract #K 11.05 (1997), unpublished

[25] M. Froeschlé, F. Mignard, and F. Arenou, in *Proceedings of the Hipparcos Venice 1997 Symposium*, in press.

[26] K. Nordtvedt, Icarus **114**, 51 (1995).

[27] T. Damour and D. Vokrouhlický, Phys. Rev. D, **53**, 6740 (1996) (gr-qc/9606076).

[28] J. O. Dickey, P. L. Bender, J. E. Faller, X. X. Newhall, R. L. Ricklefs, J. G. Ries, P. J. Shelus, C. Veillet, A. L. Whipple, J. R. Wiant, J. G. Williams, and C. F. Yoder, Science **265**, 482 (1994); J. G. Williams, X. X. Newhall, and J. O. Dickey, Phys. Rev. D**53**, 6730 (1996).

[29] J. F. Bell and T. Damour, Class. Quantum Grav. **13**, 3121 (1996) (gr-qc/9606062).

[30] J. F. Bell, F. Camilo, T. Damour, Astrophys. J. **464**, 857 (1996) (astro-ph/9512100).

[31] C. M. Will, Astrophys. J. **204**, 224 (1976).

[32] C. M. Will, Astrophys. J. Lett. **393**, L59 (1992).

[33] For information about the project, see http://einstein.stanford.edu/.

[34] I. Ciufolini, E. Pavlis, F. Chieppa, E. Fernandes-Vieira, and J. Pirez-Mercader, Science **279**, 2100 (1998).

[35] K. S. Thorne, in *Proceedings of the Snowmass '95 Summer Study on Particle and Nuclear Astrophysics and Cosmology*, edited by E. W. Kolb and R. Peccei (World Scientific, Singapore, 1995), p. 398 (gr-qc/9506086).

[36] R. A. Hulse, Rev. Mod. Phys. **66**, 699 (1994).

[37] J. H. Taylor, Rev. Mod. Phys. **66**, 711 (1994).

[38] T. Damour, 1987, in *300 Years of Gravitation*, eds. Hawking S. W. and Israel W., 128–198 (Cambridge University Press, Cambridge).

[39] M. Kramer, Astrophys. J., in press (astro-ph/9808127).

[40] A. Wolszczan, Class. Quantum Grav. **11**, A227 (1994).

[41] I. H. Stairs, Z. Arzoumanian, F. Camilio, A. G. Lyne, D. J. Nice, J. H. Taylor, S. E. Thorsett, and A. Wolszczan, Astrophys. J., in press (astro-ph/9712296).

[42] L. S. Finn and D. F. Chernoff, Phys. Rev. D **47**, 2198 (1993) (gr-qc/9301003); C. Cutler and É. E. Flanagan, Phys. Rev. D **49**, 2658 (1994) (gr-qc/9402014); E. Poisson and C. M. Will, Phys. Rev. D **52**, 848 (1995) (gr-qc/9502040).

[43] This convention holds sway, despite the fact that pure Newtonian gravity predicts no gravitational radiation. It is often the source of confusion, since the energy carried by the lowest-order "Newtonian" quadrupole radiation manifests itself in a (post)$^{5/2}$-Newtonian, or $O[(v/c)^5]$ correction in the equation of motion.

[44] M. Walker and C. M. Will, Phys. Rev. Lett. **45**, 1741 (1980); J. L. Anderson, Phys. Rev. Lett. **45**, 1745 (1980); T. Damour, Phys. Rev. Lett. **51**, 1019 (1983); D. Christodoulou and B. G. Schmidt, Commun. Math. Phys. **68**, 275 (1979); R. A. Isaacson, J. S. Welling, and J. Winicour, Phys. Rev. Lett. **53**, 1870 (1984).

[45] L. Blanchet, T. Damour, Phil. Trans. R. Soc. London **A320**, 379 (1986); L. Blanchet and T. Damour, Phys. Rev. D **37**, 1410 (1988); L. Blanchet and T. Damour, Ann. Inst. H. Poincaré (Phys. Theorique) **50**, 377 (1989); T. Damour and B. R. Iyer, Ann. Inst. H. Poincaré (Phys. Theorique) **54**, 115 (1991); L. Blanchet, Phys. Rev. D **51**, 2559 (1995) (gr-qc/9501030); L. Blanchet and T. Damour, Phys. Rev. D **46**, 4304 (1992).

[46] K. S. Thorne, Rev. Mod. Phys. **52**, 299 (1980).

[47] R. Epstein and R. V. Wagoner, Astrophys. J. **197**, 717 (1975) (EW).

[48] E. Poisson, Phys. Rev. D **47**, 1497 (1993); E. Poisson and M. Sasaki, Phys. Rev. D **51**, 5753 (1995) (gr-qc/9412027); C. Cutler, L. S. Finn, E. Poisson, and G. J. Sussman, Phys. Rev. D **47**, 1511 (1993); M. Sasaki, Prog. Theor. Phys. **92**, 17 (1994) (gr-qc/9402042); H. Tagoshi and M. Sasaki, Prog. Theor. Phys. **92**, 745 (1994) (gr-qc/9405062); T. Tanaka, H. Tagoshi, and M. Sasaki, Prog. Theor. Phys. **96**, 1087 (1996) (gr-qc/9701050).

[49] R. V. Wagoner and C. M. Will, Astrophys. J. **210**, 764 (1976).

[50] A. G. Wiseman, Phys. Rev. D **46** 1517 (1992).

[51] C. M. Will and A. G. Wiseman, Phys. Rev. D **54**, 4813 (1996) (gr-qc/9608012).

[52] A. G. Wiseman and C. M. Will, Phys. Rev. D **44**, R2945 (1991).

[53] C. M. Will and M. E. Pati, papers in preparation.

[54] L. Blanchet, T. Damour, and B. R. Iyer, Phys. Rev. D **51**, 5360 (1995) (gr-qc/9501029).

[55] L. E. Kidder, C. M. Will, and A. G. Wiseman, Phys. Rev. D **47**, R4183 (1993) (gr-qc/9211025); L. E. Kidder, Phys. Rev. D **52**, 821 (1995) (gr-qc/9506022).

[56] L. Blanchet, T. Damour, B. R. Iyer, C. M. Will, and A. G. Wiseman, Phys. Rev. Lett. **74**, 3515 (1995) (gr-qc/9501027).

[57] L. Blanchet, B. R. Iyer, C. M. Will, and A. G. Wiseman, Class. Quantum Grav. **13**, 575 (1996) (gr-qc/9602024).

[58] See http://www.ligo.caltech.edu/LIGO_web/Collaboration/manual.pdf.

[59] J. A. Lobo, Phys. Rev. D**52**, 591 (1995).

[60] M. Visser, Gen. Relativ. Gravit., in press (gr-qc/9705051).

[61] C. M. Will, Phys. Rev. D**57**, 2061 (1998) (gr-qc/9709011).

[62] C. Talmadge, J.-P. Berthias, R. W. Hellings, and E. M. Standish, Phys. Rev. Lett. **61**, 1159 (1988).

[63] C. Cutler, T. A. Apostolatos, L. Bildsten, L. S. Finn, É. E. Flanagan, D. Kennefick, D. M. Marković, A. Ori, E. Poisson, G. J. Sussman, and K. S. Thorne, Phys. Rev. Lett. **70**, 2984 (1993) (astro-ph/9208005).

[64] C. M. Will, Phys. Rev. D**50** 6058 (1994) (gr-qc/9406022).

[65] L. Blanchet and B. S. Sathyaprakash, Class. Quantum Grav. **11**, 2807 (1994); Phys. Rev. Lett. **74**, 1067 (1995).

[66] F. D. Ryan, Phys. Rev. D**52**, 5707 (1995); E. Poisson, Phys. Rev. D**54**, 5939 (1996) (gr-qc/9606024).

X-RAY EMISSION FROM COMPACT SOURCES

Lynn Cominsky*
Department of Physics and Astronomy
Sonoma State University, Rohnert Park, CA 94928

and

Stanford Linear Accelerator Center
Stanford University, Stanford, CA 94309

ABSTRACT

This paper presents a review of the physical parameters of neutron stars and black holes that have been derived from X-ray observations. I then explain how these physical parameters can be used to learn about the extreme conditions occurring in regions of strong gravity, and present some recent evidence for relativistic effects seen in these systems. A glossary of commonly used terms and a short tutorial on the names of X-ray sources are also included.

*Supported by NSF grant PHYS-9722126 and DOE Contract DE-AC03-76SF00515.1

© 1998 by Lynn Cominsky.

1 Overview

X-ray emission from compact objects was discovered in 1962,[1] with the detection of what we now know as a neutron star in the low mass binary, Sco X-1. In the past 35 years, the field has matured, and satellite based X-ray observations of neutron star and black hole binaries are conducted by scientists from many different nations, and with many different instruments. Recently, the increased collecting area and high time resolution of satellites such as NASA's Rossi X-ray Timing Explorer have produced data that have broken new ground in allowing high-energy astrophysicists to make physical measurements of compact binaries. Unobtainable in earth-bound laboratories, X-ray satellite observations remain the best way to study the effects of strong field gravity and to test the predictions of General Relativity in this regime.

In this paper, I will review the basic physics which produces X-ray emission from neutron stars and black holes, focusing primarily on accretion, which is gravitationally-driven mass transfer onto a compact object. I will then review the major observational categories of accreting neutron star and black hole binaries, and discuss the derivation from the data of important physical parameters including spin, mass, radius, binary orbital elements and the surface magnetic field strength and equation of state for neutron stars. Finally, I will close with a brief look at some new evidence and predictions of relativistic effects seen in these systems. A glossary of frequently used astronomical jargon is included, as well as a short tutorial on the names of X-ray sources.

2 X-ray Emission Mechanisms

In this section, I review the different theoretical mechanisms which are believed to produce X-ray emission in systems containing either neutron stars or black holes. Non-gravitationally driven emission mechanisms are reviewed so that they can be contrasted with accretion, the primary gravitationally-driven mechanism that produces high-energy photons.

2.1 Rotationally Powered Pulsars

Rotationally powered pulsars were discovered by Jocelyn Bell and Anthony Hewish in 1964,[2] and recognized as rotating neutron stars in work by Gold.[3] They have been extensively studied at radio wavelengths, where $\sim$ 1000 such objects are detected (for a comprehensive review, see *Pulsar Astronomy*[4]). The radio emissions are the result of a pulsar wind of relativistic plasma, powered by the loss of rotational kinetic energy as the pulsar's spin slows, and beamed along the open magnetic dipole field lines that are offset from the rotational axis of the star. The luminosity that is detected is consistent with the loss of rotational kinetic energy, $\dot{E} = I\omega\dot{\omega}$, where I is the moment of inertia of the object, and ω is its angular velocity.

At the highest gamma-ray energies, seven rotationally powered pulsars are seen. Figure 1 shows the multi-wavelength lightcurves of the seven known gamma-ray pulsars. Theoretical speculation as to the origin of the high-energy radiation and its phase relation to the emission at lower energies has settled into two main types of models: those in which the high-energy radiation is emitted from the polar caps (*i.e.*, where the open magnetic field lines emanate[5]) or those in which it arises from the "outer gaps", where the last closed field line intersects the speed of light cylinder[6] (see Figure 2).

A very successful recent outer gap model by Romani and Yadigaroglu[7] provides calculations of the rotating neutron star's magnetosphere and accounts for the origins of the different types of emission and their phase relationships, as a function of only two important parameters: the angle between the rotation axis and the magnetic dipole axis, and the angle to our line-of-sight. Figure 3 shows a typical calculation from their work.

In all cases, the neutron stars which have been carefully studied have masses consistent with 1.4 $M_\odot$, radii of about 10 km, and therefore moments of inertia $I = 10^{45}$ g cm^2. Their spin periods vary from milliseconds to tens of seconds. Changes in the pulse period can be used to derive the pulsar's spin-down age as well as its magnetic field. The spin-down age is defined as:

$$\tau = \frac{P}{2\dot{P}} \tag{1}$$

while the magnetic field can be found from:

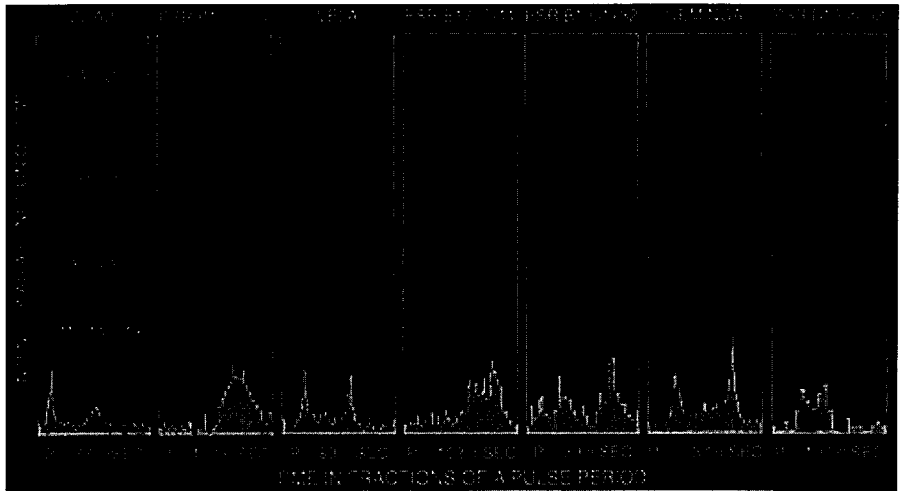


Figure 1: Multi-wavelength lightcurves showing the emission as a function of pulse phase for the seven rotational powered neutron star pulsars which are detected by the EGRET experiment on board NASA's Compton Gamma-ray Observatory.[8]

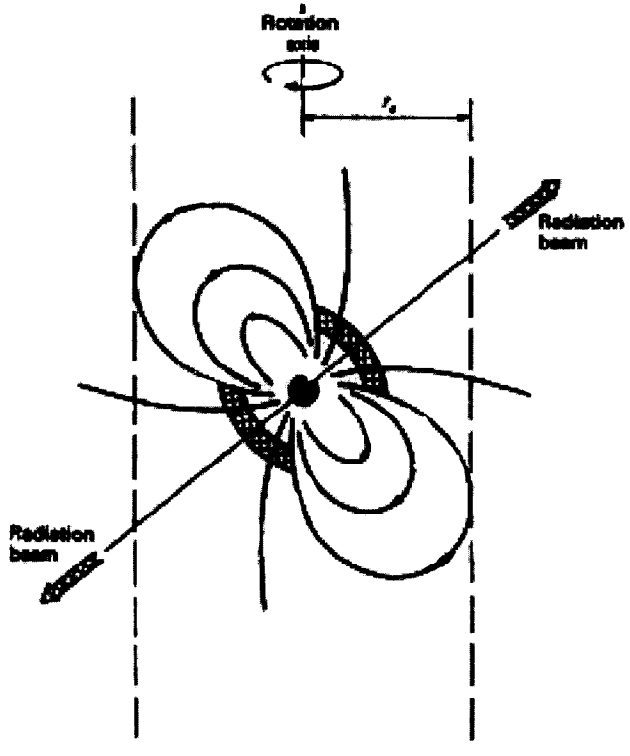


Figure 2: A schematic showing the relationship between the spin and magnetic field axes for a rotating neutron star. The polar caps are the cross hatched regions. The outer gaps are the portions of the outermost closed field lines which extend from the polar cap to the speed of light cylinder, which is denoted by the dashed lines.[4]

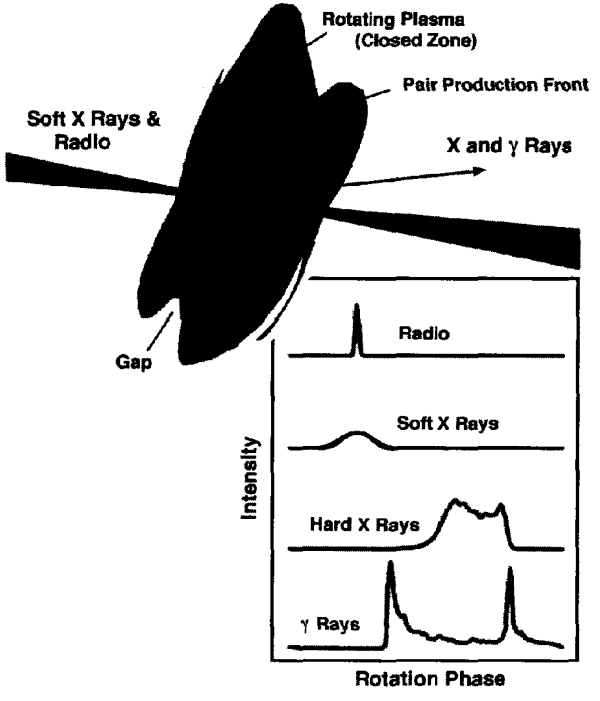


Figure 3: Neutron star magnetosphere showing the correlation between emitting surfaces and wavelength as predicted by the outer gap model calculations of Romani and Yadigaroglu.[7]

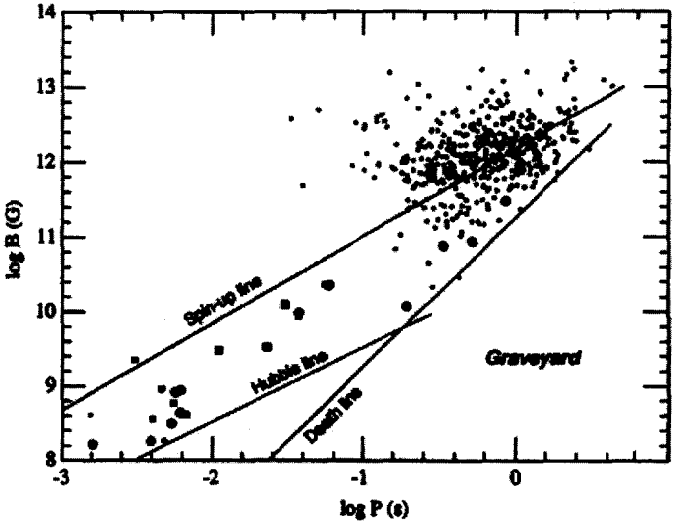


Figure 4: Derived magnetic fields for many radio pulsars as a function of pulse period.[9]

$$B = 3.2 \times 10^{19} \, (\dot{P}P)^{1/2} \text{ Gauss.} \quad (2)$$

Figure 4 shows the spin periods for many of the well-studied radio pulsars as a function of their derived magnetic field strengths. The millisecond pulsars are believed to be "recycled", *i.e.*, spun up via accretion from a companion star, after a previous evolutionary phase in which they were spinning down as isolated pulsars. The spin-up line noted on the figure indicates the minimum period that would result from accretion at the Eddington limit, while the Hubble line corresponds to a spin-down age roughly as old as the universe, or $\sim 10^{10}$ years. The death line delimits the region where the polar cap voltage is so low that pulsations are likely to switch off, and thus will not be observable.

Although X-ray luminosities in these systems can be quite high, often comparable to those produced by accretion, the steady spindown of the pulse period is in marked contrast to the much more variable spin rate behavior in the accret-

ing binaries. And, for those rotationally powered neutron stars which are not in binaries, no companion stars can be detected.

2.2 Shock Emission From Pulsar Winds

Recently, several radio pulsars have been found with non-compact companion stars. These systems, the "recycled" pulsars, are rotating much more rapidly than conventional non-binary radio pulsars, having been spun up by accretion from their companions. They are believed to be much older systems, with magnetic fields that have decayed to $\sim 10^8$ Gauss.

Binary radio pulsars are seen with different types of non-compact companions. First to be discovered at high energies were very small systems, where the effects of the pulsar wind on the companion's outer layers caused the rapid ablation of the small star. For example, in the case of the "Black Widow" pulsar[10,11] PSR 1957+20, the companion is only $\sim 0.02 M_\odot$, and is being "eaten" by the $\sim 10^{35}$ ergs s^{-1} wind from the 1.6 ms pulsar. Due to the shock-heating of the outer layers of the small star, about half of this emission occurs at higher energies (*i.e.*, X- and gamma-rays).

Cominsky, Roberts and Johnston[12] discovered X-ray emission from PSR B1259-63, a 69 ms radio pulsar[13] that orbits a Be-star (SS 2883). This unusual system has a 3.4 year, very eccentric orbit, and has now been observed throughout its orbital cycle. The high-energy emission seen from this system is most likely due to shock-heating of the Be-star's wind by that of the pulsar.[14,15] Recent work has shown that the power-law spectral index of the high-energy emission remains almost constant throughout the long orbital cycle; it is only the normalization that changes as the two stars separate.[16] Figure 5 compares the energy spectrum seen at periastron (closest approach) to that at apastron (furthest separation) in the PSR B1259-63 system. The X-ray luminosity seen in this system varies by a factor of ten from $\sim 10^{33} - 10^{34}$ ergs s^{-1} as the neutron star travels from apastron to periastron.

2.3 Magnetars (Soft Gamma Repeaters)

The presence of extremely intense magnetic fields in a class of isolated neutron stars known as magnetars, and their association with an observational class of objects known as the Soft Gamma-ray Repeaters (SGR) has been suggested[17,18]

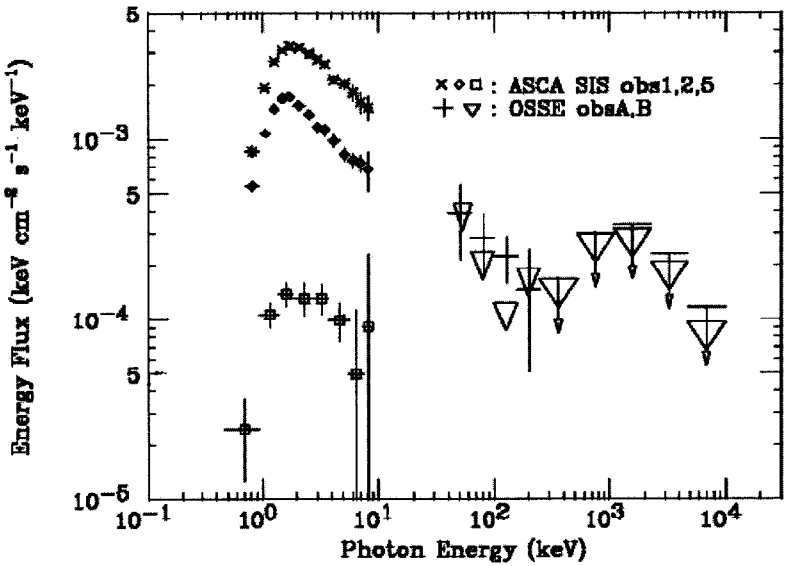


Figure 5: The spectrum of the Be-binary pulsar system PSR B1259-63 showing the changes in intensity as a function of the separation between the two stars. The two upper curves at low energies are from periastron observations,[15] while the lower curve is at apastron. X-ray emission up to 100 keV is detectable at periastron,[19] but apastron observations have resulted in only upper limits.[16]

by Thompson and Duncan. Unlike conventional gamma-ray bursts, the SGRs show repeated bursts of soft gamma-rays over many years from the same location. There are now[20] four SGR: three in our galaxy and one in the Large Magellanic Cloud. All are associated with young supernova remnants. The key difference between magnetars and rotationally powered pulsars is that the energy source powering the magnetars is derived from the rapidly decaying magnetic field, which heats the core and surface of the star.

As an example, detailed observations[21] of SGR1806-20 have shown 7.47 second hard X-ray pulsations with a spindown rate of 2.6×10^{-3} yr^{-1}, indicating a pulsar spindown age of $\sim$ 1500 – 8000 years, and a magnetic field of $\sim 2 - 8 \times 10^{14}$ Gauss. Peak burst fluences seen from SGR1806-20 are $\sim 10^{41}$ ergs s^{-1}, $\sim 10^3$ times the Eddington limit. These super-Eddington luminosities are possible because of the magnetic suppression of the Thomson cross-section for electron scattering. Quiescent X-ray luminosities following the bursts are $\sim 10^{35} - 10^{36}$ ergs s^{-1}, and are modulated at the neutron star spin period. Bursts from SGRs are easily distinguished from X-ray bursts and other gamma-ray bursts: SGR blackbody burst temperatures of 9 keV are about three times the highest temperatures seen in thermonuclear (Type I) X-ray bursts, but they are still much softer than non-repeating gamma-ray bursts.

In the magnetar models,[17,18] the sudden cracking of the neutron star crust leads to a rapid injection of Alfvén waves into the magnetosphere, particle acceleration and the formation of a optically thick pair plasma. Magnetic confinement of this plasma can produce the observed spectral similarities between bursts of greatly varying fluence. The repeated bursts from these objects are indications of starquakes caused by magnetic stresses in the neutron star crust.

2.4 Accretion In Neutron Star and Black Hole Binaries

In accreting X-ray binaries, the X-ray luminosity originates in $\sim 10^7 - 10^8$ K plasma near the neutron star's surface or from the inner accretion disk around the compact object. Accretion is extremely efficient; it liberates approximately 200 MeV per nucleon, compared to only 7 MeV per nucleon for the process of fusing hydrogen into helium. The accretion luminosity observed is given by:

$$L_{accr} = \frac{\eta \, G\dot{M}M}{R} \sim 1.2 \times 10^{38} \, \eta \text{ ergs s}^{-1} \qquad (3)$$

assuming that $M = 1.4$ M$_\odot$, $R = 10$ km, $\dot{M} = 10^{-8}$ M$_\odot$ yr^{-1}, and where η is an efficiency factor for the conversion of gravitational potential energy into radiated luminosity. The efficiency differs for neutron stars, black holes, and the assumed General Relativistic metric. Typical accretion luminosities observed in these systems vary from $10^{36} - 10^{38}$ ergs s^{-1}. Comprehensive references for this section can be found in *X-ray Binaries*[22] and *Accretion Power in Astrophysics.*[23]

The inner accretion flow occurs in a region of strong field gravity. For example, the potential well of a neutron star is 0.2 c^2 at 10 km for 1.4 M$_\odot$, and the characteristic velocity of the material in this region, $(\frac{GM}{R})^{1/2} \sim$0.5c.

In studying this inner accretion flow, one must take into account the very rapid dynamical timescales for matter to move through the emitting region:

$$\tau_{dyn} = (\frac{r^3}{G\,M})^{1/2}. \qquad (4)$$

This is about 0.1 ms for matter at $r = 10$ km, and 2 ms at $r = 100$ km. The typical orbital period of circulating matter,

$$P_{orb} = 2\pi\tau_{dyn} \sim 1 \text{ ms}. \qquad (5)$$

The two main accretion mechanisms are Roche lobe overflow, which most often occurs in low-mass binaries, and stellar wind capture, which is common for high-mass binaries with super-giant companions. A third mechanism for the initiation of mass transfer is often seen in systems with Be-star companions, as they are known to spontaneously eject matter in their equatorial regions due to their rapid rotation (even when they are not in binaries with compact objects).

2.4.1 Roche-lobe Overflow

Figure 6 shows the geometry for Roche lobe overflow. In this figure, the Roche lobe equipotential surface resembles a figure "8". The companion star completely fills its Roche lobe as the star evolves, and material in its expanding envelope is transferred through the inner Lagrangian (crossover) point to the vicinity of the compact object.

The compact object does not fill its Roche lobe; rather the accreted material swirls around it, interacting viscously to form an accretion disk (with a typical radius $R_{disk} \approx 0.8 R_{lobe}$). The intersection point of the incident stream with the

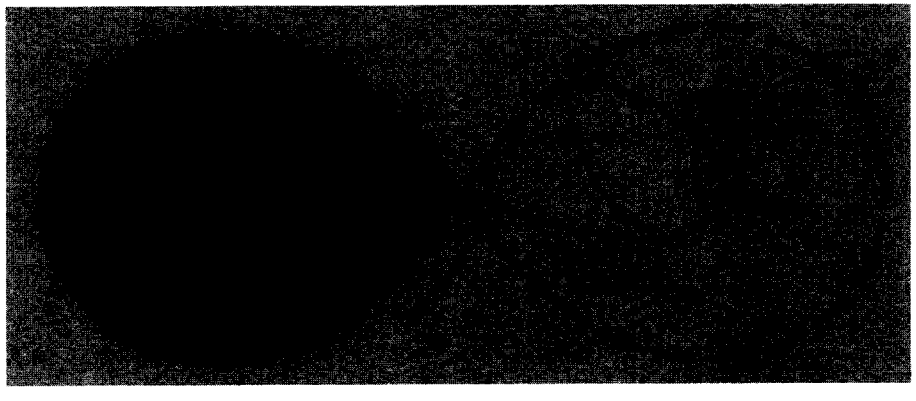


Figure 6: Roche lobe geometry, showing the incident mass stream through the inner Lagrangian point and the intersection of the returning stream with the newly arriving material. The low mass companion on the left hand side completely fills its Roche lobe, while the position of the compact object is denoted by a plus sign (+) inside the right hand lobe.[24]

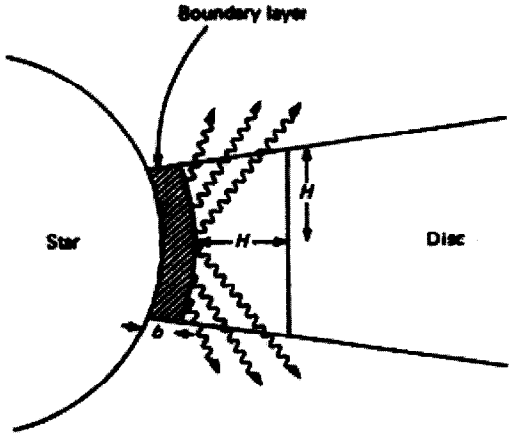


Figure 7: Accretion in a weak ($\sim 10^8$ Gauss) magnetic field. Note that the accretion disk extends to the neutron star surface.[23]

returning stream is a region of great turbulence, which can have interesting observational properties.

If the magnetic field of an accreting neutron star is relatively weak ($\sim 10^8$ Gauss), the accretion disk can extend all the way to its surface. The material then shocks to make a boundary layer, which radiates at X-ray wavelengths. This is illustrated in Figure 7.

For the much stronger magnetic fields (B $\sim 10^{12}$ Gauss) which are characteristic of pulsars, any accretion disk which is present can extend only to the surface of the magnetosphere. Matter must then travel along the field lines to the vicinity of the polar cap. The infalling material then shocks above the polar cap, producing observable high-energy emission (see Figure 8).

2.4.2 Stellar Wind Capture

Figure 9 shows a typical scenario for accretion due to the capture of stellar wind material by a compact object. Early-type, massive companion stars evolve by radiating a wind of mostly hydrogen plasma at a rate of approximately 10^{-6} $M_\odot$ yr^{-1}. A small fraction of this wind, typically 1%, can be captured as it passes

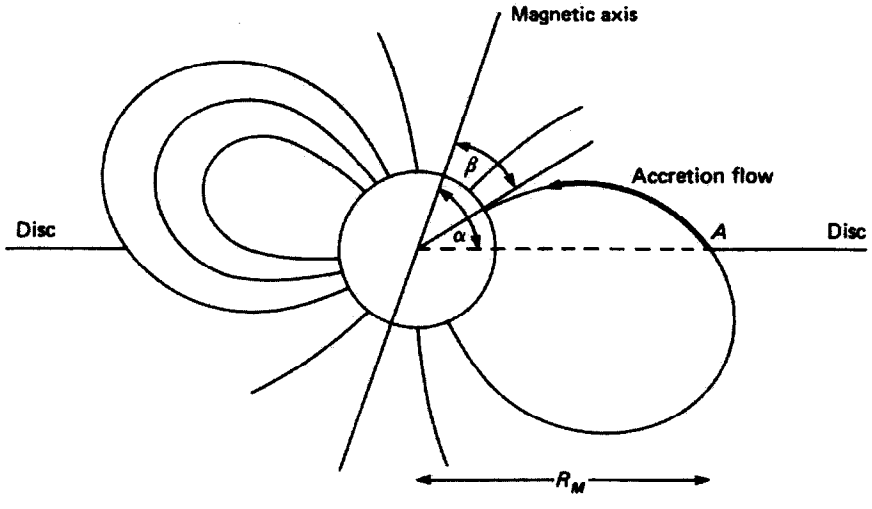


Figure 8: Accretion in a strong ($\sim 10^{12}$ Gauss) magnetic field. Note that the accretion disk is held off the neutron star surface by the centrifugal barrier formed by the rotating magnetosphere.[23]

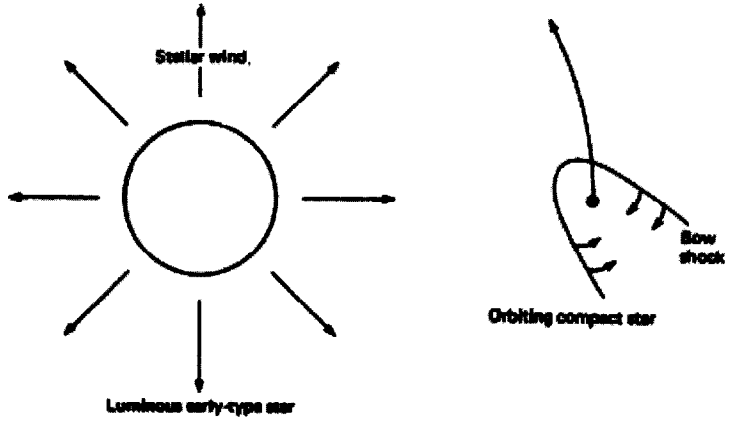


Figure 9: Accretion from a stellar wind.[23]

the compact object. A bow shock forms around the compact object, which can produce observable X-ray emission.[24]

2.5 Thermonuclear Reactions

As the matter accretes onto the neutron star surface, it settles into layers in which different elements undergo thermonuclear burning. Figure 10 illustrates the typical shell structure for neutron stars with relatively weak magnetic fields. In this case, the accretion flow can spread out over the entire neutron star surface, rather than being confined to the polar cap region by a much stronger magnetic field (as in the case for pulsars).

Following the discovery[25,26] of X-ray bursts in 1976, considerable theoretical work was done to account for the luminosity and spectral profiles of these sporadically occurring events. It is now well established that X-ray bursts are due to unstable thermonuclear burning in the surface layers of neutron stars.[27,28]

Three nuclear burning regimes have been defined, based on the mass accretion rate onto the neutron star surface.[29] For accretion rates in excess of 2.6×10^{-8} $M_\odot$ yr^{-1}, both the hydrogen and helium shells will burn stably; thus

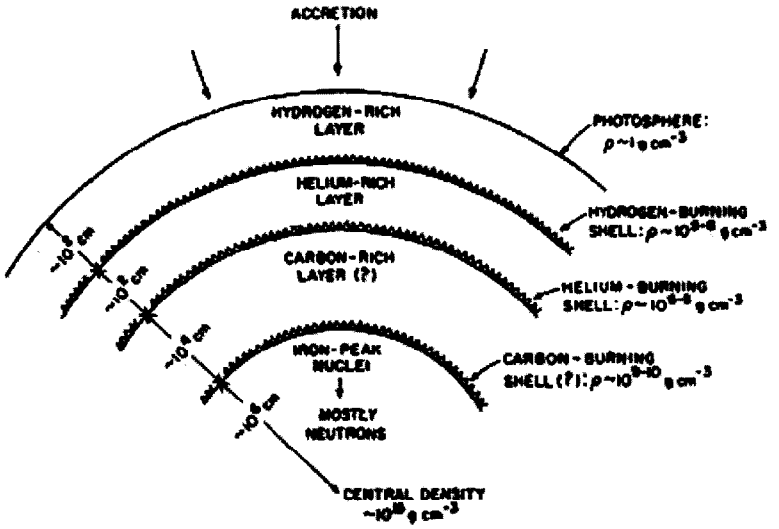


Figure 10: *Structure of a neutron star, showing the different elemental burning shells.*[30]

no bursts are seen. This is supported by considerable observational data in which the bursts are seen to grow weaker and then stop as the underlying, persistent, accretion-driven luminosity increases. For accretion rates between 2.6×10^{-8} $M_\odot$ yr^{-1} and 2×10^{-10} $M_\odot$ yr^{-1}, helium burns unstably, and runaway reactions cause helium flashes. Below about 2×10^{-10} $M_\odot$ yr^{-1}, hydrogen burns unstably, resulting in hydrogen flashes.

3 Accreting X-ray Binaries

In this section, I review the common characteristics of the major observational groups of neutron star and black hole binaries in which X-rays are produced by accretion onto the compact object.

3.1 X-ray Binary Pulsars

The primary distinguishing characteristic of X-ray binary pulsars is their extremely intense magnetic field, typically $\sim 10^{12}$ Gauss, which confines the bulk of the X-ray emission to the polar cap region. The offset magnetic dipole in these systems produces a unique pulse profile as the neutron star rotates on its spin axis. Typical pulse profiles are shown in Figure 11. Note that the pulse profile is quite variable as a function of energy. Although not totally understood, differences in pulse profile shape can be attributed to a combination of differences in viewing angle to the system, magnetic field strength, offset dipole angle and magnetic opacities along the line of sight.

There are over 30 well studied X-ray binary pulsar systems in our Galaxy. Their spin periods range from 0.025 to 840 seconds, and they are in orbits with binary periods ranging from 0.029 to 187.5 days.[31,32] Companion stars in these systems span the widest range of possible types, from stars much less massive than our Sun to supergiants. A prominent sub-group among the X-ray binary pulsars has Be-star companions. These systems tend to be in wider, often eccentric orbits and to show transient episodes of accretion, most probably triggered by sudden mass outflows by the Be-companion. Typical X-ray luminosities for these systems range from $10^{36} - 10^{38}$ ergs s^{-1}.

Studies of the relation between the spin and orbital periods in these systems have led to a correlation in only one type of sub-group, the Be-companion systems.

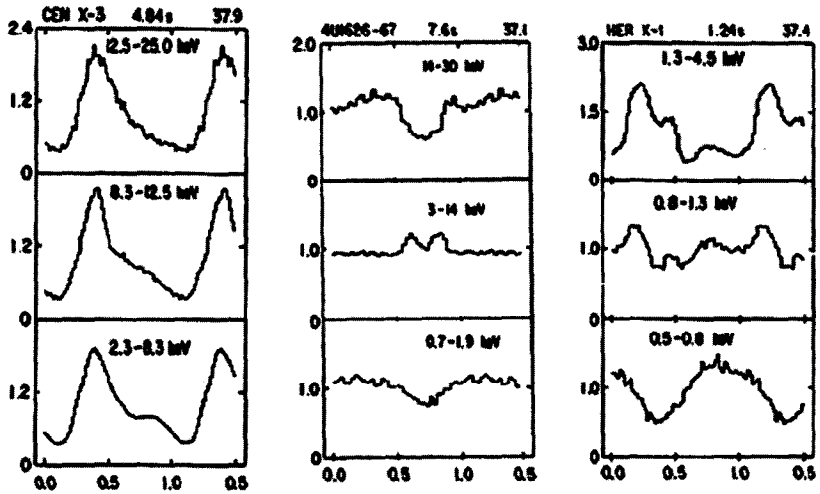


Figure 11: Pulse profiles as a function of phase and energy for three different X-ray binary pulsars.[33]

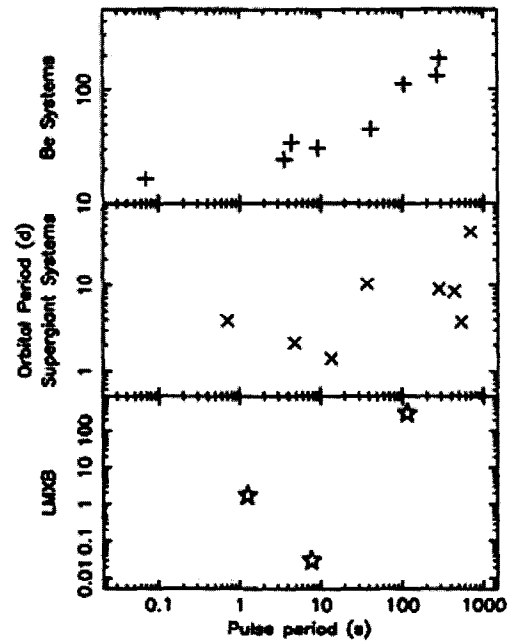


Figure 12: Orbital period as a function of pulse period for different types of accreting x-ray binaries. Note the linear relation for the Be-star systems.[31]

As can be seen in Figure 12, there is little correlation in the low-mass (Roche lobe overflow) or supergiant (stellar wind capture) systems. The roughly linear correlation in the Be-companion systems (often called the Corbet relation[34]) is taken as evidence that these systems have reached a rough equilibrium between the spin period and the orbital separation.

3.2 X-ray Bursters

Because of the presence of X-ray bursts, the magnetic field in these systems is believed to be $\sim 10^8$ Gauss. Models have shown that higher field strengths quench the bursting behavior, by narrowly focusing the accretion onto the polar cap

region, and thus greatly raising the local accretion rate (and the observed luminosity) to the range in which nuclear burning is stable.[35] Accretion occurs over a greater portion of the surface area in neutron stars with relatively weak fields, although the uneven channeling of accreted matter by the flow from the accretion disk can lead to hot spots which trigger the flashes, and to continuous low level fluctuations in the burning rate which can propagate around the star, producing the so-called "rings of fire".[36]

There are over 15 well-studied[37] bursting systems; all have low-mass companions, and most are located either in globular clusters or near the center of our Galaxy. These positional associations have led to the belief that X-ray bursters are an older population of object than the pulsars. Although their magnetic fields may once have been strong enough to cause pulsation, it is possible that the fields have decayed away in the intervening $\sim 10^{10}$ years. Typical luminosities for the persistent, accretion-driven emission in these systems are $\sim 10^{36}$ ergs s^{-1}, while the burst peaks are often about 100 times as luminous.

Orbital periods in these systems have been determined with high confidence using X-ray data in the two systems which exhibit X-ray eclipses. In 1981, Cominsky and Wood[38] discovered 15 minute long eclipses, which recur every 7.1 hours, in the transient X-ray burst source MXB 1659-29. Later observations of the source EXO 0748-676 showed that it too had X-ray eclipses.[39] All of the orbital periods for the other bursters, however, have been derived from quasi-regularly occurring X-ray dips in intensity due to the turbulent area of the gas stream blocking the line of sight, or by optical observations of intensity modulation due to X-ray heating of the companion star and/or accretion disk. Known orbital periods in the bursting systems vary from 0.19 to 398 hours.

X-ray bursts are well described by thermal black-body spectra,[40] in which the radius is roughly constant as the temperature increases to around 3×10^7 K and then cools until the resulting luminosity is undetectable below the background (most of which stems from the persistent accretion luminosity). Some bursts are accompanied by a sharp jump in radius at the beginning,[41] before settling down to the $\sim$ 10 km radii more typical of neutron stars. This radius expansion is evidence that the burst flux has reached the Eddington limit (roughly 10^{38} ergs s^{-1}), and that the photosphere of the neutron star has temporarily inflated due to excess radiation pressure. Figure 13 shows a black-body fit to four energy channels of X-ray data for a burst from a globular cluster source, in which radius expansion is present. The radius expansion at the peaks of some bursts allows one to use the Eddington limit to derive distances to the bursters, *i.e.*, they act as "standard candles". Distances derived in this manner are consistent with those derived from other means, *e.g.*, by analyzing the constituents of the globular clusters, or by noting the distribution of sources around the Galactic center.[42]

The recent discovery[43] of quasi-periodic oscillations (QPO) at a period of about 3 ms (or half that) within X-ray bursts themselves, has greatly intrigued the high-energy astrophysical community. Since the burst emission is a direct indication of conditions on the surface of the star, these QPO are taken as the first direct evidence of the neutron star's spin, and are perhaps the result of one or two hot spots of emission rapidly rotating with the star.

There is also a second type of X-ray bursting pattern - the so-called "Rapid" or Type II bursts, which, until recently, were only seen from one object, the Rapid Burster[44] (MXB 1730-335). This type of pattern, in which thousands of closely spaced, non-thermal bursts are observed daily, is attributed to instabilities in the neutron star's magnetosphere. Episodes of rapid bursting are transient, occurring for only $\sim$months out of $\sim$years. However, the Rapid Burster also shows Type I, or thermonuclear flash bursts on occasion, indicating that the neutron star is accreting matter, although very unsteadily. The rapid bursts repeat in a pattern that has been likened to that of a Relaxation Oscillator: the fluence in each burst is proportional to the waiting time to the next burst.[45] This "$E - \Delta t$" relationship is illustrated in Figure 14.

In 1996, another object (GRO J1744-28) was discovered that displays a rapid bursting pattern with inter-burst intervals as short as $\sim$minutes.[46] However, unlike the Rapid Burster, GRO 1744-28 also pulses with regular 0.47 second pulsations.[47] The observed spinup and Doppler shifts of these pulses during its transient outburst indicates that it is accreting from a low-mass binary companion. Magnetic field strength estimates for the pulsar are $\sim 3 \times 10^{11}$ Gauss.[48]

In 1998, 2.5 ms pulsations were discovered[32] from the transient source SAX J1808.4-3658. This remarkable system, the fastest accreting X-ray pulsar now known, is in a $\sim$ 2 hour binary orbit[49] with a companion star that has less mass than 0.1 $M_\odot$. SAX J1808.4-3658 also emits Type I X-ray bursts.[50] Magnetic field estimates[51] for the neutron star in this system are $\sim 10^8$ Gauss leading to the conclusion that this low-mass X-ray binary system is a progenitor to the recycled millisecond pulsars. GRO1744-28 may be a similar evolutionary "missing link".

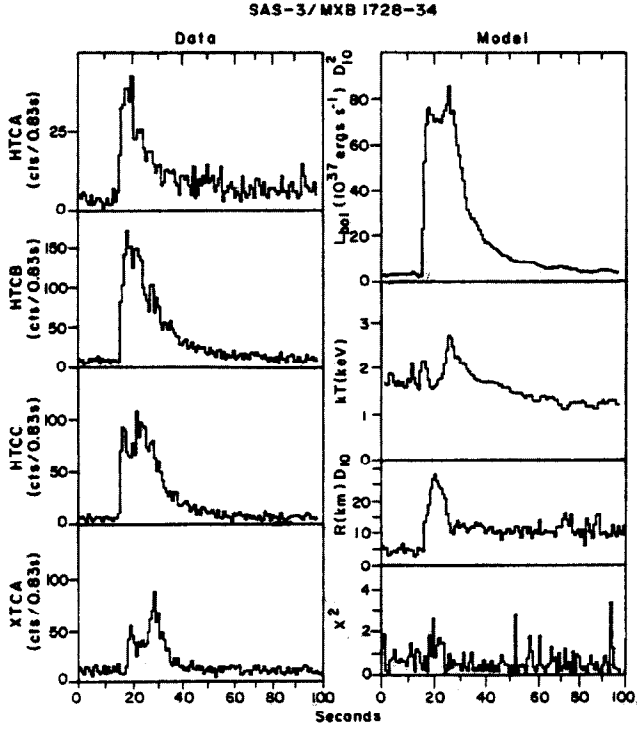


Figure 13: Blackbody fit for four X-ray energy channels. Note the jump in radius at the beginning of the burst as a result of Eddington-driven photospheric expansion.[52]

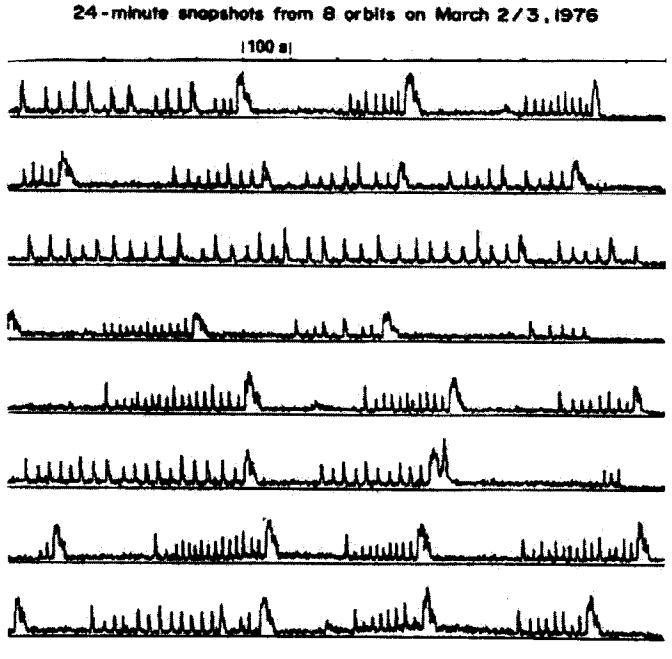


Figure 14: Type II bursts from the Rapid Burster, MXB 1730-335, showing the $E - \Delta t$ relationship, in which the fluence in a burst is proportional to the interval to the next burst.[53]

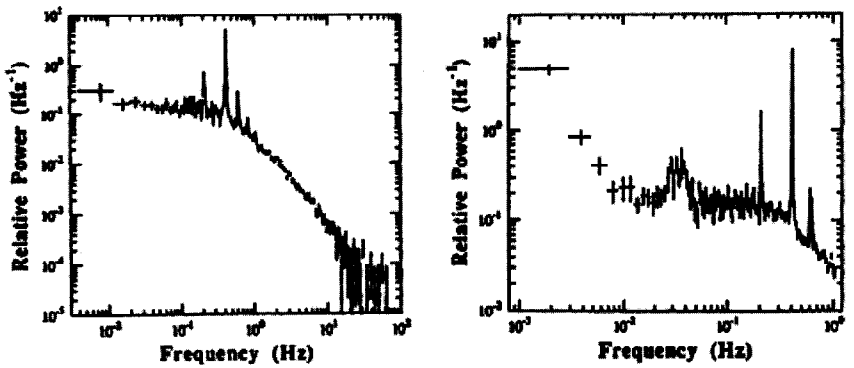


Figure 15: Power spectra from the X-ray binary pulsar Cen X-3 on two different occasions. Note that the spectrum on the right shows both the pulsing and QPO behavior, while the one on the left shows only the pulsing.[54]

3.3 Quasi-Periodic Oscillators

Unlike pulsars and bursters, quasi-periodically oscillating X-ray sources can be either neutron stars or black hole candidates.[55] Other QPO sources are the very luminous sources near the Galactic center, presumably neutron star binaries, which do not exhibit X-ray bursts as their mass accretion rate is too high. Both of the "microquasar" black hole candidates also show QPO.[56,57] Fourier spectra of pulsars often, but not always, exhibit QPO in addition to pulsations, which can be distinguished by the much-higher quality factors ($Q > 100,000$ vs. $100 - 1000$, see Figure 15).

When originally discovered[58] in observations using the EXOSAT observatory, the QPO ranged from 6 - 60 Hz. A typical example of QPO from a black hole candidate is shown in Figure 16. The most popular interpretation of the phenomena was the so-called "magnetospheric beat frequency" model.[59] In this model, the QPO emission arises from the beat between the Keplerian orbital frequency of matter circulating at the object's magnetosphere and the neutron star spin period. Many detailed studies were done of the QPO properties compared to the properties of the underlying persistent luminosity. Objects exhibiting QPO were

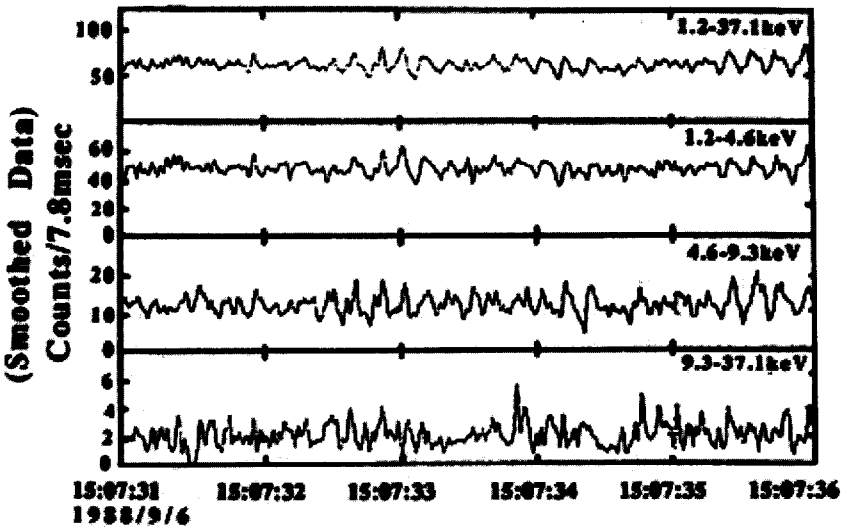


Figure 16: Time series from the black hole candidate GX 339-4, showing the 6 Hz quasi-periodic oscillations.[60]

classified as either "Z" or "atoll" sources,[61] depending on the pattern that they made in a plot of the normalized count rates in lower energy channels vs. those in higher energy channels, for a given intensity state ("color-color" diagram, see Figure 17). It is now established that the "Z" pattern is typical among the strong, non-bursting, non-pulsing galactic center sources, which have luminosities at approximately the Eddington limit, and magnetic fields of $\sim 10^9$ Gauss. In contrast, the "atoll" pattern is characteristic of the X-ray bursters, with luminosities of only about 1% of Eddington, and magnetic field strengths about ten times lower.

With the launch of the Rossi X-ray Timing Explorer (RXTE) at the end of 1995, a new observational window was opened, with both higher time resolution and greater sensitivity. Almost immediately, QPO began to be detected in the kHz range.[62] To date, there are at least 13 low-mass X-ray binaries which show QPO in the range 300-1200 Hz. Of these 13 sources, 4 are "Z" sources, and 9 are "atoll" sources. The kHz QPO, unlike the original QPO, typically occur in

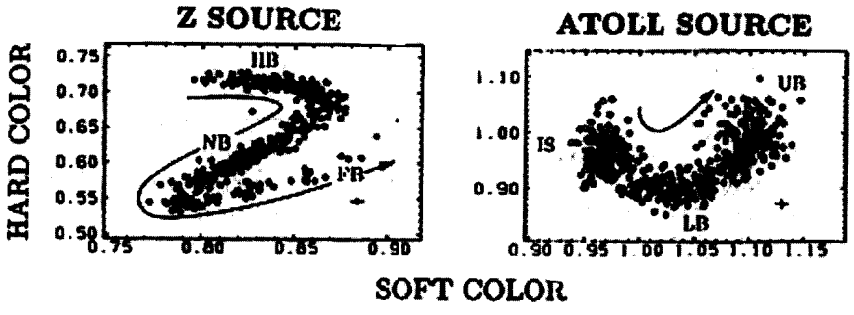


Figure 17: A comparison of the color-color diagrams for a Z vs. an atoll QPO source.[61]

pairs, known as the "twin peaks". Initial analysis of the frequency spacing of these peaks indicated that their separation was almost constant, at about 300 Hz. And, amazingly enough, this seemed to match the frequency (or half that, assuming that the bursts can emanate from two hot spots) seen in the bursts from the atoll sources.[63]

While the original magnetospheric beat frequency model can not be applied to the kHz QPO, it has been retooled to account for these new observations.[43] The frequency of the upper twin peak has now been associated with the Keplerian orbital frequency of matter in the inner accretion disk, perhaps just outside the marginally stable orbit or at the sonic point.[64] The lower twin peak then is interpreted as the beat frequency between this Keplerian orbital frequency and the neutron star's spin frequency. The 300 Hz difference frequency, which is often seen in the X-ray bursts, is that of the neutron star spin.

New data, however, and data from the strongest, and hence best signal-to-noise source (Sco X-1) have cast some doubt on this tidy interpretation. In Sco X-1, the value of the difference frequency is observed to be anticorrelated with the luminosity of the source.[65] It is not possible for the neutron star's spin to change on such rapid timescales, which leaves the door open for other possible models, such as those in which oscillating photon bubbles (PBO) in the accretion stream produce the QPO.[66] In addition, the claim of ~constant separation between the twin peaks in other, less luminous sources, may be an artifact of the statistics.[67] This is illustrated in Figure 18.

4 Physical Parameters of X-ray Binaries

In this section, I will review how X-ray data from Earth-orbiting satellites are used to determine the fundamental physical parameters of accreting neutron star and black hole binaries, and how these parameters change with time. Observations of compact X-ray sources provide information about the extreme conditions of temperature, density, magnetic field and strong gravity that are not available in earth-bound laboratory environments.

4.1 Neutron Star Spin

The spin of a neutron star is directly observable in Fourier analysis of the time series data obtained with sufficiently sensitive X-ray detectors. Extremely high-Q spikes ($> 100,000$) are seen in the power spectra of these objects, for even short stretches of data. For proportional counter detectors with areas in excess of about 1000 cm^2, individual pulses are easily seen, as pulsed fractions of the underlying accretion luminosity are typically $\sim 30\%$. A typical power spectrum, from only ~ 30 minutes of data, and less than 500 cm^2 of detecting area is shown in Figure 19. The fundamental, and the first two harmonics of the 3.6 second period are easily seen in this power spectrum for the transient Be-star system, 4U0115+63.

Changes in the spin rate from neutron stars were long believed to be well described by accretion torque theory.[68,69] Systems with accretion disks were believed to be prograde, and hence should spin up as they received angular momentum from the accreted material. In this standard accretion torque theory, the relation between the observed spin period P, the luminosity L of the X-ray source (which measures the amount of transferred matter) and the predicted change in the spin period $\dot{P}$ are related by:

$$\dot{P} \propto P^2 \, L^{6/7} \tag{6}$$

As shown in Figure 20, the theory appears to fit the long term trend towards spin up found in many systems. In this figure, the different curves refer to different

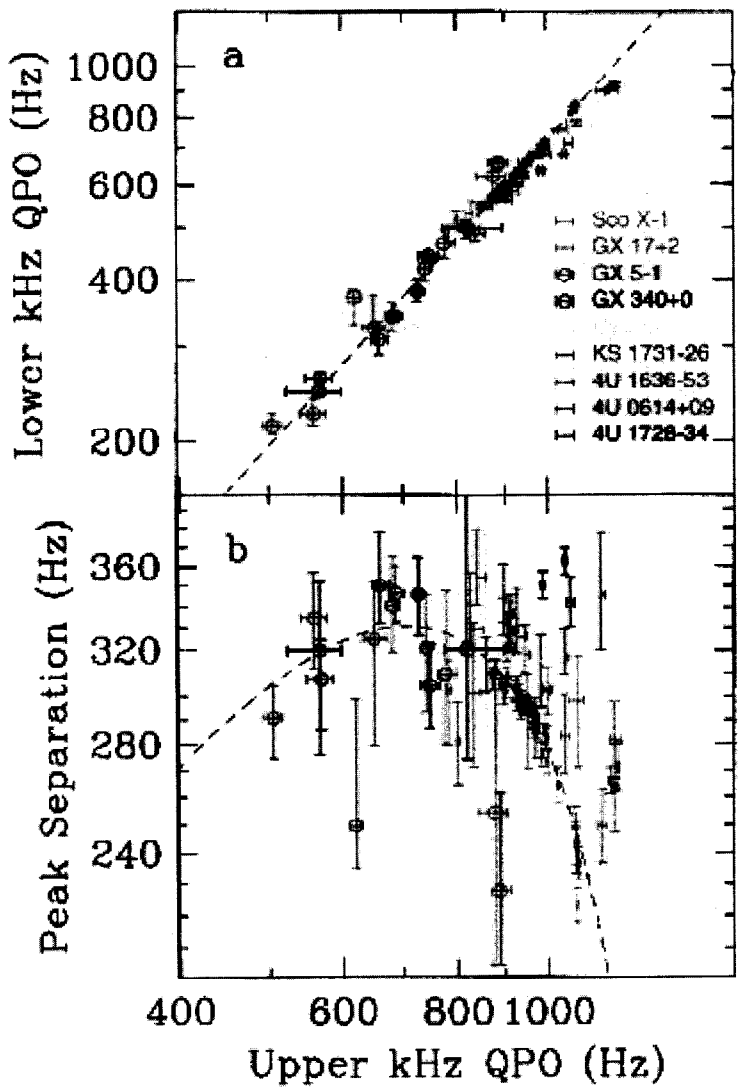


Figure 18: The frequency of the upper kHz QPO peak vs. the frequency of the lower QPO peak for many sources. Note that the peak t necessarily constant as a function of the upper kHz QPO peak frequency. This makes it difficult to associate the peak separation with the neutron star spin.[67]

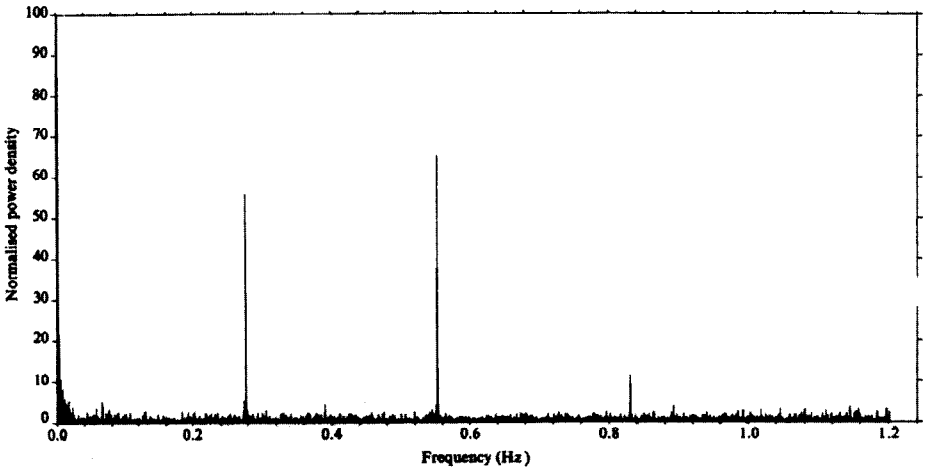


Fig. 1 Power density spectrum for 4U0115+63. The data were obtained with the SAS 3 Y-axis detectors during one satellite orbit near 10.5 January 1978 (UT).

Figure 19: Power spectrum of a typical X-ray binary pulsar, 4U0115+63, using data from SAS-3. The fundamental, at $P = 3.6$ s is easily seen, as well as the first two harmonics.[70]

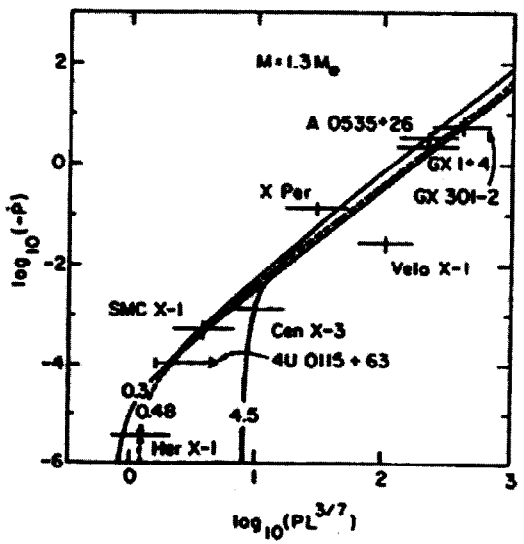


Figure 20: Period vs. luminosity according to standard accretion torque theory.[68]

values of the assumed magnetic moment μ in units of 10^{30} Gauss cm^3. Data for this figure were obtained by taking snapshots of the pulse period of the individual sources at irregular intervals, with separations of months, for up to 10 years for some sources.

However, recent continuous pulse monitoring by the BATSE experiment on board the Compton Gamma-ray Observatory (CGRO) has shown that the simple accretion torque picture must be replaced with something far more complex.[71] Long term (~years) spin up trends, based on coarsely sampled data, are the result of alternating intervals of spin up and spin down. Detailed analysis of the two different states has shown that the distribution of torques is roughly bimodal, with the spin up torque slightly larger than the typical spin down torque. In some systems, the torque changes sign as often as $\sim 10 - 100$ days, while in others, the reversal is every $10 - 20$ years. These observations have been interpreted as evidence that the accretion disks in these system may be alternating between episodes of prograde and retrograde rotation.[72] Although this type of behavior can be seen in numerical simulations of wind-fed systems, it seems difficult to achieve in systems that are accreting via Roche lobe overflow.

In transient systems, however, the standard accretion torque theory seems to hold, at least for the much shorter timescales of the outbursts. During outbursts, accretion occurs, and the neutron stars spin up in good agreement with accretion torque theory. Between outbursts, when little or no accretion occurs, the neutron stars spin down again.

4.2 Neutron Star Magnetic Field

The strength of the magnetic field on the surfaces of neutron stars can be directly inferred from the presence of cyclotron scattering resonance features in the energy spectra of (at least nine) X-ray pulsars. Figure 21 shows a comparison between the spectra of three X-ray pulsars with these features and one without. The magnetic field strength can be derived[73] as follows:

$$E = \frac{ehB}{2\pi m_e}(2n+1) + \frac{p_z^2}{2m_e} \tag{7}$$

where E is the true cyclotron line energy, B is the magnetic field strength at the neutron star surface, p_z is the angular momentum along B, and n is the quantum number specifying the Landau level. For the first cyclotron line, the electrons transit from $n = 1$ to the $n = 0$ level. Neglecting changes in the longitudinal angular momentum, we therefore find that:

$$E = 2\mu_B B \tag{8}$$

where μ_B is the Bohr magneton. However, the observed absorption energy E_{obs} is related to the true energy E by the gravitational redshift factor:

$$E_{obs} = \frac{E}{1+z}, \tag{9}$$

where the redshift z is defined as:

$$1 + z = (1 - \frac{2GM}{Rc^2})^{-1/2} \tag{10}$$

and has typical values of $z \sim 0.3$. This yields a relation between the true magnetic field strength B and the observed line energy E_{obs}:

$$B = 1.12 \times 10^{12} (\frac{E_{obs}}{10 \text{ keV}})(\frac{1+z}{1.3}) \text{ Gauss} \quad (11)$$

Cyclotron scattering resonance line features are observed between 7 and 40 keV, which corresponds to a range of surface magnetic field strengths between $(0.8 - 4.4) \times 10^{12}$ Gauss.

4.3 Neutron Star Radius

The most direct measurements of the radii of neutron stars are derived from observations of X-ray bursts in globular cluster sources, to which the distances can be determined by independent methods. These black-body spectral fits yield photospheric radii of $10 - 20$ km, consistent with the range expected for likely equations of state. However, significant uncertainties remain: there may be deviations from pure black-body spectra due to the effects of the neutron's star photosphere; the radius expansion seen in many of the bursts indicates that the observed radiation may not be emitted directly from the neutron star's surface; and the analysis technique of subtracting the persistent accretion luminosity prior to fitting the burst spectrum may introduce additional uncertainties.[37]

The more recent observations of kHz QPO in many of the accreting X-ray binaries have provided a new method for the radius estimation for neutron stars.[74,75,76] The highest observed QPO frequency is ~ 1100 Hz: this is seen in at least 8 systems. The ubiquitous nature of this highest frequency argues that it is associated with the maximum predicted orbital frequency, found at the disk termination radius. For most systems with weak magnetic fields, this radius should be approximately equal to the marginally stable orbit r_{ms}, given an assumed General Relativistic metric:

$$r_{ms} = \frac{\alpha G M}{c^2}. \quad (12)$$

In the above equation, α is a factor which varies from 1 for a maximally prograde Kerr (rotating) system, to 9 for an extreme retrograde Kerr system. For the non-rotating Schwarzschild metric, $\alpha = 6$. Therefore, for the canonical 1.4 $M_\odot$ neutron star (in a Schwarzschild metric), the assumption that the neutron star radius must be inside r_{ms} sets a firm limit of 12 km on its radius.

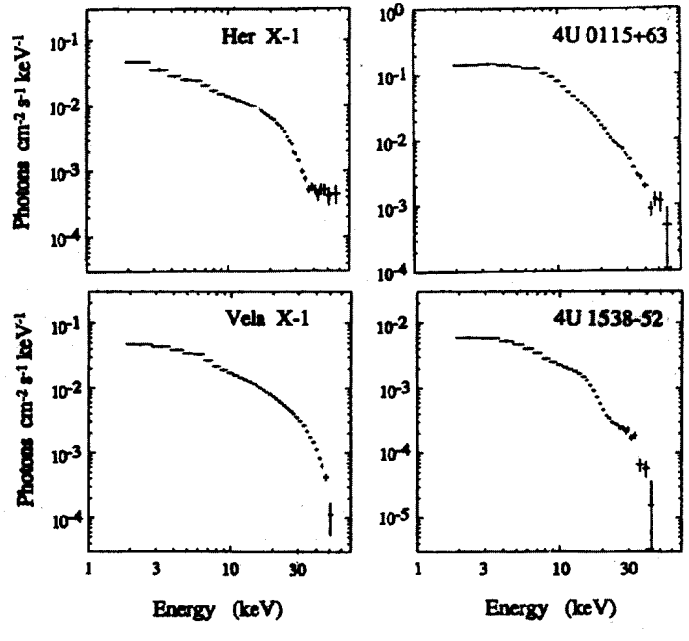


Figure 21: Energy spectra for four binary X-ray pulsars using data from the *Ginga* satellite. Three show cyclotron lines: Her X-1 at ~ 35 keV, 4U0115+63 at 12 and 23 keV, and 4U1538-52 at 21 keV. Vela X-1, in the bottom left hand panel, does not. The feature near 6.7 keV is the Iron K line.[31]

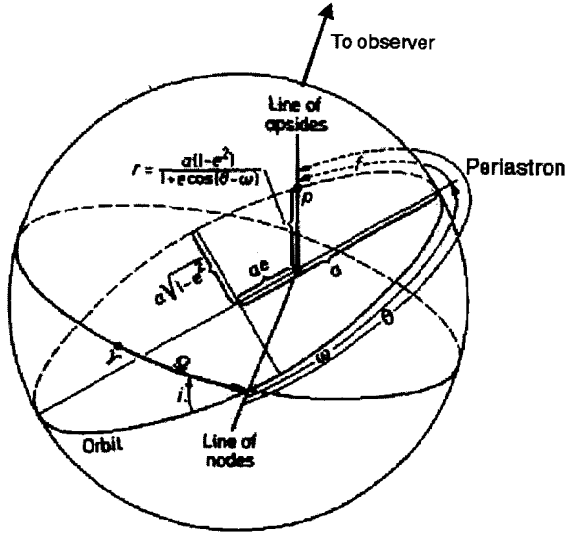


Figure 22: The definition of the orbital elements. The semi-major axis (a), and the eccentricity (e) can clearly be seen on the figure. The longitude of periastron (ω) is the angular distance between the plane normal to the observer's line of sight and the periastron point along the orbit. The inclination angle (i), is the angular distance between the plane of the orbit and the plane normal to the observer's line of sight.[77]

4.4 Orbital Determination in X-ray Binaries

The five observational parameters which uniquely describe a binary orbit are shown in Figure 22. They are: a, the amplitude of the semi-major axis; T, the time of periastron passage; ω, the longitude of periastron; e the eccentricity; and i, the inclination angle of the orbit to our line-of-sight. The accuracy to which these parameters can be determined has considerable dynamic range and depends on the type of observational data which are available for a given system.

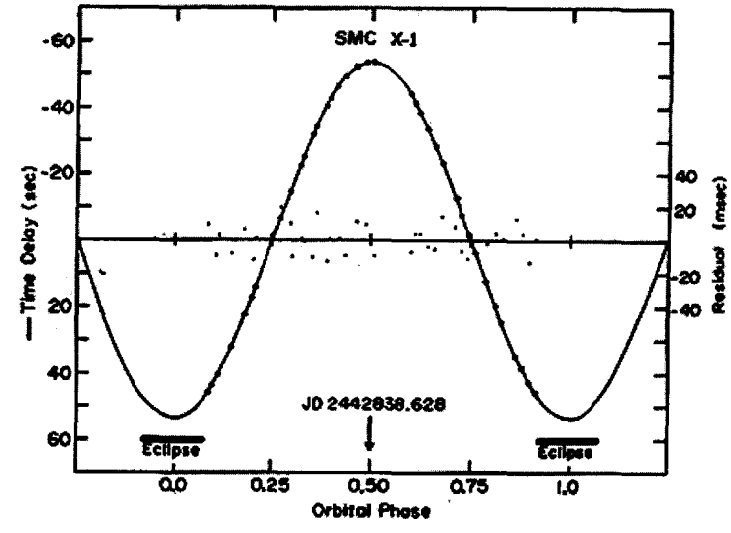


Figure 23: Doppler orbital fit for SMC X-1, a eclipsing, circular binary.[78]

4.4.1 X-ray Pulsars

Analysis of Doppler shifted pulses in X-ray binaries is used in the most accurate determination of the five orbital elements.[79] Fitting either the Doppler delays or the pulse arrival times to Kepler's equation together with additional pulse period derivative terms (to account for accretion torques) can yield results that are extremely accurate (uncertainties are typically less than 10^{-6}). Figure 23 shows an example of the orbital fit for a high-mass, eclipsing binary with a circular orbit (SMC X-1), while Figure 24 shows a Be-star transient pulsar in an eccentric orbit (4U0115+63). In both cases, Doppler delays of the order of 50 -100 seconds are seen, and residuals from the fits are ~tens of milliseconds. Figure 25 shows a summary of the orbital fits for 11 X-ray pulsars. The wide (and often eccentric) orbits seen in the Be-systems V0332+53, 2S1553-54, and 4U0115+63 contribute to the transient X-ray outbursts which are often seen in these binaries.

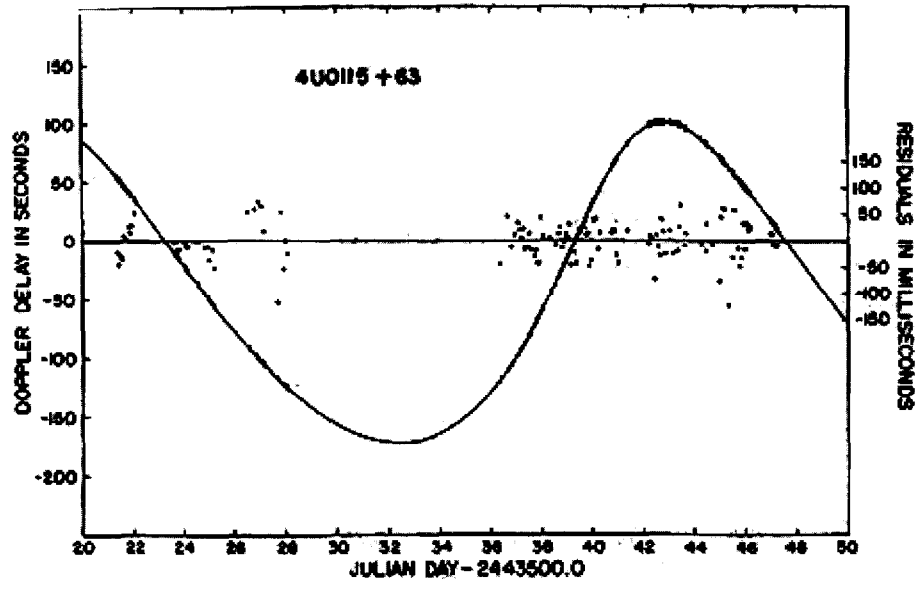


Figure 24: Doppler orbital fit for 4U0115+63, an eccentric, transient Be-star system. Note the Doppler curve is skewed by the eccentric nature of the orbit.[80]

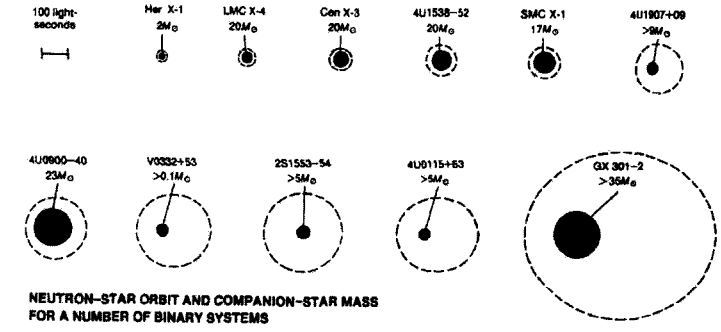


Figure 25: Comparison of the orbital dimensions for many of the X-ray binary pulsars. The five smallest systems show eclipses, as does 4U0900-40 (Vela X-1), which is in a slightly eccentric, but close orbit, with a supergiant companion. The wider, often eccentric systems with small companions on the bottom row contain Be-stars.[81]

4.4.2 Non-pulsing Low-mass X-ray Binaries

In contrast to the high-mass and/or pulsing X-ray binaries, there are only two low-mass X-ray binaries which show true X-ray eclipses by the companion star.[38,39] The orbital periods in these two systems can be determined to better than 10^{-8}, but the other four orbital elements cannot be well determined (although some information is available from optical observations). The two eclipsing sources and an additional 10 LMXBs show irregularly spaced modulations or fluctuations in X-ray intensity ("dips") that are attributed to obscuration by the gas stream from the companion star, as it joins the accretion disk. However, for the great majority of these sources, no defining orbital signatures can be observed in X-ray data. Optical modulation of the low-mass companion (or perhaps the X-ray heated accretion disk) has resulted in the majority of the orbital determinations in these types of systems, with correspondingly greater uncertainties.

The relative paucity of eclipsing low-mass X-ray binaries was explained in a classic paper by Milgrom.[82] Obscuration by the relatively large accretion disk prevents X-rays from being detected when the inclination of the binary is optimal for viewing X-ray eclipses by the companion star (*i.e.*, in the orbital plane). When the viewing angle begins to rise above the orbital plane, the companion star is so small that there are only a few systems in which eclipses will be seen. And even in these systems, the eclipses are very short, due to the grazing angle of the line-of-sight. This observational situation is depicted in Figure 26 and a summary of the orbits for many LMXBs in shown in Figure 27. Note that Her X-1 and 4U1626-67, both pulsars, have low-mass companions, as do several of the black hole candidates (represented in this figure by A0620-00). One of the two eclipsing[38] binaries, MXB 1659-29, is also shown in this figure.

4.4.3 Orbital Evolution

Changes in the period of the binary orbit have only been detected in a handful of X-ray binaries. In contrast to the Hulse-Taylor binary radio pulsar system,[83] the decay of X-ray binary orbits is not attributable to the emission of gravitational radiation. In the high-mass X-ray binary Cen X-3, successive eclipse timings over more than ten years were used by Kelley et al.[84] to show that the fractional change in the orbital period was decreasing by about 10^{-6} yr^{-1} (see Figure 28). This orbital decay is consistent with a loss of angular momentum from the system

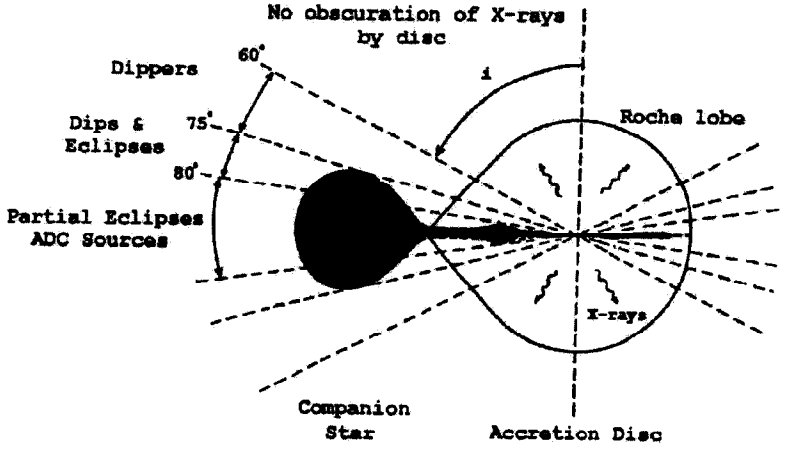


Figure 26: Observational properties of low-mass X-ray binaries as a function of line-of-sight angle to the orbital plane.[81]

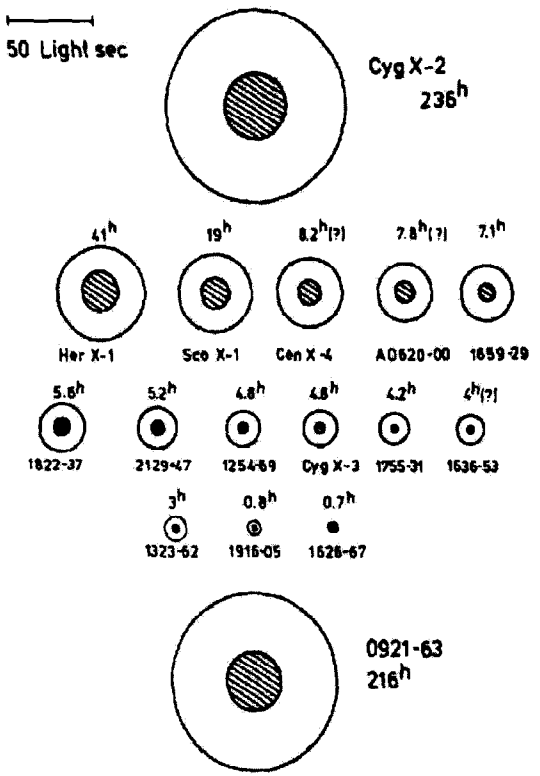


Figure 27: Comparison of the orbital dimensions for many of the low-mass X-ray binaries. Note that Her X-1 and 4U1626-67 are pulsars, A0620-00 is a black hole candidate, and MXB 1659-29 is one of two eclipsing bursters.[81]

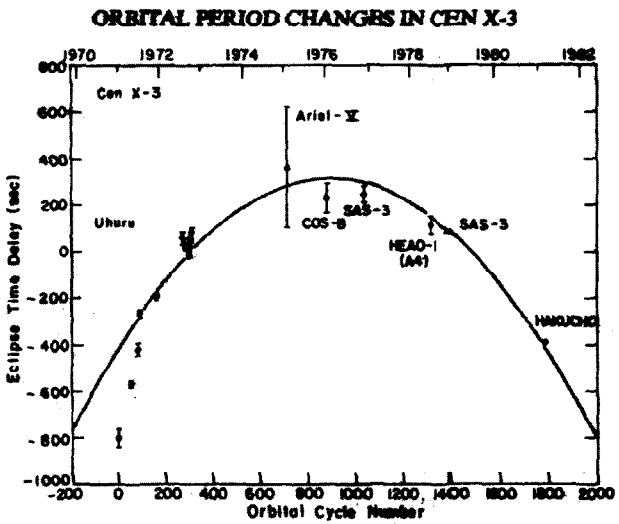


Figure 28: Decrease in the orbital period of the X-ray binary pulsar Cen X-3 as shown by the quadratic change in the arrival time of its eclipses.[84]

due to the loss of matter from the companion star in a stellar wind of 6×10^{-6} $M_\odot$ yr^{-1}. Tidal coupling between the neutron star and its companion can also produce the observed decay. In this case, the companion star is rotating asynchronously with the orbit, and the resulting tidal distortion will exert a torque which can decrease the orbital angular momentum.

By similar reasoning, orbital decay in low-mass X-ray binaries should occur at a fractional rate of $< 10^{-10}$ yr^{-1}. However, in the one eclipsing low-mass X-ray binary for which data span many years (EXO 0748-676), the observed changes in eclipse timing suggest a rate which is $\sim$ 700 times greater, and rather than decreasing, the orbital period is seen to increase! The sign and magnitude of the orbital period derivative indicates that the neutron star may be spinning down, and transferring its lost angular momentum to the binary system (see Figure 29).

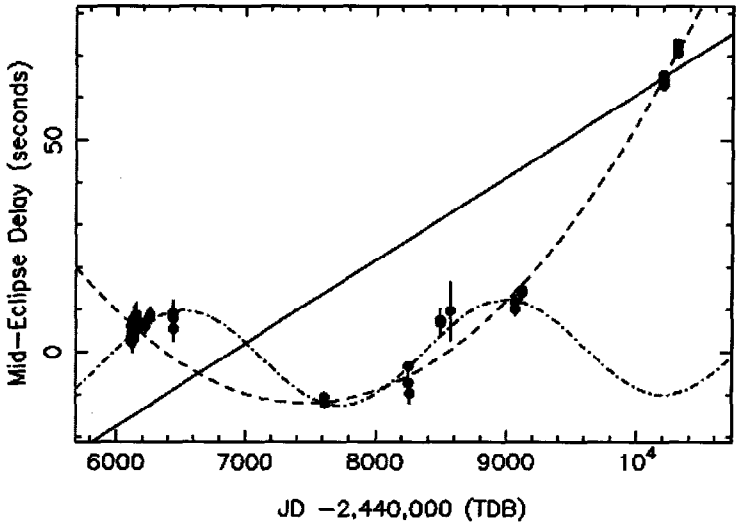


Figure 29: Change in the orbital period of the eclipsing low-mass X-ray binary EXO 0748-676 as shown by the changes in the arrival time of its eclipses. The sinusoid dot-dashed model is for a triple system. This, and the constant orbital period model (straight line) are ruled out by the long term observations. However, even the best fit, constant orbital period derivative model (dashed line) does not formally fit the data.[85]

4.5 Mass Determination in X-ray Binaries

Techniques used to determine the mass of the compact object in an X-ray binary system differ for the X-ray pulsars (six of which are in eclipsing systems) and black hole candidates. In the following sections, I will describe the methods used in each case.

4.5.1 Neutron Star Masses

For X-ray pulsars in binary systems, the derived orbital elements from Doppler shift analysis can be combined with Kepler's Third Law to derive the mass function for the compact object. The mass function is defined as:

$$f(M) = \frac{M_x^3 \sin^3 i}{M_{tot}^2} \tag{13}$$

where M_x is the mass of the compact object, i is the inclination angle to the orbital plane from our line-of-sight, and M_{tot} is the total mass in the binary system. Detailed spectroscopic studies of the companion star are then used to determine its mass by identifying its spectral type. (This process has significant uncertainties, however, in that the relationship between mass and spectral type is derived from studies of non-interacting binaries.) And for those (six) systems which show X-ray eclipses, the assumed mass-radius relationship for the companion star is then employed to find the inclination angle to the system. Figure 30 shows a summary of the mass measurements derived in this manner for the six eclipsing systems, as well as the mass measured for the Hulse-Taylor binary radio pulsar.[83] In all cases, the mass of the neutron star is consistent with 1.4 $M_\odot$.

For the LMXB QPO sources, an equation-of-state independent relation has been derived[74] that yields the mass in terms of the Keplerian orbital frequency at the marginally stable orbit, r_{ms}. Matter at r_{ms} can be related to the highest QPO frequency observed for a given source:

$$\frac{M}{M_\odot} = \frac{2198}{\nu_{max}} \tag{14}$$

Substituting the typical highest QPO frequency observed ($\sim$ 1100 Hz) yields neutron star masses of $\sim$ 2 $M_\odot$, significantly higher than those found for the neutron star pulsars. The LMXBs, however, have long been believed to be a much older population than the pulsars, and this higher mass is consistent with

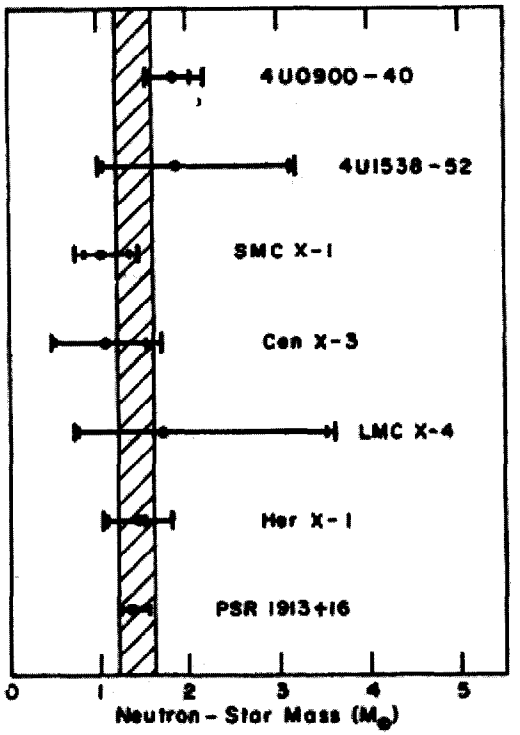


Figure 30: The masses of the six eclipsing X-ray binary pulsars and the Hulse-Taylor binary radio pulsar.[79]

accretion for $\sim 10^8$ y at $\sim 10\%$ of the Eddington limit onto the original 1.4 $M_\odot$ object created in the supernova explosion.[75,76]

4.5.2 Black Hole Candidate Masses

The determination of the orbital elements in binary systems with no detectable pulsations or no X-ray eclipses requires the use of optical data from the companion star. Radial velocity curves are obtained, with amplitude $K \sin i$ km s^{-1}, and orbital period P_{orb} (see Figure 31). These data measure the effect of the unseen compact object on its companion star, and can be related through Kepler's laws to the mass function (see Equation 13), establishing a lower limit on the mass of the unseen, X-ray emitting companion:

$$f(M) = \frac{P_{orb}\, K^3}{2\pi G} < M_x. \tag{15}$$

The lower limit to the compact object's mass M_x stems from the fact that without eclipses, it is not possible to precisely know the system's inclination angle. The angle can be estimated, however, from studies of the ellipsoidal light variations during the binary orbit. However, assuming that the companion star has a mass of zero, and that the inclination angle is 90 degrees, yields a mass function equal to the mass of the compact object. Since these are extreme values, realistic values for the mass of the companion and the inclination angle will result in a black hole mass greater than the derived mass function.

Black hole candidates are those objects for which the best estimates for the compact object's mass (after taking into account the mass of the optical companion and the inclination angle to the system) are greater than 3 $M_\odot$, the Rhoades-Ruffini upper limit to the mass of a neutron star.[86] Other common observational characteristics shared by some of the black hole candidates include two-component energy spectra, in which one component is very soft, and highly variable, transient X-ray outbursts. Roche-lobe schematics for some typical black hole candidate binaries are given in Figure 32 and Figure 33. The best black hole candidates are those in which the mass function itself is greater than $3M_\odot$. For example, in V404 Cygni[87] the mass function is 6.08 ± 0.06 $M_\odot$, and the most likely mass is $12M_\odot$.

It should be noted however, that the Rhoades-Ruffini neutron star mass limit of 3 $M_\odot$, although independent of assumptions about the specific form of the neutron star equation of state, does make certain assumptions that can be questioned. For

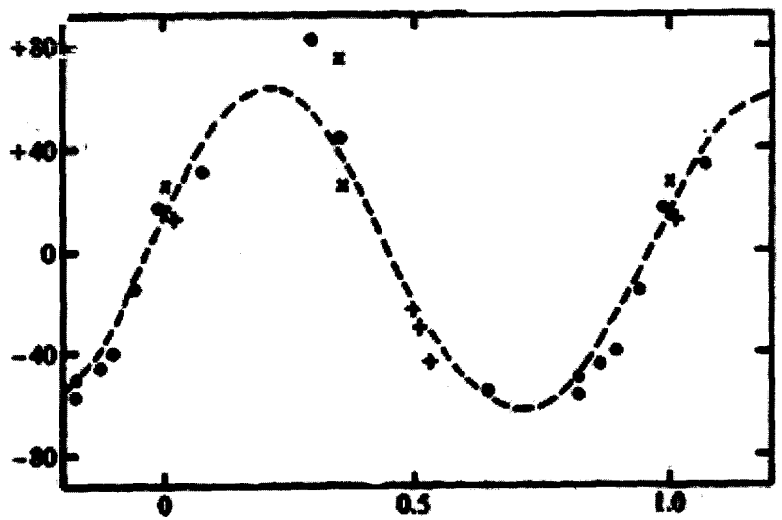


Figure 31: Optical radial velocity curve for the companion to the black hole candidate Cygnus X-1. Radial velocities are measured in km s^{-1}, and are shown as a function of phase in the 5.6 d binary orbit.[88]

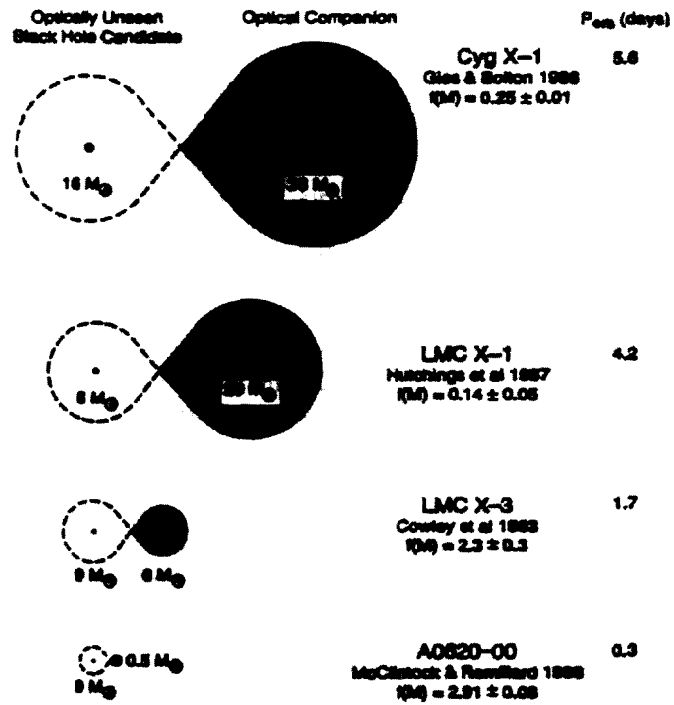


Figure 32: Roche lobe schematics for five black hole candidates.[89]

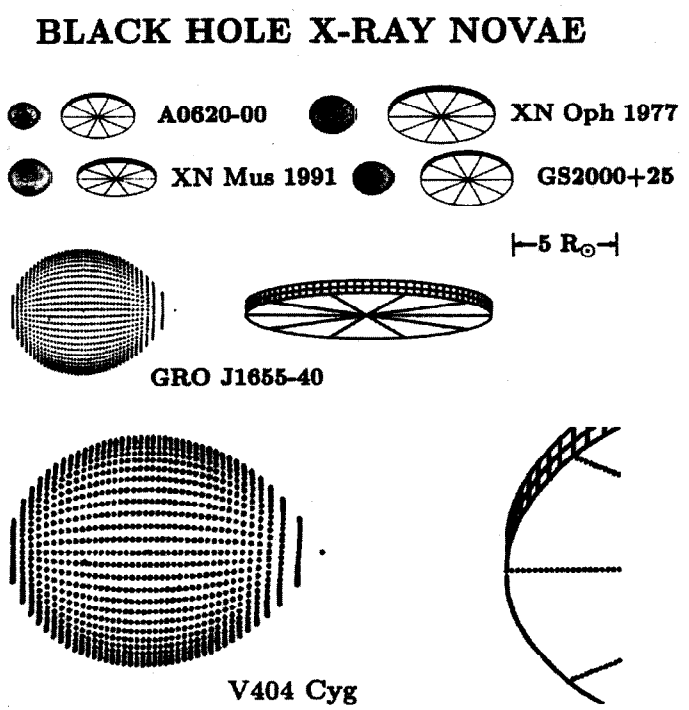


Figure 33: Roche lobe schematics for X-ray nova black hole candidates, indicating the approximate inclination angle to the line of sight to each binary system, and the size of the accretion disk around the compact object.[90] A0620-00 appears in both this figure and in Figure 32.

example, Rhoades and Ruffini assume that General Relativity provides the correct description of gravity in the strong field regime, and that gravity provides the stabilizing force which overcomes repulsive pressures in the star's core. If instead hadronic forces in the core can stabilize the star, it may be possible to increase the mass of a neutron star far above this limit.[91,92]

4.5.3 Microquasar Masses

There are two galactic black hole candidate binaries which differ markedly from those discussed in the previous section. Known as the microquasars, GRS1915+105 and GRO J1655-40 exhibit superluminal radio jets (as do extragalactic quasars), and transient, highly variable X-ray emission.[93] Because the masses of these two objects are only $\sim 10\ M_\odot$ rather than $\sim 10^6\ M_\odot$, the radii of their accretion disks and radio lobes are also about one million times smaller. The temperature of the radiation from the inner accretion disk is correspondingly higher, with the bulk of the observable emission in the X-ray bandpass. And of course, unlike their galaxy-gobbling extragalactic counterparts, the microquasars are being fueled by accreting matter from a binary companion star. This comparison is illustrated in Figure 34.

Observations of the microquasar sources with the Rossi X-ray Timing Explorer have shown QPO from both sources.[56,57] Optical observations[94] and careful analysis of ellipsoidal light variations from the eclipsed companion to GRO J1755-40 have resulted in a mass determination for the compact object of $7.02 \pm 0.22\ M_\odot$. This agrees well with the interpretation of the 300 Hz QPO from this object as resulting from matter circulating at the marginally stable orbit, r_{ms}. Applying the same argument to the 67 Hz QPO seen from GRS 1915+105 yields a mass estimate[56] of $\sim 33\ M_\odot$.

An alternative origin for the 67 Hz QPO from GRS1915+105 has been advanced by Nowak and his collaborators.[95] They interpret the QPO as diskoseismic eigenfrequencies of the microquasar's accretion disk, and derive a black hole mass ranging from 11 to 33 $M_\odot$, for systems containing Schwarzschild to maximally prograde Kerr black holes, respectively.

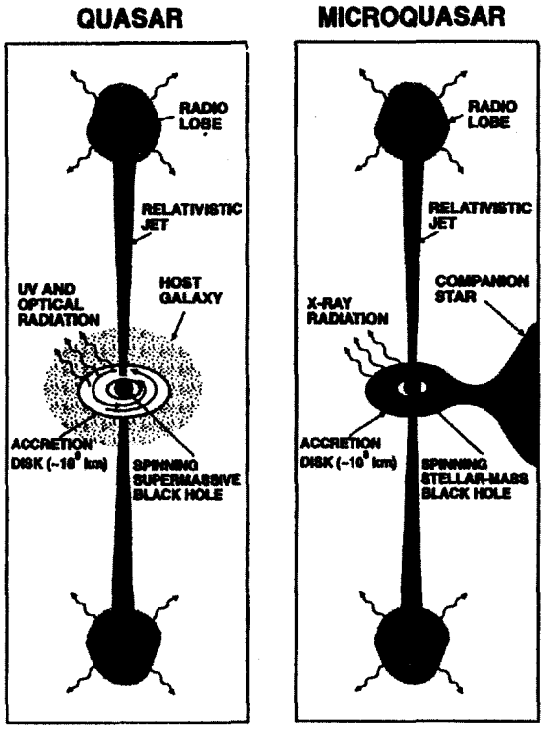


Figure 34: A comparison between extragalactic quasars and microquasar X-ray binaries.[96]

4.6 Neutron Star Equations of State

In order to use X-ray observations to learn about the neutron star equation of state, it is necessary to measure both the mass and the radius of the star for a given object.[74] Although observations of X-ray bursts have greatly constrained the mass-radius relation[31] for many of the low-mass X-ray binaries, many uncertainties remain. The energy spectra from X-ray bursts may not be perfect blackbodies; models of neutron star atmospheres have shown[97] that the color temperature T_c is greater than the effective temperature T_e by a factor of ~ 1.5. However, recent observations of the "twin-peaks" QPO, combined with X-ray burst observations, may have given us a new observational handle on the problem.[75,76]

As an example, Kaaret, Ford and Chen[75] show that the mass derived from the highest frequency QPO is typically $\sim 2M_\odot$. For the well studied X-ray burster 4U1636-53, they derive a mass of $2.02 \pm 0.12M_\odot$, where the uncertainty takes into account both the observational uncertainty in the QPO measurement, as well as additional theoretical uncertainties in the location of the emitting region responsible for the QPO, and in the moment of inertia, which depends on the equation of state. Interpreting a 4.1 ± 0.1 keV absorption feature in the spectrum of an X-ray burst from this same object as redshifted absorption from Fe XXV ions in matter accreted onto the neutron star surface, one can use Equation 10, with a mass of $2.0M_\odot$ to determine the radius. The value for the radius, found in this manner, is 9.6 ± 0.6 km. QPO have been seen in the bursts from this source, at a frequency of 581 Hz, which may be twice the neutron star spin frequency.[98] These parameters are best suited to a soft or moderately soft equation of state for a rapidly rotating neutron star, such as EOS-A or -AU, as calculated by Cook, Shapiro and Teukolsky.[99]

5 Relativistic Effects

In this final section, I review very recent, sometimes controversial evidence and ideas regarding the direct observation of relativistic effects in systems containing neutron stars and black holes.

5.1 Evidence For Event Horizons In Black Holes

Narayan and his colleagues[100,101,102] have suggested that accreting black hole binaries, when in quiescence (*i.e.*, not during a transient X-ray outburst) will have very different X-ray spectra than quiescent neutron star transients. Their advection dominated accretion flow (ADAF) models give good results for fits of the energy spectra over a wide bandpass, from optical through X-ray wavelengths. In these models, the presence of the event horizon in a black hole binary results in much lower accretion efficiency, as the advected material simply disappears over the horizon, without radiating much of the thermal energy carried in the infalling plasma. (The accretion efficiency, η, in these systems is estimated at 0.05%.) For the neutron star systems in which the infalling material shocks at the surface (or the magnetosphere) $\eta \approx 10\%$.

An additional prediction of the ADAF models is that black hole transients should have a wider dynamic range of minimum to maximum luminosities, since their quiescent states should be $\sim$ 500 times less luminous. Although hard to prove due to limited observations of non-emitting X-ray transients, there is some evidence that appears to support this prediction.[101]

5.2 Evidence For Black Hole Spin

Zhang, Cui and Chen[103] have interpreted the often seen soft-component in X-ray energy spectra from black hole candidate objects as blackbody emission from the inner accretion disk at the marginally stable orbit, r_{ms}. They have used this idea to calculate blackbody spectra for black holes with different rotation rates. Distinct spectral differences are then found in the three different spin classes: maximally prograde Kerr, Schwarzschild (slowly or non-spinning) and maximally retrograde Kerr. The accretion efficiencies η differ for these three classes of objects from 30% to 6% to 3%, as r_{ms} changes from 1 to 6 to 9 times (GM/c^2), respectively.

In this scenario, the microquasar sources contain maximally prograde Kerr black holes. The extremely high accretion efficiency ($\eta = 30\%$) coupled with the small r_{ms} raises the temperature of the inner part of the accretion disk, producing a relatively hard spectral component, in agreement with the observations.

The traditional galactic black hole candidates in binaries with a pronounced soft-component, are identified with slowly or non-rotating (Schwarzschild) black holes. And the black hole candidates which seem similar to these sources, but are lacking the spectrally-soft component are suggested to be maximally retrograde Kerr black holes. In this case, the inefficient accretion ($\eta = 3\%$) coupled with the relatively large r_{ms} combine to lower the temperature of the soft component until it falls below the standard X-ray bandpass, and is undetectable by present instruments.

5.3 Lense-Thirring Effect

Stella and Vietri[104] were the first to suggest that QPO observed in the range 20 - 40 Hz from neutron star binaries could result from the Lense-Thirring effect. If the plane of the accretion disk of material orbiting the neutron star is tilted with respect to that defined by the binary system (which is assumed aligned with the spin axis of the neutron star), then the orbiting material must precess in a prograde direction as its inertial frame is dragged along with the rotating object. The frequency of this precession in the weak field approximation is given by:

$$\nu_{LT} = \frac{8\pi^2 I \nu_K^2 \nu_s}{Mc^2} = 13.2\ I_{45}\ \mathrm{M}_\odot^{-1}\ \nu_{K3}^2\ \nu_{s2.5}\ \mathrm{Hz} \qquad (16)$$

where I_{45} is the moment of inertia of the neutron star in units of 10^{45} g cm^2, ν_{K3} is the Kepler frequency of the orbiting material in kHz, and $\nu_{s2.5}$ is the spin frequency of the neutron star in units of 300 Hz. Note that this interpretation assumes that the kHz QPO observed in these systems result from material orbiting at r_{ms}, and that the difference frequency from the twin kHz QPO peaks yields the neutron star spin period, ν_s. Some of the evidence in support of this conjecture is given by the quadratic dependence of ν_{LT} on ν_K. This quadratic dependence is observed in changes in the corresponding QPO frequencies from the X-ray burster MXB 1728-334.

Extending Stella and Vietri's suggestion to include black holes, Cui, Zhang and Chen[105] applied the Lense-Thirring formalism to the QPO seen in the two microquasar sources as well as several other candidate systems. For the black hole candidates, the weak field approximation used by Stella and Vietri was not appropriate, so Cui et al. modified the equations to apply to strong field systems. Assuming a mass for the candidate object and interpreting the observed QPO as the Lense-Thirring precession frequencies yielded spin rates which were found to be consistent with those in their earlier work.[103] In particular, the microquasar GRO J1655-40, which has a well-determined[94] black hole mass of $\sim 7\ \mathrm{M}_\odot$, is found to be

rotating at 95% of maximum, while the more typical black hole candidates, such as Cyg X-1 are only rotating at $\sim$ 50% of maximum. Further analysis, however, of GRO J1655-40 has complicated the picture as three other QPO frequencies have been found[57] in some observations of the source. It is therefore not clear which QPO frequency, if any, should be attributed to the effects of frame-dragging.

5.4 Jet Formation In Black Hole Systems

Perhaps the most dramatic results of observations using the Rossi X-ray Timing Explorer are those[106] of Eikenberry *et al.*, in which ground-based infrared data are combined with the X-ray data to show the first direct evidence for a connection between instabilities in the inner accretion disk around a black hole and the formation of jets emanating from outside its event horizon. As can be seen in Figure 35, the X-ray and infrared emissions are correlated at the beginnings of the flares, although with a $\sim$ 310 second lag at infrared wavelengths. As the emissions seen in the two bandpasses decouple, the X-ray luminosity begins to oscillate wildly (with a quasi-period of about 10 seconds), until it finally cuts off. The infrared emission persists and decays slowly, following the disappearance of the X-rays. The cycle then repeats about every 30 minutes.

Eikenberry and his colleagues interpret these striking results in the following manner: the initial linkage between the X-ray and infrared flares indicates that the emissions in both bandpasses are triggered by the same event. The rapidly oscillating intense X-ray luminosity is associated with activity in the inner disk. Some of the disk material is then ejected to form the longer-lasting non-thermal infrared[107] (and radio[108]) emission. These flares are similar to, but not as large as the previously observed[93] superluminal events from the same object. The subsequent disappearance of the X-ray emission may indicate that the material in the inner disk has left the system: either by falling over the black hole's event horizon or in the form of the jet material.

5.5 Gravity Wave Generation From Rotating Neutron Stars

The unexpected strength of a gravitational radiation driven instability in hot young rapidly rotating neutron stars has recently been shown[109,110] to quickly ($\sim$ 1 year) reduce the initially rapid rotation rate of the star to $\sim$ 100 Hz. The instability occurs in the $l = 2$ $r-$mode, and the resulting radiation couples to

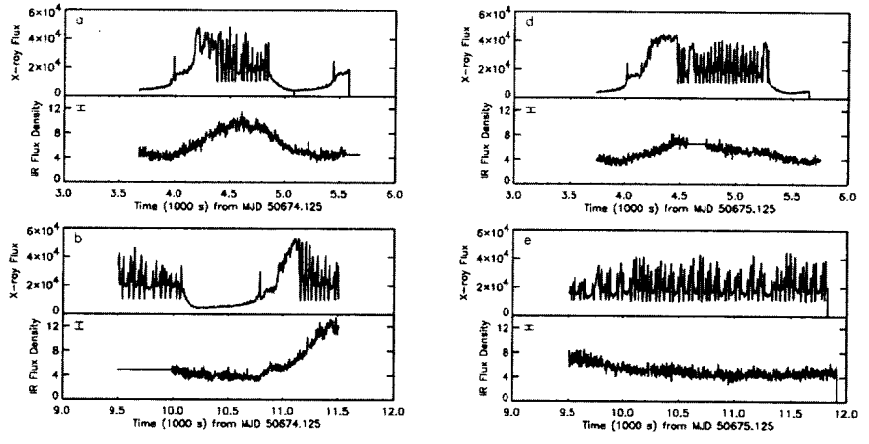


Figure 35: Coordinated infrared and X-ray observations of GRS 1915+105, a microquasar.[106]

this mode through current, rather than mass multipoles. Detectable gravitational waves may result from this process, which carries away most of the angular momentum of newly formed neutron stars.

Another mechanism for reducing the neutron star's spin rate by the emission of potentially observable gravitational radiation has been suggested by Bildsten.[111] If the thermal distribution of the neutron star's interior has a large scale asymmetry which is not aligned with its spin axis, a time varying quadrupole moment can be produced, of sufficient strength to spin down accreting neutron stars to the observed periods near 300 Hz. Unlike the previously discussed r-mode instability, this thermally driven mechanism will only affect accreting neutron stars, as the outer crust must be heated by the repeated electron captures which occur in the accreted material.

6 Summary

X-ray observations of compact objects are a powerful tool for studying gravity in the strong field regime. New information about the physical properties of black

holes and neutron stars offers the hope that predictions of General Relativity in this regime will be thoroughly examined. Evidence for relativistic effects in compact X-ray binaries has been accumulating rapidly in data from NASA's Rossi X-ray Timing Explorer. Observations of rapidly rotating neutron star systems may also lead to the first direct detection of gravitational radiation.

7 Acknowledgements

Many thanks to Tim Graves, who worked very hard getting the figures into shape, and assembling my PowerPoint presentations. Assistance from Jonathan Wong in converting one of the computer movies and in running the SLAC display projector is also greatly appreciated. This paper has greatly benefited from careful readings by J. Garrett Jernigan, whose comments were very helpful.

8 A Short Tutorial on the Names of X-ray Sources

As in ground-based astronomy, the names of objects in the X-ray sky are a combination of history and convenience, and may seem mysterious to the uninitiated. In the 1960s, X-ray observations were done with rocket-borne instrumentation. The positional information returned from these types of experiments was often crude (a few degrees), resulting in names for the newly discovered sources such as Sco X-1 (the first X-ray source found in the constellation Scorpius) or Cen X-3 (the third source found in the constellation Centaurus). As positional information improved, and particularly for sources located near the center of our Galaxy, and along the Galactic plane, the names evolved to approximate the positions in galactic coordinates. For example, GX 339-4 is located near a galactic longitude of 339 degrees and latitude -4 degrees.

With the launch of the Uhuru satellite in 1970, came the first all-sky survey. The subsequent improvement in source positions (to much less than 0.1 square degrees in most cases) required a different naming system. The Fourth Uhuru catalog[112] lists $\sim$ 350 X-ray sources, with names like 4U1617-15 (Sco X-1), where the 4U represents the catalog name, and the source location is now given in right ascension and declination (similar to the Earth's longitude and latitude). Sco X-1 is located near 16 hours, 17 minutes of right ascension, and near -15 degrees in declination. The numbers are truncated, not rounded, and usually don't change in common usage, even when the positions improve. Sometimes, additional digits will be added (*e.g.*, MXB1728-334, which is near a declination of -33.4 degrees). Many other satellites have discovered sources that are not in the 4U catalog, and lend their initials as prefixes to positional numbers. For example, the English Ariel satellite (A), European EXOSAT satellite (EXO), the MIT X-ray burst (MXB) and persistent (2S) sources discovered with SAS-3, the Vela satellites (V), or the Italian BeppoSAX satellite (SAX).

If not otherwise indicated, positionally derived names are given in 1950 coordinates (the actual positions of the sources change continually as the earth's coordinate system precesses with respect to its astronomical equivalent). As we approach the year 2000, it has become more convenient to name sources according to the positions that they will have in that Julian year. This has led to names such as GRO J1655-40 or GRO J1744-28 (Compton Gamma-ray Observatory discovered sources given in J2000 coordinates.) But when referring to radio pulsars, it is still common to use Besselian 1950 coordinates, which are designated with the letter B (*e.g.*, PSR B1259-63). Some of the other sources referred to in this paper were discovered by the Russian-French satellite Granat (GRS 1915+105), are named after their optical counterparts (V404 Cyg) or are soft gamma repeaters (SGR 1806-20).

9 Glossary

Accretion: The process of gravity-driven mass transfer. In X-ray binaries, the accretion flow leaves the companion star and ends up at the compact object.

Apastron: In an eccentric binary orbit, the distance of greatest separation between the stars.

Be-star: A main-sequence companion star often associated with neutron star pulsars in X-ray binaries. Be-stars are of spectral type B (surface temperatures $\sim$10,000 K), and also show emission lines. They have a typical mass of 5 - 10 $M_\odot$, and are rotating very rapidly. The rotation rate is near break-up, which often causes the outer layers of the star to be spontaneously ejected (even for isolated Be-stars), triggering a transient X-ray flare in those binary systems with orbiting neutron stars.

Burster or *X-ray Burst Source*: A neutron star which emits occasional very intense flashes of X-ray emission. Bursts last for 10-20 seconds, and recur on timescales of hours to days. They are interpreted as thermonuclear flashes occurring in the surface layers of the neutron star. Thermonuclear bursts are often referred to as *Type I*, to distinguish them from the very rapid *Type II* bursts which are believed to result from instabilities in the accretion flow.

Eddington limit: The critical luminosity where the outward force from radiation pressure balances the inward gravitational force on the accreting matter, and hence stops the accretion flow. The Eddington luminosity is given by:

$$L_{Edd} = \frac{4\pi G M m_p c}{\sigma_T} = 1.3 \times 10^{38} \left(\frac{\mathrm{M}}{\mathrm{M}_\odot}\right) \text{ ergs s}^{-1}. \quad (17)$$

In the above equation, M is the mass of the accreting object, m_p is the mass of a proton (assuming hydrogen plasma accretion), and σ_T is the Thomson cross-section for electron scattering.

Equation of State: The equation $P = P(\rho)$, which determines the relationship between the mass and radius for compact objects. Equations of state (EOS) vary from *soft* to *stiff*. In a soft EOS, the average system energy is attractive when nuclear density ($\rho_{nuc} = 2.8 \times 10^{14}$ g cm^{-3}) is reached. For stiff EOS, energy from repulsive interactions dominates at densities greater than nuclear. This extra pressure against gravitational collapse produces neutron stars with higher maximum masses, larger radii for a given mass, thicker crusts and lower central densities.

Lense-Thirring effect: A prediction of General Relativity, first worked out by physicists W. Lense and H. Thirring in 1918. They realized that a torque-free gyroscope will precess relative to an inertial frame at infinity. For X-ray binaries, this translates into the observational effect that particles in accretion disks which are inclined with respect to the spin axis of the compact object may precess as they orbit around it. Another way to say this is that rotating bodies have the ability to drag their inertial frame (*i.e.*, space and time) around with them as they rotate. Hence this effect is often known as *frame-dragging*.

$\mathrm{M}_\odot$: One solar mass, $\sim 2 \times 10^{33}$ grams.

Magnetar: A neutron star with an extremely intense magnetic field, $B \approx 10^{14} - 10^{15}$ Gauss. Starquakes causing magnetic reconnection in the surface layers of magnetars have recently been suggested as the cause of the bursts seen from the *Soft Gamma-ray Repeaters*.

Microquasar: An accreting black hole binary which shows superluminally expanding radio jets. The jets are similar to, but much smaller than those seen in extragalactic quasars.

Periastron: In an eccentric binary orbit, the distance of closest approach between the stars.

Prograde: A system in which the accreting material has the same angular momentum vector direction as does the spin vector for the compact object. In prograde systems, the accreting object should spin up as it accretes material.

Pulsar: A rotating neutron star that emits strictly periodic pulsations. These can be found both isolated and in binary systems.

Pulsar Wind: A relativistic positron/electron plasma which is emitted from isolated, rotating neutron stars. The wind is powered by the loss of rotational kinetic energy as the pulsar's spin slows.

Quality or *Q-factor*: This is defined in a power spectrum as $Q = \frac{\nu}{\Delta\nu}$. It is used to characterize the sharpness of the observed spectral peaks.

Retrograde: A system in which the accreting material has the opposite angular momentum vector direction to the spin vector for the compact object. In retrograde systems, the accreting object should spin down as it accretes material.

Quasi-periodic Oscillations: Regular pulses that occur at slightly changing frequencies. QPO are seen from both neutron stars and black holes in X-ray binaries.

Roche lobe: The gravitational equipotential surface in a binary system. *Roche lobe overflow* is an accretion mechanism where the slowly evolving low-mass companion star's outer layers expand outside of the Roche lobe. Matter can then be transferred through the inner Lagrangian (L_1) point to the vicinity of the compact object. Typical accretion rates with this mechanism are around 10^{-10} $\mathrm{M}_\odot$ yr^{-1}.

Stellar Wind Capture: The mass transfer mechanism typically found in high-mass X-ray binary systems. The rapidly evolving high-mass companion star sheds copious material (usually hydrogen plasma) in an outgoing wind, some of which can be captured by the compact object. Typical mass loss rates are around 10^{-6} $\mathrm{M}_\odot$ yr^{-1}; of this, perhaps 1% is captured by the compact object.

X-ray binary: A white dwarf, neutron star, or black hole with a companion star that is usually on the *main sequence*, *i.e.*, burning nuclear fuel in its core. The conventional way to refer to these systems is to classify them according to the

mass of their companion stars. A *high mass X-ray binary* has a companion star that is more massive than 5 $M_\odot$, while a *low-mass X-ray binary* has a companion star less massive than 2 $M_\odot$.

X-ray transient or *X-ray nova*: A binary system in which transient, intense X-ray emission is visible for a duration of ~months out of ~years. Transient emission can be seen from both black hole and neutron star X-ray binaries.

References

[1] R. Giacconi et al., Phys. Rev. Letters **9**, 439 (1962).

[2] A. Hewish et al., Nature **217**, 709 (1968).

[3] T. Gold, Nature **218**, 731 (1968).

[4] A. G. Lyne and F. Graham-Smith, *Pulsar Astronomy*, (Cambridge University Press, Cambridge, 1990), **16**.

[5] J. K. Daugherty and A. K. Harding, ApJ **252**, 337 (1982).

[6] K. S. Cheng, C. Ho and M. A. Ruderman, ApJ **300**, 522 (1986).

[7] R. W. Romani and I.-A. Yadigaroglu, ApJ **438**, 314 (1995). See also http://astrophys.stanford.edu/home/ion/pulsar/frames.html

[8] http://cossc.gsfc.nasa.gov

[9] D. Bhattacharya in *X-ray Binaries*, edited by W. H. G. Lewin, J. van Paradijs and E. P. J. van den Heuvel, (Cambridge University Press, Cambridge, 1995), **26**, pp. 233-251.

[10] S. R. Kulkarni and J. J. Hester, Nature **335**, 801 (1988).

[11] S. R. Kulkarni et al., Nature **359**, 300 (1992).

[12] L. Cominsky, M. Roberts and S. Johnston, ApJ **427**, 978 (1994).

[13] S. Johnston et al., ApJ **387**, L37 (1992).

[14] M. Tavani, J. Arons and V. M. Kaspi, ApJ **433**, L37 (1994).

[15] V. M. Kaspi et al., ApJ **453**, 424 (1995).

[16] M. Hirayama et al., submitted to ApJ (1998).

[17] C. Thompson and R. Duncan, MNRAS **275**, 255, (1995).

[18] C. Thompson and R. Duncan, ApJ **473**, 322 (1996).

[19] J. E. Grove et al., ApJ **447**, L113 (1995).

[20] C. Kouveliotou et al., IAU Circ. No. 6944, see also http://www.magnetars.com and http://ssl.berkeley.edu/ipn3/sgr1627-41

[21] C. Kouveliotou et al., Nature **393**, 235 (1998).

[22] *X-ray Binaries*, edited by W. H. G. Lewin, J. van Paradijs and E. P. J. van den Heuvel, (Cambridge University Press, Cambridge, 1995), **26**.

[23] J. Frank, A. King and D. Raine, *Accretion Power in Astrophysics*, (Cambridge University Press, Cambridge, 1992), **21**.

[24] J. Blondin's web site: http://www.physics.ncsu.edu/people/faculty.html. This site has excellent movies which show his numerical simulations of various types of accretion in X-ray binaries.

[25] J. E. Grindlay et al., ApJ **205**, L127 (1976).

[26] R. D. Belian, J. P. Connor and W. D. Evans, ApJ **206**, L135 (1976).

[27] S. E. Woosley and R. E. Taam, Nature **263**, 101 (1976).

[28] P. C. Joss, Nature **270**, 310 (1977).

[29] M. Y. Fujimoto, T. Hanawa and S. Miyaji, ApJ **247**, 267 (1981).

[30] P. C. Joss, Comments on Ap. **8**, 109 (1979).

[31] N. E. White, F. Nagase and A. N. Parmar in *X-ray Binaries*, edited by W. H. G. Lewin, J. van Paradijs and E. P. J. van den Heuvel, (Cambridge University Press, Cambridge, 1995) **26**, pp. 1-57.

[32] R. Wijnands and M. van der Klis, Nature **394**, 344 (1998).

[33] N. E. White, J. H. Swank and S. S. Holt, ApJ **270**, 711 (1983).

[34] R. H. D. Corbet, A&A **141**, 91, (1984).

[35] P. C. Joss and F. K. Li, ApJ **238**, 287, (1980).

[36] L. Bildsten, in *The Many Faces of Neutron Stars*, edited by R. Buccheri, A. Alpar and J. van Paradijs, (Kluwer, Dordrecht, 1998), pp. 419-449.

[37] W. H. G. Lewin, J. van Paradijs and R. E. Taam in *X-ray Binaries*, edited by W. H. G. Lewin, J. van Paradijs and E. P. J. van den Heuvel, (Cambridge University Press, Cambridge, 1995), **26**, pp. 175 - 232.

[38] L. Cominsky and K. Wood, ApJ **283**, 765 (1984).

[39] A. N. Parmar et al., ApJ **308**, 199, (1986).

[40] J. H. Swank et al., ApJ **212**, L73 (1977).

[41] J. A. Hoffman, L. Cominsky and W. H. G. Lewin, ApJ **240**, L27 (1980).

[42] J. van Paradijs, Nature **274**, 650 (1978).

[43] T. E. Strohmayer et al., ApJ **469**, L9 (1996).

[44] W. H. G. Lewin et al., ApJ **207**, L95, (1976).

[45] J. A. Hoffman, H. L. Marshall and W. H. G. Lewin, Nature **271**, 630 (1978).

[46] C. Kouveliotou et al., Nature **379**, 799 (1996).

[47] M. H. Finger et al., Nature **381**, 291 (1996).

[48] S. Rappaport and P. Joss, ApJ **486**, 435 (1997).

[49] D. Chakrabarty and E. H. Morgan, Nature **394**, 346 (1998).

[50] J. J. M. in't Zand et al., A&A **331**, L25 (1998).

[51] W. Cui, E. H. Morgan and L. G. Titarchuk, ApJ **504**, L27 (1998).

[52] L. Cominsky, Ph. D. Thesis at the Massachusetts Institute of Technology (1981).

[53] W. H. G. Lewin, Annals New York Acad. Sci. **302**, 210 (1977).

[54] T. Takeshima et al., PASJ **43**, L43 (1991).

[55] M. van der Klis in *X-ray Binaries*, edited by W. H. G. Lewin, J. van Paradijs and E. P. J. van den Heuvel, (Cambridge University Press, Cambridge, 1995) **26**, pp. 252 - 307.

[56] E. H. Morgan, R. A. Remillard and J. Greiner, ApJ **482**, 993 (1997).

[57] R. A. Remillard et al., ApJ in press or http://xxx.lanl.gov/astro-ph/9806049 (1998).

[58] M. van der Klis et al., Nature **316**, 225 (1985).

[59] M. A. Alpar and J. Shaham, Nature **316**, 239 (1985).

[60] S. Miyamoto et al., ApJ **383**, 784 (1991).

[61] G. Hasinger and M. van der Klis, A&A **225**, 79 (1989).

[62] T. E. Strohmayer, W. Zhang and J. H. Swank, IAU Circ. No. 6320 (1996).

[63] M. van der Klis,in *The Many Faces of Neutron Stars*, edited by R. Buccheri, A. Alpar and J. van Paradijs, (Kluwer, Dordrecht, 1998), pp. 337-368.

[64] M. C. Miller, F. K. Lamb and D. Psaltis, ApJ, in press (December 1, 1998).

[65] M. van der Klis and R. A. D. Wijnands, ApJ **481**, L97 (1997).

[66] R. I. Klein et al., ApJ **469**, L119 (1996).

[67] D. Psaltis et al., ApJL **501**, L95 (1998).

[68] P. Ghosh and F. K. Lamb, ApJ **232**, 259 (1979).

[69] P. Ghosh and F. K. Lamb, ApJ **234**, 296 (1979).

[70] L. Cominsky et al., Nature **273**, 367 (1978).

[71] L. Bildsten et al., ApJS **113**, 367 (1997).

[72] R. W. Nelson et al., ApJ **488**, L117 (1997).

[73] S. L. Shapiro and S. A. Teukolsky, *Black Holes, White Dwarfs and Neutron Stars*, (John Wiley and Sons, New York, 1983) p. 382.

[74] W. Kluzniak, P. Michelson and R. V. Wagoner, ApJ **358**, 538 (1990).

[75] P. Kaaret, E. C. Ford and K. Chen, ApJ **480**, L27 (1997).

[76] W. Zhang, T. E. Strohmayer and J. H. Swank, ApJ **482**, L167 (1997).

[77] Adapted from K. R. Lang, *Astrophysical Formulae*, (Springer-Verlag, Berlin, 1986), p. 541.

[78] F. Primini, S. Rappaport, and P. C. Joss, ApJ **217**, 543 (1977).

[79] S. Rappaport and P. C. Joss in *Accretion Driven Stellar X-ray Sources*, edited by W. H. G. Lewin and E. P. van den Heuvel, (Cambridge University Press, Cambridge, 1983) p. 1 - 39.

[80] S. A. Rappaport et al., ApJ **224**, L1 (1978).

[81] P. A. Charles and F. D. Seward, *Exploring the X-ray Universe*, (Cambridge University Press, Cambridge, 1995).

[82] M. Milgrom, A&A **67**, L25 (1978).

[83] R. A. Hulse and J. H. Taylor, ApJ **195**, L51 (1975).

[84] R. Kelley et al., ApJ **268**, 790 (1983).

[85] P. Hertz, K. S. Wood and L. R. Cominsky, ApJ **486**, 1000 (1997).

[86] C. E. Rhoades and R. Ruffini, Phys. Rev. Lett. **32**, 324 (1974).

[87] J. Casares and P. A. Charles, MNRAS **271**, L5 (1994).

[88] B. L. Webster and P. Murdin, Nature **235**, 37 (1972).

[89] Adapted from E. D. Bloom in *Particle Physics, Astrophysics and Cosmology*, edited by J. Chan and L. de Porcel (Stanford University, Stanford, 1996) p. 159.

[90] Adapted from J. Orosz, private communication through J. E. McClintock (1998).

[91] S. Bahcall, B. W. Lynn and S. B. Selipsky, Nucl. Phys. B **325**, 606 (1989).

[92] J. C. Miller, T. Shahbaz, and L. A. Nolan, MNRAS **294**, L25 (1998).

[93] I. F. Mirabel and L. F. Rodriguez, Nature **371**, 46 (1994).

[94] J. A. Orosz and C. D. Bailyn, ApJ **477**, 876 (1997).

[95] M. A. Nowak et al., ApJ **477**, L91 (1977).

[96] I. F. Mirabel and L. F. Rodriguez, Nature **392**, 673 (1998).

[97] W. H. G. Lewin, J. van Paradijs and R. E. Taam, Sp. Sci. Rev. **62**, 223 (1993).

[98] T. E. Strohmayer et al., ApJ **498**, L135 (1998).

[99] G. B. Cook, S. L. Shapiro and S. A. Teukolsky, ApJ **424**, 823 (1994).

[100] R. Narayan and I. Yi, ApJ **428**, L13 (1994).

[101] R. Narayan, M. R. Garcia and J. E. McClintock, ApJ **478**, L79 (1997).

[102] R. Narayan, D. Barret and J. E. McClintock, ApJ **482**, 448 (1997).

[103] S. N. Zhang, W. Cui and W. Chen, ApJ **482**, L155 (1997).

[104] L. Stella and M. Vietri, ApJ **492**, L59 (1998).

[105] W. Cui, S. N. Zhang and W. Chen, ApJ **492**, L53 (1998).

[106] S. S. Eikenberry et al., ApJ **494**, L61 (1998).

[107] R. P. Fender et al., MNRAS **290**, L65 (1997).

[108] L. F. Rodriguez and I. F. Mirabel, ApJ **474**, L123 (1997).

[109] L. Lindblom, B. J. Owen and S. M. Morsink, Phys. Rev. Lett. **80**, 4843 (1998).

[110] B. J. Owen et al., Phys. Rev. D **58**, in press (Oct. 15, 1998).

[111] L. Bildsten, ApJ **501**, L89 (1998).

[112] W. Forman et al., ApJS **38**, 357 (1978).

DARK MATTER TOMOGRAPHY

Tony Tyson
Bell Labs, Lucent Technologies
Murray Hill, NJ 07974

ABSTRACT

Over cosmic time the small fluctuations in dark matter density in the early Universe grew via gravitational instability. Until recently our view of this process has been indirect, relying on radiation from the small baryonic component. If we could measure directly the development of these mass overdensities over a range of scales and look-back times, cosmology and the nature of the dark matter would be constrained in a new way.

Gravitational lensing provides an opportunity to image directly these dark matter concentrations. This article is a short tutorial and review of recent progress in strong and weak gravitational lens tomographic imaging of dark matter. The lens-distorted images of background galaxies are inverted, yielding a two-dimensional map of the projected dark matter overdensity. The maps cover a range of scales and thus constrain the nature of dark matter. On 10 kpc scales, high resolution strong lens inversions reveal a soft core in the mass distribution, unexpected in scale-free cold dark matter N-body simulations. On 10 Mpc scales, statistical weak lensing inversion maps of mass show the outer mass profile of overdensities in the linear regime. These weak lensing studies, using mosaics of CCDs, are opening a new window on the Universe.

© 1998 by Tony Tyson.

1 Cosmological Motivation

Cosmology is in the midst of a revolution. As exciting new observational techniques and experiments loom on the horizon, the current observations have already replaced the standard cosmology of the 1980s with a new cosmology. In turn, this new cosmology raises very different questions about the nature of mass and energy in the Universe. Defining Ω as the sum over all forms of mass-energy, we can summarize the paradigm shift by listing the key parameters. Indeed, the parameter list itself is changing as old parameters are pinned down and new ones are forced on us.

Standard Cosmology (1980s):

$$\Omega = 1$$

$$H_0 = 50 - 100$$

$$\Lambda = 0$$

$$\Omega_{matter} \approx \Omega_{WIMP,axion} \sim 1 >> \Omega_{baryon}$$

Current Model (1998):

$$\Omega_{matter} \approx \Omega_{WIMP,axion} \approx 0.2 - 0.4 >> \Omega_{baryon}$$

$$H_0 = 67 \pm 10$$

Either $\Omega \neq 1$ or $\Lambda \neq 0$

Either way, we are led to new physics. The nature and distribution of dark matter plays a central role. Ω_m has been obtained mainly through extrapolating indirect observations of luminous clusters of galaxies to the rest of the Universe. If we could pin down Ω_m via a direct observation, this would significantly clarify cosmology. It may be premature to conclude that we have accounted for most all the dark matter, since until very recently we have relied on the notion that radiation traces mass.

Inflation theories are based on a flat Universe as suggested by the high degree of isotropy of the cosmic microwave background: total energy density equal to the critical density. Since ordinary matter (baryons) contributes only about 5% of the critical density, there must be something else. The leading candidate is relic, elementary particles. Generically, they fall into two classes: relativistic "hot" dark matter, and non-relativistic "cold" dark matter. Neutrinos of rest mass 30 eV or so are the prime examples of hot dark matter; they move quickly because they were once in thermal equilibrium and are very light. Axions and neutralinos are examples of cold dark matter. Neutralinos move slowly because they too were once in thermal equilibrium and they are very heavy. Axions are extremely light, but were never in thermal equilibrium (having been produced very cold).

Some candidates for the non-baryonic dark matter fit the observed large scale structure less well than others. Relativistic hot dark matter, such as neutrinos with a small rest mass (30 eV would be cosmologically significant), form structure from the top down and too late. Neutrinos also free stream [Landau damping or collisionless phase mixing] on galaxy scales, suppressing galaxy formation via gravitational instability. Observations very clearly indicate that galaxies formed at redshifts $z \sim 2 - 5$, before superclusters which are just forming today. So hot dark matter is not a major component of the dark matter.[1]

Nonrelativistic weakly interacting elementary particles (so-called Cold Dark Matter) work better. Cold Dark Matter (CDM) allows structure to form earlier via bottom-up hierarchical accretion onto dark matter overdensities. Given enough time (old enough Universe) and the right initial spectrum of density fluctuations, it ought to be possible to account for the observed structure evolution. Cold dark matter particles cannot move far enough to damp perturbations on small scales, and structure then forms from the bottom up: galaxies, followed by clusters of galaxies, and so on.[2] For a mass fluctuation spectrum normalized to the cosmic microwave background anisotropies found by the COBE satellite, most galaxies should have formed around redshifts $z \sim 2-4$, just as recent observations indicate. For example, evidence for early structure formation is shown in Fig. 1 where the star formation rate gleaned from studies of distant galaxies and quasars is shown vs look-back time or redshift.

At present, the cold dark matter + inflation scenario looks promising since it is consistent with a large body of observations: measurements of the anisotropy of the cosmic background radiation (CBR), redshift surveys of the distribution of matter today, and deep probes of the Universe. However, observations have hardly begun to discriminate between different inflationary models and versions of CDM. At the moment, we can say that the data favor a flat Universe, almost scale-invariant density perturbations, and cold dark matter with a small admixture of baryons.

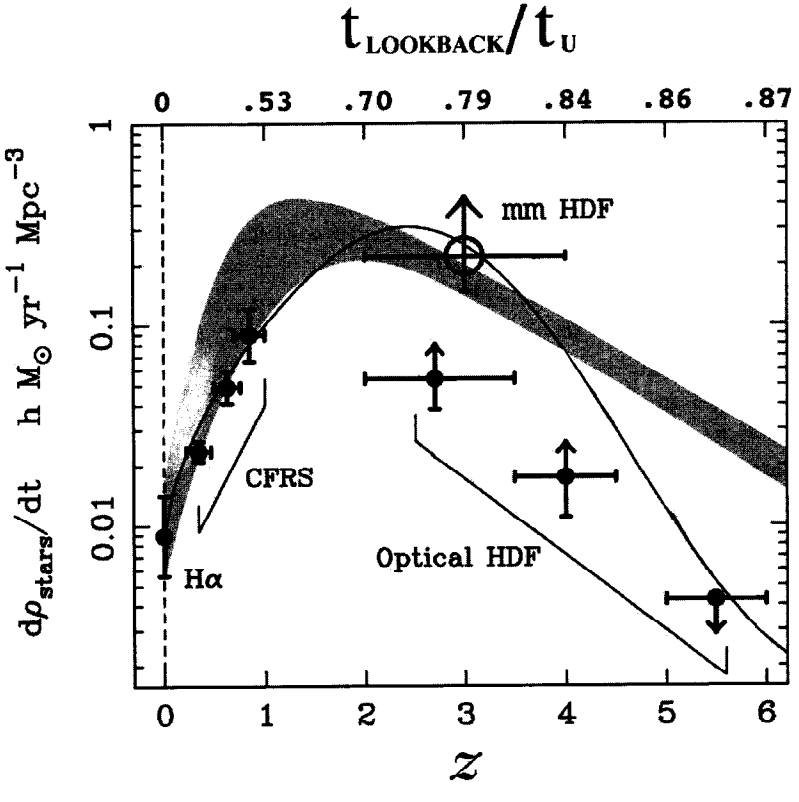


Fig. 1. Star formation vs look-back time or redshift, from studies of galaxies and quasars. The optical Hubble Deep Field data have not been corrected for dust absorption in this plot. These distant objects, if resolved, may be used as tools for measuring intervening mass concentrations.

Surprisingly, simulations of structure formation based on "standard" $\Omega = 1$ CDM with scale-free Gaussian initial fluctuations and zero cosmological constant still make galaxy clusters too late. When normalized to the CMB fluctuations on the 10^3 Mpc scale found by COBE, standard CDM produces too much power today on scales of galaxy clusters, creating relatively too few rich clusters at redshift 0.5. Indeed, a successful detailed model of structure formation must be compatible with the existence of galaxies at redshifts greater than 5, as well as the observed number densities of clusters of galaxies vs redshift.

Most of the matter in the Universe is of unknown form and dark: Stars (and related material) contribute a tiny fraction of the critical density, $\Omega_{\rm lum} = 0.003 \pm 0.001h^{-1} \sim 0.004$, while the amount of matter known to be present from its gravitational effects contributes around 100 times this amount, $\Omega_{matter} = 0.3 \pm 0.1$. The gravitational mass of dark matter is needed to hold together everything large in the Universe: galaxies, clusters of galaxies, and superclusters. A variety of methods for determining the amount of matter all seem to converge on $\Omega_{matter} \sim 1/3$; they include measurements of the masses of clusters of galaxies and the peculiar motions of galaxies. It is the prospect of directly measuring Ω_m and mass structure that is the subject of these notes.

Finally, the theory of Big Bang nucleosynthesis and the recently measured primeval abundance of deuterium[3] tightly constrain the baryon density: $\Omega_{baryon} = (0.02 \pm 0.002)h^{-2} \sim 0.05$. The factor of ten discrepancy between this number and dynamical and lensing measurements of the matter density is evidence for nonbaryonic dark matter.

2 Historical Perspective

Dark matter leads to large scale structure formation, via gravitational instability. Today, our task is to find out how much there is and where it is. Historically, the realization (not widely shared at the time) that large structures, clusters of galaxies, were dominated by non-luminous matter came in the 1930s. Fritz Zwicky, shown in Figure 2, found an amazingly high dispersion in radial velocities of galaxies in a cluster; about an order of magnitude larger than expected from the virial theorem if each solar luminosity was assigned one solar mass.[4,5] Zwicky went on to suggest a number of gravitational lens investigations using galaxies as lenses, which are just being done now. It took 30 years for the next decisive step

Fig. 2. Fritz Zwicky, late 1960s. Courtesy Wal Sargent.

and for the real beginning of acceptance by the scientific community of the notion that most of the mass in the Universe is dark.

In the 1960s Vera Rubin and Kent Ford were taking spectra of nearly edge-on spiral galaxies, using the new image intensifiers and a long slit placed along the plane of each galaxy. The resulting radial velocity as a function of position through the galaxy looked pathological; they found non-Keplerian orbits for stars and gas in spiral galaxies, implying an enclosed mass rising $\sim r$, and dark massive halos. While the nuclei of these galaxies were optically luminous, the mass seemed to be mostly in a diffuse outer halo where there was little or no light. As the debate over the amount of mass in the Universe and distribution of mass within galaxies raged, some began to ponder the observational consequences of a "plum pudding" Universe[6] and heavy galaxies. In 1973, Press and Gunn showed[7] that $\Omega \sim 1$ in dark compact objects would produce too many "Einstein rings." Today the outer extent of galaxy dark matter halos is still an open question, although the volume normalized galaxy mass in a large volume is understood to be much less than the critical mass density.

The first cosmic mirage was discovered by Walsh *et al.* in 1979:[8] they found a doubly lensed quasar Q0957+561. The lens, a foreground galaxy with a massive dark halo, produces two bright images of the same quasar (presumably in addition to a third, highly demagnified image)—an example of strong gravitational lensing. This event, happening as it did in the midst of speculation about dark matter halos, ushered in the era of gravitational lensing. In the following decade, lensing evolved from a curious rare phenomenon into an astrophysical tool. On the theoretical front, it was realized in the 1980s that dominant mass dark halos were good: they stabilize galaxy stellar disks. Technology drove new observations. Like the introduction of the image intensifier in the 1960s, the debut of the CCD in astronomy would have consequences for our understanding of dark matter. Our discovery of the faint blue galaxies (first a hint with deep photographic plates, and then their obvious presence in CCD images in the early 1980s) immediately suggested their use as a tool in studying dark matter via statistical weak gravitational lensing.

Another form of strong gravitational lensing, but this time with resolved background galaxies, was discovered in clusters of galaxies in 1987 nearly simultaneously by two groups (Lynds and Petrosian in the US, and Soucail *et al.*in France). A single background galaxy is heavily distorted into one or more long arcs con-

centric with the mass centroid of the cluster. This type of strong lensing forms the basis of the highest resolution mass maps (see below).

Today, we have two kinds of dark matter, and two problems. Hot and cold. Baryonic and non-baryonic. Most of the baryonic matter is dark, and most of the dark matter is non-baryonic. A key problem with viewing our Universe with traditional techniques (velocity dispersion, X-ray gas luminosity) is that we have been literally looking under lamp posts. So it is necessary to take what we currently know about the mass content of the Universe with a degree of caution. With the new tool of gravitational lens tomography, there are several questions that may be addressed: Have we found it all? Perhaps there is dark matter outside luminous clusters and superclusters. We already know mass-light segregation occurs on some scales. Does the distribution of mass inside clusters agree with the N-body theory predictions? Cold dark matter is scale-free. Does the mass halo cut off sharply at large radius? What about the vast volume in between clusters? Is there dark matter there and what are its morphological properties? Fundamentally, we need to weigh space. Given that no complete survey of mass by lensing has yet been done, we must remain open minded. Is $\Omega_m < 0.5$? Is $\Omega_\Lambda \neq 0$?

3 Clusters: The Place to Be and Be Seen

Our view of the Universe traditionally has been baryon-biased: we see or detect luminous objects. Yet for the past 15 years it has become more certain that the Universe is dominated by non-baryonic dark matter.[2,9] Of course, luminous objects, through their dynamics, can supply indirect estimates of the local underlying mass. On very large scales, our currently favored notion of hierarchical structure formation predicts that mass and light should be correlated.

As overdense regions in the Universe grow, some of them become sufficiently overdense that they "freeze out"; inside some radius they overcome the expansion of the Universe. As their density rises into the non-linear regime, their internal density fluctuation spectrum bears little resemblance to the primeval spectrum. However, it is hoped that at least the matter *content* of clusters is representative of the Universe. It is probably too much to hope that their mass-to-light ratio is also representative. Traditionally, clusters have been found via their radiation (optical light or X-ray). Until recently, the masses of galaxy clusters have played a central role in our understanding of dark matter problems.

Cluster masses can be estimated by three different techniques which, on scales of 1 Mpc, give consistent results. The first, which dates back to Zwicky (1933), uses the measured velocities of cluster galaxies and the virial theorem to determine the total mass (i.e., $KE_{\rm gal} \simeq |PE_{\rm gal}|/2$). Assuming further that the mass-to-light ratio of clusters is the same as the mean mass-to-light ratio in the Universe, one can combine measurements of the dispersion of velocities of hundreds of galaxies in several clusters with a measurement of the mean light density to arrive at an estimate of the mean mass density[10]: $\Omega_M \approx 0.25$. If clusters have more luminosity per mass than average, this technique would underestimate Ω_M.

The second method uses the temperature of the hot X-ray emitting intracluster gas and the virial theorem to arrive at the total mass. In this method, one must assume that the hot gas is pressure supported and in equilibrium. The density and temperature profiles of the hot gas must also be obtained. Often temperature is not mapped, so isothermality is assumed when inverting flux maps to get the potential.

The third and most direct method is utilizing the gravitational lensing effects of the cluster on much more distant galaxies. Light from distant galaxies scatters off the mass gravitationally. For small impact parameters, close to the cluster center, lensing is strong enough to produce multiple images. Farther out, lensing distorts the shape of distant galaxies. The lensing method allows the cluster surface mass density to be mapped directly. [In the following I will use h, the Hubble constant in units of 100 km s^{-1} Mpc^{-1}. Recently, h has been measured[11] to 10% accuracy, $h = 0.67 \pm 0.1$.]

X-ray measurements more easily determine the amount of hot, intracluster gas; most of the baryonic mass in a cluster resides in this form rather than in the mass of individual galaxies (this fact is also confirmed by lensing measurements). If we can assume that clusters are fair samples of matter in the Universe, we can use baryogenesis to get an indirect estimate of Ω_M. Together with the total cluster mass, the ratio of baryonic mass to total mass can be determined; a compilation of the existing data[13] gives $M_B/M_{\rm TOT} = (0.07 \pm 0.007)h^{-3/2} \sim 0.15$. Invoking the "fair sample" ansatz $\Omega_B/\Omega_M = M_B/M_{\rm TOT}$, the accurate BBN determination of Ω_B can be used to infer: $\Omega_M = (0.3 \pm 0.05)h^{-1/2} \sim 0.4$.

Traditional methods of estimating total mass of clusters are indirect: mass estimators based on kinematics or X-ray flux maps involve models of orbits or the

state of the hot gas, leading to possible systematics in the derived gravitational potential. While the dynamical evidence for dark matter is strong on scales from galaxies to superclusters of galaxies, it is worthwhile exploring independent observational techniques which do not rest on assumptions about the orbits of test particles. But luminous clusters may be a biased sample.

There are two ways in which confining the study of mass to clusters may lead to a biased measure of the mass content of the Universe. First, the mass-to-light ratio within the cluster may not be typical; i.e., a higher baryon fraction may lead to more light per mass. Second, being selected because of their luminosity, clusters may be a statistically biased sample of mass. If we examine other regions in equal detail, we will be able to examine this possibility.

4 Introduction to Lensing

Gravitational lens distortion of background galaxies enables calibrated measurements of the distribution of dark matter in the Universe. This lens phenomenon is most naturally divided into two broad classes: strong lensing and weak lensing. In weak lensing the gravitational deflection angles are very small and single sources produce single (but distorted) images. In strong lensing, sources appear highly distorted and can form multiple images. Whether a given source is weakly or strongly lensed depends on the impact parameter of the ray: whether its image appears outside or inside the critical "Einstein" radius. For an excellent lensing tutorial, see Narayan and Bartelmann.[12]

The large number of potential sources in the wide area outside the critical radius of a foreground lens offers the possibility of statistical tomographic reconstruction of the mass distribution in the outer parts of the lens. This weak gravitational lensing provides a direct measure of mass overdensity on a variety of scales (several kpc to tens of Mpc, depending on the distance and compactness of the lens).

Strong lensing analysis constrains the mass distribution in the parts of the lens which exceed the critical density for image splitting. In cases where multiple images of a source are created by the lens, the details of the position and distortion of these sub-images are highly sensitive to the projected two-dimensional mass distribution within the lens.

In weak lensing it is necessary to average over the apparent orientation of tens of sources for each mass resolution element: typically, weak lens statistical inversion uses tens of thousands of "arclets" (distorted background galaxies) over a wide field.[14,15] The largest source of noise in weak lens inversion of deep optical images is the ellipticity noise of the source population itself; galaxies are not round and they come in all orientations. A number of non-parametric algorithms have been developed for inverting the arclets, and regularized approaches have been developed to avoid edge effects and systematics near the strong lens regime.

Up to a constant sheet of mass, lensing inversion can be calibrated and has no adjustable parameters. Calibration of the mass scale can be done both through simulations and via observation of a mass standard. Realistic simulations of the whole source-lens-atmosphere-detector process must be performed, including multiple background galaxy redshift shells, masses for individual cluster galaxies, dark matter lens model, atmospheric seeing, and pixel sampling and sky shot noise. "Blank" field Hubble Space Telescope imaging data, together with seeing deconvolved ground-based data, are used to derive the source galaxy angular scales. Mass calibration via strong lensing forms an independent check on weak lensing mass scale calibration. There are now many of these mass calibrator clusters over the sky; the redshift of the arc is known so that the mass enclosed with that radius is known. Through comparison with large N-body simulations for various cosmogonies, this new window on mass in the Universe constrains the nature of dark matter.

5 The Sources

To 30th magnitude per square arcsecond surface brightness (4×10^{-18} erg sec^{-1} cm^{-2} arcsec^{-2} in 100 nm bandwidth at 450 nm wavelength, or about five photons per minute per galaxy collected with a four-meter mirror), there are about 50 billion galaxies over the sky. Nature has been particularly kind to us: Since many of the faint galaxies are resolved and are distributed up to high redshifts, they may be employed in statistical gravitational lens studies of foreground mass distributions. As we will see, if galaxies were a few times smaller or less luminous in the past, this "cosmic mirage" would be extremely difficult to use. For mass mapping by statistical gravitational lens inversion, the sources must meet several requirements: The sources must have (1) redshifts large compared with

the lens, (2) a number density on the sky sufficient to sample the lens shear field on relevant scales, (3) an intrinsic angular diameter larger than the ratio of seeing FWHM to the magnification of the lens (i.e., we must be able to detect the image distortion), and (4) other properties (blue color and unique surface brightness) enabling efficient separation of the sources from superimposed galaxies within the lens and foreground.

The extreme blue colors of the faint galaxies results from their large look-back times: we see these young galaxies as they were billions of years ago just forming a generation of hot and UV-bright stars at redshifts of 1 to 3, so that their UV excess flux is redshifted into the blue. In a redshift-magnitude plot of hundreds of galaxies, the trend to redshift ~ 1 at 25th B magnitude is clear. A typical galaxy seen at $z = 1$ may be a 0.1 L^* galaxy, so a survey at 25th magnitude would cover a wide range in redshifts extending from 1 to 3. For arclet inversion of $z < 0.3$ lenses, the lack of detailed redshift data for each of these sources produces less than a 10% mass scale error. We can do even better: if deep imaging is done in several wavelengths, there is enough low resolution spectral information to supply so-called "photometric redshifts," accurate to $\Delta z = 0.1$, so we can separate galaxies into broad foreground and multiple background redshift bins. This raises the lens inversion signal-to-noise level. The number density of galaxies on the sky vs magnitude is shown in Figure 3.

Without luminosity evolution these distant background galaxies would have prohibitively faint surface brightness ($SB \sim (1+z)^{-4}$). Even with their luminosity evolution (brighter earlier), very faint levels of surface brightness must be achieved. This is the most difficult part of this technique, and is why many orbits on HST or several nights on large ground-based telescopes are required per field in order to achieve good signal-to-noise imaging in several wavelength bands. At any surface brightness limit, there is a redshift selection function, as shown in Figure 4. Two limiting surface brightnesses are used in this simulation of the number count vs redshift of 25-26th magnitude galaxies: 29 mag arcsec^{-2} (typical of four hours of large ground-based telescope CCD exposure) and 26 mag arcsec^{-2} (typical of several orbits of the HST CCD). As may be seen, there is a huge increase in the number of galaxies behind a redshift 0.5 lens by going from 26 to 29 surface brightness (or about ten photons per minute per galaxy collected with a four-meter mirror). The largest background is the Poisson noise from the "dark" no-moon night sky diffuse light (mostly low level aurora) at a surface brightness of about

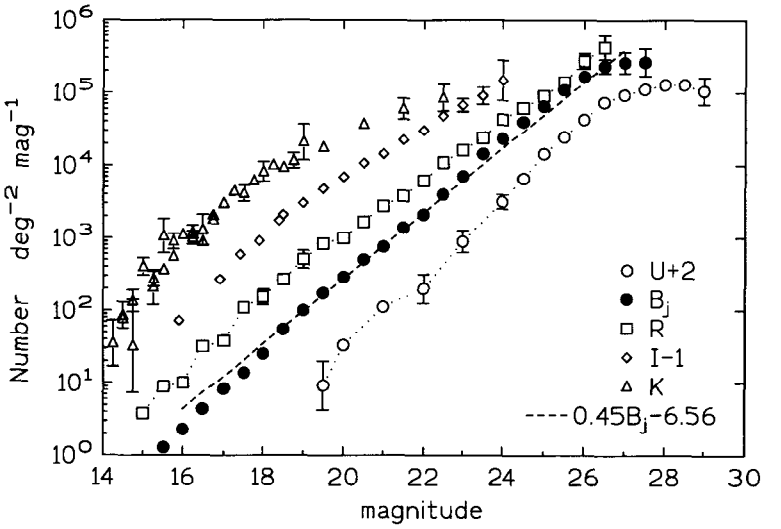


Fig. 3. Number density of galaxies (the sources) on the sky, as a function of magnitude (-2.5 log flux) for several wave-bands, U (360 nm), Bj (450 nm), R (600 nm), I (800 nm), and K (1200 nm). There is a faint galaxy every few arcseconds on the sky, in deep imaging on large telescopes, sufficient to define the shear field of a foreground gravitational lens.

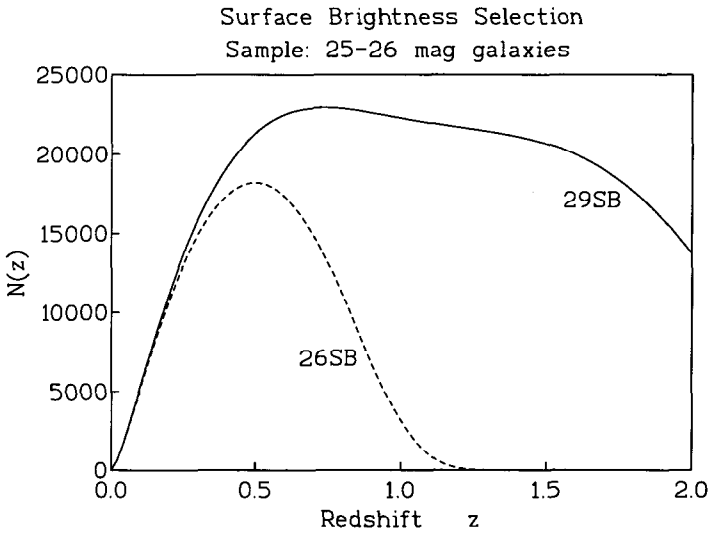


Fig. 4. The effects of a surface brightness limit on the distribution of redshifts found in a simulated survey of 26 magnitude galaxies. Due to the Hubble expansion, higher redshift galaxies appear to have fainter surface brightness, thus systematically dropping out of a surface brightness limited sample. In this simulation, the higher star formation rate of galaxies in the past was taken into account.

23 mag arcsec^{-2} at 400 nm, rising by a factor of three at red wavelengths.

Galaxies fainter than 26 B magnitude also cover a wide range of angular scales; typical half-light diameters are 1 arcsecond.[16] Image quality is critical; galaxies of less than 1 arcsecond diameter must be resolved, in order to measure their ellipticity. Four- to six-meter class telescopes can image one square degree to a magnitude limit of 27 B (blue) mag and 29 B mag arcsec^{-2} surface brightness in several nights per wave-band. Sufficiently deep multi-wavelength imaging can provide the input data for studies of mass via gravitational lensing.

6 Light Bending

The path of a photon from a distant source is bent as it passes by the foreground mass (gravitational lens), making the source appear at an altered position (for a good review, see Ref. 17). This light-bending is accompanied by another first-order effect: systematic image distortions. If the background source image is resolved, then this image stretching is observable. A galaxy of angular size one arcsecond may appear to be moved by an angle $\beta = 4GM(r)/rc^2$ and distorted into an arc many arcseconds long by sufficient foreground mass $M(r)$ interior to projected radius r. The scattering time is small compared with the period of orbiting test particles, thus avoiding orbit assumptions.

The light deflection is proportional to the mass in the lens and is about two arcsec for a typical galaxy.[18] If ϕ is the intrinsic source position in the absence of the lens, the observed source appears at position θ:

$$\vec{\phi} = \vec{\theta} - \vec{\delta}(\vec{\theta}), \tag{1}$$

$$\vec{\delta}(\vec{\theta}) = (D_{LS}/D_S)\ \vec{\beta}(\vec{r}),$$

where r is the impact parameter for the ray at the lens, and D_S and D_{LS} are the observer-source and lens-source angular diameter distances. For a thin lens the light bending angle $\vec{\beta}$ is given in terms of the projected two dimensional mass density Σ:

$$\vec{\beta} = (4G/c^2) \int d^2u\ \Sigma(\vec{u})\ (\vec{r} - \vec{u})\ /\ |\ \vec{r} - \vec{u}\ |^2\ . \tag{2}$$

We would like to invert this relation and solve for the mass density $\Sigma(x, y)$. Since we do not know how the sources were distributed on the sky prior to lensing, this deflection itself is not observable. But its gradient, the shear, is directly observable. In the simple case of a pont mass, $\beta = 4GM/rc^2$, so a source exactly behind the mass appears as an "Einstein ring" image of radius $\theta_E = (M\ /\ 10^{11}\ M_\odot)^{1/2}\ (D\ /\ 10^6\ \mathrm{kpc})^{-1/2}$ arcseconds, where $D = D_L D_S/D_{LS}$. If the lens mass is elliptical in shape or has multiple mass components, this circular ring symmetry is broken, so that in nature circular rings are rare. If the source angle ϕ is less than θ_E, then multiple images of the source are formed. Galaxy lenses ($M \approx 10^{12}\ M_\odot$) will produce multiple images with separations ≈ 3 arcseconds,

while galaxy cluster lenses ($M \approx 10^{14}\ M_\odot$) can produce image separations of one arcminute. Astronomers often use velocity dispersion as a proxy for mass: for an isothermal distribution $M/r \sim v^2$ is independent of r. In the case of an isothermal distribution of gravitationally bound masses of line-of-sight velocity dispersion σ_v, $\beta(r) = 4\pi(\sigma_v/c)^2$, $\theta_E = 29\ (\sigma_v\ /\ 10^3\ \mathrm{km\ sec^{-1}})^2\ (D_{LS}/D_S)$ arcseconds. Many clusters have measured velocity dispersions $\approx 10^3$ km sec^{-1}. Most have $\sigma_v \approx 700$, and a few have $\sigma_v \approx 1500$ km sec^{-1}. As an aid in understanding the observational consequences of gravitational lensing, particularly strong lensing, it is useful to visualize the effects of such a massive cluster in a movie of 1000 realizations of lensing.

6.1 Simulation Movies

MPG MOVIES of our view, over the next few billion years, of a cluster of galaxies and the effects of its dark matter are shown in
"`ftp://netlib.bell-labs.com/cm/physics/who/tyson/MPEG/LENS0.mpg`"
and
"`ftp://netlib.bell-labs.com/cm/physics/who/tyson/MPEG/LENSMAX.mpg`".
These simulations were made by our group, principally Ayana Holloway. They show the cluster A1689 as viewed by Earth orbiting space telescopes over the next few billion years; as it moves through the Universe its alignment with the background faint blue galaxies constantly changes. The data for the cluster (the orange bright galaxies) are from the Hubble Space Telescope, and the mass model is one which we are developing for this particular cluster, based on very deep multi-color HST and ground-based imaging of A1689, via strong lens parametric techniques. (For a description of parametric strong lens mass determination, see Tyson, Kochanski, and Dell'Antonio.[19])

These first two movies are based on two extreme mass models: LENSMAX.mpg has much of the mass in the halos of the cluster galaxies, while LENS0.mpg has very little mass in the cluster galaxies (most mass residing in a single dark mass).

The background galaxies are placed in six redshift sheets, out to $z = 3$. Fainter and bluer galaxies are at higher redshift. Some galaxies just behind the cluster (coded brighter and less blue) are only slightly distorted by the concentration of dark matter. The highest redshift faint galaxies easily form large arcs (perturbed Einstein rings) at larger radii.

Our mass distribution is composed of dark matter halos for individual cluster galaxies as well as the dominant diffuse mass of the cluster. This largest mass has a soft core. A non-singular lens mass produces an odd number of images of a background galaxy. When the galaxy is at large projected radius (weak lens regime) it is slightly distorted orthogonal to the radius vector. As the galaxy moves into the critical region, it splits into an odd number of images, usually two arcs on opposite sides of the lens and a third, demagnified image near the center. The details of these images—their parity and magnification—are sensitive to the gradient of the lens mass. For regions near the core where the lens mass density exceeds the critical mass density, the magnification determinant is negative—the lens-focused light bundle comes to a focus before reaching Earth. Images inside this radius (one of the odd images) appear demagnified and with opposite parity. These faint radial spokes may be seen for many of the background galaxies over a finite range of impact parameter, and are a sensitive indicator of the size of the soft mass core. Another observed feature of our lens model for this cluster is the unusually thin arcs (images of high-z galaxies); this is due to a mass profile which is steeper than isothermal.

We spent some effort to mpeg encode these movies in a way that is maximally compatible with Windows and Unix mpeg players. If you experience problems, please let us know. We use MPEG-1 format.

7 Tomographic Reconstruction of Galaxy Cluster Mass

7.1 Weak Lens Statistical Inversion

Historically, the first solution to this inverse problem was via statistical weak lens inversion, in which the mean shear field is estimated from the tangential alignment of stretched single images of many independent background galaxies projected near the cluster.[15] It was logical to apply this new technique first to massive clusters where there was already indirect evidence for a high mass contrast. Weak lens statistical inversion still plays a dominant role in most gravitational lens examinations of cluster mass, since few clusters have the unlikely favorable alignment with a background source that gives rich multiple images of that single source. Pedagogically, however, strong lensing presents a striking demonstration of the

effects of lensing, as seen in the above movies. In principle, strong lensing also gives higher resolution mass maps than statistical weak lensing. I postpone discussion of the more generally useful technique of statistical weak lensing to a later section.

7.2 Strong Lens Parametric Mass Models

Multiple images of the same resolved source galaxy contain rich information related to the mass distribution in the lens. In cases where portions of the lens exceed critical density, producing multiple images of some sources, it is possible to obtain higher resolution in the mass map by combining weak lens inversion at large radius with the strong lensing constraints in the inner region. There are perhaps a few hundred moderate redshift clusters which have super critical core densities, and some of them have the chance alignment required to give obvious multiple resolved images of a bright single source. One approach is to use a parametric lens model in which multiple mass components are parametrized by their centroid, mass profile, and ellipticity. An iterative approach, regularized by the weak lens inversion solution for the mass at large radius, may then be used to obtain a unique solution for the lens mass map at high resolution.

This has the desirable property that the resulting mass map has high resolution in regions of high mass density. In practice we have found that a reliable way of converging to the solution is to demand that the strong lensed source images, when unlensed by the estimated mass distribution in the lens, reconstruct to an identical single image,[20] and that this image of the source maps to the observed arcs (and only these arcs!) in the image plane when lensed by this same mass. It is necessary to use the lens *mass* distribution, rather than the approach of starting with a two-dimensional lens potential, because non-circular lens potentials can be unphysical in their corresponding mass reconstructions.

With my colleagues Ian Dell'Antonio and Greg Kochanski, we have developed a code for this weak-strong mass reconstruction and applied it to complex lens simulations as well as deep HST imaging of two massive clusters, the first being the $z = 0.4$ galaxy cluster CL 0024+17. The projected mass density contrast at large radius is obtained from over a thousand arclets, excluding the bright long arcs, in a wide field using a ground-based telescope. The mass density contrast at intermediate radius (up to two arcminutes) is obtained from weak lens inversion of 350 arclets found in the HST deep images. Finally, the mass map in the inner region is reconstructed from the multiple strongly lensed images of a single source galaxy via the procedure outlined above.

As an example, Figure 5 shows a high resolution mass map of the $z = 0.39$ cluster 0024+1654, based on parametric inversion of the associated gravitational lens.[19] This lens creates eight well-resolved whole or partial sub-images of a background galaxy, seen in deep imaging with the Hubble Space Telescope.

Of course we do not know what the source galaxy looks like; so we cannot start with that prior. The source galaxy optical image must be part of the model to be optimized along with the mass distribution in the lens. We parametrized the source as 58 smooth disks of light. Each of these disks was characterized by an intensity, a scale radius, and the (x, y) position on the source plane (four parameters). A source plane resolution of seven milli-arcseconds per pixel was chosen to allow sufficient evaluations of the model to be done within a reasonable time (12 months), and to allow the model to represent almost all details of the observations. The source is then ray-traced through the lens plane, and the resulting image is compared on a pixel-by-pixel basis with the HST image. We parametrize the mass distribution as a cluster of mass concentrations ("mascons"). Each mascon is based on a power law model[21] ("PL"), for the mass density $\Sigma(r)$ vs projected radius r, with both an inner core radius and an outer cutoff radius

$$\begin{aligned} \Sigma(x) &= \frac{K_1(1+\alpha x^2)}{(1+x^2)^{2-\alpha}} \qquad x < X_o \\ \Sigma(x) &= K_2 x^{-3} X_o^3 \qquad x \geq X_o, \end{aligned} \tag{3}$$

where $x = r/r_{\rm core}$, $X_o = r_{\rm cutoff}/r_{\rm core}$, and α is the PL model index. ${K_1}^{0.5}$ is proportional to the central line-of-sight velocity dispersion. We build up elliptical mass distributions by superposing a line of overlapping circular mascons. In principle, each mascon is described by nine parameters, although most do not require more than a few parameters. The first four come directly from the equation above (K_1, an inner mass core radius $r_{\rm core}$, an outer mass cutoff $r_{\rm cutoff}$, and the slope of the mass profile α). For elliptical mass distributions, there are three parameters describing the ellipticity (the position angle θ, the length of the line of mascons l_{core}, and the uniformity of the spacing of the mascons along the line). For mass components not associated with optically observed galaxies, the x and y position in the lens plane are also free.

The mass and linear scale sensitivity of this parametric lens inversion technique vary with position in the cluster; cluster mascons projected near a long arc have the effect of their mass distribution highly magnified. For galaxies that are more than about five arcseconds from the arcs, only their total mass matters, and we parametrize this by the cutoff radius (because $M \propto r_{\rm cutoff}$). Galaxies farther than about 20 arcseconds from the arcs are parameterized in groups.

In all, the mass and source models are determined by 512 parameters. However, we have over 3800 significantly nonzero (3σ) pixels in the arcs which supply sufficient information. Because the optical point spread function of the WFPC2 is smaller than one pixel, the signal is nearly uncorrelated even on neighboring pixels; thus, we have many more independent constraints than model parameters. In addition, pixels with no signal serve as constraints, because they prevent the model from putting flux in areas of the image where there is none. The resulting mass model is over-constrained. The optimizer uses simulated annealing and bootstrap resampling to avoid false minima.

As we developed the model, it had enough power to predict the central image, based on the three major arcs, then to correctly predict the multiple subimages near the outer arcs. One of these was buried under a cluster galaxy and was recovered as a result of this prediction by modeling and subtracting the foreground galaxy's light. The galaxy masses in the model were initialized using the Faber-Jackson relation between a galaxy's optical luminosity and mass. When the model evolved to a low χ^2, we performed robustness tests by multiply perturbing the position and/or mass of mascons and observing the reconvergence to the solution (see Figure 5). Over 2×10^6 models were searched to reach the solution. In addition to the large diffuse mascons, other mascons were added to our fit to allow it to match the complexity of the cluster's mass distribution. Several "dark" galaxies were found.

The vast majority of the dark matter is *not* associated with the galaxies, and appears as a smooth elliptical distribution centered near the position of the brightest galaxy and elongated in the SE-NW direction. The elongation is in the same direction as that of the X-ray isophotes. This dark matter not associated with galaxies shows no evidence of infalling massive clumps: other than these two major clumps, we find no dark mascons with total mass greater than $5 \times 10^{12} M_\odot$ (1.5% of the cluster mass), out of the 25 in the fit.

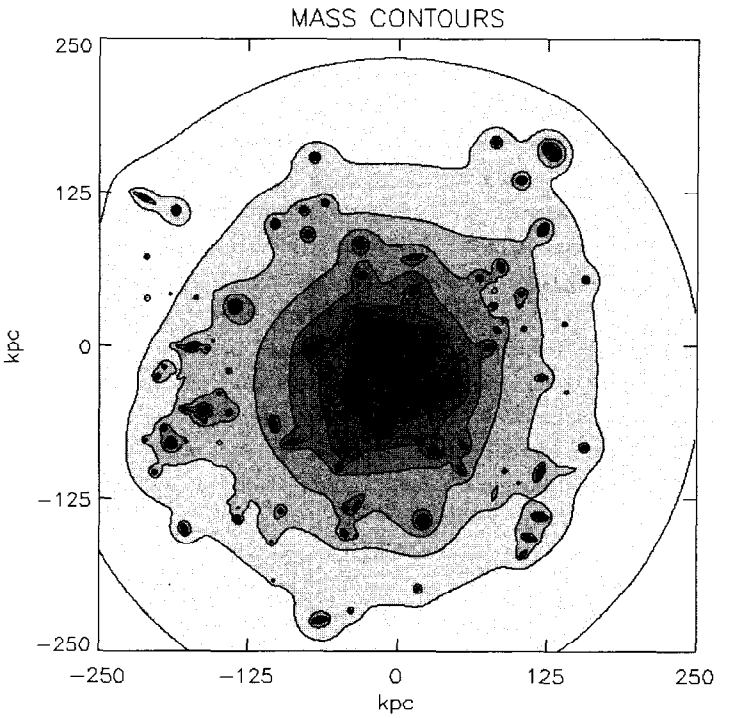


Fig. 5. A map of the projected mass density in the redshift 0.4 cluster CL0024+1654, obtained by fitting the observed distorted images of a $z = 1.6$ source galaxy. Dark matter halos of individual galaxies in the cluster are seen, but the major dark matter component is relatively smooth compared with CDM simulations.

Excluding mass concentrations centered on visible galaxies, more than 98% of the remaining mass is represented by a smooth concentration of dark matter centered near the brightest cluster galaxies, with a 33 h^{-1} kpc *soft core*. The asymmetry in the mass distribution is <3% inside 107 h^{-1} kpc radius. The dark matter distribution we observe in CL0024 is far more smooth, symmetric, and nonsingular than in typical simulated clusters in either $\Omega = 1$ or $\Omega = 0.3$ CDM cosmologies. Integrated to 107 h^{-1} kpc radius, the rest-frame mass to light ratio is $M/L_V = 238 \pm 16\ h\ (M/L_V)_\odot$, rising with radius (see Figure 6). Because galaxies were brighter in the past, this translates to a mass-to-light ratio of 400 h now. For scale, we would need around 1200 to extrapolate to $\Omega_M = 1$.

The mass distribution in CL0024 is remarkably relaxed. If one assumes Gaussian density fluctuations in an $\Omega = 1$ Universe, the fluctuation that seeded CL0024 must have had a very large amplitude quite early (rare) to have become virialized by $z = 0.39$. One is led to consider non-Gaussian fluctuations or $\Omega << 1$. An important open question is whether galaxy cluster formation is still continuing at recent times. That is, do a significant fraction of galaxies fall into a pre-existing deep cluster potential? Given that galaxies are detected at redshifts as high as 4, simple top-down (large scales first) structure formation is ruled out. But there may be several components to the mass-energy of the Universe, and thus multiple contributors to the density profile of clusters.

A key result of this first high resolution mass map is the existence of a 33 h^{-1} kpc soft core. Any possible singularity must be quite small, contributing less than $2 \times 10^{11}\ h^{-1} M_\odot$ to the total mass within 33 h^{-1} kpc (10% of the mass of one of the central elliptical galaxies). Because cold collisionless particles have no characteristic length scale, the soft core suggests either nongravitational interactions or gravitational coupling to a dissipative baryonic component of significant mass fraction. While HDM can produce soft cores, HDM is not consistent with the high density of dark matter that we find in the individual cluster galaxies. Because of the relatively low X-ray luminosity (0.7 L_X^*), it is unlikely that this can be attributed to hot gas alone.

High resolution mass maps of clusters of galaxies will be useful to compare with future N-body/gas-dynamical simulations. None of the recent simulations show evidence of a soft core, in disagreement with these observations. Indeed, as the resolution of simulations increases from 100 to 30 kpc the central mass becomes more singular.[22] At this time it is too early to conclude whether this

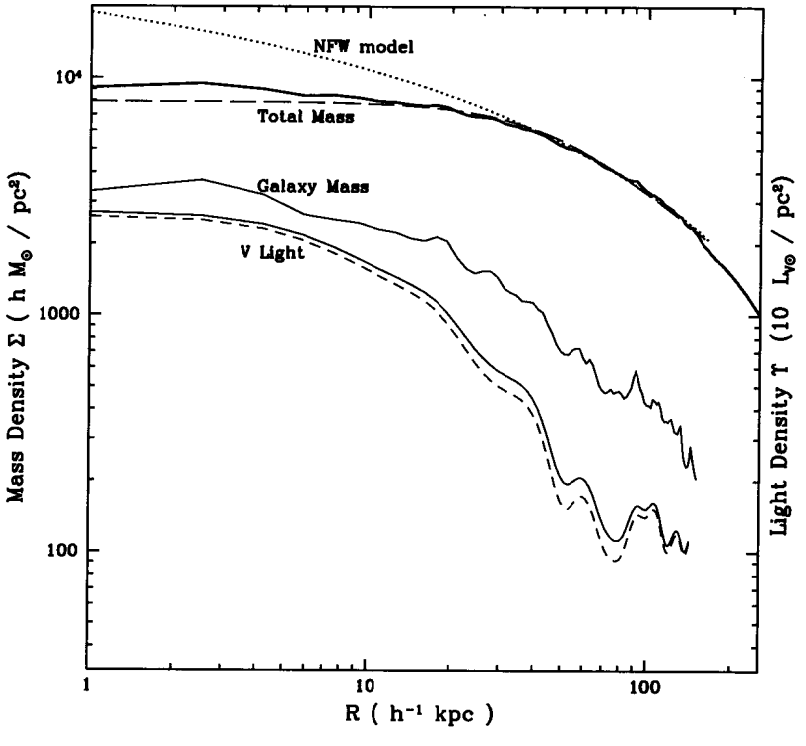


Fig. 6. The projected mass density of CL0024 and the projected light density are plotted vs radius. The diffuse dark matter component dominates, and has a soft core. Note also that the mass-to-light ratio rises with radius.

soft mass core is a result of baryon interactions or a feature of non-baryonic dark matter. Another technique, the Sunyaev-Zel'dovich effect, which makes use of the Compton up-scattering of CMB photons when viewed through a cluster containing hot gas, may soon prove helpful in resolving this question. While the angular resolution achievable in this CMB decrement imaging is lower than the strong lens inversion maps, "SZ" imaging of the projected hot gas density depends on redshift and other cosmological parameters in the same way as gravitational lens mass imaging, so that the cluster gas to total mass ratio may be derived directly from the combined lens/SZ data.

8 Statistical Weak Lensing

If the source angle relative to the lens ϕ is larger than θ_E, multiple images of a source will not be formed, but the single image of the source will be elongated as shown in Figure 7. Since there are roughly 100 independent resolved sources per square arcminute, this enables a different kind of solution to the inverse problem, called "statistical weak lensing." This method has the advantage that it is generally applicable, and gives good signal-to-noise inversions in regions of low shear which are much more representative of random regions.

The huge mass associated with clusters of galaxies distorts all the background galaxies many arcminutes from the cluster. Foreground galaxy clusters at redshifts 0.2(0.5) with radial velocity dispersions above 700 km sec^{-1} have sufficient mass density to significantly distort background galaxies of redshift greater than 0.4(1). Lensing preserves the surface brightness and spectrum of the source, so that arcs tend to have the very faint surface brightness and blue color of the faint blue galaxies. A simulation of both weak and strong lensing is shown in Figures 8 and 9. First, a long cone is populated with galaxies of size and luminosity appropriate to their epoch, out to redshift 3. Then a cluster mass is placed at redshift 0.3, distorting the images of these background galaxies.

8.1 Inverse Problem: Mass from Image Distortion

This gravitational lens distortion is quantified using the intensity-weighted second moment of the galaxy image orthogonal and along the radius relative to the lens center.[15] A dimensionless scalar alignment T, calculated from these principle axis

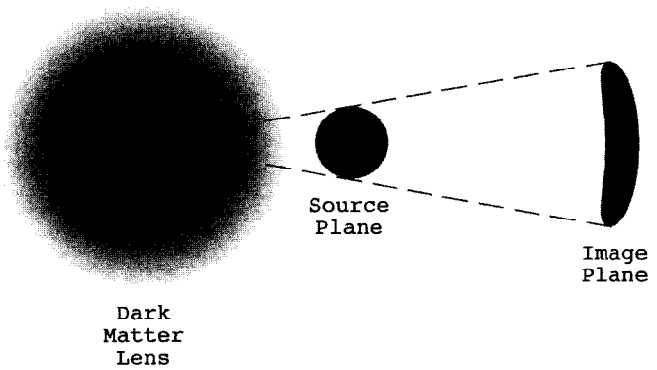


Fig. 7. Resolved background galaxies (source plane) are mapped into distorted images (image plane) by the light bending due to the foreground mass (left).

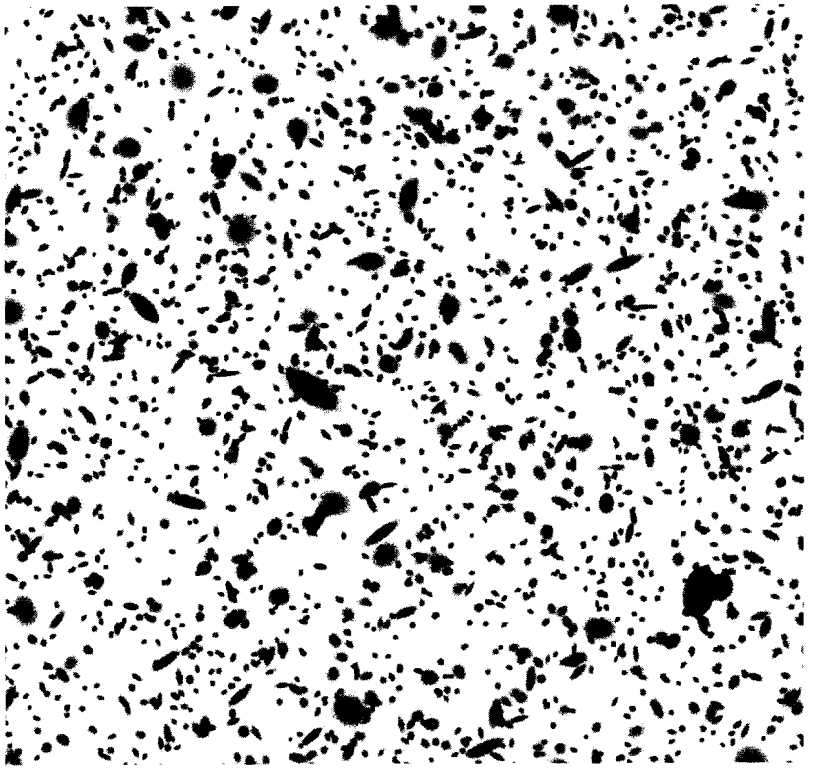


Fig. 8. A view of the sky, several arcminutes wide, looking through the forest of galaxies out to redshift 3. This simulation was made to simulate Hubble Space Telescope deep imaging data. No large dark matter concentration (lens) was included.

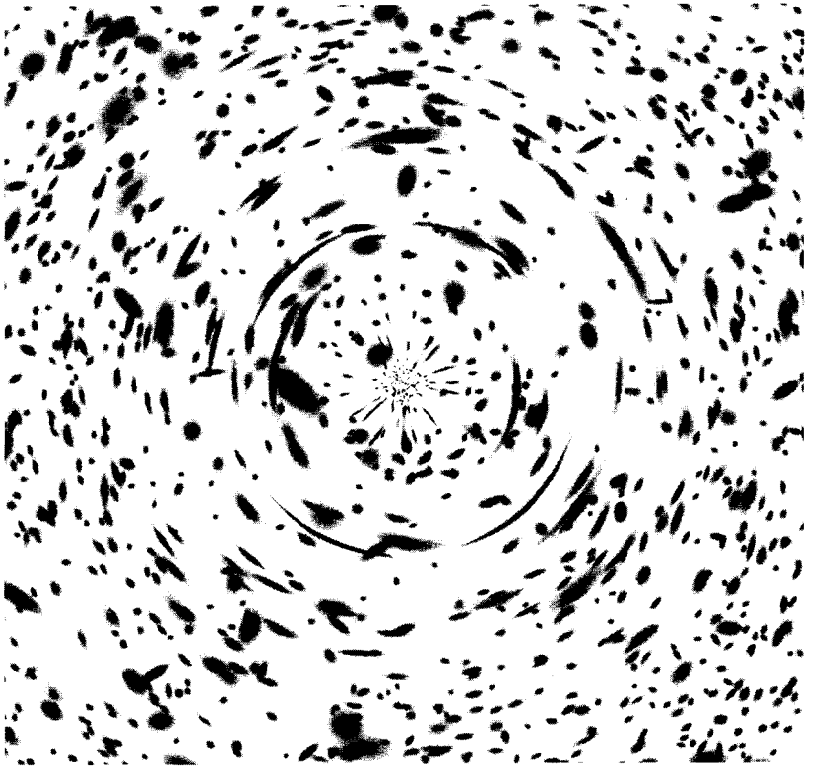


Fig. 9. The effects of a dark matter concentration of the type found in rich clusters of galaxies on the simulated sky shown in the previous figure. Foreground galaxies are unaffected; background galaxies are increasingly distorted at higher redshift. Both strong lensing (large arcs, some multiple images of a single source) and weak lensing (mildly distorted single source galaxy image) are seen. The radial spikes close to the center are a diagnostic of a soft mass core.

transformed source ellipticities, is related to the projected mass density clumping, and is defined at each point $(\vec{r})$ in the image plane via the (r, θ) principal-axis transformed second moments of the background galaxy image:

$$T(\vec{r}) = \frac{i_{\theta\theta} - i_{rr}}{i_{\theta\theta} + i_{rr}} = \frac{2(1-\kappa)\gamma}{(1-\kappa)^2 + \gamma^2} \simeq \frac{2\gamma(\vec{r})}{1-\kappa(\vec{r})}, \tag{4}$$

where the convergence[23] $\kappa(r) = \Sigma(r)/\Sigma_c$ and the shear $\gamma(r) = [\bar{\Sigma}(r) - \Sigma(r)]/\Sigma_c$, and where Σ_c is the critical surface mass density, related to the distance ratio: $\Sigma_c = c^2/(4\pi G D)$. The distance ratio for a foreground-background pair is[17]

$$D = \frac{(1 - q_o - d_1 d_2)(d_1 - d_2)}{(1 - q_o - d_2)(1 - d_2)(1 + z_{fg})}, \tag{5}$$

where $d_1 = \sqrt{1 + q_o z_{fg}}$ and $d_2 = \sqrt{1 + q_o z_{bg}}$.

Introducing a galaxy light distribution prior, the tangential second moments are

$$i_{\theta\theta} = M_{20} \sin^2\phi + M_{02} \cos^2\phi - 2M_{11} \sin\phi \cos\phi, \tag{6}$$

$$i_{rr} = M_{20} \cos^2\phi + M_{02} \sin^2\phi + 2M_{11} \sin\phi \cos\phi,$$

where ϕ is the position angle of the vector from the point (x, y) to the background galaxy, relative to the x-axis. The intensity-weighted second moment $M_{lm,g}$ of the background galaxy g is defined by

$$M_{lm,g} = M_{0,g}^{-1} \int (\delta x)^l \, (\delta y)^m \, W(\delta x, \delta y) \, [I_g(\delta x, \delta y) - I_0] \, dx \, dy, \tag{7}$$

$$M_{0,g} = \int W(\delta x, \delta y) \, [I_g(\delta x, \delta y) - I_0] \, dx \, dy,$$

where $\delta x = (x - < x >_g)$ and similarly for δy; the sky intensity near this background galaxy is given by I_0, and optimal normalized Gaussian weights (adaptive kernel) $W(\delta x, \delta y, I(x, y))$ are calculated from the half luminosity radius. An unlensed population of galaxies randomly placed and oriented will give a net distortion $T(\vec{r})$ of zero at every point in the image plane, while a population of lensed galaxies will give a positive value at the point corresponding to the lens center. At some point off the lens center $T(\vec{r} - \vec{r}_0)$ is proportional to the projected mass at that new point, as shown in Figure 10.

In three-dimensional simulations of deep shear surveys, it is apparent that in addition to the shear noise arising from intrinsic source galaxy random ellipticities there are errors arising from occasional chance projections of one arclet very near

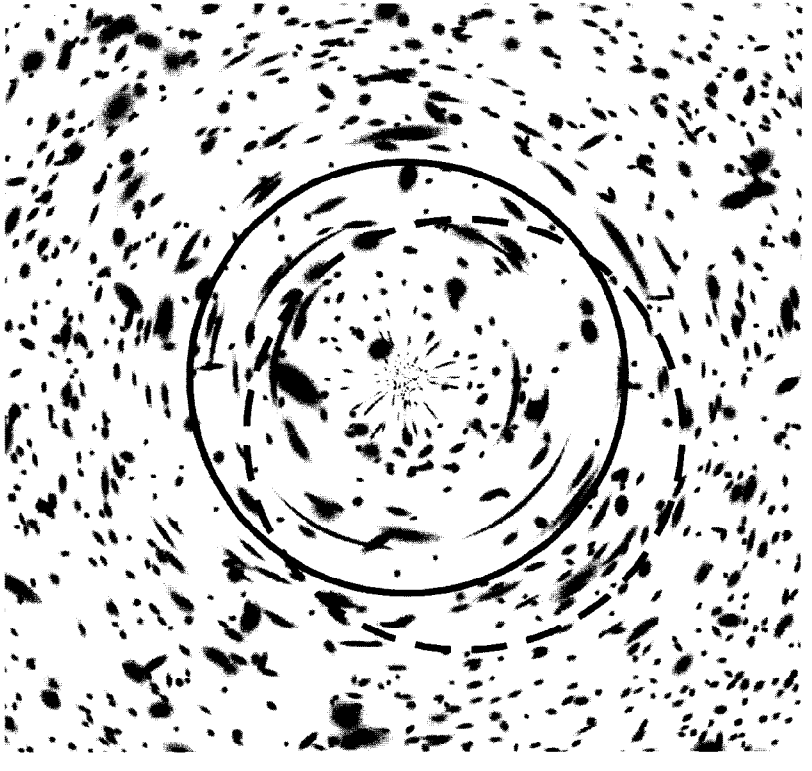


Fig. 10. The sum of the tangential component of the shear around the solid circle is simply related to the average mass density interior. This is also true for the offset (dashed) circle. This provides a method for mapping the projected mass density on the sky, producing images of the dark matter. The resolution is given by the size of the circle. In practice, a weighted average of an array of sizes is used in deriving the mass map.

or on top of another. We have found that a regularized iteration, in which the kernel $W(\delta x, \delta y)$ is elliptical and the maximum likelihood is matched to each arclet, is even more effective in reducing the tail of the arclet ellipticity noise.

9 Cluster Mass Distribution from Arclets

To construct an image of the gravitational lens projected mass distribution, the distortion statistic $T(\vec{r})$ may be computed over a grid of positions as candidate lens centers. In the weak lensing limit it can be shown that the tangential alignment T is a measure of the mass contrast:[14,23]

$$T(\vec{r}) \;=\; 2\,[\Sigma_{av}(<r) - \Sigma(r)]\;\Sigma_c^{-1}. \tag{8}$$

At any point $\vec{r}$ in the image plane we can sum over the tangential alignment of all source images about that point, creating a continuous scalar distortion statistic $\bar{T}$:

$$\bar{T}(\vec{r}) = \int K(\vec{u})\; T(\vec{r} - \vec{u})\; d\vec{u}, \tag{9}$$

where the apodization kernel $K(\vec{u})$ weights source images at large radius less, and is generally of the form $K(\vec{u}) \;=\; (u^2 \;+\; u_0^2)^{-1}$. For $\bar{T}$ to be simply related to the mass[24] $K(s)$ must asymptotically approach the power law s^{-2} at large s. The distortion $\bar{T}$ at any point is related to the mass density contrast, since any mass can be represented as a sum of cylindrically symmetric distributions, and light bending angles from different mass components add. The solution to this inverse problem for the contrast of the projected lens mass density Σ is then given by a simple integral of the shear over the radius from the lens center. The average projected mass density interior to radius r is given by

$$\bar{\Sigma}(r) \;=\; \Sigma_c C B(r) \int_r^{r_{\max}} T(r)\; dlogr \;+\; \bar{\Sigma}(r, r_{\max}), \tag{10}$$

where C is the seeing correction, obtained via simulations. The function B is given by $B(r) = (1 - r^2/r_{max}^2)^{-1}$, and $\bar{\Sigma}(r, r_{max})$ is the average density in the annulus between r and $r_{\max}$. For a sufficiently large field, $\bar{\Sigma}(r, r_{max})$ is small compared with peak density. This arclet inversion is called "aperture mass densitometry." No map is obtained; just the radial profile. A radial plot of the projected surface mass density found using Eq. 10 is shown in Figure 11 for $z_l = 0.3$, $\sigma_v = 1000$ km sec^{-1} simulations with two different mass profiles.

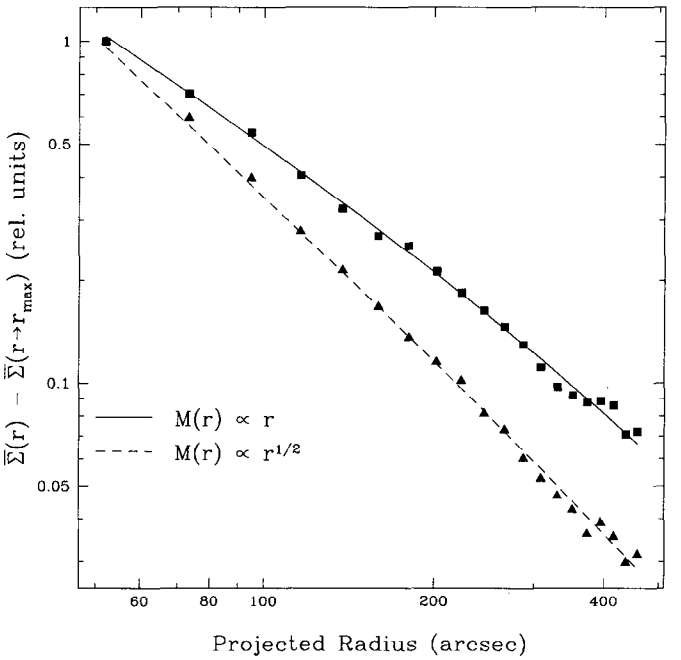


Fig. 11. A radial plot of the projected mass density contrast obtained by inversion of the arclets in two 29 B mag arcsec^{-2} noisy simulations of a cluster of galaxies dominated by dark matter with velocity dispersion $\sigma_v = 1000$ km sec^{-1}. Two test mass distributions are shown, arbitrarily normalized at 50 arcsec radius. The lines are the input mass density functions.

The distortion image $\bar{T}(x,y)$ uniquely locates the lens mass, obtains $M(<r)/r$, and gives its morphological shape on the sky. A useful check of this procedure uses the mass map (Eq. 9) as input to a Bayesian search in model mass distribution space, solving for the dark matter mass and core size. For the known source ellipticity distribution, a given source redshift distribution, and a test radial mass distribution model, a maximum likelihood calculation yields the lens M/r (or equivalent velocity dispersion) and core radius. From the inversion of 6000 arclets surrounding the rich $z = 0.18$ cluster A1689, we found a steeper than isothermal profile beyond 300 h^{-1} kpc radius.[25] Some clusters at high redshift are nearly as compact in mass. In rich compact clusters mass appears to trace the cluster red light, on scales of a few $\times 100\ h^{-1}$ kpc, with rest-frame V band mass-to-light ratios of a few hundred h in solar units. Weak lens tomographic studies of the outskirts of clusters with the Big Throughput Camera (see below) show that mass density falls rapidly at large radius as expected in models of structure formation, but there are often other interesting mass concentrations nearby.

The mass core radius is smaller than most observed X-ray core radii in nearby clusters, suggesting that the X-ray gas may be less relaxed dynamically than the dark matter. Over the past decade, theoretical CDM N-body simulation mass core sizes have evolved steadily downwards, past the observed mass cores, towards a singular distribution. In 1990 we reported a (30-50 kpc) mass core radius much smaller than the X-ray core (and then current N-body simulation cores). By 1998 the N-body simulations have high resolution and give mass cores much smaller than our observed mass cores (still 30-40 kpc). At least for rich clusters these lens studies appear to confirm the large mass which was implied by virial calculations using velocity dispersions.

Because faint blue galaxies become more common at redshifts above 0.5, most studies of weak lensing have been confined to lenses around redshift 0.2 to 0.4. However, clusters exist at higher redshift and two of them near $z = 0.8$ have been studied recently by weak lensing.[26] As multiband deep imaging becomes more common in lensing work, color-redshifts will enable weak lens studies of clusters at higher redshift. Figure 12 shows an example of weak lensing reconstruction of mass, with 1st order strong lensing corrections included. See also Squires *et al.*1997.[27]

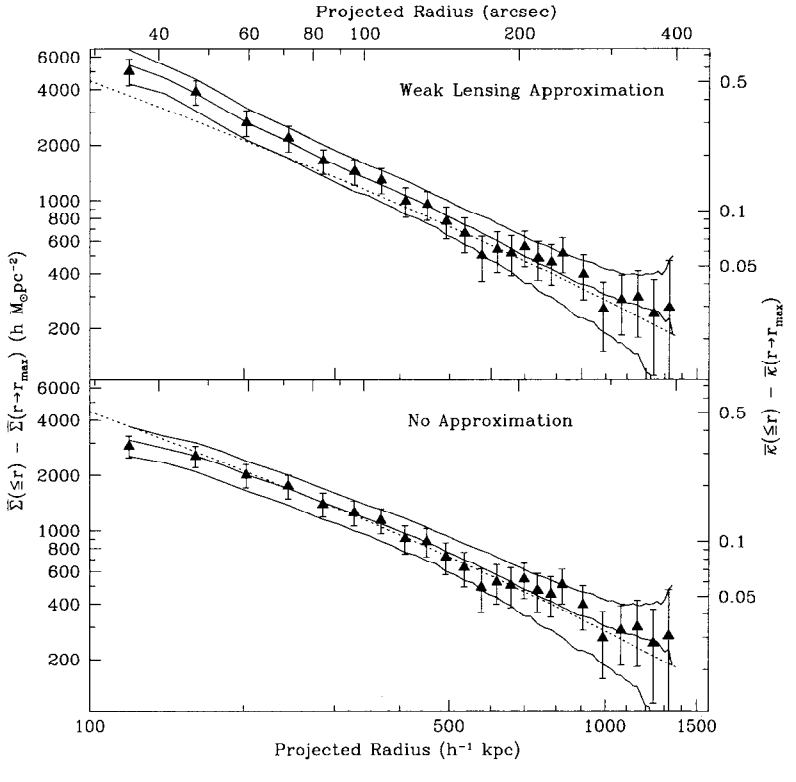


Fig. 12. The radial mass distribution of the cluster RXJ1347, with and without strong lensing corrections,[28] as in Equation 3. The cluster RXJ 1347.5-1145, at a redshift of $z = 0.451$, has an X-ray luminosity $L_X = 2 \times 10^{46}$ erg s^{-1}. $\kappa = \Sigma/\Sigma_{crit}$.

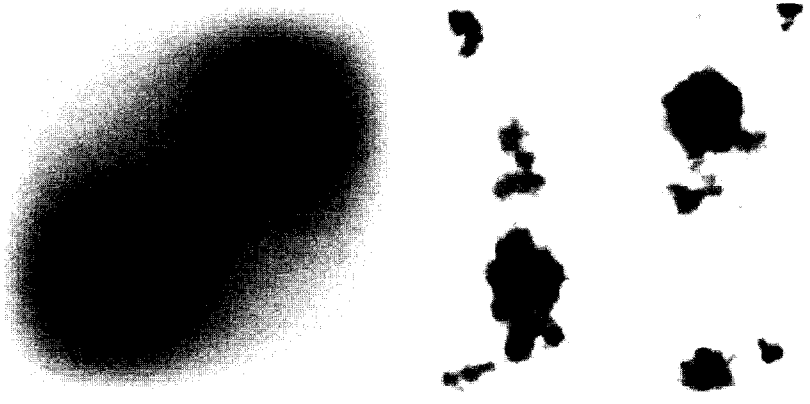


Fig. 13. A test of the low level shear detection. In this simulation, two dwarf mass concentrations [*left*] cause a small shear perturbation in the background galaxies. Weak lens inversion using 11,000 arclets recovers the masses [*right*] at high significance. Shear noise in this simulation is 0.003 and scales as $N_{arclets}^{-1/2}$.

9.1 Detection Limits

Cold dark matter models predict a filamentary structure of mass emerging by redshift 0.5, with peak densities of $10^5 \rho_c$. But most of this dark matter is distributed in less spectacular mass concentrations, with projected (two-dimensional) densities less than $\Sigma_c \sim 1$ g cm^{-2} and rms fluctuations around 0.02 Σ_c. What are the chances of detecting these low level mascons with gravitational lens tomographic searches?

Simulated low level shear in a CCD survey for dark matter is shown in Figure 13, where two dwarf cluster masses (ten times smaller than those discussed above) were placed at redshift 0.3 in a three-dimensional distribution of galaxies extending out to redshift 3. These mascons had 40 arcsec wide soft cores, 600 km sec^{-1} velocity dispersion, and an outer mass cutoff at 500 arcsec. This corresponds to a projected density of $0.06\Sigma_c$.

This simulates a small 2000 times 2000 pixel part of the field of the Big Throughput Camera, in moderate seeing (0.9 arcsec FWHM) and including all blue galaxies to 26th magnitude. This is representative of most of the data we already have, so we wanted to see how far we could push the mass detection. In this inversion we also wanted to push the angular resolution, so we used an inner radius (Gaussian filter) of two arcminutes. Random noise in the inversion is about 0.003 Σ_c and the peak densities in the reconstructed mass map are 0.014 Σ_c. In other words, due to depolarization due to seeing and ellipticity scatter and noise, our efficiency of mass reconstruction was $0.014/0.06 = 0.23$ and the resulting signal to noise ratio for recovery of the two mascons was 5. Dividing by the shear-to-mass efficiency, the rms mass noise is recalibrated to $0.013\Sigma_c$, smaller than the rms in CDM theories. No false mascons with the density of these two test masses were found in the reconstructed mass density image or in repeated realizations. The input diffuse mascons were thus detected at high significance. Larger smoothing filters in the inversion produce somewhat lower shear-to-mass efficiency but much lower false mascon rates and higher significance for overdensities covering large angles.

What will ultimately limit tomographic mass detection? There are two sources of noise and one source of systematic error. The photon shot noise from the night sky diffuse surface brightness may be overcome with longer integration times. In the same way, higher redshift (fainter) galaxies may be reliably photometered. This raises the efficiency, especially for higher redshift lenses. But the source galaxy random ellipticity orientation noise will remain about $e_{rms} = 0.3$, so lower level mass density detections will have to use lower angular resolution, capitalizing on the root N averaging of source ellipticity noise over larger resolution elements. There is also a more fundamental limit due to systematic shear error due to the finite number density of point-spread-function calibration stars. Rms shear variation of stellar images has been reduced to levels of 0.0007 averaged over a 45-arcminute field. Most of this is systematic and fixed, and is due to the optics. This may be calibrated to better than 0.001 per square arcminute, but there remain induced systematics from focus variations. This ultimately limits one arcminute resolution mass reconstructions to about 0.002 Σ_c, even in multi-color deep imaging to 27th magnitude. This is over ten times lower than the rms mass density fluctuations in blank fields predicted in any CDM model. Lower limits may be reached by varying the camera rotation angle.

10 Large Scale Dark Matter

By concentrating on the largest luminous structures known, clusters of galaxies, we are biased to radiation. We have in effect been looking under the lamp post for dark matter. Does light trace mass under all conditions? Larger scale applications of this dark matter mapping may eventually find clumped dark matter unrelated to galaxies or clusters of galaxies. Mosaics of CCDs make such a large scale search for coherent alignment in the distant faint galaxies particularly attractive, and dark matter on angular scales up to degrees can in principle be studied in this way. Dark matter may exist in places where there is no current star formation activity. Larger scale applications of this dark matter mapping may eventually find clumped dark matter unrelated to galaxies or clusters of galaxies; I will show a deep image of a good candidate.

If we wish to extract only information on the statistics of the foreground mass overdensities, rather than map them, there is a tensor statistic analogous to the scalar two-point correlation function.[29,30,23,31,32] All pairs of galaxies separated by some angle on the sky are summed separately in bins of mutual orientation. This integral tensor statistic called the orientation correlation function (OCF) is then built from these summed orientation data. The OCF has both parallel and orthogonal orientation components, each a function of angle on the sky. The OCF detects an excess over random for background galaxies to be oriented parallel or orthogonal to one another. In this way much larger areas of the sky may be covered and a smaller amplitude statistical mass fluctuation spectrum may be detected.

The observational challenge in low shear measurement is systematic errors due to variations across the detector of the point spread function, optics astigmatism, and variable field distortion due to atmospheric refraction. These systematics can be overcome by chopping and trailing techniques, and optics calibration in rich star fields. Preliminary measurements of the arclet orientation correlation function in random fields yield shear correlations which are below the SCDM prediction. The key in finding the rest of the dark matter, even in places where there are no clusters, is to measure shear over much larger areas. This also means that systematics will have to be controlled at the 0.1-0.5 percent shear level. With Gary Bernstein at U. Michigan, we have built a large area blue-sensitive CCD mosaic camera for the four-meter Blanco telescope at CTIO. With this instrument we have a chance of viewing mass in ordinary places, rather than in the inner parts of massive rare overdensities.

11 Large CCD Mosaic

The Big Throughput Camera is so named because it was the only operating large mosaic of back-illuminated CCDs, giving it much more throughput than other mosaics with comparable field of view, as well as good blue sensitivity.[33] [Note: the Sloan CCD mosaic just had its "first light"; this is a much bigger array of the same type of CCDs as the BTC.] The format is a 2×2 array of thinned, broadband antireflection-coated 2048×2048 SITe CCDs. The four-CCD mosaic in the BTC covers an 11.7×11.7 cm area. Mounted at the prime focus of the four-meter Blanco Telescope at CTIO, the 24 micron pixels subtend 0.43 arcseconds on the sky, providing a 34.8 arcmin field of view. Each CCD is separately adjustable in position to $\sim 5\,\mu m$ accuracy and cooled to 170 K.

The BTC's large field of view is ideal for searches for rare objects or events, statistical studies of large numbers of objects, and studies of anything which covers large angular scales. In its year of operation the BTC has been used by a group searching for faint quasars and two groups searching for high-redshift supernovae (20 discovered per night!), which can be used to constrain fundamental cosmological parameters. For the first time, large numbers of high-redshift supernovae are being discovered: 32 identified in first-year BTC images have been confirmed and announced by one team, and a similar number has been found by the other team. A search for Kuiper belt objects is being conducted.

Our studies of large-scale dark matter distribution using weak gravitational lensing exemplify the other two typical uses of the BTC, amassing large-number statistics and covering large angular scales. In weak lensing, the signature of intervening concentrations of dark matter is a small nonrandom component induced in the otherwise random orientations of background galaxies, hence large-number statistics are required. In addition, large angular fields are required to trace the full extent of the dark matter.

We require a large contiguous field of view. The gaps are filled in with "shift-and-stare" exposures, with large shifts of about 1000 pixels (7 arcmin). After about 18 exposures, a contiguous field of 47 arcminutes is covered. The unique problem of large focal-plane mosaics is field distortion, since they cover most of

the usable field of view of large telescopes. In the case of the BTC at the Blanco telescope, the prime focus corrector induces a radial distortion of about 60 pixels, or a few percent, at the corners of array. In addition, the corrector lenses over each of the CCDs introduce a small additional distortion, differential refraction over 35 arcmin is significant and constantly changing in direction, atmospheric pressure variations can cause changes in the focal plane scale from exposure to exposure, and the small field rotation caused by the slight polar misalignment of the telescope is evident over this wide a field. Without correction for these effects, multiple shifted exposures cannot be combined into a single deeper image.

These effects are corrected using a parametric model of the entire optical system and observations of astrometric standard fields from the United States Naval Observatory at a variety of airmasses with each filter used in the observing run. The observed star positions provide constraints on our model of the optical system in which parameters such as the amplitude of the radial distortion and its center relative to the BTC center are allowed to vary. The fit generally converges to a solution with an rms scatter of 0.3 pixels, or 0.13 arcseconds. We then apply the model with its best-fit parameters to our science observations, and for some applications such as surveys, tenth-arcsecond accuracy is more than sufficient. However, higher accuracy is required for our gravitational lensing studies. This is because gravitational lensing analyses are based on image shapes, and shapes on a combined frame can be significantly altered if the image positions differ by even a tenth of an arcsecond on the individual frames. Therefore, for each set of images to be combined, we derive an additional third-order polynomial distortion to each exposure by forcing the coordinates of high signal-to-noise objects to agree among exposures. The residuals from this fit are about 0.06 pixels, or 26 milliarcseconds.

In gravitational weak lens shear measurements one must be particularly careful that the point-spread function (PSF) is not systematically out of round. We have found that astigmatism in the telescope optics is the dominant cause of stellar profile ellipticity systematics. We convolve the image with a flux-conserving kernel designed to circularize stellar images. Astigmatism combined with phase diversity across the chips means that the ellipticity and position angle vary with position, so we construct a position-dependent kernel by fitting a fourth-order polynomial function of x and y to the observed moments of those objects which appear to be stellar. The convolution step adds a minimal amount to the stellar image size, on the order of a few percent, while reducing stellar ellipticities by a factor of ten or more, to less than 0.001. The flow diagram for this pipeline software is shown in Figure 14.

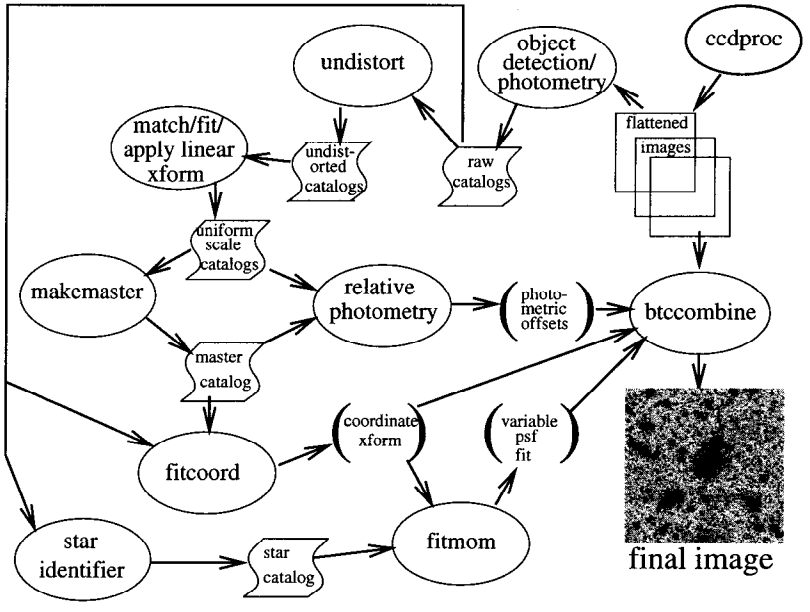


Fig. 14. The data flow in the Big Throughput Camera data reduction pipeline. The observed shapes of stars in the field are used to correct for low-level residual astigmatism.

11.1 Prospecting for Dark Matter

Different CDM cosmologies naturally give rise to differing mass clustering over time. For example, ΛCDM in a flat universe (consistent with current CMB observations) produces an older, larger universe today. These contrasting scenarios can now be tested.

Until recently there have been no tools which could map mass directly on 0.1 - 10 Mpc scales, without reference to luminosity, in an unbiased way. Gravitational lens inversion is now a mature field, and has yielded mass maps over a variety of mass and length scales, from 5 kpc to 2 Mpc. So far, lens search programs

have concentrated on obvious clusters of galaxies where one knows there is a deep gravitational potential well coincident with the luminous baryons. But this is "searching under the lamp post"; in so doing we build in a bias between mass and luminosity. Weak lensing analysis can now be applied to blank fields to explore mass generally, without resort to light. Furthermore, evidence is mounting from strong lensing observations that the input physics for the N-body work is incomplete. It is fair to say that the process of mass clustering is not yet well understood.

The recent primordial deuterium measurement together with the observed high cluster baryon density (the "baryon catastrophe"), indicate $\Omega_{\rm matter} = 0.4$; this is also predicted by large-scale peculiar velocities. This additional mass in dark structures, over what has been inferred from studies of clusters, will be clustered throughout a deep probe out to high redshift. The clustered mass lenses the high redshift galaxies and is visible in gravitational lens tomographic maps. Such a mass probe of random areas relies heavily on wide-field deep imaging in superb conditions and over a wide wavelength range. While deep lensing probes of many blank fields will be necessary in order to test this and measure the cosmic variance, a hint of what may be in store is shown in Figure 15. A large mass structure with peak mass density of $0.02\Sigma_c$ is found in a BTC field chosen for its absence of obvious clusters of galaxies. In some blank fields examined in this way with the BTC, we have found high redshift clusters by first detecting the mass overdensity via weak lens inversion. In a BTC survey of "normal clusters" chosen by their below average X-ray luminosity, multiple mass concentrations with occasional mass bridges connecting them are found.

12 Cosmic Complementarity

The observational situation in the $\Lambda - \Omega_M$ plane, mid-1998, is reviewed in Figure 16. Three complementary experiments are plotted. Recently, two groups (the Supernova Cosmology Project and the High-z Supernova Team) using Type Ia supernovae as standard candles (objects of known luminosity) and assuming that their flux measurements were not contaminated by sample selection, evolution, or dust systematics, both conclude that the expansion of the Universe is accelerating rather than decelerating (i.e., $q_0 < 0$). If correct, this implies that much of the energy in the Universe is in an unknown component. The simplest explanation is

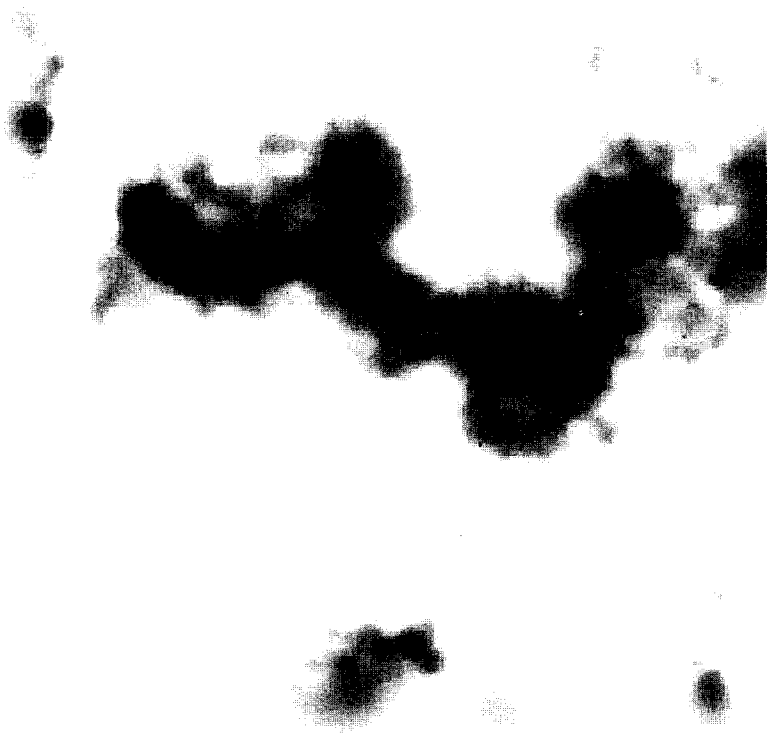


Fig. 15. The mass density reconstructed from 8700 blue arclets in moderately deep BTC integrations in a field devoid of known clusters. A large low-level (shear 0.02) mass is discovered. Note other mass concentrations in this clusterless "blank" field. Keck spectroscopy has just been obtained for 28 galaxies in the mass concentration shown in this field.

a cosmological constant with $\Omega_\Lambda \sim 2/3$. My estimate of the likelihood contours are shown.

The pattern of anisotropy in the CMB depends upon the total energy density in the Universe. The first acoustic peak in the multipole power spectrum is $l_{\rm peak} \simeq 200/\sqrt{\Omega_M + \Omega_\Lambda}$. These current likelihood contours are shown as dashes. The CMB anisotropy does not measure mass at $z = 1000$ directly; a model relating three mass-derived effects on the radiation must be fit to the CMB anisotropy data. Note that the likelihood contours for these first two experiments are roughly orthogonal; as the new generation of satellite CMB observations map out the fine scale structure of the microwave sky, the position our Universe occupies in this plane will be strongly constrained.

Finally, as direct lensing observations of dark matter explore more of the Universe, the vertical bar marked "known matter" will move to the right. The black-grey shading going up to higher Ω_m denotes the situation if future explorations for dark matter discover significant high M/L structures in the vast regions between clusters. It will be interesting, and a test of our cosmological model, if the results of all three tests (SN, CMB, and Lensing) are consistent. It will be even more interesting if they are not. Such a dark matter and dark energy survey is feasible, using current technology.

13 Dark Matter Telescope

One might ask why we need any new facilities in astronomy, given the fantastic advances made possible with the Hubble Space Telescope, the Keck Telescopes, and the imminent Gemini telescopes. The Hubble Space Telescope is a relatively small (2.4 meter) and expensive telescope offering the finest possible resolution, making it ideal for studies of stars and tenth-arcsecond features in galaxies. The Keck Telescopes offer unparalleled collecting area, focused on small-area imaging and spectroscopy of the faintest and most distant objects. The Gemini telescopes will combine superb light gathering power with IR capability. All of these telescopes, however, have relatively small fields of view and are operated in the traditional mode for astronomy: a given project is pursued for at most a few nights at a time, after which a different group of astronomers uses a different instrument to pursue a different scientific goal. This mode of operation is necessary for projects which serve large communities with varied scientific interests, but it requires compro-

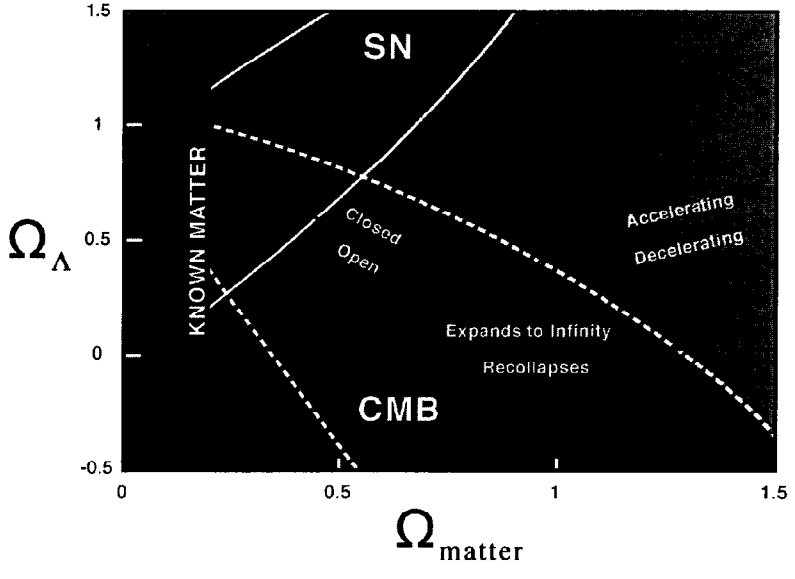


Fig. 16. The current situation in the cosmological constant-matter plane. Three key experiments are indicated: Supernova measurements of the acceleration (SN), cosmic microwave background anisotropy measurements (CMB), and direct measurements of mass (shading). All three will see vastly decreased errors in the next few years. The "known matter" is a lower limit based on extrapolating cluster masses to the field, assuming identical M/L ratios. As more mass is found, this vertical black bar will move to the right. Note that the recent determination of the cosmic abundance of deuterium suggests $\Omega_{matter} = 0.4$.

mises in the design of the telescope, and precludes the pursuit of projects which require very large amounts of data and telescope time.

Several recent projects present a different model for the pursuit of astronomical science goals, whereby a telescope is completely dedicated to one or two important projects for a period of years. An example is the Sloan Digital Sky Survey, for which a dedicated 2.5-meter telescope and instrumentation have been built. This project will yield a high-quality digital map of the entire northern sky, plus redshifts of several million galaxies and quasars.

A definitive measure of the density spectrum of the dark matter on supercluster scales is technically within reach. Deep imaging of millions of distant galaxies at several wavelengths (to get statistical redshifts from their colors) repeated in many different directions would have the precision to detect a fraction (less than ten percent) of the statistical weak lensing shear predicted by current Cold Dark Matter theory. An accurate measurement in 25-degree areas of the sky, to get the cosmic variance, would provide a direct measure of mass content and the mass fluctuation spectral index. With a dedicated six-meter "Dark Matter Telescope," an advanced mosaic of CCDs covering over a degree, and deep imaging in multiple wavelength bands, this project would take five to ten years to complete. The payoff in our understanding of dark matter and cosmology would be well worth the effort.

We propose, therefore, the construction of a six-meter class telescope with a single instrument optimized for the pursuit of a single goal: measurement of the distribution of dark matter in the Universe. The massive clusters, however, are like the whitecaps on the ocean. To really answer the central question in cosmology, "where is all the mass in the Universe?", we must study the "ocean" directly. The proposed Dark Matter Telescope will do this, showing where the dark matter is, how it clumps, and thus limiting the possible candidates for its constitution. If the fields are chosen appropriately, these same data may be used to map the inner part of the Kuiper belt of comets, and high-redshift supernovae, discovering thousands of these objects in a few years. Indeed, the supernova database will be an important constraint on cosmology and dark matter. So the dark matter mapping and supernova survey would use the same imaging data and yield related cosmological data on both the dark energy and mass budget of our universe.

The distant star-forming galaxies form half of the toolbox required for this unique cosmological probe. The rest of the tools are a wide-field CCD mosaic and a special wide-field low-aberration optical telescope with state-of-the-art light gathering capability and 350-900 nm wavelength coverage. To undertake this tomographic gravitational lens reconstruction of dark matter images at high redshift and large look-back times requires superb imaging of distant background galaxies. At 29th magnitude, there is a distant blue galaxy every arcsecond on the sky. Because the surface brightness of these distant galaxies falls as the 4th power of the redshift, these distant galaxies are so faint that existing telescopes with their relatively small fields of view would spend weeks of telescope time to produce only a narrow pencil beam probe. One requires superb angular resolution over a two to three degree field, coupled with the light gathering power of a large six to eight meter mirror. The Dark Matter Telescope would require superb blue-red optics and a state-of-the-art wide-field CCD array camera with a total photon throughput well in excess of even the Keck telescopes. Innovative software will be required to handle the flood of data. Unfortunately, present-day telescopes suffer especially from small fields of view, but also from poor image quality and over-commitment, allowing only tantalizing forays into dark-matter mapping. The wide spectral coverage, together with low aberration requirements, are a challenge for normal telescopes which have multi-element glass correctors. A design being seriously discussed is the Paul-Baker three-mirror telescope. This corrects for aberrations and is intrinsically achromatic and non-dispersive. A new-technology Dark Matter Telescope of six-meter aperture would cost approximately 30M 1998 dollars, and the camera would cost an additional 8-15M 1998 dollars. Operations over the period of this survey (ten years) would cost considerably less than ten percent of the capital cost per year because of no instrument changes and no visitor support. The data would be uniformly acquired by a small crew and released to the community.

The Dark Matter Telescope would be much more cost effective and optically efficient than other telescopes because it would be optimized for wide-field surveys. In terms of "throughput" (collecting area times imaging area on the sky), it would be 200 times more efficient than any of the wide-field mosaics now being built for existing or new telescopes, and tens of thousands times the throughput of the Hubble Space Telescope. A special wide-field low-aberration optical telescope in the six-meter class is required in order to take the next step in our understanding of the origin of structure, the dark energy, and mass in our Universe.

Acknowledgments

My principal collaborators in this research are Gary Bernstein, Ian Dell'Antonio, Phil Fischer, Raja Guhathakurta, Ayana Holloway, Greg Kochanski, and Dave Wittman.

References

[1] White, S. D. M., Frenk, C., & Davis, M., 1983, Astrophys. J. 274, L1.

[2] Blumenthal, G., Faber, S., Primack, J. & Rees, M., 1984, Nature 311, 517.

[3] Burles, S. & Tytler, D., 1998, Astrophys. J. 499, 699; astro-ph/9712265.

[4] Zwicky, F., 1933, Helv. Phys. Acta 6, 124.

[5] Zwicky, F., 1937, Phys. Rev. 51, 290.

[6] Dyer, C. C. & Roeder, R. C., 1973, Astrophys. J. 180, L31.

[7] Press, W. & Gunn, J. E., 1973, Astrophys. J. , 185, 397.

[8] Walsh, D., Carswell, R. F., and Weymann, R. J., 1979, Nature 279, 381.

[9] Peebles, P. J. E., 1993, *Principles of Physical Cosmology* (Princeton University Press).

[10] Carlberg, R. G., Lee, H. K. C., Ellington, E., Abraham, R., Gravel, P., Morris, S., & Pritchet, C. J., 1996, Astrophys. J. 462, 32.

[11] Garnavich, P. *et al.*, 1998, Astrophys. J. 493, 53.

[12] Narayan, R, & Bartelmann, M., 1996, astro-ph/9606001.

[13] Evrard, A. E., *et al.*, 1996, Astrophys. J. 469, 474.

[14] Tyson, J. A., Valdes, F., Jarvis, J. F., & Mills, A. P., 1984, Astrophys. J. Lett. 281, 59.

[15] Tyson, J. A., Valdes, F., & Wenk, R. A., 1990, Astrophys. J. Lett., 349, L1.

[16] Im, M., Casertano, S., Griffiths, R. E., Ratnatunga, K. U. & Tyson, J. A., 1995, Astrophys. J. 441, 494.

[17] Blandford, R. D. & Narayan, R., 1992, Ann. Rev. Astron. Astrophys, 311.

[18] Turner, E. L., Ostriker, J. P., & Gott, J. R., 1984, Astrophys. J. 284, 1.

[19] Tyson, J. A., Kochanski, G., & Dell'Antonio, I. P., 1998, Astrophys. J. Lett. 498, 107.

[20] Colley, W. N., Tyson, J. A., and Turner, E. L., 1996, Astrophys. J. Lett. 461, 83.

[21] Schneider, P., Ehlers, J., & Falco, E. E., 1993, *Gravitational Lenses*, (Springer, NY), 244.

[22] Syer, D. & White, S. D. M. 1998, Mon. Nat. Royal Astron. Soc. 293, 337.

[23] Miralda-Escudé, J., 1991, Astrophys. J. 370, 1.

[24] Kaiser, N. & Squires, G., 1993, Astrophys. J. Lett. 404, 441.

[25] Tyson, J. A. & Fischer, P., 1995, Astrophys. J. Lett. 446, 55.

[26] Clowe, D., Kaiser, N., Luppino, G., Henry, J. P., & Gioia, I. M., 1998, Astrophys. J. 497, 61.

[27] Squires, G. *et al.*, 1997, Astrophys. J. 482, 648.

[28] Fischer, P., & Tyson, J. A., 1997, AJ, 114,14.

[29] Gunn, J. E., 1967, Astrophys. J. 147, 61.

[30] Valdes, F., Tyson, J. A. & Jarvis, J. F., 1983, Astrophys. J. 271, 431.

[31] Blandford, R. D., Saust, A. B., Brainerd, T., & Villumsen, J. V., 1991, Mon. Not. Royal Astron. Soc. 251, 600.

[32] Seljak, U. & Zaldarriaga, M., 1998, astro-ph/9811123.

[33] Wittman, D., Tyson, J. A., Dell'Antonio, I. P., Bernstein, G., Smith, D., Lee, R. W., & Blouke, M. M., 1998, SPIE, 3355, 626.

PHYSICS OF GRAVITATIONAL WAVE DETECTION: RESONANT AND INTERFEROMETRIC DETECTORS

Peter R. Saulson*
Department of Physics
Syracuse University, Syracuse, NY 13244-1130

ABSTRACT

I review the physics of ground-based gravitational wave detectors, and summarize the history of their development and use. Special attention is paid to the historical roots of today's detectors.

*Supported by the National Science Foundation, under grant PHY-9602157.

© 1998 Peter R. Saulson.

1 The Nature of Gravitational Waves

1.1 What Is a Gravitational Wave?

There is a rich, if imperfect, analogy between gravitational waves and their more familiar electromagnetic counterparts. A nice heuristic derivation of the necessity for the existence of electromagnetic waves and of some of their properties can be found in, for example, Purcell's *Electricity and Magnetism*.[1] A key idea from special relativity is that no signal can travel more quickly than the speed of light. Consider the case of an electrically charged particle that is subject to a sudden acceleration of brief duration. It is clear that the electric field cannot be everywhere oriented in a precisely radial direction with respect to its present position; otherwise information could be transmitted instantaneously to arbitrary distances simply by modulating the position of the charge. What happens instead is that the field of a charge that has been suddenly accelerated has a kink in it that propagates away from the charge at the speed of light. Outside of the kink, the field is the one appropriate to the charge's position and velocity before the sudden acceleration, while inside it is the field appropriate to the charge subsequent to the acceleration. In each of those regions, the field looks like a radial field for the appropriate charge trajectory, but the kink itself has a transverse component, necessary to link the field lines outside and inside. The transverse field propagating at the speed of light away from the accelerating charge is the electromagnetic wave.

Heuristically, we can use similar reasoning to think about the case of gravitational fields. If the gravitational field of a mass were radial under all circumstances, then instantaneous communication would be possible. The transmitter could be a mass whose position could be modulated, and the receiver would be a device that determined the orientation of its gravitational field. In order that gravity not be able to violate relativity, it must be the case that the gravitational field of a recently accelerated body contains a transverse kink that propagates away from it at the speed of light. This transverse kink is the gravitational wave that carries the news that the body was recently accelerated.

It is helpful to consider the radiation from an extended system, a more realistic description than the point particle moved by prescription assumed in the foregoing paragraphs. In the electromagnetic case the lowest order moment of a charge distribution involved in radiation is the dipole moment, radiation by a time varying monopole moment (i.e., electric charge itself) being forbidden by the law of conservation of charge. The law of conservation of energy plays a similar role in forbidding monopole gravitational radiation, since that would require the total mass of an isolated system to vary. In addition, though, there can be no time-varying mass dipole moment of an isolated system; conservation of linear momentum forbids this. The difference between the gravitational case and the electromagnetic one is that the "gravitational charge-to-mass ratio" is the same for all bodies, by the Principle of Equivalence; the equal and opposite reaction that accompanies any action on a given body precisely cancels second time derivatives of the mass dipole moment of any isolated system. Similarly, the law of conservation of angular momentum forbids the gravitational equivalent of the magnetic dipole moment from varying in time. The lowest order moment of an isolated mass distribution that can accelerate is its quadrupole moment

$$I_{\mu\nu} \equiv \int dV \left(x_\mu x_\nu - \frac{1}{3}\delta_{\mu\nu} r^2 \right) \rho(\mathbf{r}) .$$

The second time derivative of the mass quadrupole moment plays the same role in gravitational radiation as does the first time derivative of the charge dipole moment in electromagnetic radiation, that of leading term in typical radiation problems.

But what is the analog of the electric field in gravitational wave problems, that is to say, the wave field amplitude itself? An answer comes from study of the Einstein field equations in the weak field limit. Specifically, one considers the case where the space-time metric can be approximated as $g^{\mu\nu} = \eta^{\mu\nu} + h^{\mu\nu}$, where $h^{\mu\nu}$ is a small perturbation of the Minkowski metric $\eta^{\mu\nu}$. Then, if one chooses a particular coordinate condition (the so-called "transverse traceless gauge"), the Einstein equations become a wave equation for the perturbation $h^{\mu\nu}$. Furthermore, the perturbation takes on a particular form: for a solution to the wave equation corresponding to a wave travelling along the z axis, h must be a linear combination of the two basis tensors

$$h_+ = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

and

$$h_\times = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} .$$

Notice that for this wave along the z axis, the effects of the wave are only in the x and y components. In other words, the wave is transverse.

1.2 The Effect of Gravitational Waves on Test Bodies

The statement that the gravitational wave amplitude is the metric perturbation tensor h is probably hard to visualize without considering some examples. Imagine a plane in space in which a square grid has been marked out by a set of infinitesimal test masses (so that their mutual gravitational interaction can be considered negligible compared to their response to the gravitational wave). This is a prescription for embodying a section of the transverse traceless coordinate system mentioned earlier, marking out coordinates by masses that are freely-falling (i.e., that feel no non-gravitational forces).

Now imagine that a gravitational wave is incident on the set of masses, along a direction normal to the plane. Take this direction to be the z axis, and the masses to be arranged along the x and y axes. Then, if the wave has the polarization called h_+, it will cause equal and opposite shifts in the formerly equal x and y separations between neighboring masses in the grid. That is, for one polarity of the wave, the separations of the masses along the x direction will decrease, while simultaneously the separations along the y direction will increase. When the wave oscillates to opposite polarity, the opposite effect occurs.

If, instead, a wave of polarization $h_\times$ is incident on the set of test masses, then there will be (to first order in the wave amplitude) no changes in the distances between any mass and its nearest neighbors along the x and y directions. However, $h_\times$ is responsible for a similar pattern of distance changes between a mass and its next-nearest neighbors along the diagonals of the grid.

There are several other aspects of the gravitational wave's deformation of the test system that are worth pondering. Firstly, the effect on any pair of neighbors in a given direction is identical to that on any other pair. The same *fractional* change occurs between other pairs oriented along the same direction, no matter how large their separation. This means that a larger *absolute* change in separation occurs, the larger is the original separation between two test masses. This property, which we can call "tidal" because of its similarity to the effect of ordinary gravitational tides, is exploited in the design of interferometric detectors of gravitational waves.

Another aspect of this pattern that is worthy of note is that the distortion is uniform throughout the coordinate grid. This means that any one of the test masses can be considered to be at rest, with the others moving in relation to it. In other words, a gravitational wave does not cause any absolute acceleration, only relative accelerations between masses. This, too, is fully consistent with other aspects of gravitation

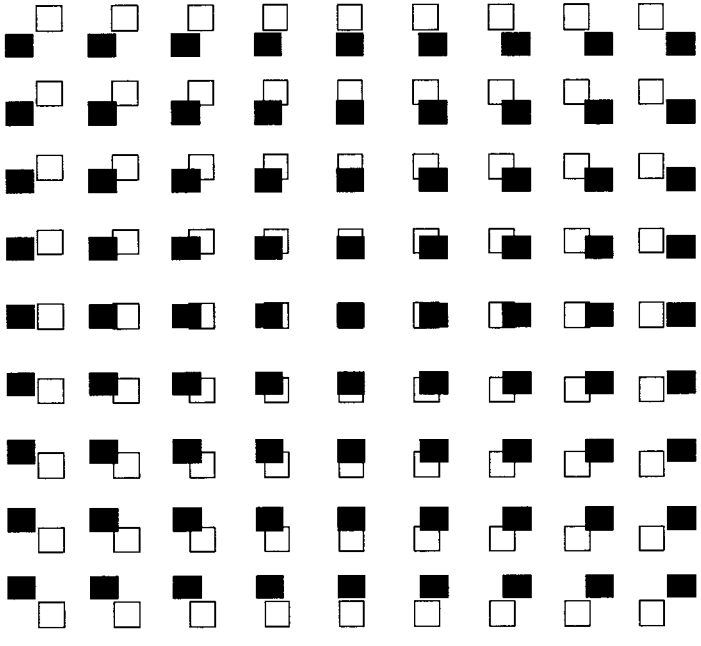


Fig. 1. An array of free test masses. The open squares show the positions of the masses before the arrival of the gravitational wave. The filled squares show the positions of the masses during the passage of a gravitational wave of the plus polarization.

as described by the general theory of relativity: a single freely-falling mass cannot tell whether it is subject to a gravitational force. Only a measurement of relative displacements between freely-falling test masses (the so-called "geodesic deviation") can reveal the presence of a gravitational field.

1.3 A *Gedanken* Experiment to Detect a Gravitational Wave

In the discussion in the preceding section, we took it for granted that the perturbations h_+ and $h_\times$ to the flat-space metric were, in some sense, real. But it is only by considering whether such effects are measurable that one can be convinced that a phenomenon like a gravitational wave is meaningful, rather than a mathematical artifact that could be transformed away by a suitable choice of coordinates.

To demonstrate the physical reality of gravitational waves, consider the example system of the previous section. We will concentrate our attention on three of the test masses, one chosen arbitrarily from the plane, along with its nearest neighbors in the $+x$ and $+y$ directions. Imagine that we have equipped the mass at the vertex of this "L" with a lamp that can be made to emit very brief pulses of light. Imagine also that the two masses at the ends of the "L" are fitted with mirrors aimed so that they will return the flashes of light back toward the vertex mass.

First, we will sketch how the apparatus can be properly set up, in the absence of a gravitational wave. Let the lamp emit a train of pulses, and observe when the reflected flashes of light are returned to the vertex mass by the mirrors on the two end masses. Adjust the distances from the vertex mass to the two end masses until the two reflected flashes arrive simultaneously.

Once the apparatus is nulled, let the lamp keep flashing, and wait for a burst of gravitational waves to arrive. When a wave of $\hat{h}_+$ polarization passes through the apparatus along the z axis, it will disturb the balance between the lengths of the two arms of the "L." Imagine that the gravitational wave has a waveform given by

$$h^{\mu\nu} = h(t)\hat{h}_+.$$

To see how this space-time perturbation changes the arrival times of the two returned flashes, let us carefully calculate the time it takes for light to travel along each of the two arms.

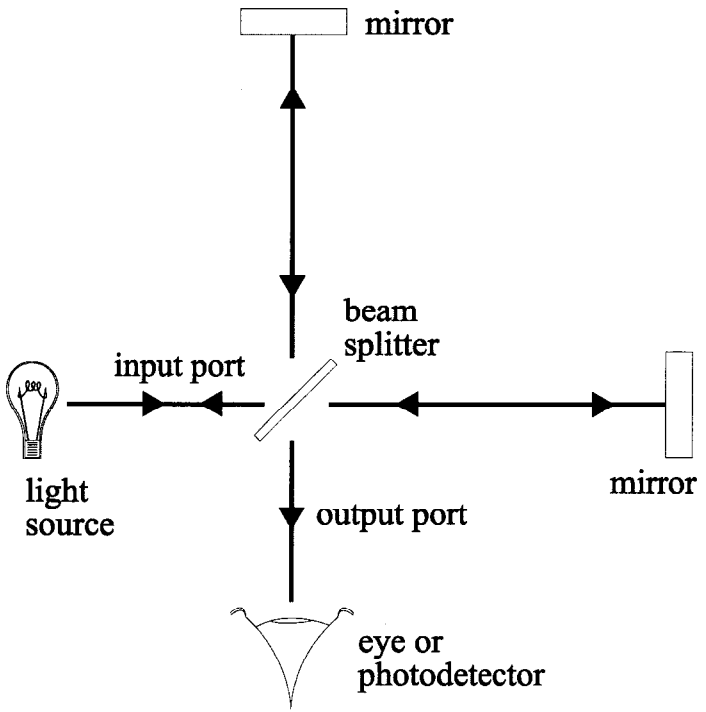


Fig. 2. A schematic diagram of an apparatus that can detect gravitational waves. It has the form of a Michelson interferometer.

First, consider light in the arm along the x axis. The interval between two neighboring space-time events linked by the light beam is given by

$$\begin{aligned} ds^2 = 0 &= g_{\mu\nu}dx^\mu dx^\nu \\ &= (\eta_{\mu\nu} + h_{\mu\nu})\, dx^\mu dx^\nu \\ &= -c^2dt^2 + (1 + h_{11}(2\pi ft - kz))\, dx^2. \end{aligned} \tag{1}$$

This says that the effect of the gravitational wave is to modulate the square of the distance between two neighboring points of fixed coordinate separation dx (as marked, in this gauge, by freely-falling test particles) by a fractional amount h_{11}.

We can evaluate the light travel time from the beam splitter to the end of the x arm by integrating the square root of Eq. (1)

$$\int_0^{\tau_{out}} dt = \frac{1}{c}\int_0^L \sqrt{1 + h_{11}}dx \approx \frac{1}{c}\int_0^L \left(1 + \frac{1}{2}h_{11}\,(2\pi ft - kz)\right) dx, \tag{2}$$

where, because we will only encounter situations in which $h \ll 1$, we've used the binomial expansion of the square root, and dropped the utterly negligible terms with more than one power of h. We can write a similar equation for the return trip

$$\int_{\tau_{out}}^{\tau_{rt}} dt = -\frac{1}{c}\int_L^0 \left(1 + \frac{1}{2}h_{11}(2\pi ft - kz)\right) dx. \tag{3}$$

The total round trip time is thus

$$\tau_{rt} = \frac{2L}{c} + \frac{1}{2c}\int_0^L h_{11}(2\pi ft - kz)dx - \frac{1}{2c}\int_L^0 h_{11}(2\pi ft - kz)dx. \tag{4}$$

The integrals are to be evaluated by expressing the arguments as a function just of the position of a particular wavefront (the one that left the beam-splitter at $t = 0$) as it propagates through the apparatus. That is, we should make the substitution $t = x/c$ for the outbound leg, and $t = (2L - x)/c$ for the return leg. Corrections to these relations due to the effect of the gravitational wave itself are negligible.

A similar expression can be written for the light that travels through the y arm. The only differences are that it will depend on h_{22} instead of h_{11} and will involve a different substitution for t.

If $2\pi f_{gw}\tau_{rt} \ll 1$, then we can treat the metric perturbation as approximately constant during the time any given flash is present in the apparatus. There will be equal and opposite perturbations to the light travel time in the two arms. The total travel time difference will therefore be

$$\Delta\tau(t) = h(t)\frac{2L}{c} = h(t)\tau_{rt0}, \tag{5}$$

where we have defined $\tau_{rt0} \equiv 2L/c$.

If we imagine replacing the flashing lamp with a laser that emits a coherent beam of light, we can express the travel time difference as a phase shift by comparing the travel time difference to the (reduced) period of oscillation of the light, or

$$\Delta\phi(t) = h(t)\tau_{rt0}\frac{2\pi c}{\lambda}. \tag{6}$$

Another way to say this is that the phase shift between the light that traveled in the two arms is equal to a fraction h of the total phase a light beam accumulates as it traverses the apparatus. This immediately says that the longer the optical path in the apparatus, the larger will be the phase shift due to the gravitational wave.

Thus, this *gedanken* experiment has demonstrated that gravitational waves do indeed have physical reality, since they can (at least in principle) be measured. Furthermore, it suggests a straightforward interpretation of the dimensionless metric perturbation h. The gravitational wave amplitude gives the fractional change in the difference in light travel times along two perpendicular paths whose endpoints are marked by freely-falling test masses.

1.4 Another Way to Picture the Effect of a Gravitational Wave on Test Bodies

In standard laboratory practice, it is not customary to define coordinates by the world-lines of freely-falling test masses. Instead, rigid rulers usually are used to do the job. The forces that make a rigid ruler rigid are something of a foreign concept in relativity, appearing ugly and awkward after the gravitational force has been made to disappear by expressing it as the curvature of space-time. On the other hand, non-gravitational forces are not only a fact of nature, but part of the familiar world of the laboratory. For many purposes, it is convenient to retreat from a purely relativistic picture and instead use a Newtonian picture in which gravity is treated as force on the same level as other forces.

What we are seeking is not a different theory of gravitational waves, but a translation of the theory discussed in the previous section into more familiar language. So let us reconsider the same *gedanken* experiment as before, but imagine that we have augmented the equipment with a rigid ruler along each axis. We saw that when a gravitational wave passed through our set of test masses, the amount of time it took for light to travel from the vertex mass to the end mass and back was made to vary. How can we

describe how this came about in the standard language of the laboratory? If we imagine (notwithstanding their fixed coordinates in the transverse traceless gauge) that the test masses have moved in response to the gravitational wave, we can have a consistent picture of the effect. What is necessary is that the gravitational wave give a tidal force across the pair of masses that will cause them to move apart by the amount necessary to account for the change in light travel time through the system. It is as if the far masses felt forces whose magnitude were given by

$$F_{gw} = \frac{1}{2} m L \frac{\partial^2 h_{11}}{\partial t^2}, \tag{7}$$

where m is the mass of each of the two test bodies, and L is the separation between their centers of mass.

There are several features of this expression that are worthy of note. The force is proportional to the mass of the test bodies, as required by the Principle of Equivalence. The force is also proportional to the separation between the two test masses, making it akin to conventional gravitational tidal forces. The dependence of the gravitational wave force on the second time derivative of h is reminiscent of Newton's Second Law $F = m\ddot{x}$. A natural interpretation follows: in a conventional laboratory coordinate system, free masses actually change their separation by an amount $\Delta L = hL/2$.

Note how in two different coordinate systems the same phenomenon (and particularly the same measurement) is described in completely different language. In transverse traceless coordinates, the free test masses still just fall freely, each marking out its own coordinate (by definition) under any gravitational influence, but the light travel time between them changes as the metric of space-time varies. In standard laboratory coordinates, light travel time changes because the test masses move. Neither of these pictures is more "correct" than the other. The laboratory-coordinate picture is markedly more convenient for seeing how to combine the effect of a gravitational wave with the effects of noise forces of various kinds. The transverse traceless coordinates offer the most clarity when one wants to consider extreme cases, such as test masses separated by distances comparable to or longer than the gravitational wave's wavelength.

2 Generating Gravitational Waves

As mentioned above, the second time derivative of the mass quadrupole moment I plays the same role in gravitational wave emission as does the first derivative of the charge dipole moment in electromagnetic radiation problems, that of strongest source term in most situations. More specifically, the expression for gravitational wave generation is

$$h_{\mu\nu} = \frac{2G}{Rc^4} \ddot{I}_{\mu\nu}, \tag{8}$$

usually referred to as the "quadrupole formula."[2]

Something about this expression should immediately give one pause—the prefactor of $2G/c^4$. In SI units, this has the value 1.6×10^{-44} $\mathrm{sec^2 kg^{-1} m^{-1}}$. It will take tremendously large values of $\ddot{I}/R$ in order for even modest values of h to be generated. *A priori*, one can think of two strategies that might work: make R small, or make $\ddot{I}$ large.

2.1 Laboratory Generators of Gravitational Waves

To construct a source of gravitational waves in the laboratory would allow one to have the benefit of placing it as close as possible to one's detector, thus exemplifying the first strategy in the previous paragraph. Of course, it would have other benefits as well. Control of the waveform, polarization, and other features would enable the detector to be carefully optimized to match the signal. At an even deeper level, confidence in the detection of gravitational waves could be assured by the requirement that they must be seen when, and only when, they were being emitted.

What one would really like to do is to replicate for gravitational waves what Hertz was able to accomplish for electromagnetic ones. His experiments of 1886–'91 not only conclusively demonstrated the existence of electromagnetic waves, they validated Maxwell's theory of electromagnetic radiation by exploring the rich phenomenology of polarization, reflection, and interference. They also began the process of harnessing the phenomenon for practical use. Marconi's work started by his following closely in Hertz's footsteps, and real long-distance communication via radio was not long in coming.

Unfortunately, no practicable way has been conceived to replicate Hertz's success in the gravitational domain. Assume we could construct a dumbbell consisting of two masses of one ton each, at either end of a rod two meters long. Spin this quadrupole about an axis orthogonal to the connecting rod passing through its midpoint, at an angular frequency $f_{rot} = 1$ kHz. Neglecting for simplicity the contribution of the connecting rod, we have a system very similar to a binary star system. The amplitude of the gravitational waves generated by this device will be

$$h_{lab} = 2.6 \times 10^{-33}\mathrm{m} \times \frac{1}{R}. \tag{9}$$

Before we rush to plug in a distance R of a few meters, as Hertz was able to do for his experiment, we need to remember that wave phenomena are only distinguishable from near-field effects in the "wave zone," that is, at distances from the source comparable to or larger than one wavelength. With $\omega_{rot} = 2\pi \times 1$ kHz, we have $\lambda = 300$ km! The receiver for our Hertzian experiment must be at least that far away from the transmitter. Hertz's electromagnetic experiments involved waves of six meters down to 60 cm in length, so the distance across the lab was fine for him.

At a distance of one wavelength, our laboratory generator gives gravitational waves of amplitude

$$h_{lab} = 9 \times 10^{-39}. \tag{10}$$

This is pretty small.

Even creating such a strong source as this may not be practicable. Consider the stress in the connecting rod of the dumbbell. It must supply the centripetal force necessary for the masses to move in a circle. If the rod were made of good steel, it would need a cross-sectional area substantially greater than that of a one ton sphere in order not to fail under the stresses in a device with the parameters we have assumed. So we'd have to reduce the rotation frequency to keep the generator from flying apart, with a consequent reduction in the transmitted wave amplitude.

2.2 Astrophysical Sources of Gravitational Waves

Even if a gravitational version of the Hertz experiment is not feasible, all is not lost for the detection of gravitational waves. The best reason for optimism that detectable levels of gravitational radiation exist comes from the presence in the universe of objects with truly remarkable values of $\ddot{I}$. These systems are so extreme that even though their distances from our detectors are quite large, they still generate gravitational waves with amplitudes that exceed by almost 20 orders of magnitude the signal strengths from laboratory generators of the type described above.

It would be beyond the scope of this review to describe in detail all of the many astronomical objects that might be important sources of gravitational waves. Readers are urged to consult the article by Finn in these proceedings for further information on the variety of possible sources. But for the sake of a self-contained treatment, we show here how to estimate the magnitude of the strongest gravitational waves arriving at the Earth.

For the case of a binary star, there is an elegant way (due to Kafka[3]) of writing the amplitude of the quasi-sinusoidal gravitational wave strain. We can massage the quadrupole formula into a manifestly dimensionless form by recognizing that the mass dependence can be rewritten as a proportionality to the product of the Schwarzschild radii $R_s = 2GM/c^2$ of the stars. The frequency dependence and all of the stray factors remaining collect nicely as the separation r of the two stars. The gravitational wave amplitude is

$$h_{ns} = R_{s1}R_{s2}/rR. \tag{11}$$

If the binary consists of two neutron stars, then the Schwarzschild radii are both about 4 km. Astronomers estimate that within a sphere of radius 200 Mpc, roughly one of these sytems will coalesce each year. When the stars have a separation of ten diameters (or around 200 km), then the signal we would receive from that distance will have an amplitude of almost 10^{-23}. The stars can probably approach closer still before the system is destroyed.

A glance at this expression shows why a neutron star binary is a good choice as a strong source of gravitational waves. The substantial masses of the two stars make the numerator large. The fact that they are compact objects means that their separation r can be quite small. We could always wish that the distance R to the nearest example of such a system were smaller, but even so our estimated signal strength, while small in absolute terms, is certain dramatically larger than we were able to produce in our model laboratory generator.

Perhaps the only sort of astronomical system we can imagine that might generate stronger gravitational waves would be a binary system consisting of two black holes. Although it may be hazardous to treat such dramatically relativistic objects with the quasi-Newtonian physics used to derive Eq. (11), it will probably still give a good order of magnitude estimate. The possible advantages of black holes as sources of gravitational waves are twofold. Firstly, it is possible that the masses of black holes may be substantially in excess of the 1.4 $M_\odot$ typical of neutron stars. Secondly, black holes can approach to a separation r as close as their Schwarzchild radius R_s without disruption; instead the two will coalesce into a single larger black hole. Thus we guess that the gravitational signal from a black hole coalescence could be as large as

$$h_{bh} \sim R_s/R. \tag{12}$$

For a pair of 10 $M_\odot$ black holes at 200 Mpc, this expression would indicate a signal of $h \sim 5 \times 10^{-21}$.

Of course, we do not have nearly such secure knowledge of the existence of such binary systems as we do for the neutron star case. There is strong evidence for the existence of 10 $M_\odot$-class individual black holes in binary systems with main sequence stars.[4] The abundance of black hole binaries in this mass range is unknown, so the validity of our choice of 200 Mpc as a fiducial distance is uncertain at best.

One weakness of the elegant expressions Eqs. 11 and 12 is that they do not explicitly refer to the frequency of the gravitational wave. As we will see later, the various kinds of detectors of gravitational waves will each function best only in a certain band of frequencies. For reference, we anticipate the results of that discussion with the following summary: resonant-mass detectors work best in the vicinity of 1 kHz, interferometric detectors from around 10 Hz to a few kHz, and space interferometers between around 10^{-4} and 10^{-1} Hz. The strongest black hole signal will come at the frequency f_{qnm} of the lowest quasi-normal mode of the resulting combined black hole,

$$f_{qnm} \approx 0.7 \frac{c^3}{2GM}.$$

This makes resonant-mass detectors best suited for looking for black holes with masses of 10 $M_\odot$. Terrestrial interferometers will look for 3 $M_\odot$ to 1000 $M_\odot$ objects, while interferometers in space will search for black holes in the range near 10^6 $M_\odot$.

2.3 Summary

The contrast between terrestrial and astronomical sources of gravitational waves is striking. In spite of the fact that astronomical generators will only be found at distances far beyond the optimum, their other physical parameters are such that they will provide the strongest available gravitational wave signals. The contrast is so strong that this advantage alone must outweigh the ability to control the character and timing of the signal that would be inherent in a laboratory generator. Instead, even the most basic experiments on the nature of gravitational waves will necessarily involve astronomical observations.

3 Weber and the Birth of Gravitational Wave Detection

3.1 Weber's Original Vision

It is reasonable to ask the question whether even the astronomical signals are large enough to be detected by any conceivable device. On the face of it, the odds are daunting. A strain of 10^{-20}, for example, only generates relative motion of 10^{-20} meters between two test masses separated by one meter. Compare this with other characteristic length scales, and the challenge is clear: wavelength of visible light of 10^{-6} meter, atomic diameter of 10^{-10} meter, nuclear diameter of 10^{-15} meter. Nevertheless, it appears that gravitational wave signals from astronomical sources will probably soon be detected. It is the purpose of the rest of this review to show how this is possible.

There are at least two questions that might be raised by consideration of the numbers listed in the previous paragraph. The first is whether measurements of a macroscopic body can be capable of resolving motions substantially smaller than nuclear diameters. The second question that one might ask is whether success might be easier if the scale of the apparatus were made substantially larger than a meter, in order to take advantage of the fact that test masses separated by a larger distance will move by a proportionally larger amount in response to a gravitational wave of a given strength. Pondering the answers to these questions will lead to understanding of the most promising techniques for gravitational wave detection.

The postwar atmosphere of optimism about astronomical progress must have swept up Joseph Weber in the late 1950s. Weber became convinced that the time was right to try to extend the astronomical revolution beyond the electromagnetic spectrum. At that time, it was not obvious that strain sensitivities of 10^{-22} should be the goal. It was equally plausible that objects such as we discussed above might possibly be abundant enough that their typical distance might be the few kpc associated with galactic dimensions instead of 200 Mpc. So strains of 10^{-17} or perhaps even stronger might have been the proper goal to aim for. (Weber knew of a very optimistic estimate of wave strength by Wheeler,[5] which would allow an energy density of order of the cosmological closure density.) Weber's thinking showed the way to achieve such strain sensitivities; indeed, devices following directly in a line of development from his first instrument have in the past few years approached rms sensitivities of 10^{-19}, with the prospect of extension to a new generation of detectors sensitive to waves with amplitudes of order 10^{-20} or better.

Weber's early thinking is described in his *Physical Review* article of 1960,[6] and placed in the larger context of his thinking about general relativity in a small monograph published in 1961.[7] He describes a conceptual detector, in reciprocal relationship to a gravitational wave emitter, as a simple "mass quadrupole," sketched as two masses connected by a spring. Weber extends the general relativistic equation of geodesic deviation to include the non-gravitational forces applied by the elastic restoring force and the mechanical dissipation in the spring. The equation of motion of the system

then becomes that of a simple harmonic oscillator, with the driving term given by the effective force from the gravitational wave [our Eq. (7)].

Weber next shows how an extended elastic body behaves in such a way that each of its normal modes of vibration can be studied independently. (The gravest mode of a cylinder has a large quadrupole moment, and is the one that is usually used for detection.) He focuses attention on the use of a piezoelectric crystal as the detecting body, partly because he hopes that the electric field will make it a detector with effective size larger than half an acoustic wavelength, but also in large measure because the electric fields generated by the gravitational-wave-induced stress will give an integrated voltage between its ends that may be "large enough to be observed with a low-noise radio receiver." Weber calculates the amount of mechanical power that a sinusoidal gravitational wave can dissipate in the resonant detector as a function of frequency, then invokes the standard electrical network theorems to show what fraction of this power can be transferred to the input impedance of an amplifier.

A simple discussion of sensitivity follows. Weber first remarks that "in microwave spectroscopy it has been found that all spurious effects other than random fluctuations can be recognized." Then Weber states that the excitation of the detector must exceed the noise power associated with its thermal excitation.

Finally, Weber discusses possible practical experimental arrangements. In most of the discussion the devices are supposed to be made of large blocks of piezoelectric material. But in a footnote Weber states that the experimental work he is carrying out with David Zipoy and Robert L. Forward will probably make use of a large block of metal instead. (This is justified on the grounds that a half-wavelength at the 1 kHz frequency being contemplated is already large; thus the piezoelectric length-enhancement effect may not be necessary, and in any case such a large block of piezoelectric material "may not be obtainable as a single crystal".)

Two experimental strategies are foreseen: use of a single detector with examination of its output for a diurnal cycle associated with the scanning of its sensitivity pattern across the sky, and the cross-correlation of a pair of detectors so that external influences (presumably gravitational waves) can be distinguished from "internal fluctuations." He notes the necessity of preventing the excitation of the detector by "earth vibrations," and discusses an "ingenious" idea of Zipoy's for what is now called active vibration isolation.

Weber's very concise discussion is remarkable for the prescience with which it foreshadowed not only his own work, but that of so many others. It also marks a watershed in the history of general relativity. In a single blow, Weber wrested consideration of gravitational waves from theorists concerned about issues such as exact solutions, and appropriated the subject instead for experimentalists trained in issues of radio engineering. The boldness and brilliance of this move are remarkable.

3.2 The Logic of Weber's Idea

Weber sweeps quickly over a variety of issues that are worthy of more leisurely consideration. We'll give an overview of the important issues in this section, then devote the rest of this review to discussing their implications.

The detector Weber outlined can be divided into several subsystems: a set of test masses that respond to the gravitational wave, a transduction system that converts this mechanical response to a convenient electrical signal, a low-noise pre-amplifier, and a post-amplification averaging and recording mechanism. Notwithstanding the cleverness of Weber's original version, many variations on his basic scheme are possible, and indeed are responsible for much of the progress since he first announced the results of gravitational wave observations in 1969.[8]

Let's see how to analyze the original Weber design into these canonical subsystems. Weber explicitly pointed out how one could construct an analog of a pair of lumped test masses by monitoring an internal mode of vibration of an extended block of elastic material. In the version where this block is made of piezoelectric material, the same material serves both as test masses and as transducer from mechanical to electrical signal form. In the version in Weber's footnote (the one he actually built) a large aluminum cylinder serves as the set of test masses; piezoelectric strain gauges glued about the girth of the cylinder perform the transduction. The pre-amplifier is Weber's low-noise radio receiver. No averaging filter is shown in Weber's diagrams, but is implicit in his discussion.

Perhaps the most interesting choice that Weber made was to connect his test masses in a resonant system. It appears that Weber, at least in 1961, thought this was a necessity. In a footnote, he cites previous work by Pirani[9] in which the latter considered "measurement of the Riemann tensor by comparing accelerations of free test particles," but Weber continues, "The results of this chapter indicate that interacting particles must be used, in practice." In fact, it is not required either in principle or in practice, but it is interesting to consider why Weber may have thought so then, and what advantages still accrue to the use of resonant masses.

Weber couches a good deal of his discussion in terms of steady sinusoidal signals, still a common practice in much of engineering and even more so around 1960. If a gravitational wave did have this form, then masses connected by a spring give a resonant amplification of the response to a signal at the resonant frequency. The amount of this amplification is given by the resonator's quality factor Q, and can be substantial; Weber quotes an estimated $Q \sim 10^6$, still not a bad ballpark number.

On the other hand, there is essentially no resonant amplification if one has a sinusoidal signal whose frequency does not closely match the resonant frequency of the detector, or if the signal has a broad-band frequency content, as it would if it were a brief burst. Resonant amplification only comes about when the input force drives the resonant system with the proper phase for a substantial number of cycles; this can only occur when there is a good match between the signal frequency and the mechanical resonant frequency.

But even though the search for gravitational waves has come to focus mostly on burst-like signals, the resonant-mass configuration can still give a powerful advantage, albeit one not discussed by Weber in his 1961 book. A weak signal must compete for visibility against the noise in the pre-amplifier stage. This is why Weber made a point of calling for the lowest noise levels possible in this component. The noise in such amplifiers is generally of a broad-band character, best characterized by its power spectral density $S_v(f)$ which is typically roughly constant (or "white") over a wide range of frequencies. Usually there is an additional $1/f$ component that dominates at low frequencies.

The extent to which this noise competes with a signal depends in an essential way on the duration of the signal. We use the term "burst" to refer to a signal of limited duration in time; call its length τ_s. A fundamental theorem of signal detection states that the optimum contrast between a given signal and white noise can be attained when the time series containing the noise plus any possible signals is convolved with a template of the same form as the signal. This is called the *matched filter* when it is implemented in real time by an analog device. The heuristic idea behind such an optimum is that the matched filter rejects all components of the noise that do not "look like" the signal for which one is searching. Still, some noise passes through the matched filter. How much? If the Fourier transform of the signal waveform $v(t)$ is given by $V(f)$, then it passes noise power of

$$N^2 = \int_{-\infty}^{\infty} |V(f)|^2 S_v(f) df.$$

Another very general theorem of Fourier analysis takes the form of a classical "uncertainty relation." It states that there is an inverse relationship between the duration of a signal in the time domain and its width in the frequency domain:

$$\Delta f \Delta \tau \approx 1.$$

See what this implies for the question at hand. If we are looking for a brief signal, then its matched filter passes noise of a wide bandwidth. Thus, brief signals compete much less well against broad-band pre-amplifier noise than does a long-duration signal of the same amplitude.

Here is where a resonant detector of gravitational waves can make a difference even when one is looking for a broad-band signal. If the gravitational wave signal $h(t)$ contains substantial power in the vicinity of the detector's resonant frequency, then it will excite the motion of the detector's mode at an amplitude $\Delta L \sim hL$, not much different from if the test masses were free. But the subsequent behavior of the resonant detector is quite different from if the detector were made of free masses. The motions of free test masses only persist for the duration τ_s of the gravitational wave signal. But the resonant detector "rings" for a time of order of the damping time of the resonance,

$$\tau_d = \frac{Q}{\pi f_0} \gg \tau_s,$$

where f_0 is the resonant frequency of the detector.

It is the motion of this resonant system, converted to electrical form by the transducer, that is presented to the input terminals of the pre-amplifier. So it is an electrical signal of long duration τ_d that competes with the amplifier noise. As a long duration deterministic signal, its matched filter has a much narrower width in frequency than had the original signal $h(t)$, and so passes a much smaller proportion of the amplifier noise. Thus, the resonant response of the test mass system allows a weak signal to compete more effectively against amplifier noise than would be the case with free masses.

3.3 The Cost of Resonant Detection

This advantage of resonant-mass detectors is substantial; it is responsible for the continued vitality of the Weber style of detector over 30 years after it was first proposed. Still, it comes with a price that is not negligible. Implementing the matched filter described above, which is essential to attaining the advantage of the resonance, is tantamount to averaging the output of the amplifier for times of order τ_d. In the jargon of the field,

such a system has a low *post-detection bandwidth* (usually shortened simply to "bandwidth.") The averaging washes out any details of the waveform $h(t)$ on time scales short compared to τ_d. What one gains in signal-to-noise ratio, one gives up in temporal resolution. Whether this is a price one ought to be willing to pay or not depends on the stakes: if it is absolutely necessary even to detect the signal, averaging with a matched filter is certainly worthwhile. If the signal could be detected anyway, averaging simply throws away information, and should be avoided. In the high signal-to-noise case, the resonance does not help, but neither does it hurt much—a simple filtering operation could remove the resonant signature and allow reconstruction of the original signal waveform.

(N.B.: As we will show below, the actual choice of matched filter for a resonant detector is more subtle than that just described. Instead of τ_d, a shorter averaging time is almost always the optimum choice. Nevertheless, the qualitative thrust of the argument given in the previous paragraph still applies.)

3.4 Free-Mass Detectors as an Alternative

Given the trade-off between sensitivity and bandwidth that resonant systems tempt one to make, it is worth exploring whether there are other entirely non-resonant detection schemes that can achieve high sensitivity to gravitational waves without sacrificing signal bandwidth. In fact, such free-mass detectors have been developed by a variety of workers, including the same Robert Forward who worked with Weber on the original resonant detector.[10,11] The essential advantage of free-mass detectors comes from the fact that the farther apart their test masses are placed, the larger is the relative displacement between them caused by a given gravitational wave amplitude $h(t)$. (This scaling relation holds true up to the point that the light travel time between the masses becomes comparable to the period of the wave; that is when separation of the masses becomes comparable to the wavelength of the wave.) But the resonance in a resonant detector comes roughly when the sound travel time across the bar matches the period of the wave. That is to say, resonant detectors reach their optimum sensitivity when the separation of the test masses is of order of the acoustic wavelength at the gravitational wave frequency. Since the speed of sound in materials is of order 10^{-5} of the speed of light, a free-mass detector at its optimum length can have an advantage in signal size of 10^5 over a resonant-mass detector at its optimum length.

Another advantage is that no resonance is used to boost the signal. Thus, in principle a free-mass detector can have a completely white frequency response. This ideal can not be completely achieved in practice, since some of the noise sources discussed below have strong frequency dependences of their own. Still, it is possible to achieve useful bandwidths measured in decades rather than in fractions of an octave.

This signal size advantage would be a hollow one if there were no sensitive way to measure the relative displacement of test masses separated by many kilometers. Fortunately, there are such ways. As we saw above, the travel time of electromagnetic signals between the test masses can be measured with great precision. Interferometry using visible or near-infrared light to measure the separation of free masses has become a well-developed technology that now is completely competitive with the best resonant-mass detectors, and which is about to undergo a great leap in sensitivity as new instruments of multi-kilometer scale come on line in the next couple of years. Radio ranging between interplanetary space probes separated by many millions of kilometers has been used for some time; optical interferometers in solar orbit, with million kilometer baselines, are now being planned.

The conceptually simpler free-mass detectors are in practice substantially more complicated devices; the freedom of the test masses must be tamed by servo systems to keep them operating properly. This is in part what is responsible for the time lag in their development, even though they were conceived not much later than resonant-mass detectors. In the remainder of the review, we will discuss both styles of gravitational wave detector.

4 Noise Sources

In this section, we will focus our attention on understanding the most fundamental noise sources with which the practice of gravitational wave detection has to contend. Perhaps not surprisingly, the list will seem to have little to do with general relativity or with gravitational waves, as such. The chief concerns of gravitational wave detector designers are those that would confront anyone attempting to measure the effect of a very weak force on a mechanical system: Brownian motion (also known as thermal noise), and noise from the readout system (both in its direct influence on the output of the system and through its "back-reaction" on the mechanical front end). A ubiquitous but non-fundamental noise source, seismically-induced vibration, is treated as well.

It is pedagogically simpler to introduce the topics first in the context of interferometers. Then, we will describe how similar considerations apply to resonant-mass

detectors, where the signal processing issues are a bit more subtle.

4.1 Thermal Noise

The first recognition of a classical physical limit to measurement precision occurred when thermal noise was discovered in galvanometers, in the early 1930's.[12] The premier current-measuring instrument of their day, galvanometers typically consist of a coil of wire suspended from a fine fiber so that it rests between the poles of a strong permanent magnet. Leads from the coil are attached to the external source of current, which generates a torque about the vertical axis. The resulting angular displacement of the coil can be read through an optical lever arrangement that uses a small mirror fixed to the coil.

From our point of view as students of mechanical instruments, the components of a galvanometer that make it function as a current-measuring device are less interesting than its basic mechanical features as a single degree-of-freedom oscillator: an inertia element (characterized by the moment of inertia I of the coil, mirror, and connecting fixtures), an elastic element (represented by the torsional spring constant κ of the fiber from which the coil was suspended), mechanical damping (from some combination of air friction, electrical resistance, and internal friction in the suspension), and a means of observing the angular coordinate θ (the optical lever).

Attempts to make current measurements of the highest precision confronted the fact that, scrutinized carefully, the angle of the coil was not strictly fixed, but instead jittered about its mean position. At first, it seemed natural to attribute the motion to seismic excitation of the galvanometer and its surroundings. But careful mounting of galvanometers to rigid piers isolated from excess building vibrations only went so far in minimizing the noise. Furthermore, seismic noise is typically strongly variable with time, depending on the violence of the weather and on the diurnal cycle of human activity. But well-constructed galvanometers exhibit a noise whose amplitude does not vary in time.

It was not long after Einstein's explanation of Brownian motion as the result of random impacts upon the observed object by molecules from the surrounding fluid that physicists recognized that galvanometers were exhibiting the same phenomenon. According to the Equipartition Theorem, each degree of freedom of a system in thermodynamic equilibrium at temperature T should have an energy whose expectation value is $k_BT/2$. Applying this to the potential energy $\frac{1}{2}I\omega_0^2\theta^2$ associated with the angular displacement θ of the galvanometer, one finds

$$\theta_{rms} = \sqrt{\frac{k_BT}{I\omega_0^2}}, \tag{13}$$

where $\omega_0 = \sqrt{\kappa/I}$ is the resonant frequency of the galvanometer. The heuristic explanation of this theorem is the essentially atomic nature of all mechanisms of dissipation; in this case by the random impact of air molecules on the coil (in cases where air friction dominates the damping) or the random motion of electrons through the coil driving a noisy magnetic torque (if electrical dissipation dominates).

Because it is rooted in thermodynamics, this result has a generality far beyond the details of any particular system. There is a natural analogy between the galvanometer's torsional oscillator and the fundamental longitudinal mode of a Weber-style resonant-mass detector of gravitational waves. Note one striking feature of Eq. (13): even though the origin of the fluctuations lies in the mechanism that is responsible for the dissipation, the rms displacement does not depend on the magnitude (let alone on the mechanism) of the dissipation.

Subsequent work brought recognition of analogous phenomena in other systems. Perhaps the most important was the discovery of the electrical analog of Brownian motion by Johnson and Nyquist in 1928.[13,14] But full recognition of the essential unity of all thermodynamic fluctuation phenomena awaited the formulation of the Fluctuation-Dissipation Theorem. A particularly useful form was established in 1951 and 1952 by Callen and co-workers.[15] We will quote it in a form most directly applicable to mechanical systems, but the derivations in the original papers make it clear how it can be applied to any linear system in thermodynamic equilibrium.

In Callen's formulation, it is convenient to describe the dynamics of a physical system in terms of the network functions called the impedance and admittance, as measured at the point of interest in the system. These are defined in terms of the steady-state response of the system to sinusoidal excitation. The impedance Z is defined as the complex ratio of the force applied at the point of interest to the resulting velocity at that point. That is, if the system is driven with a force $F_0e^{i\omega t}$ and it responds in the steady state with $v = v_0e^{i(\omega t+\phi)}$, then the impedance is

$$Z(\omega) = \frac{F_0}{v_0}e^{-i\phi}.$$

A point mass has an impedance $Z_m(\omega) = i\omega m$, a Hooke's Law spring has $Z_k(\omega) = k/i\omega$, and a dashpot that supplies a force $F = bv$ thus has an impedance $Z_b = b$. The

related concept called the admittance Y is defined by

$$Y(\omega) \equiv Z^{-1}(\omega) = \frac{v_0}{F_0} e^{i\phi}.$$

With these preliminaries, Callen's Fluctuation-Dissipation Theorem can be succinctly stated. The thermodynamic fluctuations analogous to Brownian motion have a magnitude given by the application at the point of interest of a random force with a power spectrum

$$S_F(\omega) = 4k_B T Re(Z). \tag{14}$$

The strength of the applied force power spectrum is proportional to the dissipative (real) part of the impedance; hence the name "fluctation-dissipation" theorem. Note that this expression has the same form as the more familiar power spectrum for the Johnson noise voltage, $S_V(\omega) = 4k_B T R$, where the resistance R is the real part of the electrical impedance. The similarity is not accidental, but is only one example of the many phenomena unified by the theorem.

An alternative form of the theorem, more useful in some situations, directly gives the displacement fluctuation power spectrum instead of the equivalent applied noise force. It states

$$S_x(\omega) = \frac{4k_B T}{\omega^2} Re(Y). \tag{15}$$

Again, the power spectrum scales with the amount of dissipation in the system.

Clearly, this description of fluctuation phenomena is richer than the Equipartition Theorem, since here we have expressions for the entire power spectrum of the fluctuations, not just their rms amplitude. But are the two descriptions even consistent? The rms fluctuation, for example, the expression in Eq. (13), has no dependence on the magnitude of the dissipation. But Eq. (15) shows that the fluctuation power spectrum is proportional at each frequency to the amount of dissipation at that frequency. How can both be true? An oscillator with low dissipation shows a very pronounced peak in its response at the resonance frequency, while one with larger dissipation exhibits a less dramatic peak. So, although the driving noise force is smaller when the dissipation is smaller, the response on resonance is greater. The two effects precisely cancel, as can be verified by direct integration, thus guaranteeing that the integral of the power spectrum Eq. (15) is equal to what one would predict from the Equipartition Theorem.

These two faces of thermal noise, rms magnitude, and power spectrum, are each important in the appropriate context. In a broad-band gravitational wave detector, such as one using an interferometer, the power spectrum carries the most valuable information.

This insight is embodied in the universal choice to suspend the test masses as pendulums. Pendulums are chosen because they are the best way known to create a low frequency oscillator with very low dissipation. Heuristically, most of the restoring force in a pendulum comes from the tension in its wires (due in turn to the gravitational force on the mass); this process has no dissipation associated with it. The only unavoidable dissipation is that associated with the flexure of the wires, but in a properly designed pendulum the fraction of restoring force associated with flexure is small. Hence, the internal friction in the wires is "diluted" by a large factor (perhaps of order 10^3).

Similarly, one wants to minimize the thermal noise associated with internal vibrations of the test masses. This can be achieved only by making the masses out of a material with very low dissipation. Fortuitously, fused silica has very low mechanical dissipation at acoustic frequencies at room temperature.

A standard design rule in those devices is to attempt to place all resonances (such as those associated with the pendulum suspension of the test masses or those involving internal vibrations of the test masses themselves) outside of the frequency band in which signals will lie. When this is done, only the off-resonance amplitude of the power spectrum is important. The off-resonance transfer function of an oscillator to a given force is controlled by the compliance of the resonator in the low frequency limit, and by the inertia of the oscillator above resonance. If the dissipation that sets the driving force can be made low, so can the power spectrum of thermal noise at the frequencies of interest.

4.2 Readout Noise and the Quantum Limit

All experiments need readout and recording systems to register the effects for which we are searching. If the effect is large enough, then these functions can be carried out essentially perfectly. But in the case of the tiny mechanical effects we expect from gravitational waves, even to make the mechanical system's response large enough to record requires very carefully designed readout systems. It is not possible in all cases to ensure that the noise in the readout system is small compared to the mechanical noise in the test masses.

Readout noise has two faces, either one of which may dominate depending on the circumstances. The most familiar is additive noise that competes with a fair copy of the mechanical signal in the output of the measuring system. But measurement systems also unavoidably add mechanical noise to the front end; this "back reaction" noise

must also be kept small if the highest possible precision is to be attained.

The trade-off between additive noise and back-reaction noise is most familiar to physicists from discussions of the quantum mechanical Uncertainty Principle. And, indeed, the Uncertainty Principle governs the ultimate precision of a large class of measurements.

4.2.1 The Heisenberg Microscope as a Prototype Measuring Instrument

It is convenient to make a mental division of a gravitational wave interferometer into two parts. Call the nearly freely-falling mirrored test masses (and the space-time between them) the "system to be measured," and the laser, light beams, and photodetector the "measuring apparatus." There is a deep analogy here with the archetypal quantum mechanical measurement problem called the "Heisenberg microscope." Bohr gave a particularly clear description of it, using a semi-classical treatment. In his 1928 essay "The Quantum Postulate and the Recent Development of the Quantum Theory," Bohr wrote:[16]

> In using an optical instrument for determinations of position, it is necessary to remember that the formation of the image always requires a convergent beam of light. Denoting by λ the wave-length of the radiation used, and by ϵ the so-called numerical aperture, that is, the sine of half the angle of convergence, the resolving power of a microscope is given by the well-known expression $\lambda/2\epsilon$. Even if the object is illuminated by parallel light, so that the momentum h/λ of the incident light quantum is known both as regards magnitude and direction, the finite value of the aperture will prevent an exact knowledge of the recoil accompanying the scattering. Also, even if the momentum of the particle were accurately known before the scattering process, our knowledge of the component of momentum parallel to the focal plane after the observation would be affected by an uncertainty amounting to $2\epsilon h/\lambda$. The product of the least inaccuracies with which the positional co-ordinate and the component of momentum in a definite direction can be ascertained is just given by [the uncertainty relation].

In a gravitational wave interferometer, we are hardly dealing with a microscopic system: the test masses will have masses of 10 kg or more. Yet because we aspire to such extreme precision of measurement, it is crucial to consider the sort of quantum effects usually relevant only for processes on the atomic scale. Note that we are not satisfied to know our test masses' precise positions at one moment only; we want to know the history of the path length difference of the interferometer. Perturbations of the momenta of the masses cannot be ignored, therefore, since the value of the momentum at one time affects the position later.

In the Heisenberg microscope, the phenomenon conjugate to the registration of the arrival of a photon that has bounced off an atom is the recoil of the atom caused by the change in the photon's momentum upon reflection. In a gravitational wave interferometer we register an arrival rate of photons that depends on the difference in phase between electromagnetic fields returning from the two arms. We can recognize the conjugate phenomenon by looking for a fluctuating recoil that can affect the same degree of freedom that we measure. Fluctuating radiation pressure on the test masses causes them to move in a noisy way. The resulting fluctuation in the length difference between the two arms shows how this effect can alter the phase difference between light arriving from the two arms; this identifies it as the conjugate phenomenon.

4.2.2 Shot Noise in an Interferometer

The general principles discussed above can be made much clearer by consideration of specific cases. In an interferometer, all of the physics involved in the fundamental noise limits can be made explicit.

First, consider the question of the direct or "additive" noise, which sets the limit to how small a strain can be recognized. A Michelson interferometer with free mirrors can be described as a transducer from gravitational wave strain to output light power. Part of the reason for its good sensitivity comes from the fact that the output power changes from zero to its maximum and back again with a change in path length difference of only one optical wavelength, of order 1 μm. But it would clearly not be sufficient to only register the difference between one fringe and the next; with a precision of one optical wavelength, even if the arm lengths were as long as useful (of order of half the gravitational wavelength, or 150 km for waves of frequency 1 kHz) one could only register strains of about 10^{-11}.

The key to reaching strain sensitivities of 10^{-21} lies in determining the path length difference to a tiny fraction of a fringe, say 1 part in 10^{10}. Is this possible?

First, recall that the output power is given by

$$P_{out} = P_{in} \cos^2(k_x L_x - k_y L_y). \tag{16}$$

Consider in particular the behavior of the interferometer near an operating point at

which half of the maximum possible power exits the output port. At this point, the change in output power is maximized for a given change in path length difference. If we want to observe a very small change in arm length difference, then we must be able to recognize a very small change in the output power of the interferometer. In other words, the readout precision of an interferometer is limited by the precision with which we can measure optical power.

The fundamental limit to this power measurement is the so-called "shot noise" in the light. We can model the light flux at the photodetector as a set of discrete photons whose arrival times at the photodetector are statistically independent, although with a deterministic mean rate $\bar{n}$. Whenever we count a number of discrete independent events characterized by a mean number $\bar{N}$ per counting interval, the set of outcomes is characterized by a probability distribution $p(N)$ called the *Poisson distribution*,

$$p(N) = \frac{\bar{N}^N e^{-\bar{N}}}{N!}. \tag{17}$$

(This is also colloquially referred to as "counting statistics.") When $\bar{N} \gg 1$, the Poisson distribution can be approximated by a Gaussian distribution with a standard deviation σ equal to $\sqrt{\bar{N}}$.

We are trying to determine the rate of arrival of photons $\bar{n}$ (with units of $\sec^{-1}$), by making a set of measurements each lasting τ seconds. The mean number of photons in each measurement interval is $\bar{N} = \bar{n}\tau$. The Poisson fluctations of the measurement process mean that the fractional precision of a single measurement of the photon arrival rate (or, equivalently, of the power) is given by

$$\frac{\sigma_{\bar{N}}}{\bar{N}} = \frac{\sqrt{\bar{n}\tau}}{\bar{n}\tau} = \frac{1}{\sqrt{\bar{n}\tau}}. \tag{18}$$

This says that if we were to try to estimate $\bar{n}$ from measurements for which $\bar{n}\tau \sim 1$, then the fluctuations from instance to instance will be of order unity. If $\bar{n}\tau$ is very large, then the fractional fluctuations are small.

Let's carry through the calculation for the power fluctuations, and thence to the noise in measurements of h. Each photon carries an energy of $\hbar\omega = 2\pi\hbar c/\lambda$. If there is a power P_{out} at the output of the interferometer, then the mean photon flux at the output will be

$$\bar{n} = \frac{\lambda}{2\pi\hbar c} P_{out}. \tag{19}$$

At the half-power operating point,

$$\frac{dP_{out}}{dL} = \frac{2\pi}{\lambda} P_{in}. \tag{20}$$

We can also consider this to be the sensitivity to the test mass position *difference* δL, since the interferometer is equally sensitive (with opposite signs) to shifts in the length of either arm.

Now consider the fluctuations in the mean output power $P_{out} = P_{in}/2$, averaged over an interval τ. The mean number of photons per interval is $\overline{N} = (\lambda/4\pi\hbar c)P_{in}\tau$. Thus we expect a fractional photon number fluctuation of $\sigma_{\bar{N}}/\bar{N} = \sqrt{4\pi\hbar c/\lambda P_{in}\tau}$. Since we are using the output power as a monitor of test mass position difference, we would interpret such statistical power fluctuations as equivalent to position difference fluctuations of a magnitude given by the fractional photon number fluctuation divided by the fractional output power change per unit position difference, or

$$\sigma_{\delta L} = \frac{\sigma_N}{N} \Big/ \frac{1}{P_{out}} \frac{dP_{out}}{dL} = \sqrt{\frac{\hbar c\lambda}{4\pi P_{in}\tau}}. \tag{21}$$

Recall that we can describe the effect of a gravitational wave of amplitude h as equivalent to a fractional length change in one arm of $\Delta L/L = h/2$, along with an equal and opposite change in the orthogonal arm. The net change in test mass position difference is $\delta L = Lh$, so if we interpret brightness fluctuations in terms of the equivalent gravitational wave noise σ_h, we have $\sigma_h = \sigma_{\delta L}/L$, or

$$\sigma_h = \frac{1}{L}\sqrt{\frac{\hbar c\lambda}{4\pi P_{in}\tau}}. \tag{22}$$

There is no preferred frequency scale to this noise; the arrival of each photon is independent of the arrival of each of the others. Note also that the error in h scales inversely with the square root of the integration time. These facts can be summarized by rewriting Eq. (22) as the statement that the *photon shot noise* in h is described by a white amplitude spectral density of magnitude

$$h_{shot}(f) = \frac{1}{L}\sqrt{\frac{\hbar c\lambda}{2\pi P_{in}}}. \tag{23}$$

4.2.3 Radiation Pressure Noise in an Interferometer

A hint at where quantum mechanics might have some deep relevance comes when we consider how shot noise scales with the optical power used in the interferometer. As shown in Eq. 23 above, the shot noise readout precision improves as the square root of the optical power. Taken at face value, this would suggest that we could achieve arbitrarily good measurement precision, so long as we were able to use an arbitrarily powerful laser to illuminate the interferometer.

That conclusion is encouraging, but ought to cause some unease. After all, didn't Einstein fail to find a way to defeat the Uncertainty Principle's limit to the precision of mechanical measurements?[17] Or did he just blunder not to consider using an interferometer to make the measurements?

Of course not. An interferometer is in fact very much like the "Heisenberg microscope" of Bohr. The lesson we should draw from that example is that we must not neglect the effects of recoil in an interferometer. In effect, an interferometer with free masses is a Heisenberg "macroscope": the role of the small object whose position is to be determined is played by the set of macroscopic test masses. Their large size gives an obvious advantage against recoil, but can't make the effect vanish entirely.

The recoil effect does have one extra bit of subtlety, though. Recall that the measurement we make in a Michelson interferometer has to do with the difference in length of the two arms. So the recoil force we need to calculate is not the common mode force (the component that is equal in the two arms), but only the noise in the difference of the recoil forces in the two arms. If we use a naive picture in which photons independently choose which arm to enter, then the origin of a differential recoil force noise is clear. Each photon enters either one arm or the other; whenever one arm gains a photon, the other loses one.

To estimate the size of this effect, first recall that the force exerted by an electromagnetic wave of power P reflecting normally from a lossless mirror is

$$F_{rad} = \frac{P}{c}. \tag{24}$$

The fluctuation in this force is due to shot noise fluctuation in P. That is

$$\sigma_F = \frac{1}{c}\sigma_P, \tag{25}$$

or, in terms of an amplitude spectral density

$$F(f) = \sqrt{\frac{2\pi\hbar P_{in}}{c\lambda}} \tag{26}$$

independent of frequency.

This noisy force is applied to each mass in an arm. For now, let us consider a simple "one-bounce" interferometer. We will allow the mirrors at the ends of the arms to be free masses, but in this example assume that the beam splitter is much more massive than the other mirrors. The fluctuating radiation pressure from the power $P_{in}/2$ causes each mass to move with a spectrum

$$x(f) = \frac{1}{m(2\pi f)^2}F(f) = \frac{1}{mf^2}\sqrt{\frac{\hbar P_{in}}{8\pi^3 c\lambda}}. \tag{27}$$

The power fluctuations in the two arms will be anti-correlated. The radiation pressure noise is then

$$h_{rp}(f) = \frac{2}{L}x(f) = \frac{1}{mf^2 L}\sqrt{\frac{\hbar P_{in}}{2\pi^3 c\lambda}}. \tag{28}$$

4.2.4 The Standard Quantum Limit

Thus we have two different sources of noise associated with the quantum nature of light. Note that they have opposite scaling with the light power-shot noise declines as the power grows, but radiation pressure noise grows with power.

If we choose to, we can consider these two noise sources to be two faces of a single noise that we can call optical readout noise, given by the quadrature sum

$$h_{o.r.o.}(f) = \sqrt{h^2_{shot}(f) + h^2_{rp}(f)}. \tag{29}$$

At low frequencies, the radiation pressure term (proportional to $1/f^2$) will dominate, while at high frequencies the shot noise (which is independent of frequency, or "white") is more important. We could improve the high frequency sensitivity by increasing P_{in}, at the expense of increased noise at low frequency. At any given frequency f_0, there is a minimum noise spectral density; clearly, this occurs when the power P_{in} is chosen to have the value P_{opt} that yields $h_{shot}(f_0) = h_{rp}(f_0)$.

When we solve for P_{opt} and insert it into our formula for $h_{o.r.o.}$ we find

$$h_{QL}(f) = \frac{1}{\pi f L}\sqrt{\frac{\hbar}{m}}. \tag{30}$$

We have renamed this locus of lowest possible noise $h_{QL}(f)$, for "quantum limit," to emphasize its fundamental relationship to quantum mechanical limits to the precision of measurements. Note that the expression does not depend on P_{in} or λ, or any other feature of the readout scheme, even though such details were useful for our derivation. Thus, this examination of the workings of our Heisenberg microscope provides an instrument-specific derivation of the Heisenberg Uncertainty Principle. And it reminds us of the truth Bohr's remarks expressed, that in any measurement the Uncertainty Principle emerges from the specific mechanism of the measurement.

There was a moment when some physicists believed, on seemingly sound physical grounds, that this picture of how photons interact with a beam splitter was so flawed that interferometers could perhaps evade the Uncertainty Principle.[18] The argument can be made based on quotation from quantum mechanical Scripture, Dirac's *The Principles of Quantum Mechanics*.[19] There one can read that photons in an interferometer travel down both arms simultaneously; furthermore, it is written that interference can only take place between a photon and itself, so the very existence of interference in a quantum mechanical world is proof of this picture. If this were taken as absolute and literal truth, then it would appear to rule out any differential radiation pressure at all, since the number of photons, and hence the recoil forces, would be identical in the two arms. Without the resulting differential recoil of the test masses, there is no quantum limit. Gravitational waves could in principle be measured with arbitrary precision. Some physicists defended this as gospel, despite the fact that the argument appeared to use quantum mechanical reasoning to disprove quantum mechanics.

The stubbornly naive were untroubled by this argument, and expected the Uncertainty Principle to hold. Some physicists read a few pages further in Dirac's book, to the passage explaining that allowing the possibility of energy measurements, say by observation of recoil of the mirrors, causes collapse of the wave function in such a way that photons end up either in one arm or the other. (Dirac's first discussion refers to an interferometer with rigidly fixed mirrors.) The learned were saved from error by the work of Caves,[20] who invoked the concept of vacuum fluctuations to explain the quantum mechanical behavior of photons at a beam splitter. A vacuum electromagnetic field with zero-point fluctuations enters the interferometer through the output port; its superposition with the field from the laser causes the light to behave in the way expected from semi-classical reasoning.

4.3 Seismic Noise

We have neglected to consider above another source of noise in gravitational wave detectors that is so common and important as to be essentially ubiquitous. This is what is commonly called seismic noise, the continual shaking of the terrestrial environment due to a variety of contingent causes, ranging from small earthquakes to ocean waves driven by large weather systems to automobiles striking potholes in poorly paved streets. Such a complex phenomenon can have no simple explanation from basic physics, yet dealing with it forms a substantial part of the challenge to designers of gravitational wave detectors. (Only moving the whole detector into space suffices to remove it entirely from consideration.)

At a reasonably quiet location, the spectrum of seismic noise from 1 Hz to several hundred Hz can be approximated as

$$x(f) = \begin{cases} 10^{-7}\mathrm{cm}/\sqrt{\mathrm{Hz}}, & \text{from 1 to 10 Hz} \\ 10^{-7}\mathrm{cm}/\sqrt{\mathrm{Hz}}(10\mathrm{Hz}/f)^2, & \text{for } f > 10 \text{ Hz}. \end{cases} \tag{31}$$

The magnitude of this mechanical noise background is distressingly large. The rms amplitude of the noise over this interval is of order 1 μm. The good news is that the spectrum falls with increasing frequency f. But even so, throughout the range of frequencies of interest to gravitational wave detectors, it involves motions many orders of magnitude larger than would be driven by any conceivable incident gravitational wave. There is no possibility of success unless the effects of seismic noise can be strongly attenuated.

It is straightforward to see the way in which seismic noise mimics a gravitational wave signal in an interferometer. As long as the separation between the mirrors is not very small, then the seismic inputs to each mirror are effectively independent; the difference in arm lengths is driven by the quadrature sum of the noise at all mirrors. The situation is a bit more subtle for a resonant mass detector. If it is suspended at its midpoint, it would appear that its internal modes should not be excited by any motion of the suspension point. However, this argument assumes perfect symmetry of the resonator about the suspension. The approximate symmetry of real systems may give several orders of magnitude of effective isolation, but the seismic spectrum is so large that additional isolation is always required.

Fortunately, the design of seismic isolators is a well-developed art. One can construct mechanical multi-pole low-pass filters that provide outstanding attenuation at frequencies well above those of the filter poles. The art of doing so was introduced to the field of gravitational wave detection by the founder, Weber.[21]

4.3.1 A Simple Two-Pole Isolator

The essence of vibration isolation can be understood using only ideas from freshman physics. Imagine that the object to be isolated has mass m. Assume that it is a rigid body, and that we are only interested in its motion x_m in a single direction. Then we can treat the object as a point mass. If it rests on the ground, it shares the ground's motion x_g, so $x_m = x_g$. To isolate the mass, replace the rigid connection to the ground

with a compliant connection, that is, attach it to the ground through a spring with spring constant k. Then the equation of motion of the mass is

$$m\ddot{x}_m = -k(x_m - x_g). \tag{32}$$

With the usual frequency domain ansatz $x_m(t) = x_m \exp(i\omega t)$, $x_g(t) = x_g \exp(i\omega t)$, Eq. (32) can be solved for the vibration transfer function

$$\frac{x_m}{x_g} = \frac{\omega_0^2}{\omega_0^2 - \omega^2}, \tag{33}$$

where $\omega_0^2 \equiv k/m$. The asymptotic behavior of Eq. (33) reveals the essential features of an isolator—although for low frequencies $\omega \ll \omega_0$, $x_m/x_g \approx 1$ (the mass moves rigidly with the ground), and at high frequencies $\omega \gg \omega_0$, $x_m/x_g \approx \omega_0^2/\omega^2$ (the amplitude of the mass's motion falls steeply with increasing frequency.)

(The transfer function of Eq. (33) has the unphysical feature that it predicts an infinite response at the resonant frequency ω_0. This problem is remedied if we add a damping force to our oversimplified model. Even so, an isolator does have the unfortunate feature that it gives a resonant amplification of the input noise for frequencies near ω_0.)

So, a strategy for isolation is as follows: to isolate an experiment for signal frequency ω, construct an isolator with a low enough resonant frequency ω_0 so that the isolation factor ω_0^2/ω^2 is sufficiently small to reduce the seismically driven value of x_m to a tolerable level.

4.3.2 Isolation Stacks

If it proves difficult to make ω_0 sufficiently low (it often does), another straightforward idea can often be made to work instead. Make two or more isolators with resonant frequencies as low as is convenient, then cascade them with the system you want to isolate farthest from the noisy ground. Solving for the resonant frequencies of the coupled chain of isolators is a normal mode problem that may get complicated. But the high frequency limit of a chain of N isolators, each with resonant frequency ω_0 by itself, has the simple form

$$\frac{x_m}{x_g} \approx \left(\frac{\omega_0^2}{\omega^2}\right)^N. \tag{34}$$

In Weber's original experiment, a vibration isolation "stack" of alternating steel plates and rubber sheets was used. No details are given in the text of his papers, but the idea was clear, and was adopted by those who followed Weber as well. Among the nice features of this isolator are: rough equality of the isolation in all degrees of freedom (since the rubber is compliant in both compression and in shear), and a small level of resonant amplification (since the rubber is rather lossy).

4.3.3 Isolation for Interferometers

The pendulum that provides a low-dissipation suspension for the test mass in an interferometer also provides two poles of seismic isolation. Typically, the resonant frequency is close to 1 Hz. This single stage might almost provide enough isolation for gravitational wave measurements in the vicinity of 1 kHz, but certainly requires substantial augmentation for good sensitivity at lower frequencies. This additional isolation needs to be at least roughly isotropic, since there are a variety of cross-coupling mechanisms that can make vertical noise couple to the phase of the light beam.

The simplest way to add isolation is to suspend the pendulum from a stack of the sort used to isolate a resonant-mass detector. Since interferometers are usually aimed at having good sensitivities down to the lowest possible frequencies, there is a strong incentive to make a stack with the lowest possible resonant frequencies.

More dramatic re-engineering of the basic stack concept can yield greater rewards. The strongest effort in this direction has been made by the VIRGO group in Pisa.[22] The basic idea is to build a stack based on many cascaded pendulums, but to build enough vertical compliance into each pendulum stage to give effective vertical resonant frequencies comparable to the 1 Hz resonances of the pendulum stages. In its original version, the vertical springs were air springs made of flexible bellows. This yielded impressive performance, which was improved even further when the vertical compliance was enhanced by magnetic anti-springs arranged to cancel some of the vertical stiffness. The original version would have made seismic noise negligible (compared with shot noise) for all frequencies above 10 Hz, while the enhanced version would work down to 4 Hz. The air springs did suffer from one serious drawback; the temperature coefficient of the spring force is determined by the ideal gas law, a much stronger sensitivity than elastic springs. Drifts proved very hard to control. To remedy this problem, the system was completely redesigned, with the air springs replaced by pre-stressed steel cantilevers in the form of narrow triangular blades, supplemented with magnetic anti-springs.

4.4 Noise in Resonant-Mass Detectors

Key features The essential complication in understanding resonant-mass detectors (as compared to interferometric detectors) is that the degree of freedom of interest is that of a simple harmonic oscillator (or a collection of them, as we'll see in the next section). So in addition to any intrinsic frequency dependence in the noise, there is a deliberately constructed resonant transfer function in the detector itself. As we saw earlier in this review, the resonance was introduced as part of a strategy for overcoming wide-band noise in the amplifier.

The use of this strategy involves different heuristic concepts than are appropriate for interferometers. In particular, optimizing the sensitivity of a resonant detector to short bursts almost always involves choosing to average the output over times that are long compared with the length of the burst itself. Then, the measurable quantity is no longer $h(t)$, but is instead net change in the vector amplitude (magnitude and phase) of the resonator's oscillation. This in turn can be expressed in terms of the energy that the wave would have deposited in a resonator at rest.[23] If the gravitational waveform $h(t)$ has a Fourier transform $H(f)$, then that excitation energy E is (for an optimal orientation between bar and wave)[24]

$$E = \frac{M v_s^4}{L^2} \left| H\left(f_0\right) \right|^2 , \tag{35}$$

where M is the total mass of the bar, v_s is the speed of sound in the material, L is the overall length of the bar, and f_0 is the resonant frequency.

The distinctive features are twofold: characterization of all candidate events by a single number (usually its "energy" or else $T \equiv E/k_B$), and a signal-to-noise optimization that involves choosing the right averaging time (or bandwidth).

Resonant transducers The second generation of resonant-mass detectors replaced Weber's piezoelectric transducer with a kind of a bridge circuit, in which the mechanical motion unbalanced the bridge by modulating the inductance or capacitance of one leg of the bridge. As with piezoelectric transducers, achieving a high level of coupling has proven difficult to achieve. A standard measure of the coupling is the Gibbons-Hawking parameter β, defined as "the proportion of elastic energy of the detector that can be extracted electrically from the transducer in one cycle."[25] In principle, the excitation of the bridge could be increased without limit, but in practice large fields usually lead to excess dissipation in the transducer even before electrical breakdown occurs. Transducers have been limited to working values of β of around 10^{-2}.

A heuristic way of understanding the design problem is to think of the issue as an attempt to design a transducer that makes a reasonable electrical impedance at its output appear to the mechanical system as a mechanical resistance sufficient to supply appreciable damping to the bar. With bar masses in excess of one ton, this may seem inordinately difficult. The good values mentioned in the previous paragraph avoided this problem by making use of so-called resonant transducers, which have been adopted almost universally since the idea was proposed by Paik in 1976.[26]

Paik's design called for a smaller mechanical resonator to be attached to the main resonant mass M. The resonant frequency of the smaller resonator itself (i.e., with the larger resonator "clamped") is chosen to match that of the main resonator. The actual coupled system then has two normal modes. If the mass ratio $m/M \equiv \alpha \ll 1$, then it is easy to show that the ratio of the amplitude of motion of the small mass, compared with that of the main resonator, is

$$\left| \frac{x}{X} \right| = \frac{2}{\sqrt{\alpha}}.$$

When a gravitational wave burst interacts with such a resonant system, it will at first mainly excite the vibration of the large bar. (The Paik resonator is a small device at one end of the bar, so the gravitational wave strain has only a small baseline for creating a stretch in its spring.) The free motion of this two-mode system then exhibits "beats," during which the mechanical energy of the main resonator's original motion is transferred into excitation of the small resonator. During this phase of the beat cycle, the effect of the gravitational wave has been transformed into a motion $2/\sqrt{\alpha}$ times larger than in a detector without the resonant transducer. The electro-mechanical transducer is mounted so as to measure the motion of the small mass with respect to the end of the main resonator, thus presenting this larger motion to the rest of the signal processing system.

The advantage this offers in detecting weak signals is probably obvious. The larger motion generates a comparably larger electrical output from the transducer, reducing the importance of a given level of electrical amplifier noise. Another way of seeing the advantage is to recognize the much smaller mechanical impedance required to damp the motion of the smaller mass, which means that β is increased by a factor of order α^{-1}. In present day designs, the mass ratio α is typically of order a few times 10^{-3}. The value is set in an optimization that involves not only thermal noise and additive amplifier noise but the back-action noise as well.

4.4.1 Thermal Noise in Resonant-Mass Detectors

Even though resonant-mass detectors work in a rather different way from free-mass detectors, they still benefit from low levels of dissipation. Here, one makes exactly the opposite design choice in regard to the resonant frequency—rather than trying to put all resonances outside the signal band of interest, the whole detection strategy is based on placing the resonance at a frequency where one expects signal power to be large. So, on the face of it, all of the $\frac{1}{2}k_BT$ of energy contributes to the rms scatter of the instrument's output. But it would be too hasty to conclude that there is no dependence of the effective thermal noise on the dissipation level. The typical resonant-mass detector can have a damping time of order 1000 sec, while the resonant frequency is near 1 kHz (i.e., a quality factor Q of order 10^6). The signal being sought might be a brief burst of radiation, lasting only a millisecond. In this brief interval, the bar's amplitude and/or phase of oscillation are altered. The characteristic time for such changes due to thermal noise is the damping time. So only a small fraction of the thermal noise power is relevant in the detection process. For this reason, low dissipation can be as valuable in resonant-mass detectors as in interferometers.

To make the analogy with free-mass detectors clearer, we can make a frequency domain version of the argument in the previous paragraph. It is most convenient to consider the comparison between the thermal noise driving force, Eq. (14), and the force exerted by the gravitational wave on the detector. A brief burst of gravitational waves will have a broad spectrum; its matched filter will admit power over a wide band. The thermal noise force power spectrum that may obscure the effect of the gravitational wave is broad-band as well; if the dissipation in the bar has the form of velocity damping, then the spectrum is white. The amplitude of this broad band noise spectrum is proportional to the amount of dissipation. So the signal to noise ratio will be better, the smaller is the amount of dissipation.

An expression for the minimum detectable noise $\Delta E_{\rm min}$ (at the 1 σ level) is[27]

$$\Delta E_{\rm min} \approx 2k_BT\frac{\tau_s}{\tau_d}, \tag{36}$$

where T is the temperature, τ_s is the averaging time, and τ_d is the damping time of the bar. In this case (where we assume only thermal noise is important), it is clear that the shorter the averaging time the better the sensitivity.

4.4.2 Readout Noise in a Resonant-Mass Detector

Most of the issues of readout noise discussed above for interferometers were first faced in the context of resonant-mass detectors. We have inverted the historical order solely for pedagogical reasons; the physical origin of readout noise in an interferometer is transparent, and nominally free masses make a more straightforward analogy to the Heisenberg microscope than does the complex amplitude of a resonant mode. A key role was played by Braginsky[28] in calling attention to the necessity of understanding these issues in resonant-mass gravitational wave detectors.

Transducer model The transducer is treated as a black box, in the form of a two-port network, emphasizing the fact that the system has a single input port and a single output port. Unlike the two-ports usually studied in electrical engineering classes, the transducer changes the dimensions of the signal: a mechanical motion at the input is converted to an electrical signal at the output.

One such device is a seismometer pickup: a permanent magnet attached to the end of the bar, arranged so that it moves back and forth in the vicinity of a conducting coil whenever the bar vibrates. Motion of the mechanical system generates an electromotive force in the coil that can be electrically amplified and recorded.

In gravitational wave detectors, the transducers usually involve a reactance (i.e., an inductance or a capacitance) whose value is modified by relative motion between the transducer mass and the end of the bar. This generates an electrical signal by virtue of the fact that the reactive component is placed in one leg of a bridge circuit. This kind of transducer is a passive device. It can add noise if its level of dissipation is large enough so that thermal noise is important.

There has also been development of parametric transducers, in which a balanced bridge (two counter-varying reactances) is excited at microwave frequencies. This is the kind of transducer used on the UWA detector.[29] There is a deep analogy between this kind of transducer and an interferometer.[30] This means that there are both extra benefits (gain in the transducer) and extra complications (such as noise in the oscillator that excites the bridge.)

Pre-amplifier The remaining essential part of the readout system is the low noise pre-amplifier, whose job is to transform the tiny electrical output of the transducer into a signal large enough so that it can be recorded. One needs the noise to be as low as

possible. To describe the noise, and to understand how it affects the detection process, it is valuable to consider a general black box representation of a noisy amplifier as a kind of a two-port network, just as we did for the transducer. Other than the trivial difference that this two-port is an all-electric device, there are two key differences between this kind of two-port and the model transducers discussed above. One is the fact that the amplifier has gain, i.e., it can supply more energy at its output than is supplied by its input. The other is the existence of two sources that represent the generation of noise. They are usually represented as a voltage source and a current source at the input of the two-port network; this is especially convenient, and has wide generality, but is only one of several equivalent ways of representing the noise. More on the general theory of noisy two-port networks can be found in the pioneering paper by Rothe and Dahlke.[31]

It is interesting to pause to inquire why two noise sources are necessary. Recall the Helmholtz theorem, often known in specialized forms as Thévenin's theorem or Norton's theorem.[32] The essence of the theorem is that an arbitrary network of sources and passive components can be represented, as far as its behavior at a given port is concerned, by a single source and a single impedance. But we are dealing here with a network in which two ports are relevant. At each port, one needs an impedance and a source, or their equivalents elsewhere in the circuit. It is traditional to replace the source at the output with an equivalent noise source at the input, whose strength is smaller than the output noise by a factor of the amplifier gain.

These two noise sources play different roles in the measurement process. There is one noise source that is physically present at the input, causing an influence on the system (here the electromechanical transducer) that is attached to the amplifier input. In the jargon of gravitational wave detection, this noise is responsible for "back action," since noise at the input of the pre-amp is thereby applied to the output end of the transducer,where it can cause a mechanical noise force at the transducer input; this is in turn attached to the resonant-mass detector proper. More on this below.

The second noise source (the one replacing the output noise source) is usually referred to as "additive noise": it is added to the amplified signal by the time it appears at the output, without causing any physical effect on the system hooked up to the input.

Amplifier noise in resonant-mass gravitational wave detectors, and the "amplifier limit" In contrast to our discussion of detection strategies in the case where thermal noise dominates, here we discuss the case when additive amplifier noise is the only important noise source. Then, we can best search for a brief burst of gravitational radiation by performing a cross-correlation between the system output and a template consisting of a sinusoid at the mechanical resonant frequency that is damped with the same time constant as the resonance itself. In other words, we look for responses that look like the test mass system suddenly set into resonance. The signal can arrive with any phase of course, so we need to keep track of both sine and cosine components with the bar's damping time. A two-phase lock-in amplifier can be set up to perform exactly this form of averaging.

In this case the energy sensitivity of the detector is given by[27]

$$\Delta E_{\rm min} \approx k_B T_n \frac{\lambda + \lambda^{-1}}{\pi \beta \tau_s f_0}, \tag{37}$$

where T_n is the noise temperature of the amplifier, λ is the ratio of the transducer output impedance to the amplifer noise impedance, β is the Gibbons-Hawking coupling parameter, and τ_s is the averaging time.

Consider the post-detection bandwidth implied by this prescription. The output of the cross-correlation described above is hardly affected if we displace the template with respect to the signal time series by one or even several cycles of oscillation. For there to be a substantial change in the value of the cross-correlation, the template must be displaced by a duration of order the damping time of the mechanical resonance. This means that, if the signal-to-noise ratio is not large, the arrival time of the impulsive gravitational wave signal will be uncertain by of order the damping time. In other words, the post-detection bandwidth Δf of such a signal extraction system is narrow, of order

$$\Delta f \approx 1/\tau_d.$$

For a quality factor of 10^6, this bandwidth is very narrow indeed. Increased bandwidth could be achieved, of course, at the expense of the signal-to-noise ratio, by averaging the output time series for a shorter time than the bar's damping time τ_d.

Combined optimum in the presence of thermal noise and additive amplifier noise When both kinds of noise are present at substantial levels, the best strategy is neither the rapid readout appropriate to thermal noise nor the long averaging time that would be best for amplifier noise. A broad-band output filter will admit too much amplifier noise. Using the narrow-band prescription appropriate to the amplifier-dominated case filters out much thermal noise, but also most of the signal power. Obviously, the optimum in the combined noise case lies somewhere between the extremes, where the net

noise contributions from thermal noise and amplifier noise are equal. Then, the energy sensitivity of the detector is given by[27]

$$\Delta E_{\min} \approx 2k_B T \frac{\tau_s}{\tau_d} + k_B T_n \frac{\lambda + \lambda^{-1}}{\pi \beta \tau_s f_0}, \tag{38}$$

where the averaging time is now to be chosen in such a way as to minimize the sum of the two terms.

The larger is the transducer's coupling β, the smaller is the importance of amplifier noise in units of gravitational wave strength. In the limit of large β, the bandwidth of the resonant-mass gravitational wave detector can approach that given by the thermal noise limit, where the matched template is the signal waveform itself.

For detectors with resonant transducers, bandwidth is also limited by the splitting between the coupled modes of the bar-transducer system. There is a competition between two effects: a small transducer mass gives a large mechanical amplification but a small mode splitting. In the time domain, this would be described as a long beat period; after a signal excites the bar one has to wait for a substantial fraction of the beat period for energy to be transferred to the motion of the transducer mass. This competition could be evaded by the construction of multi-mass transducers with a gradation from intermediate to tiny masses.[33]

Resonant-mass gravitational wave detectors built to date have almost always been substantially under-coupled to their amplifiers. First-generation bars of Weber's type, in which the transduction was performed by piezoelectric transducerss glued around the belly of the cylindrical mass, could only achieve coupling factors of order 5×10^{-6}. One variant constructed by the Glasgow group did achieve a large β.[34] It consisted of a cylindrical bar that had been split in half, then reassembled with a layer of piezoelectric material in effect forming the spring attaching the masses at either end. This design is the closest approach ever to Weber's original vision of a single massive piezoelectric mass as an antenna. In spite of these impressive parameters, the Glasgow bar was never even the most sensitive gravitational wave detector of its day, because the poor mechanical Q of the piezoelectric material gave a large amount of thermal noise.

4.4.3 The Quantum Limit in Resonant-Mass Detectors

One of the key features of Rothe and Dahlke's model of a noisy two-port network is the necessity to describe the noise with two generators, one directly limiting measurement precision and the other disturbing the system being measured. This is reminiscent of concepts most students of physics learn about in the context of quantum mechanics. In fact, quantum mechanics sets a minimum level of amplifier noise that can be expressed in terms of this pair of noise generators.

Quantum mechanical noise in an electrical amplifier A paper by Heffner in 1962 derived the quantum mechanical limit on amplifier noise;[35] its relevance for the gravitational wave detector problem was discovered by Giffard in 1976.[36] What Heffner showed was that, unless an electrical amplifier had a minimum level of noise, it could be used to make measurements of an electrical signal at frequency f that evaded the Uncertainty Principle

$$\Delta n \Delta \phi \geq \frac{1}{2},$$

where Δn is the uncertainty in the number of photons at the frequency f, and $\Delta\phi$ is the uncertainty in the measurement of the phase of the signal.

The amplifier noise that just barely allows measurements at the Uncertainty Principle limit is given by

$$\sqrt{S_v(f) S_i(f)} = \left(\ln \frac{2 - 1/G}{1 - 1/G} \right)^{-1} 2\pi\hbar f.$$

In the limit of large gain G, this limit becomes

$$\sqrt{S_v(f) S_i(f)} = \frac{1}{\ln 2} 2\pi\hbar f. \tag{39}$$

The combination of voltage and current noise power spectra represents yet another pair of conjugate variables governed by an uncertainty relation. This product is proportional to the noise temperature T_n of the amplifier, defined by $k_B T_n \equiv \sqrt{S_v(f) S_i(f)}$.

In the earliest days of the search for gravitational waves, the possible implications of such an amplifier sensitivity limit for the sensitivity of gravitational wave detectors were ignored. Neither Weber's papers nor the canonical theoretical analysis by Gibbons and Hawking mention such a limit. Eventually, cooler heads came to contemplate the fact that gravitational wave detectors ought to obey the Uncertainty Principle.

The electro-mechanical Uncertainty Principle Giffard gave the adaptation of this argument to the electro-mechanical case. His argument is so brief that it is almost hard to state. He points out that the functions of transducer and pre-amplifier together can be thought of as a so-called "mechanical amplifier," a two-port network whose input terminal receives a mechanical signal, but whose output is electrical.

From this, Giffard shows that the Uncertainty Principle requires that unless a gravitational wave signal has a minimum size, no linear gravitational wave detector will be able to register its arrival. He expressed the minimum size in terms of the energy U_s that the wave would deposit in a resonant detector initially at rest. The quantum mechanical limiting sensitivity is

$$U_s > 4\pi\hbar f.$$

As worded above, this argument can seem rather abstract. A heuristic description makes it as vivid as any of Bohr's *gedanken* experiments. In a mechanical amplifier, a crucial quantum mechanical role is played by the back action from the noise source that the Helmholtz Theorem places at the amplifier input. The force noise generator at the mechanical amplifier input (caused by electrical noise acting backwards through the transducer) perturbs the delicate mechanical system, here the resonant-mass system itself. Just as in the Heisenberg microscope, any design trade-off made in an attempt to reduce the position noise ends up increasing the momentum impulse applied to the system being measured.

5 History of Resonant-Mass Detectors

We now turn from a discussion of physics *per se* to a review of the way in which one assembles working gravitational wave detectors in light of the physical principles governing them. We will take a quasi-historical framework for this discussion, as a pedagogically sensible way of grappling with the issues involved. In this section, we will take the chronologically-motivated choice of treating resonant-mass detectors first; then we will start from scratch the overlapping history of interferometric detectors.

By 1966, Joseph Weber had constructed a complete working detector, and by 1968 was reporting coincident observations between detectors separated by 1000 km. The detector contained versions of every essential feature in resonant-mass gravitational wave detectors today, except for the facts that it operated at room temperature and that it used non-resonant strain transducers for its readout. The story of the development of the field since then can be seen as embodying a few key accomplishments: replication of Weber's detectors accompanied by a failure to confirm his claimed detection, clarification of the optimum way to detect gravitational wave signals in a noisy detector and of the sorts of technological developments that could lead to improved detector sensitivity, and the staged implementation of new generations of detectors embodying the improved technology.

5.1 Weber's Detectors as Gravitational-Wave Detection Systems

All of the important elements that make up a working gravitational wave detector are described in Weber's 1966 *Physical Review Letters*.[21] Looking at the signal chain from the front end, we see first the large $1\frac{1}{2}$ ton aluminum cylinder whose fundamental longitudinal mode at 1657 Hz interacts with any incoming gravitational wave. Around its midsection are glued the quartz transducers that give, through the piezoelectric effect, electrical signals proportional to the strain in the aluminum cylinder. Signal leads from those transducers pass through acoustic filters and through the wall of the vacuum chamber, then are connected via another acoustic filter to a superconducting inductance that serves as a "tank circuit" at the input of the low-noise pre-amplifier (whose noise temperature is 50 K.) The output of the pre-amp is connected to further amplification. There follows a rectifier for generating a positive-definite signal proportional to the power out of the amplifier. The end of the signal chain is a recording device, which in 1966 consisted of a pen-and-ink chart recorder.

Another aspect of the detector is the means used to prevent its being excited by influences other than gravitational waves. Isolation against mechanical influences in the form of acoustic or seismic noise is shown clearly in the diagram. Direct acoustic excitation is prevented by the placement of the key parts of the experiment inside a vacuum chamber. Transmission of vibration through the signal leads is attenuated by the acoustic filters mentioned above. The path for vibrations from the floor must pass through a pair of isolation stacks each consisting of three stages of rubber pad/steel block isolators; the aluminum bar is further isolated by a pendulum suspension consisting of a single wire sling that supports the bar about its middle. The top ends of the wire are attached to a beam that spans the space between the isolation stacks.

Further progress in reducing the sensitivity of his bar to spurious external influences enabled Weber to make his 1969 claim of "Evidence for discovery of gravitational radiation".[8] One aspect of this progress consisted of augmenting the electromagnetic shielding of his devices, after tests revealed some sensitivity. A seismometer array was also used to check for correlations between strong vibration of the ground and large detector outputs. But by far the most important system element in this regard was the construction of multiple copies of the complete detector system, and their deployment at spatially separated locations. The network in 1969 consisted of one bar of 66 cm

diameter and 153 cm length (for a resonant frequency of 1657 Hz) located at the Argonne National Laboratory in Illinois, and five other resonant-mass detectors located at the University of Maryland. One bar at Maryland, whose properties matched closely the Argonne bar, was used together with it to look for coincident excitation. Some use of the outputs of the additional bars was made, but the procedure is poorly described.

5.2 Weber's Observational Program

Weber's ongoing research program was concerned with looking for further evidence that the coincident events were real (i.e., not statistical accidents) and that they were in fact due to gravitational radiation. The *Physical Review Letters* of 9 February 1970[37] discusses the statistics from several interesting points of view. The claim that the events are real is buttressed by calculation of the rate of accidental coincidences. For the whole range of threshold levels considered, the number of coincidences exceeds the expected accidental rate by about an order of magnitude. Another even more convincing piece of evidence comes from carrying out a search for coincidences between the output of the two detectors using a scheme in which one of the detectors' outputs was delayed by two seconds, enough to remove any correlation due to excitation by gravitational wave impulses. Just as one would hope, only a small number of coincidences occurred between the offset data streams, at a level consistent with accidentals.

Still another convincing piece of evidence was revealed in the *Physical Review Letters* of 20 July 1970.[38] In it, Weber uses the classic astronomical test of testing for correlation of his event rate with sidereal time. Jansky's claim for the extraterrestrial origin of his observed radio noise had been buttressed in just this way; more recently, Jocelyn Bell had convinced her reluctant thesis advisor Hewish of the extraterrestrial nature of her pulsar signals by a similar test. Weber took advantage of the fact that the response of a cylindrical bar is anisotropic, and that the axes of his bars were oriented East-West, so that their sensitivity pattern was swept across the sky by the rotation of the Earth. The histogram of events versus sidereal time reveals a substantial anisotropy (of order of a factor of two) that peaks when the right ascension of the galactic center crosses the meridian at the mid-point between Maryland and Argonne, and again 12 hours later (as would be expected from the symmetry of the antenna pattern and the transparency of the Earth to gravitational waves). A similar histogram plotting coincidences against the solar time of their occurrence is flat within statistical errors.

One disturbing consequence follows from this observation, if one adopts the natural hypothesis that the sidereal anisotropy indicates that gravitational wave events are originating in the galactic center. Making other reasonable assumptions about the bandwidth of the signals, Weber is led to the conclusion that perhaps as much as 1000 $M_\odot$ per year is being converted into gravitational waves. It is not clear what to make of the inference that the galaxy would be entirely consumed in much less than a Hubble time at such a rate of gravitational luminosity. Weber mentions as possible ways out anisotropic emission from the sources (or focusing of the waves) in such a way that we see higher than average flux, or the possibility that his detector is not working in a linear regime, but is instead being stimulated by very weak gravitational waves to release energy stored in metastable states.

5.3 Early Response to Weber's Claims

Weber's claims attracted a great deal of attention. The experimental evidence appeared very strong. The problems with the energetics of the events could be taken as a strong counter-argument, but they could also be interpreted as a demonstration that a truly remarkable discovery was being made. The review "Gravitational-Wave Astronomy" by Press and Thorne in 1972's *Annual Reviews of Astronomy and Astrophysics*[39] gives testimony to the significance that was attributed to Weber's claims, with its extensive bibliography (completed in December 1971) of papers grappling both with the theoretical consequences of Weber's claims and with experimental ways to improve upon the sensitivity of Weber's detectors. The perceived stakes in the controversy are shown by the three-paragraph introductory section, which is mostly devoted to balancing the possibility of breakthroughs in the making against the alternative that Weber's results are mistaken. This first section concludes with a paragraph, remarkable in a review article, questioning the very urgency of its subject:

> We (the authors) find Weber's experimental evidence for gravitational waves fairly convincing. But we also recognize that there are as yet no plausible theoretical explanations of the waves' source and observed strength. Thus, we feel we must protect this review against being made irrelevant by a possible "disproof" of Weber's results. We have done this by relegating to the end of the article (Section 6) all ideas, issues, and discussions that hinge upon Weber's observations.

Both the theoretical speculation about sources and the quest for better ways to detect the waves are in evidence in one of the most influential papers of this or any other

epoch in the search for gravitational waves. Gibbons and Hawking, both renowned for work rather far removed from experimental physics, wrote in late 1970 the very practical "Theory of the Detection of Short Bursts of Gravitational Radiation."[25] The introduction presents Weber's results as established facts. Section II of the paper discusses possible sources, discussing with equanimity but at some length the extreme gravitational luminosities required. The rest of the paper is devoted to the theory of detecting weak gravitational wave bursts in the presence of noise, and to a clear and original set of proposals for maximizing the signal-to-noise ratio of detectors. Some of one's surprise at the authorship of this paper may be alleviated upon reading the thanks for insights and ideas given (at two places in the text and in the Acknowledgment) to P. Aplin of Bristol University, a very original experimentalist who published little on this subject under his own name.[40]

Gibbons and Hawking pointed out that Weber's own treatment of the theory of gravitational wave detectors, written before he started his observations, had been aimed at their response to steady sinusoidal signals. As such, it gives a misleading idea of the value of the high-Q resonance that characterizes Weber-style detectors. This lack was remedied by Gibbons and Hawking, who (with thanks to Aplin) point out that the low dissipation of a high-Q system means that the level of thermal noise power is low. They go on to show that the high-Q resonance is also of crucial importance in minimizing the effect of Johnson noise in the transducer, here playing the role we would generally call amplifier noise. (A weakness of this light treatment of amplifier noise is that it leaves out the back action effects that enforce the Uncertainty Principle.)

As we saw previously, Gibbons and Hawking noted that there is in fact a competition between these two benefits of low dissipation. To minimize thermal noise, one wants to integrate the output for as short a time as possible, to give the random walk of the resonator's complex amplitude the least opportunity to mask a signal. On the other hand, the importance of transducer/amplifier noise is minimized by integrating for as long as practicable, so that the gravitational wave signal competes with as small a bandwidth of the broadband noise as possible. Gibbons and Hawking showed how to derive the optimum averaging time that minimizes the total noise from these two sources. In so doing, they noted that the averaging time depends on the dimensionless coupling parameter they called β, the definition of which we gave above in Section 4. For Weber's detector they give an estimate of $\beta \approx 5 \times 10^{-6}$.

Gibbons and Hawking go on to note that a large value for β would have two benefits: improving the signal-to-noise ratio by making the gravitational wave signal appear as a larger electrical signal, while simultaneously changing the balance between thermal noise and electrical noise in the direction that causes the optimum sensitivity to be obtained with shorter integration times. In other words, a larger β would yield better sensitivity and high bandwidth. To obtain these benefits, they discuss a novel configuration, proposed by Aplin, that has come to be known as a "split bar." It consists of two large masses (two "ends" of a bar split in half) joined to each other by connection to either face of a layer of piezoelectric material. The benefit comes from the fact that, in this configuration, the piezo is actually functioning as the dominant spring in the system; by storing the bulk of the elastic energy, it is able to produce a larger amount of electrical energy. (In many ways this harks back to Weber's original proposal to make the bar entirely of piezoelectric material.)

Gibbons and Hawking sketch the details of a detector of this sort, in which lead zirconate titanate (PZT) is substituted for crystalline quartz (used by Weber) because of its larger piezoelectric coupling constant. Then, in spite of the fact that the thermal noise power is increased because the piezo is a rather lossy spring, the sensitivity should be increased by more than a factor of ten in energy compared with Weber's detector. At the same time, the optimum sampling time is shortened to 1 msec, so that more detailed information can be extracted from the signal.

In passing, Gibbons and Hawking also note that Weber uses a less than optimal way of searching for gravitational wave events. His definition of an event is a noticeable increase in the energy in the bar's fundamental mode. But a gravitational wave impulse will only increase the energy if it arrives with a particular phase relationship to the bar's previous excitation. If the wave arrives with a different phase, the bar's energy may be decreased, or the effect may instead primarily change the phase of the bar's vibration. They estimate that this means Weber saw only about 1/4 of the events exciting a given bar. And, since the two bars being used for a coincidence have independent phases, only $(1/4)^2 = 1/16$ of the detectable coincidences would have been registered by Weber's technique. This makes the question of the source of the gravitational luminosity that much more difficult to resolve. But it also means that, if Weber's results were real, even more events should be detectable.

Whatever mysteries there may have been regarding Weber's claims that he was detecting pulses of gravitational waves, if they were true they represented one of the most important astronomical and physical discoveries of the 20th century. So it is no surprise that a number of other workers decided to construct gravitational wave detectors. And quite naturally given Weber's apparent success, most of these detectors were built quite

a bit like Weber's. Success looked as if it should come quickly. Although it had taken Weber about a decade, working alone, to design, build, debug, and operate his detectors, those who wanted to follow the recipe he had developed needed much less time. As Weber himself described it:[41]

> The scheme that works is not very difficult to set up and costs about $25,000, excluding pay for the senior physicists involved. Once set up, the instrumentation has run without problems for a little longer than a year and a half. Furthermore, the instrumentation requires a skilled technician about a month to construct. It requires a physicist about two days, starting with an open vacuum chamber to put crystals on the apparatus and make the necessary adjustments to see coincidences. In view of the relative simplicity of the working instrumentation, it seems reasonable to suggest that others start in this way.

A group at Frascati[42] built a detector as close to Weber's as possible, including a close match with his resonant frequency of 1660 Hz. The Munich group of Billing and Winkler[43] also built a quite similar bar. At Moscow State University, a group led by Braginsky constructed another bar that matched Weber's rather closely. At Bell Labs, Tyson[44] began a program of construction along Weber's lines, with less concern to match all of his parameters exactly—the resonant frequency of his largest bar was 709 Hz. At IBM Labs, Garwin and Levine[45] built a small bar of Weber's style; they were convinced that they could improve upon Weber's statistical methods sufficiently to make up for poorer noise levels. Aplin,[40] on the other hand, began to construct a split bar along the lines he had suggested to Gibbons and Hawking. Although he did not succeed in carrying that work through to completion, the idea (and some of the equipment) was taken up by Drever's group in Glasgow.[34] There were also efforts started at Reading and Regina, and in addition, the second generation detectors of much greater sensitivity were planned by groups at Stanford, LSU, and Rome. (More about these later.)

As Weber had predicted, results began to become available very quickly. Unfortunately, it also soon became apparent that no one else was seeing results that looked anything like Weber's. By the time of the 6th Texas Symposium on Relativistic Astrophysics (held in New York December 18-22 1972),[41] several groups were ready to come forward with claims of null results at sensitivities comparable to or better than Weber's. Tyson made the strongest claims for the unreality of Weber's signals, based both on his own failure to find them in his single detector (at sensitivities better than Weber's) as well as his claims that Weber's electronics were too noisy even to record Brownian motion in his bar. Kafka, representing the Frascati and Munich groups (who performed observations in coincidence), was able to report only preliminary results (delivered to him by telegram) of five days of observations, but the failure to find any statistically significant number of coincidences was in strong contradiction to their expectations if Weber's claims were correct. Drever also reported negative results from the high-bandwidth split bar at Glasgow; he, however, tempered his claims of conflict with Weber by noting that his detector would only be expected to see Weber's events if they were of a few milliseconds' duration (an assumption that was plausible but by no means certain).

5.4 Resolution of the Weber Controversy

Over the next year and a half, there was an intensive concentration of effort to attempt to resolve the controversy. The groups listed above continued to accumulate data, none of which confirmed Weber's claim of a substantial flux of strong gravitational-wave pulses. New negative results from Garwin and Levine of IBM Labs were reported in a strongly-worded series of articles in *Physical Review Letters*.[45] Their work was carried out in the explicit belief that Weber's results were spurious, and could be shown to be so by the use of a single detector of well-characterized performance, even if that detector was small. (The titles of the three articles are miniature masterpieces of polemical style: "Absence of Gravity-Wave Signals in a Bar at 1695 Hz," "Single Gravity-Wave Detector Results Contrasted with Previous Coincidence Detections," and "New Negative Result for Gravitational Wave Detection, and Comparison with Reported Detection.") The first results to be reported came from a bar with a mass of 118 kg, the second from a 500 kg bar. Compare with the mass of Weber's principal detectors, 1500 kg.

The letters of the IBM group present the statistics of the noise output of their detectors, along with a careful tutorial on the theory of extracting impulsive signals from noise. This latter theme can be thought of as an explication of the signal processing scheme proposed (in rather telegraphic style) in Gibbons and Hawking's 1971 paper. The key idea was not to rely on net increase in the energy of the bar, but to look for sudden changes, of either sign, in either the magnitude or phase of the detector's complex excitation. The thrust of the papers is that if Weber's results were correct, then even their less massive single detector ought to show substantial departures from the

Gaussian distribution of excitation expected on the basis of thermal and amplifier noise alone. The near-perfect Gaussian fit to their data then constitutes an apparent contradiction of Weber's results.

An important argument used to buttress the claim that the IBM detector was well understood was the application of electrostatic calibration forces to one end of the bar, and the successful detection of those events (within the statistical limits set by the signal-to-noise ratio) by the data processing system. Tyson's 1972 presentation to the Texas meeting had previously emphasized the importance of this fundamental practice of experimental physics, as did the remarks of both Kafka and Drever. Levine and Garwin take Weber to task for having failed to use any calibration method, either as a check of his instruments' front ends or of his data analysis procedure.

A more subtle implicit argument against Weber's work is suggested to the reader of these papers by their admirable clarity, as contrasted with the rather Delphic pronouncements that fill Weber's own contributions to *Physical Review Letters*. Levine and Garwin make this explicit at one point when they compare their results with their best guess at how Weber's would be expressed in similar (sensible) units, complaining "We are thus forced to estimate these quantities, while noting that such information is easily obtained by the experimenter and is normally provided in the publication of a detection experiment."[46]

Garwin led a crusade against Weber's claims at the Fifth Cambridge Conference on Relativity (CCR-5), held at MIT on 10 June 1974.[47] Among the topics discussed was 1) an error in the computer program used by Weber to identify coincidences, shown to generate nearly all of the coincidences in the one data tape shared by Weber with other researchers, and 2) the puzzling feature of Weber's histogram of coincidences versus time delay showing a peak at zero delay in only the central 0.1 second wide bin, in spite of the fact that a 1.6 Hz wide bandpass filter was said to be part of the signal processing chain. But the most spectacular event of the discussion was what even those sympathetic to Garwin's cause might have felt was a trick that bordered on unsportsmanlike conduct. Weber had been given data from the detector of Douglass's group at the University of Rochester, to search for excitations in coincidence with Weber's own detectors; Weber reported at previous meetings that he had detected an excess of coincident events at a level of 2.6 standard deviations above the expected chance rate. According to Garwin's account in a letter to the editor of *Physics Today*,[47] "At CCR-5 Douglass revealed, and Weber agreed, that the Maryland Group had mistakenly assumed that the two antennas used the same time reference, whereas one was on Eastern Daylight Time and the other on Greenwich Mean Time." No stronger way can be imagined of impressing the community with the possibility that Weber was able, by some means, to find coincidences among any two data streams, whether the coincidences actually existed or not.

A panel discussion with almost precisely the same cast of characters as that of the 1972 Texas Symposium was staged at the 7th International Conference on General Relativity and Gravitation in Tel Aviv, June 23-28, 1974.[48] The plot, Weber's lonely claims of detections contradicted by the null results of the others, was also unchanged—the only substantial difference is that Weber's critics had had time to carry out more extensive searches and more careful data analysis. By this time, the Bell Labs group had carried out a coincidence run with an identical bar at the University of Rochester, operated by Douglass. The Munich group (which had by then incorporated the previously independent Frascati group) reported on the results of 150 days of coincident observations. Drever gave a report of a more extensive data run, seven months that had concluded in April 1973, yielding only one candidate coincidence; although this event could not be ruled out as a possible gravitational wave detection, neither could it be positively established as such in spite of the low probability that was estimated for it to have occurred by chance. (The detectors were only 50 m apart, and so may have both been driven by some other kind of influence.) In any event, Drever was able to show that the Glasgow experiment did not show the sort of event rate predicted by Weber's experiment, except under rather implausible assumptions about the nature of the individual gravitational wave pulses. Tyson also briefly reported on the negative results from Garwin and from Braginsky.

How did the physics community deal with these contradictory results? This is an almost classic example of attempted replication of an important claim, but with both opposing camps standing firm in their beliefs that their own results were correct. Valuable insight into the difficulties this situation posed to scientists can be found in the work of sociologist of science Harry Collins, who interviewed many of the principal actors during this period. His results are well worth consulting, even though the quotes from the interviews are reported without identifying the individual speakers.[49]

The other key resource in the written record is the transcripts of the open discussions at the 1972 and 1974 panels. Both Kafka and Tyson point out strongly that Weber (usually) uses a far-from-optimal statistical method to look for signals. Tyson also comes close to accusing Weber of fraud; the method by which Weber has deluded himself and others is said to be continual "tuning" of the statistics used to search for

coincidences, with choice of algorithm and threshold being chosen for each data set in such a way as to maximize the apparent statistical significance. (A similar suggestion was also made by Levine and Garwin.[47]) Weber denies this, but some of his remarks appear to support the accusation by using *ex post facto* reasoning to justify particular choices of pulse-detection algorithms: his test of which method is optimal is which gives a larger number of coincidences, and the choice can vary from one data set to another.

For his part, Weber appears to have believed that the results that contradicted his own did not constitute fair tests of his work, since they were not carried out in an identical way. This may strike some readers as disingenuous, since several of the other detectors were quite close copies indeed of Weber's apparatus. But it appears to be an honest reflection of Weber's belief that he was in fact detecting gravitational waves, and if others couldn't see them there must be something subtly wrong with their detectors. To this day, he continues to claim ongoing detections of gravitational wave pulses with his apparatus.

5.5 Beyond the Weber Era

One of the most interesting features of the Panel Discussion at GR7 was the time spent predicting the future of the field. Both Tyson and Drever gave optimistic predictions for future progress; both have been borne out faithfully, although at slower rates than either would have hoped to see. Tyson described the efforts already begun in 1974 (by groups at Stanford, Louisiana State University, and the University of Rome) to construct Weber-style detectors that were cooled by liquid helium to temperatures of a few degrees Kelvin. The obvious benefit, to reduce the Brownian motion noise (with power proportional to $k_B T$), is important. So, too, is the less obvious benefit that much quieter amplifiers are available at low temperature—these are the Superconducting QUantum Interference Devices, or SQUIDs.

Drever's prediction for future progress focused primarily on inteferometry, the other direction that subsequent history has in fact proven fruitful. It is a brief sketch, without any details or attribution, but still remarkably prescient. Drever starts from the premise that Weber's claimed detection of gravitational wave signals is most likely wrong. Then, he says, the point is no longer to try to verify or extend Weber's results, but instead to ask from first principles what might be the best way to probe the possible existence of gravitational waves, over the widest possible range of properties. Not only is improved sensitivity important, but the ability to look for a variety of signal types over a broad range of frequencies. Drever's own belief in this strategy can be inferred from his having adopted the wide-bandwidth Aplin-style split bar, and from having used the Glasgow detectors not only to search for brief pulses but also for a stochastic background of gravitational waves.

But it was difficult to push this strategy much farther, since the thermal noise in this kind of detector was very large; the losses came predominantly from the piezoelectric material that here serves not only as a transducer but as the primary "spring" for the resonant system, and hopes for reducing them would depend on a program of materials research with uncertain prospects at best. (Note that in the original Gibbons and Hawking paper, the split bar is argued for as the quickest way to give a substantial step in sensitivity, with the explicit assumption that since Weber was already detecting gravitational waves in a sub-optimal way, any sizable increment in sensitivity would immediately yield important dividends.) Instead, Drever recognized the kindred broadband sensitivity inherent in the proposal to build interferometric detectors and, with the prospect of placing the test masses very far apart, the possibility of substantially improved sensitivity as well.

Drever expanded on the themes mentioned briefly at GR7 in a graceful review first given as a talk to the Royal Astronomical Society in 1976, published as an article in 1977.[50] By 1978, when *Annual Reviews of Astronomy and Astrophysics* published its second review of the subject "Gravitational-Wave Astronomy," the authors Tyson and Giffard could write a much more mature piece than could Press and Thorne in 1972.[51] The style is more formal and less heuristic. There is less discussion of rather far-fetched possible sources of gravitational waves, but in its place a much more complete treatment of the theory of extracting weak impulsive signals from the output of a resonant detector. After a summary of the history of the work of Weber and of those who followed him, the authors write

> It must be concluded that the interpretation of the Weber events as gravitational wave pulses is erroneous, since there is no corroborated evidence to date either for an excess number of coincident events or any sidereal correlation.

Similar thoughts, more precisely focused, were expressed in a 1978 paper (written too late to be mentioned by Tyson and Giffard) by Kafka and Schnupp, giving the "Final Result of the Munich-Frascati Gravitational Radiation Experiment."[52] They write that

"Although the non-existence [of Weber's pulses] became obvious a long time ago, it still seems appropriate to publish our final negative result, because our experiment was as similar to Weber's as possible, whereas all other coincidence experiments deviated in one way or the other.... Moreover, we think we have set the lowest upper limits obtained by Weber-type experiments over a reasonable long period of observation," spanning 580 useful days of common observations of the two detectors. The main result of this paper is the null result that the statistics of the coincident excitation of the two detectors was just what would be expected from the laws of chance, given the noise levels in the detectors. Without mentioning Weber specifically, Kafka and Schnupp do remark that

> Scanning our whole data, we could, of course, find periods of a few days, for which at some pair of thresholds the number of coincidences was up to more than 3 standard deviations higher than the average over the various time delays. However, the same was true for arbitrary delays, and zero delay did not seem to be distinguished in any obvious way. However, one should not forget: If one searches long enough in our finite sample of data, one must find some complicated property which distinguishes zero delay signficantly from the others. (Again this is true for an arbitrary delay, but with a different property.)

The paper goes on to pay special attention to two periods, totaling 67 days in length, when the operation of the Munich-Frascati experiment overlapped with times for which the Weber group claimed to have detected substantial rates of coincidences with its own detectors. The authors write: "These results do not give the slightest hint of a simultaneous influence on both detectors. If the significant observations reported by Weber's group for these two periods had been due to gravitational radiation of any kind, they should have shown even more significantly in our experiment." The mention of "any kind" of signals refers to the fact that the present authors used not only the vector-difference algorithm that is optimal for short pulses, but also used for these 67 days the algorithm preferred by Weber, which would be more sensitive for very long wave trains that gradually excited the antenna. Kafka and Schnupp conclude this section by remarking that "we do not have an explanation for Weber's observations," although they suggest the possibility that there might have been some undiagnosed electrical feedback from signals on the telephone line from Argonne into the Maryland bar itself.

The final section of the paper compares the likely strengths and rates of gravitational wave signals from core collapse in supernovae with the then current and possible future sensitivities of gravitational wave detectors. In a dramatic figure, they superpose a model of the rate of supernovae at various distances from the Earth on the natural phase space for gravitational wave searches, event rate versus event strength. The authors point out that, even if one were able to improve the performance of gravitational wave detectors of the Weber type to the limit set by the Uncertainty Priniciple (by cooling, improving Q, or whatever other trick), one would still not have the sensitivity to detect events at the rate of several per year or greater. They conclude, "Because of the difficulties arising from this problem and because one would certainly like to measure more details than just the spectral density of pulses, the Munich group decided not to continue with (low temperature/high quality) Weber-type experiments, but rather with a Weiss-Forward type experiment, i.e., a laser-lighted Michelson interferometer."

In spite of the considerations that moved the Munich group to abandon resonant-mass detectors, the groups that had decided in the early 1970s to build cryogenic versions of the Weber bar pushed ahead. A number of strong reasons can be given to justify this strategy, including the dubious value of relying (as the Munich group did) on signal-strength predictions which necessarily must be ignorant of truly novel astronomical phenomena, as well as the belief that evolutionary development is often a more rapid and reliable strategy for progress than a revolutionary approach. And, although progress was slower than the hopes expressed for it in Tyson's 1974 remarks in Tel Aviv, this route did in fact lead to substantial increases in sensitivity well before the interferometric detectors began to catch up.

The first complete operating cryogenic resonant-mass detector was the one built at Stanford University by the group led by William Fairbank.[53] In addition to the obvious reduction of thermal noise by cooling with liquid helium to a temperature of 4.3 K, and the use of the Josephson junction SQUID as a low noise preamp, there was another technical innovation that helped the Stanford bar reach a new level of sensitivity. This was the introduction by Paik[26] of a resonant transducer, tuned to the same frequency as the bar's resonance, mounted on the end of the bar. Both Tyson and Garwin had used end-mounted transducers, but neither realized the advantages that would accrue to the tuned configuration—the ability to better "impedance-match" the mechanical excitation of the bar to the electrical system, thus increasing the coupling parameter β. (See the discussion above.) The Paik transducer represented a new generation in another sense—it made no use of the piezoelectric effect, but instead used the motion of the resonant proof mass in the transducer to modulate the value of an inductance in

the circuit that carried the persistent current through the Josephson junction.

As the first detector to operate at this new level of sensitivity, no coincidence observations were possible, so the first paper contains only the results of single-detector operation.[53] (The group at LSU, led by former Stanfordian W. O. Hamilton, was collaborating with the Stanford group, but was developing several subsystems independently, including an alternative transducer design.) The 1982 paper in *Astrophysical Journal Letters* included a histogram of excitations of the Stanford detector, showing that even without coincidences it set a new upper limit that, at least at large event rates (greater than of order 10^{-2} per day), was orders of magnitude more stringent than that achieved by the Munich group. An even better result was predicted once a second comparable detector would operate in coincidence.

The noise temperature of 20 mK was certainly a milestone, but doesn't entirely characterize the performance of the bar as a detector of rare events. The histogram of excitations recorded during the 74 days of data presented in the paper show a substantial excess of large excitations, beyond what would be expected from purely Gaussian statistics. The cause was unclear, but one possible explanation was acoustic emission within the bar from sudden stress relaxations. Without coincident operation with a comparable detector, gravitational wave pulses could not be ruled out either.

It wasn't until 1986 that coincident operation of all three cryogenic detectors started in the 1970s (Stanford, LSU, and Rome) was finally achieved. The results of that run were less than spectacular, because none of the detectors was operating as well as it might. Indeed, the event rate-event strength plot that characterized this coincidence run was hardly better than the single-detector plot from the Stanford 1981 solo paper.[24]

The problems with the LSU and Rome detectors were soon sorted out; the rms noise levels in these detectors have now fallen below 10^{-18}. Meanwhile, the Stanford bar met an untimely demise when it was irreparably damaged in the Loma Prieta earthquake of 1989. The LSU-Rome axis has carried out coincident observations since 1991. A first look gave a new upper limit on the flux of gravitational wave pulses that was presented in an otherwise unpublished conference report.[54] Another cryogenic detector that had been under development for some time, the niobium bar of Blair's group in Perth, recently achieved a comparable sensitivity. Coincidence observations continue as this review is being written.

The 1986 coincidence run was a consummation of years of work, whose occurence had been devoutly wished for by many friends of the field for some time. Aside from the fact that it may have been somewhat too late (from the point of view of Stanford,

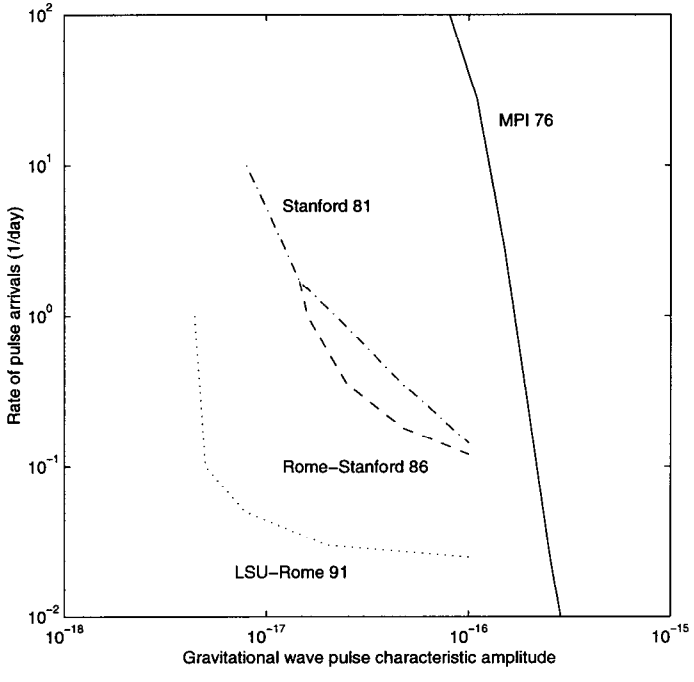


Fig. 3. Upper limits set by various experiments on the rate of gravitational wave impulse arrivals, as a function of the characteristic amplitude of the impulse.

whose detector wasn't working as well as in 1981) or somewhat too early (from the point of view of LSU or Rome, who had not yet shaken all of the bugs out of their systems), there was another accidental side effect of its timing. That was the complete coordination of the time when all three systems went off the air to fix the problems that running together had made obvious. This is the reason that none of the state-of-the-art detectors was on the air on February 23, 1987, when Supernova 1987A appeared. The closest observed supernova in centuries was a chance no one would have chosen to miss, although in fairness at a distance of order 50 kpc it is unlikely, according to standard estimates of the gravitational luminosity, that it would have been seen.

There were, however, non-state-of-the-art gravitational wave detectors observing at the time. Weber has kept a room temperature bar in operation nearly continuously since the '70s, as has the Rome group. An unusual chain of reasoning was constructed, involving a suspect time for the supernova collapse, an unorthodox signature for the gravitational wave event, *ad hoc* assumptions about neutrino physics, and tremendous gravitational luminosity, but leading to a claim of significant detections of a large flux of gravitational wave pulses.[55] This claim has attracted much less attention than did the original claims of Weber in the early 1970s. A serious effort has been made to demonstrate that the statistical significance of the analysis has been overstated, due to construction of the signature to match the data stream.[56]

6 History of Interferometers

6.1 The Work of Gertsenshtein and Pustovoit

Almost as soon as Weber had begun work on the first gravitational wave detector of the resonant-mass style, the idea arose to use interferometry to sense the motions induced by a gravitational wave. Weber and a student, Robert Forward, considered the idea in 1964.[10] We will discuss below how Forward later went about implementing the idea. But the first discussion of the idea is actually due to two Soviet physicists, M. E. Gertsenshtein and V. I. Pustovoit. They wrote in 1962[57] a criticism of Weber's 1960 *Physical Review* article, claiming (incorrectly) that resonant gravitational wave detectors would be very insensitive. Then, they make a remarkable statement justified only by intuition, that "Since the reception of gravitational waves is a relativistic effect, one should expect that the use of an ultrarelativistic body—light—can lead to a more effective indication of the field of the gravitational wave."

Gertsenshtein and Pustovoit followed up this imaginative leap by noting that a Michelson interferometer has the appropriate symmetry to be sensitive to the strain pattern produced by gravitational waves. They give a simple and clear derivation of the arm length difference caused by a wave of amplitude h. Next, they note that L. L. Bernstein had with ordinary light measured a path length differences of 10^{-11} cm in a 1 sec integration time. The newly invented laser, they claim, would "make it possible to decrease this factor by at least three orders of magnitude." (The concept of shot noise never appears explicitly here, so it is not clear what power levels are being anticipated.) They assume that one might make an interferometer with arm length of 10 m, thus leading to a sensitivity estimate of $10^{-14}/\sqrt{\text{Hz}}$ for "ordinary" light, or as good as $10^{-17}/\sqrt{\text{Hz}}$ for a laser-illuminated interferometer. This, Gertsenshtein and Pustovoit claim, is 10^7 to 10^{10} times better (it isn't clear whether they mean in amplitude or in power) than what would be possible with a Weber-style detector. Putting aside their unjustified pessimism about resonant-mass detectors, their arguments about interferometric sensing are right on the mark, even conservative.

For improvements beyond the quoted level, they make suggestions that are somewhat misguided. They say that observation time could be lengthened beyond 1 sec, which would be obvious for some sources (such as "monochromatic sinusoidal signals" or signals of long period) and hopeless for short bursts. Their other suggestion is to use "known methods for the separation of a weak signal from the noise background"; this suggestion is curious because known methods appear to be already built into their estimates that are referenced to a specific observing time. The other lack that is obvious in hindsight is any mention of mechanical noise sources. Still, the gist of the idea of interferometric detection of gravitational waves is clearly present, as is a demonstration that the idea can have interesting sensitivity.

6.2 The Origins of Today's Interferometric Detectors

For a variety of reasons, not least of which must have been the fact that it was written too early (before Weber's work had progressed beyond design studies), the proposal of Gertsenshtein and Pustovoit had little influence. The activity that began the by-now flourishing field of interferometric gravitational wave detection started independently in the West. In fact, it began semi-independently at several places in the United States at around the same time. The roots of this work can be seen in a pair of papers, written in 1971-2, by two teams linked in an unusual collaboration that is acknowledged in

the bodies of the papers, although not in the author lists. The first to be published was that of the Hughes Research Lab team, whose most committed member was Robert L. Forward, the former Weber student mentioned above. Later to appear, and not in a refereed journal, was the work of Rainer Weiss, an MIT physicist who had spent an influential postdoctoral stint with Robert H. Dicke at Princeton. Linking the two groups was someone who never published anything on the subject under his own name, but whose activity is mentioned in both papers—Philip K. Chapman, who had earned a doctorate in Instrumentation at MIT's Department of Aeronautics and Astronautics before joining NASA as a scientist-astronaut.

6.2.1 Interferometer Studies at Hughes Research Lab

An account of the idea for an interferometric detector of gravitational waves, and of the performance of an early-model prototype, is found in the 1971 paper of Moss, Miller, and Forward.[10] The authors cite a program to develop "long wideband" gravitational wave detectors that had started at Hughes in 1966, around the time of Weber's first account of a working resonant detector. The motivations for a wideband detector were 1) to allow detailed meaurement of waveforms which would in turn give insight into the nature of the sources, 2) "to allow the phasing of spaced antennas to form a phased array" (in other words, to allow good temporal resolution so that the direction of the wave can be determined by arrival time differences), and 3) to allow matched filters to be used to optimize the signal to noise ratio of a complex waveform "in addition to the use of standard narrowband frequency filtering for sinusoidal signals, which is the natural filtering action of a resonant antenna." The motivation for the use of a long detector is the larger test mass displacement, which, all else being equal, should directly translate into improved signal to noise ratio.

The authors credit the original idea for this way to achieve a long wideband detector to P. K. Chapman, and go on to state that "Our work has benefited from many discussions with Dr. Chapman as well as R. Weiss, who is involved in the design and construction of his own design at MIT."

The interferometer described by Moss *et al.* was "constructed to set experimental limits on the various noise sources in the laser transducer." It is a classic one-bounce Michelson interferometer, in which both output beams are detected "in a balanced bridge to reduce sensitivity to laser amplitude noise." The operating point for this arrangement was equal photocurrents from the photodetectors at the two output ports of the interferometer. (This is to be distinguished from the use of a photodetector at a single output port of the interferometer that is dithered about the dark fringe. See our discussion of this alternative below.) The interferometer is to be held at this balanced operating point by "slowly acting servo loops... so that the effects of laser amplitude and phase noise are minimized." The flat mirrors were rigidly mounted to an optical table, which was in turn supported on air-filled rubber tubes to give a resonant frequency of 2 Hz. (Other isolation schemes for the rigid interferometer were tried without success, causing the authors to lament that "vibration isolation is still an art rather than a science.") One of the mirrors was mounted on piezoelectric elements, which provided control of the operating point as well as a means of calibration.

The main result presented in the paper is the noise level that was achieved in this prototype interferometer, equivalent to a mirror displacement sensitivity of 1.3×10^{-14} $\mathrm{m}/\sqrt{\mathrm{Hz}}$ at a signal frequency of 5 kHz. This was about a factor of $\sqrt{2}$ larger than the calculated shot noise sensitivity, which the authors state is consistent with other indications that "the sensitivity limits were set by acoustic and ground noise." This noise level was "to date... the smallest vibrational displacement directly measured with a laser." No translation of the sensitivity into strain units was given, either directly or by specification of the arm length. This is perhaps appropriate, since the rigid mirror mounts made this test instrument ill-suited for actually searching for gravitational waves.

The last section of the paper lists the improvements intended to follow this initial work. Firstly, a more powerful laser was proposed; at the then impressive power level of 75 mW, the displacement sensitivity due to shot noise would be "close to that obtained in the present resonant antennas" that were at the date of writing appearing to give significant detections. A final paragraph listed the other proposed improvements: mirrors attached to masses that were large (to reduce thermal noise) and independently suspended from vibration isolation mounts, placed in a vacuum system whose initial length of several meters could be extended "to several kilometers by adding additional evacuated tubes." This section can be read as a telegraphic summary of the plans described at greater length by Weiss in his report written the following year.

6.2.2 The Vision of Rainer Weiss

The other paper that gave birth to the massive worldwide effort to detect gravitational waves using interferometers, Rainer Weiss' 1972 "Electromagnetically Coupled Broadband Gravitational Antenna," appeared only as an unpublished research progress report

of the organization at MIT that administered the umbrella research grant supporting his work.[11] Weber's claimed detection of gravitational waves was very much on Weiss' mind in 1972, reported as possibly correct but with the recognition that the energy flux the waves appeared to carry would dominate the luminosity of the Galaxy. Weiss states that he had been inspired by a 1956 paper by F. A. E. Pirani (that discussed the identification of measurable quantities in general relativity)[9] to consider the possibility that measurements of the light travel time between freely-falling test masses would make the best probes of spacetime structure. He further states that he had realized several years prior to writing (while teaching an undergraduate seminar) that the newly developed lasers could turn Pirani's *gedanken* experiment into a practical measurement strategy. Weiss also notes that the idea "has been independently discovered by Dr. Philip Chapman of the National Aeronautics and Space Administration, Houston."

Many of the ideas that appear in the breathless final paragraph of Moss *et al.* are elaborated at substantially greater length in Weiss' report, which should be considered the first serious design study of the concept of interferometric gravitational wave detection. After the review of Weber's claims, Weiss continues with a clear summary of the physical meaning of gravitational waves in general relativity, and an examination of the possible strength of gravitational waves from the then newly discovered pulsars. He then gives a summary of the key ideas of the proposed system:

- a Michelson interferometer used as a sensor of "differential strain induced in the arms,"
- operated "on a fixed fringe by a servo system" in a modulated system very much in the tradition of Dicke's improved Eötvös experiment,[58]
- "mirrors and beam splitter mounted on horizontal seismometer suspensions" that "must have resonant frequencies far below the frequencies in the gravitational wave" and "a high Q,"
- arms that "can be made as large as is consistent with the condition that the travel time of light in the arm is less than one-half the period of the gravitational wave," in part by being arranged as "optical delay lines"of the style described by Herriott.

Weiss is quite clear about the advantage that accrues from the last point. He says

> This points out the principal feature of electromagnetically coupled antennas relative to acoustically coupled ones such as bars; that an electromagnetic antenna can be longer than its acoustic counterpart in the ratio of the speed of light to the speed of sound in materials, a factor of 10^5. Since it is not the strain but rather the differential displacement that is measured in these gravitational antennas, the proposed antenna can offer a distinct advantage in sensitivity relative to detecting both broadband and single-frequency gravitational radiation. A significant improvement in thermal noise can also be realized.

This last sentence points out one of the key insights of this report, expanded upon at much greater length in the remainder of the text. As a sensitive mechanical measurement, the interferometric detection of gravitational waves is prey to a host of mechanical noise sources whose strengths need to be minimized if success is to be achieved. By far the largest section of the paper is devoted to estimates of the magnitudes of a long list of noise sources of various kinds. They include: amplitude noise in the laser (the only place where the work of the Hughes group is cited, as an example of a shot noise limited measurement), phase noise in the laser, mechanical thermal noise, radiation pressure noise, seismic noise, thermal gradient ("radiometer effect") noise, cosmic ray impacts, "gravitational-gradient" noise, and fluctuating forces from electric and magnetic fields. This looks almost (with a few omissions) like the list of noise sources that contemporary workers are grappling with as they strive to make the new kilometer scale interferometers work; by contrast, the other earlier treatments of the subject look myopic and unbalanced. And this insight is what led to the recognition that interferometers of the greatest practical length, with the resulting dilution of displacement noise terms as compared with a strain signal, would be the way to achieve the promise of good gravitational wave sensitivity, and would be worth the substantial investments needed to build them.

6.3 Interferometer as an Active Null Instrument

The agreement between Weiss and the Hughes group on the basic features of an interferometric detector must have something to do with the fact that they and Chapman were engaged in a remote three-way collaboration. But the fact that the key features of the design remain current to this day (with a few important additions) is evidence that they responded thoughtfully to an inherent logic of experimental design. Interferometric gravitational wave detection represents an extreme example of the application of design principles of wide validity in experimental physics. It is worthwhile to examine those principles here.

6.3.1 How to Maximize the Signal to Noise Ratio

Firstly, it is important to recognize where, in the spectrum of physics experiments, this one falls. To successfully detect gravitational waves, it will be necessary to attain AC strain sensitivity that is completely unprecedented. The concept of sensitivity is the key one—the signal to noise ratio for small strains is (essentially) the only figure of merit of interest here. Precision is primary, while accuracy of calibration or absence of other kinds of systematic errors definitely takes a back seat at this stage in the development of the field. Also, it is important to remember that one is perfectly content to measure signals only above some cut-off frequency; no absolute measurements of lengths or even of differences in length between two interferometer arms are necessary.

In experiments where precision is the primary desideratum, certain general principles apply. In particular, there are two fundamental strategies for maximizing a signal to noise ratio: one can make the response of the system large, and one can make the noise small. To make the response large, one can try to "capture" the largest possible effect from an external source. Here we want to make the largest possible apparatus, since relative displacements of two masses caused by a gravitational wave are proportional to the separation of the masses. This strategy is quite common—large telescopes can be more sensitive than small telescopes, for example. Another example is the improvement Michelson and Morley achieved in 1887 over Michelson's solo result of 1881, achieved by increasing the expected ether drift fringe shift through the addition of mirrors to lengthen the interferometer's optical path length.[59]

Another way to make a large response is to arrange for the external influence to drive an indicator that changes dramatically in response to a small effect. The implementation of that strategy here is to illuminate the interferometer with visible light, whose wavelength of order 1 μm sets the scale for motions that change the output port from dark to bright. This strategy too is common, as in such simple choices as making the indicator needle on an ammeter as narrow as practicable, or constructing a galvanometer with a fine fiber so that it will swing as much as possible or in using the longest possible lever arm for its optical lever.

Not all of the measures listed above are guaranteed to improve the signal to noise ratio; that depends on the nature of the noise. But one always improves the signal to noise ratio by reducing the magnitude of the noise. The noise sources in mechanical experiments can be loosely grouped in two classes. One class is noise in the measurement of the response of the test system. For interferometers, this is the shot noise in the fringe readout, or whatever other effect swamps it (excess amplifier noise, for example). In resonant mass detectors this class is represented by noise in the preamplifier that responds to the transducer output, or by excess transducer noise such as Johnson noise. We saw above that there may be quantum mechanical limits to the reduction of this kind of noise.

The other class of noise sources are those that also affect the test system in a way that mimics the effect being sought. This includes all of the noise forces on the test masses, like seismic noise and thermal noise. (This is the class of noise effects usually called, when it wouldn't cause confusion with the principle of an interferometer's operation, "interference"). Over a large range of frequencies, these are the dominant limits to sensitivity of a gravitational wave interferometer. Where that is true, measurement strategies that maximize the strength of the external effect on the test system (such as, here, the physical separation of test masses) can be helpful, but those that just maximize the response of the measuring instrument to any external effect just magnify the response to the noise as well. Reducing the strength of these kinds of noise is always a good idea, whenever it is possible.

6.3.2 What to Do when $1/f$ Noise Dominates

The considerations listed above are mostly obvious enough to be considered common sense. But there are some subtle aspects to consider as well. The most important non-obvious fact to consider is that almost every noise source has a pronounced $1/f$ character to its spectrum, at least at low frequencies. A laser whose power stability is limited at high frequencies by fundamental processes like shot noise almost invariably will show dramatically enhanced levels of noise at low frequencies. And, in many situations even such a fundamental process as thermal noise shows a $1/f$ or steeper power spectrum, as we saw above. Furthermore, seismic noise almost invariably has a very steep displacement power spectrum.

The difficulties caused by $1/f$ noise can come in several different forms. Sometimes the difficulty is simply that the noise has a larger magnitude than the minimum level it might have (laser power noise in excess of shot noise, for example), or that the noise is just inconveniently large (such as seismic noise at any frequency!). A more subtle kind of difficulty is the large integrated magnitude of the noise at frequencies that are low compared with the signal frequency. Many measuring devices have only a limited dynamic range over which they respond linearly to external influences; it seems a shame

to waste it dealing with noise outside the band in which one expects to find signals.

Measures to deal with $1/f$ noise often dominate the design of a high-sensitivity experiment. No one explored these measures more systematically than did Robert H. Dicke, who was for a few years in the early 1960s Weiss's mentor. Dicke was moved to think deeply about these problems while working on the development of microwave radar at the Radiation Laboratory at MIT during World War II. Devices called radiometers, receivers that measure the total power emitted by a broadband (often thermal) source of microwaves, would have had a variety of uses, if they hadn't been rendered so insensitive by the large amount of $1/f$ noise in the RF preamps. Dicke invented what came to be called the "Dicke radiometer" specifically to solve this problem.[60] The heart of the scheme was a device to periodically (at 30 Hz in the original case) interrupt the flow of RF power from the antenna to the preamp, replacing it instead with a thermal source of radiation. At the "back end" of the instrument, the electronics were arranged to give a measure of the difference between the detected power from the antenna and from the reference load. The reason this defeats the $1/f$ noise in the amplifier is that the comparison between measure and reference is made rapidly enough that the preamp's output can't wander much in the interval. Or, described in the frequency domain, the signal has been translated from DC up to a high enough frequency that the preamp's noise is not dominated by excess noise with a $1/f$ character.

Dicke's invention had the immediate effect that microwave radiometry became practical even for sources with rather low antenna temperatures; this made a substantial contribution to the Rad Lab's mission[61] as well as to the practice of radio astronomy.[62] But the greatest impact came from Dicke's realization that this modulation technique would have broad applicability, wherever $1/f$ noise was a problem. This insight led him to invent the lock-in amplifier, a universal back end that can control the chopping of an experiment, calculate the difference in output between the "on" and "off" states, and average the result to further reduce the noise. By now, "lock-in amplification" (also referred to as "phase sensitive detection") has become a nearly universal practice in the fight against $1/f$ noise.

A second classic measurement carried out by Dicke illustrates further insights in the battle against $1/f$ noise. The test of the equivalence of inertial and gravitational mass carried out by Roll, Krotkov, and Dicke[58] is considered one of the great examples of a *null experiment*. As championed by Dicke,[63] this term refers to a measurement where an answer of zero carries tremendous meaning. Precise equivalence of inertial and gravitational mass (or in other words a zero value for their difference) means that gravity can be described by a metric theory.

Null experiments play a special role for experimentalists as well as theorists, because an instrument that reads zero is immune to many of the sorts of problems that plague non-null measurements. Among these are calibration drifts and limited dynamic range of an instrument (whether from noise or from non-linear response.) Of course, turning a theoretical zero into an idea for an instrument that yields a zero output takes deep insight. One could argue that the torsion balance used in the improved Eötvös experiment, whose motion would track the Sun's if the aluminum and gold masses on opposite sides had differing ratios of inertial to gravitational mass, is among the most elegant instruments ever invented.

Maintaining the integrity of a null measurement takes insight that goes beyond the design of the front end of the experiment. For example, it would be a good idea for the null position of the test masses to be arranged to correspond to a null response from the sensor. Then, one can ignore (to first order) fluctuations in the drive level of the sensor (such as the light power in the optical lever), since zero is still zero even if it is multiplied by, say, 1.01 instead of 1.00. There are a variety of ways to create a null output from an optical lever at one particular operating point. One way would be to use a matched pair of photodetectors, placed so that the light beam falls equally on each detector when the balance is at the null position; as the beam moves to follow the balance's motion, one photodetector receives more light while the other receives less, and a differential amplifier will reveal the motion. This method is essentially a DC technique.

Dicke's team implemented a clever variation that let them make use of the advantages of a lock-in amplifier. A narrow light beam fell on a single photodetector, after passing by a wire of comparable width that cast a shadow on the photodetector. The wire was caused to vibrate from side to side by driving a current through it at the frequency of one of its "violin" resonances; as it did so, its shadow also moved from side to side across the light beam. If the beam were centered on the wire's position, then the light received by the photodetector would increase equally due to the wire's vibration to the left or the right. But if the beam were off center, then one direction of wire vibration lets more light pass than the other. So in the centered case the photocurrent varies only at twice the frequency of the wire's vibration, while in the off-center case the current contains a component at the wire's vibration frequency, whose amplitude and phase carries the information of the position of the light beam with respect to the wire. A lock-in amplifier converts the modulated signal into one at DC. The wire vibration can

easily be arranged to be at a frequency high enough to avoid $1/f$ noise.

Noise will always cause a sensitive instrument to depart from the precise null reading, and noise of a $1/f$ character will do so in an especially vexing way. Once it has drifted from the null, the instrument is once again sensitive to drive level fluctuations because its output has a non-zero value. The especially elegant feature of the Dicke torsion balance was the provision to apply a correction torque to hold it at the null position. This was accomplished by means of a servo loop that fed back a torque to null the output reading of the sensor. The greater the gain of this loop, the more tightly was the balance constrained near the null position, and the smaller was the sensitivity to light level fluctuations. Of course, the output has to be recalibrated for the effect of the loop, but the same information is present. One convenient place to read the output is to measure the control torque needed to hold the balance at null, which will just be equal to the external torque applied to the balance.

6.3.3 Application of these Principles in an Interferometer

Maximization of response in an interferometer is achieved by the obvious choices: placing the far mirrors as far away as possible from the beam splitter, and by using the shortest practicable wavelength for the illumination of the interferometer. Weiss also describes increasing the optical path length by use of a delay line that causes the light to make many trips through an arm; this is effective against sensing noise (like shot noise), but is just a wash for noise effects that cause actual mirror motions since N encounters with mirrors necessary import a factor N larger amount of these kinds of noise along with the extra signal.

Minimization of the noise starts with obvious choices as well: largest possible light power to make the shot noise small, and the minimization of the multitude of mechanical noises as discussed in Weiss' report. But it also involves replicating many of the features of the Roll *et al.* experiment to minimize the deleterious effects of $1/f$ noise of various kinds. This is the reason for the application of high frequency modulation of the fringe phase (by PZT transducer at Hughes, or by Pockels cell according to Weiss), for the arrangement of the measurement to give a null reading (by differential measurement with two photodetectors at the mid-point of a fringe in the Hughes case, and by modulation about the dark fringe at MIT), and by the provision for actively holding the instrument at the null state (by PZT again at Hughes, and by a combination of Pockels cell and force applied to test masses at MIT).

The fringe modulation and locking techniques mentioned just above have as their goal elimination of excess noise from low frequency fluctuations of the laser's output power. Other ways to make the instrument null are used to make it insensitive to other sorts of fluctuations. For example, laser wavelength fluctuations can be rendered harmless if the interferometer is operated at or near the "white light fringe," that is, with well-matched arm lengths. If the light travels through arms of equal length, then the phase relationship between them is hardly altered as the wavelength varies. But if the arms have a substantial difference in their lengths, wavelength fluctuations cause spurious relative phase shifts at the output, even without fluctuations in arm length difference. Note that this sets a different condition on declaring the instrument to be null—against laser power noise, it is the operating point within a fringe that matters, but for laser wavelength noise, it is the choice of which fringe on which to sit that is relevant.

6.4 Wrap-Up of the Hughes Program

A summary of the results eventually achieved by the Hughes group was reported in 1978 in a paper in *Physical Review D* whose sole author was R. L. Forward.[64] It represents a retrospective analysis of a completed experiment, somewhat in the same style as Kafka and Schnupp's paper of the same year. But there are also strong stylistic differences between the two papers—while Kafka and Schnupp want to make the strongest possible point with the data they analyzed (concerning whether Weber's events could be gravitational waves) with hardly a pause to describe the measurement apparatus, Forward's paper is primarily concerned with explicating the basic physics of the detector, and is somewhat cavalier about the observational data that is presented. The contrast is due in part, of course, to the distinction between publication about a mature technology (bars at room temperature) and one that was clearly immature, although with great promise. It is also a reflection of two facts about the Hughes interferometer as a scientific instrument: it was not as sensitive to bursts as were the bars of 1972, and its data stream was also much harder to analyze because of its broadband nature.

The first section of the 1978 paper is a detailed derivation of the antenna pattern (sensitivity versus angle) of an interferometer. Next comes a quite detailed description of the apparatus, with a strong emphasis on the optical aspects of the interferometer that determine the noise above 1 kHz: circuit diagrams of the photodiode biasing network and of the 1-20 kHz bandpass filter, part numbers for the photodiodes and the

front-end amplifier, and a painstaking derivation of the shot noise. Mechanical aspects of the interferometer, which mainly determine performance at lower frequencies, get shorter shrift: test mass suspensions are described in a single paragraph, as neoprene and brass stacks of "the desired height" with "typical frequency of 10 Hz," without any discussion of the mechanical transfer function or of the thermal noise of the suspension. The three paragraph section devoted to the "isolation system" gives information on both seismic isolation tables and the vacuum system enclosing the interferometer, and includes the following remark, almost in passing: "The vacuum system and isolation tables were designed so that after an initial checkout and operation with 2-m sections of aluminum irrigation pipe (8.5 m total interferometer pathlength), those sections could be replaced with longer sections (up to 1 km) with a substantial increase in interferometer-gravitational radiation-strain sensitivity for the same photon-noise-limited displacement sensitivity."

A careful discussion of the calibration of the instrument and of its linearity is provided, although without including any but the most cursory details about the servo system (whose actuator was a PZT stack on which one of the mirrors was mounted) used to keep the interferometer in lock. One presumes that ignoring the behavior of the servo was justifiable on the assumption that its bandwidth was smaller than 1 kHz. Indeed, given the strategic decision only to consider the output of the interferometer at frequencies above 1 kHz, most of the omissions that strike a modern eye as surprising can be seen to make sense.

The discussion of the operation of the interferometer as a gravitational wave detector begins with a paragraph that will evoke much sympathy from present-day readers. It repeats the "ultimate plan" of operation at a remote site with long arms, but concludes with the remark, "The funding for this next move proved to be unavailable so we concluded the program by operating the system as it was, despite the high level of acoustic, electromagnetic, and vibrational noise from the other activities in the building." Operation was evidently difficult, since as the Abstract notes, "The laser interferometer was operated as a detector for gravitational radiation for 150 h during the nights and weekends from the period 4 October through 3 December 1972," a duty cycle of a bit over 10%. Environmental noise was a serious problem, and was taken seriously: a set of monitors of seismic, acoustic, optical, and electrical noises was installed, and a measure of their outputs was recorded along with the interferometer output.

Part of what made taking data so difficult was the decision to take advantage of the high bandwidth available; although the band below 1 kHz was abandoned as useless because of high noise levels, the upper frequency cut-off was taken to be 20 kHz. This choice was never discussed in the paper—one might have expected some sort of argument on astrophysical grounds that such high frequencies might contain signals, but it is just as likely that the cut-off was chosen to match the bandwidth of the "high quality stereo tape recorder" that was used as the primary data storage medium. Dealing with this much data was a tremendous burden, given the state of computer technology in the early '70s. In fact, the data processing was performed almost entirely by listening to the audio tape—one section of the paper is called "Calibration of Ear." (One 10 msec digitized chunk of data is shown in the paper, both in the time and the frequency domain.)

This method of data analysis was a clever solution to a vexing problem, and indeed continues to be a model for qualitative analysis and debugging of interferometers today. But it showed its weaknesses in what might otherwise have been the most interesting section of the paper, "Comparison of Data with Other Observers." Here, Forward looks for coincidences between the unexplained events in his data set (not coincident with environmental signals in the monitor channel) and events in the resonant mass detectors that were in operation at the same time, at Frascati, Glasgow, and the Maryland group's detectors at College Park and Argonne. In every case, Forward found no event in his detector at the time of a candidate event from another detector. He notes ruefully that "The one 'distinctive signal' reported by the Glasgow group occurred at 13 h 07 min 29 sec GMT 5 September 1972, which was prior to the start of the Malibu data collection period."

The data were of course most interesting for their comparison to the results reported by Forward's former mentor Weber, since the latter was continuing to report coincident events between his various detectors. Forward notes that there were seven time blocks during which unexplained events in the interferometer occurred in close proximity to Weber coincidences. However, he further states, "Both raw power and derivative power-squared digitized data plots digitized to 0.1-sec accuracy were obtained from the Maryland group and compared with the 0.2-sec accuracy Malibu data. None of the audible Malibu signals fell within 0.6 sec of a Maryland-Argonne coincidence."

Fair enough, but what can be concluded from this lack of coincidences? Not much, according to Forward, since "It is difficult to compare the relative detection capabilities of the various antennas since their amplitude sensitivities, bandwidths, and signal processing techniques differ widely." He goes on to state that "at the time one of the bar-antenna systems produced an event or coincidence corresponding to a gravitational-

radiation signal with an amplitude of 0.1 fm/m [10^{-16}] due to spectral components in a narrow band around the bar resonance, the amplitude of the gravitational-radiation spectral components in the entire band from 1-20 kHz was definitely less than 10 fm/m and was probably less than 1 fm/m."

6.5 System Designs for More Sensitive Interferometers

6.5.1 The Magnitude of the Challenge

Forward's paper describing the state of the art of interferometric gravitational wave detectors in 1972 appeared in 1978. It was not until ten years later still that the next improvement in the state of the art was considered significant enough by its authors to warrant another comparable paper in the *Physical Review*. It is worth understanding what factors might have contributed to this slow pace of progress. Certainly the collapse of the credibility of Weber's "events" and the consequent redirecting of the effort at the much smaller signals predicted by astrophysical theory removed a lot of the motivation for trying to achieve incremental improvements in sensitivity. Instead, ultimate success came to be seen to require heroic improvements; neither the 10^{-16} of the room temperature bars nor the 10^{-18} sensitivity of the cryogenic bars would guarantee that gravitational wave astronomy could be established. It looked like frequent signals weren't likely to appear with strengths in excess of 10^{-21}, and perhaps not even much above 10^{-22}. If that were truly the case, then every possible advantage that could be squeezed out of the ideas detailed in Weiss' report would be required. And to achieve them, a great deal more effort was required than could be obtained from the Hughes interferometer, with its sensitivity in the range of 10^{-14} to 10^{-15}interferometer's arms from 2 meters to 1 km would hardly have begun to meet the need.

A sober look at the prospects for the field pushed its practitioners to even greater ambitions than simple improvement of sensitivity at or above 1 kHz. Much of the hope for detectable signals near 1 kHz had come to be focused on the radiation from the "bounce" at the end of the gravitational collapse of a stellar core at the onset of a supernova of Type II. But the theoretical predictions of the strength of such events were very uncertain. If one wanted to maximize the potential for discoveries, one had to imagine building instruments sensitive to the widest possible varieties of signals. And, since very few of the imaginable signals had frequencies above 1 kHz, the best way to improve one's chances was to improve the sensitivity to low frequency signals. This meant minimizing the magnitudes of many of the noise sources that Weiss had listed, a task that the Hughes group had chosen (for sensible reasons) to defer.

- Improved isolation from seismic noise is straightforward in principle, but very demanding in practice. The essential idea was employed by Weber: a cascade of mass-spring oscillators can provide a very steep slope to the frequency dependence of the isolation, making it acceptable for frequencies higher (by of order a decade or more) than the characteristic resonance frequency. It is a real trick to move the whole set of resonances to low enough frequencies (a few Hz) to extend the useful band down as low as 10 Hz or so.
- On the assumption that seismic noise can be controlled, another difficult noise source looms at low frequencies. The thermal noise of the test mass about its equilibrium position can become the dominant noise source, unless very specific measures are undertaken. In place of Forward's rubber mounts for the test masses, one needs to use what Weiss described as "horizontal seismometer suspensions." By this he meant suspensions that had two attributes: low resonant frequency and high Q. The latter was to be achieved by an arrangement whereby most of the restoring force on the test mass comes from gravity (which is nearly dissipationless), as in a pendulum. His system diagram is marked with the compliant directions for the seismometer suspensions, on the assumption that they would be relatively rigid in the orthogonal directions. Present thinking has moved to simple pendulums, partly because of worries about alignment but mostly because the rigid members of single d.o.f. seismometer suspensions tend to have internal resonant modes at inconveniently low frequencies.
- Shot noise minimization needs the whole bag of tricks, including a laser whose power was rated in tens of Watts (not tens of milliWatts), and an optical path length comparable to the gravitational wavelength. The latter would require Forward's desired kilometer-scale separation plus folding of the optical path, such as by the delay line described in Weiss' report. The Herriott delay line has served well in meter-scale prototypes, but most of the large projects have adopted long Fabry-Perot cavities, proposed for this purpose by Drever. (See the next section.) These have the great advantage of allowing the use of the smallest possible test masses, and taking up correspondingly small cross sections in the vacuum pipes. But their operation is substantially trickier than corner reflectors or delay lines, since each Fabry-Perot cavity is itself an interferometer that only performs like a beam-folding element when it is locked to the laser wavelength, using a servo of some

subtlety. Fabry-Perot cavities also appear to be less subject to excess noise from light scattered into unanticipated paths from mirror imperfections, a problem not suspected by either Weiss or Forward.

- Achieving the high sensitivities to which we now aspire requires vacuum of a quality much better than the Hughes interferometer. Pressures of 10^{-6} torr or better are required. The vacuum pipes are themselves much larger in diameter, due partly to the great care needed to keep scattered light effects at low levels. Scattered light also demands that baffles be properly placed in the interior of the pipes.

So it was probably going to take more than adding a few kilometers of irrigation pipe to the Hughes interferometer to detect gravitational waves with an interferometer. The realization that all of these features would be necessary was daunting, and caused the character of work on interferometers to change. Instead of quick demonstrations, it was considered necessary to try to engineer the variety of subsystems that high sensitivity would require. Instead of a device that Forward could honestly describe as a "gravitational-radiation experiment," workers conceived of their apparatus as "prototype gravitational-wave detectors."

6.5.2 Ron Drever's Bag of Tricks

Ronald W. P. Drever was one of the leaders of the generation of experimenters who followed Weber, only to find no signals that matched his claims. Rather than build a faithful copy of Weber's original bar, he chose to follow the path invented by Aplin (and publicized by Gibbons and Hawking) of the split bar, which maximized the bandwidth of the detector. When it became clear that much greater sensitivity would likely be required, he (like the German group) chose to switch to work on interferometers. His work in this period is again marked by an enthusiastic exploration of clever ideas. It is not marked, however, with many conventional papers in refereed journals. Instead, his most stimulating work is to be found in conference proceedings and lectures at physics summer schools.

Several of Drever's most important contributions are described in the text of the lectures he gave at the NATO Advanced Study Institute on Gravitational Waves held at the Les Houches Center of Physics in 1982.[65] Cast as an overview of the interferometric method of gravitational wave detection, it is dominated by an account of three crucial improvements on the basic scheme of Weiss, Forward *et al.* Each of these variations has come to play an important role in the design of the large detectors now under construction.

The Introduction gives an astoundingly brief account of the history of the field:

> An obvious way one might consider detecting gravity waves is through the changes in separation of free test particles, and the idea of using optical interferometers for observing this has certainly occurred to many physicists: indeed one might wonder why so few searches for gravity waves have been made this way.

The work of the Hughes group is mentioned in passing as demonstrating that a simple gravitational wave interferometer could achieve the shot noise limit. Weiss's work is referred to later as having contributed the idea of the delay line as an "important practical method for improving photon-noise limited sensitivity."

Drever goes on to describe the optimization of the parameters of a delay line, from the point of view of shot noise reduction. He then remarks on a "practical difficulty" that "became apparent in early experiments at Munich and at Glasgow—the potentially serious effect of incoherent scattering of light at the multireflection mirrors or elsewhere in the system." The interference between scattered light and light following the intended paths (which is non-stationary because the path followed by the scattered light can vary in length both on long and short time scales) proved to be a very troubling noise source in delay lines. The German group, Drever notes, proposed a way of modulating the laser light that would minimize the problem. Drever then suggests that "another approach would be to make the path traveled by scattered light equal to that of the main beam, and this may in fact be achieved if another type of optical system, a Fabry-Perot cavity, is used instead of a Michelson interferometer with many discrete reflections in each arm."

The basic idea was that light traveling between parallel mirrors can be, in effect, trapped for many round trips, until it is either absorbed, scattered, or leaked out by transmission through one of the mirrors. Thus, such a cavity can play the same role as a delay line with its many spatially separate reflections. The classic Fabry-Perot cavity used flat mirrors, usually equivalent to one another, usually closely spaced compared with their diameters or with the diameter of the beam of light, and usually operated in transmission (that is, with the interesting light emerging from the mirror opposite to the one into which the light was injected). What Drever proposed was rather different: a pair of small mirrors (no larger than necessary to keep diffraction losses small), spaced

apart by kilometers. The distant mirror has as nearly perfect a reflectivity as possible; a finite transmission is used on the near mirror, which functions both as an input and an output coupler for the light. Thus modified, a Fabry-Perot cavity can serve as an arm for a Michelson interferometer.

Drever is explicit about both the advantages and the drawbacks of this beam-folding scheme.

> The diameter of the cavity mirrors can be considerably smaller than that of delay-line mirrors.... This reduces the diameter of the vacuum pipe required, and also may make it easier to keep mechanical resonances in the mirrors and their mountings high compared with the frequency of the gravity waves, thus minimizing thermal noise. The Fabry-Perot system has however, some obvious disadvantages too—particularly the requirement for the very precise control of the wavelength of the laser and of the lengths of the cavities. Indeed with long cavities of the high finesse desirable here exceptional short-term wavelength stability is required from the laser.

The heart of the difficulty is that, unlike a delay line, a Fabry-Perot cavity stores light because it is in itself an interferometer—the trapping of the light for many round trips comes about only by careful adjustment of the phases of the superposed beams. This can only occur when the wavelength of the light and the length of the cavity are in resonance, that is, matched so that an integer number of waves fits into the cavity. Very near the resonance condition, the phase of the output light varies with mirror separation in the same way as the light that has traveled through a delay line. To achieve this condition, the light and the arm have to be locked together by a servo system. Drever's lecture goes on to describe the style of servo required, one that he and his group developed in conjuction with John Hall's group at the Joint Institute for Laboratory Astrophysics in Boulder, Colorado.[66] This servo design has its roots in an analogous microwave device developed by Robert Pound.[67]

While the essence of the difficulty was thus solved, in practice the use of Fabry-Perot cavities has additional complications. One is due to the fact that when a cavity is not very close to resonance, the phase of the output light has almost no dependence at all on the separation of the mirrors, thus making it very hard to generate the sort of signal necessary to acquire the lock on resonance in the first place. An additional level of complication comes when the arm cavities are assembled into a complete Michelson interferometer, since the interferometers within the interferometer need to be separately controlled without degrading the function of the main instrument. Solving these sorts of problems robustly has proven to be challenging work.

As a stopgap measure until those challenges could be met, Drever proposed a kind of interferometer that functioned without the light from the two arms being recombined at the beam splitter for the comparison between their phases. In the context of an ordinary Michelson interferometer this may seem impossible; but, since Fabry-Perot cavities are themselves interferometers, they can each be used to make an independent measurement of the difference between the wavelength of the light and the length of the cavity. A laser illuminates the two orthogonal arms of an interferometer containing Fabry-Perot cavities. Arm 1 is read out by the Pound-Drever-Hall method to generate an error signal reflecting the mismatch between the laser light and the length of the arm. This error signal is used in a servo loop to cause the wavelength of the laser to resonate with Arm 1. The other half of the laser's light is directed by the main beam splitter to Arm 2. There, the mismatch between the light and the arm generates another error signal, reflecting the difference in length (modulo N wavelengths) between Arm 2 and Arm 1. Since a fluctuating arm length difference is the signature of a gravitational wave, this second error signal constitutes the scientific output of the interferometer. (It also of course contains the various noise sources that cause or mimic arm length differences, but in that sense it is no different from the dark fringe output of a more conventional interferometer.) Drever included the use of that error signal being used to close a second servo loop that forces the length of Arm 2 to follow the wavelength of the light. This is necessary to keep the arm at a functional operating point; it means that the error signal has to be calibrated to take the servo action into account, which sounds more complicated at the heuristic level than it actually is in practice.

The scheme described in the previous paragraph formed the basis for almost all of the work on Fabry-Perot-based gravitational wave interferometers until well into the 1990s. But none of its proponents ever expected it to play more than a stopgap role. This is mainly because, without recombination of the light from the two arms to generate "dark" and "bright" output ports, the system cannot be used as the basis of what came to be known as *power recycling*, described for the first time in this lecture, featured there as one of the "possibilities for future enhancement in sensitivity."

Drever introduces the idea of recycling (modestly referred to as a "possibility for more efficient use of light") in the context of a delay-line Michelson interferometer, for pedagogical simplicity. The key idea is as follows. If an interferometer has long arms and if it is constructed from mirrors with high enough reflectivity, then the light exiting

from the bright port may be nearly as bright as the light entering the interferometer from the laser. (This insight represents a profoundly different "take" on the issue than can be found in Weiss' work; he instead worried about optimizing the shot noise versus number of bounces on the assumption, good for short arms and poor mirrors, that substantial losses would eventually occur.) The light exiting the bright port is every bit as good as "fresh" light from the laser, so it seems a shame to waste it. Drever's proposal is to arrange by an appropriate set of mirrors to redirect the used light into the interferometer, in coherent superposition with light arriving directly from the laser. (This has to be done using a beam splitter or other partially reflecting mirror of a carefully chosen reflectivity.) This arrangement has in effect made the whole interferometer into a single Fabry-Perot resonant cavity, whose back mirror is the Michelson interferometer, and whose input/output coupler is the partially-reflecting recycling mirror.

In principle, the advantages that could be achieved with this technique are quite large. Drever quotes rms shot noise in a search for 1 msec pulses of 10^{-22}, far superior to what could be achieved without recycling. He also gives a diagram showing how the technique could be applied to an interferometer whose arms were made of Fabry-Perot cavities. The components necessary to sense and control the various internal degrees of freedom are drawn in with dashed lines, as an indication of the provisional nature of the design. In fact, more subtle schemes have had to be developed to implement such as system. But, given the quality of the mirrors available today (and the lack of commensurate progress in the power levels available from stabilized lasers), power recycling has been adopted as an essential feature of every large interferometer under construction today.

The other "possibility" described here is one "for enhancing sensitivity for periodic signals." This one is again introduced in the pedagogically simpler delay line interferometer. And again, the aim is to find a way to make use of the fact that, with good mirrors, the light would not be significantly attenuated after it has spent one half of a gravitational wave period in an interferometer arm. A periodic gravitational wave persists (by definition) for much longer than one half period; why not find a way to accumulate a phase shift on the interferometer's light for a much longer interval? The scheme proposed here does just that, by arranging for light that exits one arm after one half cycle of the gravitational wave to enter the other arm, where it stays for another half cycle. The light changes arms at the same time that the gravitational wave changes sign, at least for the signal frequency that matches the length of the interferometer. A partially reflecting mirror governs how long the light repeats this cycle before finally exiting the interferometer. As with power recycling, Drever goes on to show how a similar effect can be achieved in an interferometer that uses Fabry-Perot cavities.

It has been shown recently that the scheme can actually be implemented in a much more elegant way, using a single partially reflecting mirror at the nominally dark port. In analogy with power recycling, this scheme (called *signal recycling*) can be thought of as forming a single large Fabry-Perot resonant cavity out of the interferometer, this one resonant at the frequency of the signal sidebands on the laser light that have been created by the action of the gravitational wave.[68] This version of the idea will almost certainly also find application in the next generation of large interferometers.

6.6 The Garching 30-Meter Prototype Gravitational-Wave Detector

The 1988 paper to which we referred above was the account by the group at the Max-Planck-Institut für Quantenoptik, the successor to the bar group of the early '70s. Shoemaker *et al.*[69] provided a beautifully detailed account of the best-characterized interferometer prototype yet built. It can be thought of as the work that brought to fruition, on the meter scale at least, the ideas embodied in Weiss's 1972 design study. Through the '70s and '80s, a number of groups (including Weiss's at MIT, Drever's at Glasgow and at Caltech, and Brillet's at Orsay) worked in parallel with the MPQ group to develop prototypes of kilometer-scale working interferometric detectors.[70] The MPQ paper makes a nice example, though, since it is an especially complete account of a well-functioning instrument. So for pedagogical purposes we let it here stand for the large body of work done worldwide through the 1980s.

The interferometer described here had test masses 30 meters from the beam splitter; light made 45 round trips, for a total light travel time in an arm of 9 μs. The folding of the optical path was achieved with a Herriott delay line. The interferometer was illuminated with an Argon-ion laser at $\lambda = 514.5$ nm, capable of supplying up to 0.23 W to the photodetector (at a bright fringe) after all optical losses in the interferometer are included. The test masses consisted of simple glass mirrors with a radius of curvature 31.6 m; they were suspended from single-wire slings of free length 0.72 m, giving a resonant frequency of about 0.6 Hz. Each of these was in turn suspended from a metal plate hung from coil springs. This upper level of the suspension not only added isolation along the optic axis, but gave isolation in the other degrees of freedom that might cross-couple into the sensitive direction.

The many compliant degrees of freedom of such a system needed damping, but damping of a sort that would not add substantial amounts of noise, in particular thermal noise. The solution was a set of optical shadow sensors that detected the positions of small vanes attached to the mirrors; the outputs of these sensors, suitably filtered, drove currents through coils that were in turn positioned near magnets integrally mounted with the vanes. This is a very robust kind of servo, and it can be made sufficiently noiseless, at least at intermediate and high frequencies. In the Garching 30-meter interferometer, 16 degrees of freedom were damped using servos of this kind.

The Garching group paid careful attention to the influence of fluctuations in laser power, wavelength, position, and angle. Laser power fluctuations are dealt with primarily by implementation of the modulated dark-fringe servo system. Sensitivity to variations in the input beam's transverse position and injection angle was minimized by using an optical fiber as a spatial filter for the laser, an idea the authors credit to Weiss. Frequency fluctuations enter in two ways:

- Mismatches in the curvature of the mirrors cause the two interferometer arms to differ in length by an amount ΔL, when the mirrors are spaced so as to give the same number of round trips N through the arms. The spectral density of effective arm length noise $x(f)$ due to a frequency fluctuation spectrum $\delta\nu(f)$ is

$$x(f) = \frac{\Delta L}{N}\frac{\delta\nu(f)}{\nu}.$$

- Light that scatters out of the intended path can nevertheless find its way to the beam splitter again, where it interferes with the rest of the light. Although good mirror surfaces and coatings minimize the fraction of the light scattered, the very large mismatches in path length that can result made this a comparable mechanism for conversion of laser frequency noise into apparent mirror motion.

In order to make laser frequency noise small enough, the Garching group stabilized the laser using a combination of two servo loops. In the first loop, a rigid Fabry-Perot resonant cavity serves as a length reference; the light's wavelength is held to resonance in the cavity, by inspection of the transmitted light intensity. A second error signal is derived by interfering the light that exits the bright port of the interferometer (whose phase depends on the average of the lengths of the two arms) with some of the new light entering. Feedback was applied to a mirror at the end of the laser cavity, to a Pockels cell within the laser cavity, and to the common mode positions of the two end mirrors of the interferometer.

The noise spectrum of the system was in reasonable agreement with a noise budget prepared from estimates of the noise sources described above. The rms noise in a 1 kHz bandwidth near 1 kHz was about 3×10^{-18}, only a factor of three poorer than the best that had been achieved by a resonant-mass detector. This was one of the great triumphs of the German work. The other was the robustness of the rather elaborate active system that this interferometer had become, of order two dozen servo loops. It typically stayed in lock for 30 minutes to an hour, and upon losing lock would shortly reacquire lock automatically.

The 30-meter interferometer was run for 100 hours in March 1989, in coincidence with the ten-meter interferometer at Glasgow that had achieved comparable sensitivity. Even then, the enthusiasm for looking for gravitational waves at "low" sensitivity was such that an analysis of the results of the run was not published until 1996.[71]

6.7 Designs for Kilometer-Class Interferometers

As noted above, laboratory work on interferometers was almost from the beginning considered an engineering exercise preparatory to the construction of instruments with arms of kilometer scale. With several such devices now under construction, it is worth reviewing their distinctive features. Here, even more so than for the other cases we have been discussing, the refereed literature is a poor source of information, and so are conference proceedings. For most projects the only detailed descriptions are those contained in funding proposals. (The one redeeming feature of this form of publication is that, since the reviewers of such documents were typically not expected to be experts in the field, they contain an abundance of carefully written tutorial material, and well-reasoned justifications for most design choices.)

The three largest approved projects today, (LIGO,[72] VIRGO,[73] and GEO[74]) all went through similar parallel processes of design study, proposal, and now construction. This was a fruitful period, with a rich interchange of ideas. For pedagogical purposes, we choose to focus in this review on a single line of development, that of the U.S. LIGO Project.

6.7.1 The "Blue Book"

In almost the same sense as the early table-top interferometers were prototypes of larger instruments, so too did the proposals for kilometer-scale interferometers have a prototype. This was the report called "A Study of a Long Baseline Gravitational Wave An-

tenna System," submitted to the U.S. National Science Foundation in October 1983.[75] (It has since its presentation been called the "Blue Book" because of the color of the cheap paper cover in which it was bound.) It was prepared primarily by Weiss and two colleagues at MIT (Paul S. Linsay and the present author), as the product of a planning exercise funded by the NSF starting in 1981. The report also contained a section by Stan Whitcomb of Caltech on Fabry-Perot systems (as a partial counter to Weiss' emphasis on Herriott delay lines), as well as extensive sections written by industrial consultants from Stone & Webster Engineering Corporation and from Arthur D. Little, Inc. These latter contributors were essential, because this document contains, for the first time anywhere, an extensive discussion of the engineering details specific to the problems of the construction and siting of a large interferometer. The report was presented, by both Weiss's MIT group and that of Drever at Caltech, at a meeting of the NSF's Advisory Council for Physics late in 1983. While not a formal proposal, it served as a sort of "white paper," suggesting the directions that subsequent proposals might (and in large measure did) take.

The first half of the report is devoted to the physics of gravitational wave interferometers. This section reads much like Weiss' 1972 design study, except that many issues only touched on briefly in the first paper are here discussed at substantially greater length. In the 11 years that elapsed between the two documents, there had been real progress on several fronts. There are chapters on sources of gravitational waves, the basic physics of the response of a free-mass interferometer to a gravitational wave, a discussion of beam-folding schemes and a summary of the current prototype interferometers, and another extensive discussion of noise sources. The report is bracketed by an introductory section outlining a history of the field to 1983 and by a pair of appendices, one of which compares the quantum limits of bars and interferometers and the other showing why the interferometer beams must travel through an evacuated space instead of through optical fibers.

The main emphasis of the Blue Book was less a discussion of physics *per se* than it was a consideration of the practical aspects of the experiment as an engineering and construction project. The completely new material appears in the second half of the Blue Book, in the chapters summarizing the work of the industrial consultants. Weiss believed that the only significant impediment to achieving astrophysically interesting sensitivity was the expense of building an interferometer with long arms (the issue that had brought the Hughes group's progress to a halt). The industrial study was undertaken with the aim of identifying what design trade-offs would allow for a large system to be built at minimum cost, and to establish a rough estimate of that cost (along with cost scaling laws) so that the NSF could consider whether it might be feasible to proceed with a full-scale project.

Before such an engineering exercise could be meaningful, though, it was necessary to define what was meant by "full-scale." The Blue Book approaches this question by first modeling the total noise budget as a function of frequency, then evaluating the model as a function of arm lengths ranging from 50 meters (not much longer than the Caltech prototype) to 50 km. The design space embodied in this model was then explored in a process guided by three principles:

- "The antenna should not be so small that the fundamental limits of performance cannot be attained with realistic estimates of technical capability." This was taken to mean that the length ought to be long enough that one could achieve shot noise limited performance for laser power of 100 W, without being limited instead by displacement noise sources, over a band of interesting frequencies. The length resulting from this criterion strongly depended on whether one took that band to begin around 1 kHz (in which case $L = 500$ m was adequate), 100 Hz (where $L = 5$ km was only approaching the required length), or lower still (in which case even $L = 50$ km would not suffice). Evidently, this strictly physics-based criterion was too elastic to be definitive.
- "The scale of the system should be large enough so that further improvement of the performance by a significant factor requires cost increments by a substantial factor." In other words, the system should be long enough so that the cost is not dominated by the length-independent costs of the remote installation.
- "Within reason no choice in external parameters of the present antenna design should preclude future internal design changes which, with advances in technology, will substantially improve performance." This was a justification for investing in a large-diameter beam tube, and for making sure that the vacuum system could achieve pressures as low as 10^{-8} torr.

In an iterative process, rough application of these principles was used to set the scope of options explored by the industrial consultants. Then at the end of the process, the principles were used again to select a preferred design. Arm lengths as long as 10 km were explored, and tube diameters as large as 48 inches. An extensive site survey was also carried out by the consultants. It was aimed at establishing that sites existed that were suitable for a trenched installation (which put stringent requirements

on flatness of the ground) of a 5 km interferometer. The survey covered federal land across the United States, and a study of maps of all land in the Northeastern United States, along with North Carolina, Colorado, and Nebraska. Thirteen "suitable" sites were identified. Evaluation criteria also included land use (specifically that the site not be crossed by roads, railroads, or oil and gas pipelines), earthquake risk, drainage, and accessibility.

The site survey also attempted to identify possibilities of locating an interferometer in a subsurface mine, which would give a more stable thermal environment and perhaps also reduced seismic noise (if it were located deep enough, and if it were inactive). No mines were found in the United States with two straight orthogonal tunnels even 2 km in length.

The conclusion of the exercise was a "proposed design" with the following features:

- Two interferometer installations separated by "continental" distances.
- Interferometer arm length of $L = 5$ km.
- Beam tubes of 48-inch diameter made of aluminum (chosen for an expected cost savings over stainless steel) pumped by a combination of Roots-blowers for roughing and ion-pumps for achieving and maintaining the high vacuum. A delay line interferometer would require a diameter of almost the proposed size. Drever's beam-interchange scheme for improved narrow-band sensitivity was also listed as one justification for preferring large tubes, as was the possibility of multiple interferometers (presumably based on Fabry-Perot cavities) side by side.
- The proposed installation method was to enclose the tube in a 7' by 12' cover constructed of a "multi-plate pipe-arch," in turn installed four feet below grade in a trench that was subsequently back-filled with soil.

The total estimated cost for such a system was given as $58M.

Note that there were no specific recommendations for the design of the interferometers themselves, beyond the "straw man" used for estimating the noise budget.

The Blue Book was received respectfully by the NSF's Advisory Committee for Physics. As a result, the MIT and Caltech research groups were encouraged to combine their forces to develop a complete specific design. Subsequently, both groups received funding with the eventual goal of a joint proposal for construction of a large interferometer system.

6.7.2 The LIGO Proposal of December 1987

By 1987, substantial progress had been made on lab-scale interferometers by research groups around the world. And, encouraged by the NSF's reception of the Blue Book, more thinking had gone into the best way to construct and exploit large interferometers. Much of this progress is evident in the proposal submitted to the U.S. National Science Foundation in December of 1987 by a formal Caltech-MIT collaboration that had adopted the name of LIGO, for Laser Interferometer Gravitational wave Observatory.[76] Since September 1987 it has been led by Rochus E. ("Robbie") Vogt as Project Director, with Drever and Weiss as science team leaders. The proposal requested funds for a three-year program of R&D and engineering studies, the outcome of which was intended to be another proposal (to be submitted in 1989) requesting authorization to build a pair of 4-km long interferometers.

Prototype interferometers were functioning at respectable sensitivities, after years of assembly, debugging, analysis, and redesign. The Caltech 40-meter interferometer was the showpiece of the proposal. It employed high-finesse Fabry-Perot cavities, arranged in the simplified non-recombined configuration for ease of testing. An impressive graph shows the improvement in strain sensitivity since its first operation in May 1983, by a factor of over 10^3 in amplitude. In that interval, ultra-low loss "supermirrors" were installed, first on an early optical-bench style of test mass and then on separated compact test masses (eventually ones made of fused silica), and a variety of improvements were made to the Argon laser and to the locking servos. The noise spectrum above 1 kHz was nearly white, with a level of $h(f) \approx 2 \times 10^{-19}/\sqrt{\text{Hz}}$. This was consistent with the expected level of shot noise.

Results were also presented from the table-top (1.5 meter arm length) prototype interferometer at MIT. It employed Herriott delay lines of 56 bounces between conventional high quality mirrors that were clamped to aluminum test masses. Unlike the Caltech instrument which could not function without sophisticated frequency stabilization of its laser, MIT used an unstabilized Argon laser to which phase noise was actually added to help suppress scattered light. Pointing, damping, and slow feedback were accomplished with electrostatic actuators. Strain sensitivity could of course not compete with the much longer Caltech instrument, but even the displacement noise was nearly an order of magnitude worse, given as 4.6×10^{-17} m$/\sqrt{\text{Hz}}$, for frequencies above 4 kHz. That level was a factor of two in excess of the expected shot noise in the 60 mW of light, diagnosed as insufficiently suppressed noise from scattered light. At lower fre-

quencies acoustic noise drove the interferometer via a variety of coupling paths through the injection optics as well as the test mass suspensions.

The proposal records substantial progress toward design of a full-scale interferometer. It states that the collaboration had adopted the Fabry-Perot beam-folding system. A preliminary design is presented in an appendix of the proposal. It envisioned use of 5 to 6 W of light at 514 nm from an Argon laser employed in a power recycled configuration. An elaborate schematic diagram gave a hint of the complexity of the servos necessary to control the large number degrees of freedom that need to be kept locked for such an instrument to function. These include lengths of the arm cavities, the separation of their input mirrors from the beam splitter, the location of the power recycling mirror, and the lengths of various "mode cleaning" resonant cavities used for spatial filtering of the laser beam. In addition to these lengths, control of a number of angular degrees of freedom also needs to be included. Four separate RF modulation frequencies are specified to drive these servos. Special features are designed into the main cavity locking servo so that the phase modulation can be injected with a small Pockels cell without the inevitable losses dominating the performance of the recycling system.

Another appendix describes an alternate optical configuration based on Herriott delay lines. It employed 86 cm diameter silicon test masses of 450 kg. The simplicity of the servos was listed as one of its major advantages. A "closed-path" variation of this design was also presented, in which light leaving one arm is injected into the other. This is like a single-interchange version of Drever's system for improving sensitivity to periodic waves. Here it was employed mainly to relax the tolerances on matching the curvature of the large mirrors. (The virtues of this design have recently been explored again by the Stanford group.[77])

On the engineering and site issues, there had also been some progress since the Blue Book study, mostly made by engineers at Caltech's Jet Propulsion Laboratory, but this was not considered complete enough to highlight in the proposal. Instead, one of the first proposed tasks was to complete a preliminary engineering design. Nevertheless, a mature understanding had been achieved of what LIGO ought to *be*. This insight was expressed in a list of "Essential Features of the LIGO":

1. "Two widely separated sites under common management." Two sites had been a feature of Weiss' earliest thinking, to allow coincidence observations to search for transient signals. The new feature was the commitment to truly have them managed as a single entity, "to guarantee that two receivers of nearly equal sensitivity are on line simultaneously at two sites, with a high live time."
2. "Arm lengths of order 4 kilometers at each site," a slight scaling back of the 5 km considered previously, but still long enough to strongly dilute the effects of displacement noise.
3. "The ability to operate simultaneously several receiver systems at each site." In a way, this was the most ambitious feature of the LIGO concept. In part it grew out of a kind of conservatism that was not clearly spelled out, but that was nevertheless real. The early LIGO interferometers, if they were not to be extremely risky extrapolations from known technology, were unlikely to have sufficient sensitivity to be assured of detecting astrophysical signals. Even if that weren't so, the project would have had to wrestle with the competition between time devoted to observation and time devoted to improving the performance of the instrument. This competition had bedeviled workers on resonant-mass detectors. The key new idea for LIGO was that the precious commodity, an evacuated beam pipe, might be available with abundant cross-sectional area since the Fabry-Perot geometry had been adopted. All that was required was an arrangement of tanks at the ends of the pipe to install the test masses of various interferometers, both operational and experimental. This actually called for substantial cleverness in developing an airlock system, so that installation and operation could take place with "a minimum of mutual interference."
4. "The capability for receivers of two different arm lengths." Drever urged the adoption of this feature, to allow a clean test of the gravitational origin of candidate signals, which should show up as the tidal signature that a longer interferometer sees twice the signal.
5. "A vacuum tube diameter of order 48 inches." This had the conservative justification that it would be necessary if one had to switch from Fabry-Perot cavities to delay lines, and the great benefit of allowing multiple Fabry-Perot interferometers, as mentioned above.
6. "The capability of a vacuum level of 10^{-8} torr." This would be needed, not for the first LIGO instrument, but to avoid having fluctuations in the index of refraction of the residual gas dominate shot noise in a more sensitive "advanced" receiver.
7. "A minimum lifetime of the facilities of 20 years." This was to be not just a one-shot discovery experiment, but a laboratory to exploit the gravitational wave

window in astronomy.

8. "Adequate support instrumentation."

The heritage of the Blue Book should be evident in the above list, but so too should be the progress in thinking beyond that point.

Because of a Federal budget crisis, this proposal was not funded. However, the two groups were encouraged to continue their work, and to submit a more complete proposal in a subsequent year.

6.7.3 The LIGO Proposal of December 1989

Many of the features only hinted at in the 1987 proposal are spelled out in much greater detail in the proposal for engineering design and construction of LIGO (now with a hyphen in Laser Interferometer Gravitational-Wave Observatory) that was submitted to the NSF in December 1989.[78] It fleshes out the design of a system that would embody the eight Essential Features of LIGO first described in the 1987 proposal. (The only differences are that the vacuum tube diameter is now specified as a "clear optical diameter of approximately 1 meter," and the vacuum spec is given as "10^{-9} torr of hydrogen and 10^{-10} of other gases.")

A single detector system would require three interferometers—a 4 km interferometer at each of two widely separated remote sites, plus a 2 km interferometer at one of the sites. This shorter interferometer played two related roles: a check that a candidate signal had the tidal signature of a gravitational wave, and the simpler but crucial requirement that a real signal should appear in all three interferometers. A calculation presented here shows that an accidental event rate of around 100/hour/interferometer can be tolerated without accidental three-way coincidences occurring more frequently than once in ten years, an improvement of about two orders of magnitude over what could be tolerated with only a single interferometer at each site.

The features described above are part of a plan that is aimed at accomplishing "three primary objectives":

- "observation," or a continuous "gravitational-wave watch,"
- "development," or "full functional testing of new and advanced interferometer-based detector concepts," and
- "special investigations" using detectors optimized for "particular phenomena."

These missions are "to be conducted without mutual interference."

It would take a substantial investment (of money, ideas, and energy of scientists) to accomplish all of these goals, and so for this reason full implementation of this strategy was to be accomplished in a series of phases.

- "Phase A, The Exploration/Discovery Phase," with one three-interferometer detector system, suitable for observation or development, but not both simultaneously.
- "Phase B, The Discovery/Observation Phase," with two three-interferometer detector systems, allowing "concurrent observation *and* development or specialized search."
- "Phase C, The Observatory Phase," room for three full detector systems, "allows concurrent observation, development, special investigations, and optimal access for the scientific community at large. It completes the LIGO evolution to its presently conceived full-design capability."

Because of the cost involved in elaborate vacuum chambers with airlocks, the 1989 proposal asked only for the funds to complete Phase A. The single important investment in the capability to upgrade to Phases B and C was the design of buildings large enough to accommodate all of the vacuum tanks that would be eventually required.

It should be emphasized that, to a large degree, this aggressive planning for an elaborate facility was a necessary consequence of a simple fact—that it looked difficult to build an interferometer that would have sufficient sensitivity to assure detection of gravitational waves. Hence the need to plan for ongoing interferometer development and specially optimized instruments. Of course, these activities would also be useful even if some signals proved easier to detect than expected, since carrying out gravitational wave astronomy would call for the highest achievable signal to noise ratios. For example, angular position errors are inversely proportional to the signal to noise ratio, and are as large as ten arcmin or more when the SNR = 10, even when observations are made with a three-detector U.S.-Europe network.

The 1989 proposal is much more explicit about the details of the design of the first interferometer. The design is based on the Fabry-Perot interferometer of the 1987 proposal's appendix. It has been fleshed out with specifications of laser power, finesse, test mass parameters, and vibration isolation performance, so that a specific noise budget could be presented. The 1989 proposal also contains a preliminary discussion of what sorts of improvements would be necessary to push the noise to levels low enough

to guarantee detection of signals. This includes laser power of 60 W recycled by a factor of 100, a much more aggressive vibration isolation system, and final pendulum suspensions with a quality factor of 10^9 carrying one-ton fused silica mirrors.

6.7.4 The Situation Today

Construction of LIGO was approved in 1991. By mid-1998 (the time of the writing of this review), construction of the two facilities in Hanford, WA and Livingston, LA was over three-quarters complete. The schedule calls for construction to be completed soon. Roughly speaking, 1999 is to be devoted to installation of the scientific equipment in the completed facilities, 2000 to shakedown of the interferometers, and 2001 to engineering activities to bring the performance up to the design specifications. Then, data will be collected during 2002-3. Beginning in 2004, upgrades to improve the performance will be carried out, interspersed with additional periods of observation.

The first instrument to be installed is expected to have a noise spectrum like that shown in Figure 4. The high frequency noise spectrum should be dominated by shot noise, as determined from an input power of 6 W, multiplied by a power recycling gain of 30. Thermal noise from the 1 Hz pendulum mode will dominate the intermediate frequency band; the oscillations of the 10 kg test masses should achieve a quality factor of 1.6×10^5. (Internal thermal noise will dominate the spectrum only in a narrow band, due to test mass modal quality factors of about 10^6.) Low frequency noise will be governed of course by seismic noise that passes through the multi-layer stack.

Performance of the VIRGO 3 km interferometer will be similar at medium and high frequencies. At low frequencies, seismic noise should be much lower in VIRGO than in LIGO, since a much more aggressive filter has been designed. This should allow the noise to be dominated by pendulum thermal noise down to 10 Hz or a bit lower.

The GEO 600 meter interferometer is not expected to reach quite such low levels, but it will be surprisingly close. To make up for the shorter length, advanced technologies (including signal recycling) will be pursued aggressively from an early date. Thus this instrument will play a dual role as part of the global network of interferometers and as a prototype for features that will later be incorporated into other interferometers.

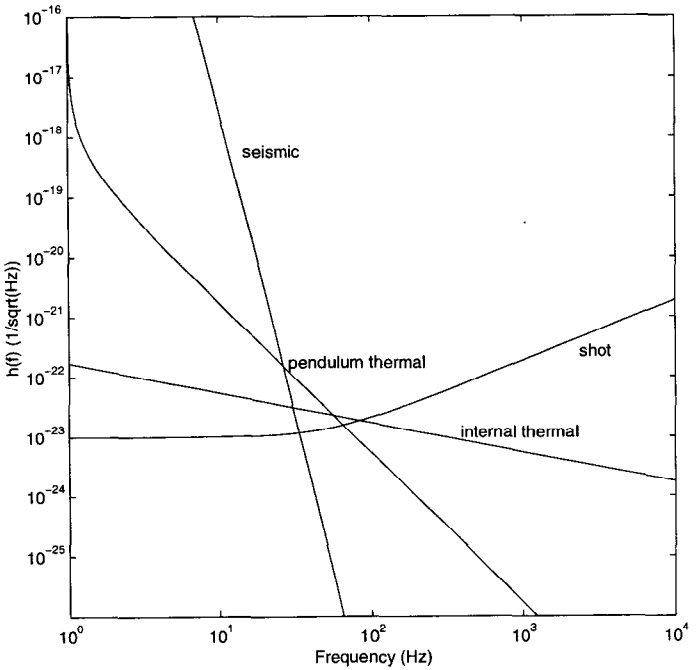


Fig. 4. An estimate of the noise spectrum of the LIGO I interferometers. The four most important noise sources are shown: seismic noise, pendulum mode thermal noise, thermal noise of internal vibrations of the mirror, and shot noise.

References

[1] E. M. Purcell, *Electricity and Magnetism* (McGraw-Hill, New York, 1965).

[2] L. D. Landau and E. M. Lifschitz, *The Classical Theory of Fields* (Pergamon, Oxford, 1975).

[3] P. Kafka, in *Space Science and Fundamental Physics, Proceedings of the Summer School*, Alpbach, Austria, 28 July - 7 August 1987 (ESA SP-283, May 1988).

[4] A. P. Cowley, *Ann. Rev. Astron. Astrophys.* **30**, 287 (1992).

[5] J. A. Wheeler, cited in Weber (1961).

[6] J. Weber, *Phys. Rev.* **117**, 306 (1960).

[7] J. Weber, *General Relativity and Gravitational Waves* (Interscience, New York, 1961).

[8] J. Weber, *Phys. Rev. Lett.* **22**, 1320 (1969).

[9] F. A. E. Pirani, *Acta. Phys. Polon.* **15**, 389 (1956).

[10] G. E. Moss, L. R. Miller, and R. L. Forward, *Appl. Opt.* **10**, 2495 (1971).

[11] R. Weiss, *Quarterly Progress Report, MIT Research Lab of Electronics* **105**, 54 (1972).

[12] E. Kappler, *Ann. Phys. (Leipzig)* **11**, 233 (1931); R. B. Barnes and S. Silverman, *Rev. Mod. Phys.* **6**, 162 (1934).

[13] J. B. Johnson, *Phys. Rev.* **32**, 97 (1928).

[14] H. Nyquist, *Phys. Rev.* **32**, 110 (1928).

[15] H. B. Callen and T. A. Welton, *Phys. Rev.* **83**, 34 (1951); H. B. Callen and R. F. Greene, *Phys. Rev.* **86**, 702 (1952).

[16] N. Bohr, *Nature* **121**, 580 (1928).

[17] N. Bohr, in *Albert Einstein: Philosopher-Scientist* (Library of Living Philosophers, Evanston, Ill., 1949).

[18] R. Weiss, in *Sources of Gravitational Radiation*, ed. L. Smarr (Cambridge University Press, Cambridge, 1979), p. 7.

[19] P. A. M. Dirac, *The Principles of Quantum Mechanics* (Clarendon Press, Oxford, 1958).

[20] C. M. Caves, *Phys. Rev. Lett.* **45**, 75 (1980).

[21] J. Weber, *Phys. Rev. Lett.* **17**, 1228 (1966).

[22] R. Del Fabbro, A. diVirgilio, A. Giazotto, H. Kautzky, V. Montelatici, and D. Passuello, *Phys. Lett. A* **132**, 237 (1988).

[23] J. L. Levine and R. L. Garwin, *Phys. Rev. Lett.* **31**, 173 (1973).

[24] E. Amaldi, O. Aguiar, M. Bassan, P. Bonifazi, P. Carelli, M. G. Castellano, G. Cavallari, E. Coccia, C. Cosmelli, W. M. Fairbank, S. Frasca, V. Foglietti, R. Habel, W. O. Hamilton, J. Henderson, W. Johnson, K. R. Lane, A. G. Mann, M. S. McAshan, P. F. Michelson, I. Modena, G. V. Pallottino, G. Pizzella, J. C. Price, R. Rapagnani, F. Ricci, N. Solomonson, T. R. Stevenson, R. C. Taber, and B.-X. Xu, *Astron. Astrophys.* **216**, 325 (1989).

[25] G. W. Gibbons and S. W. Hawking, *Phys. Rev. D* **4**, 2191 (1971).

[26] H. J. Paik, *J. Appl. Phys.* **47**, 1168 (1976).

[27] J. P. Richard and W. M. Folkner, in *The Detection of Gravitational Waves*, ed. D. G. Blair (Cambridge University Press, Cambridge, 1991).

[28] V. B. Braginsky and Yu. I. Vorontsov, *Sov. Phys. Uspekhi* **17**, 644 (1974).

[29] M. E. Tobar and D. G. Blair, *J. Phys. D: Appl. Phys.* **26**, 2276 (1993).

[30] P. R. Saulson, *Class. Quantum Grav.* **14** , 2435 (1997).

[31] H. Rothe and W. Dahlke, *Proc. IRE* **44**, 811 (1956).

[32] W. R. LePage, *Complex Variables and the Laplace Transform for Engineers* (McGraw-Hill, New York, 1961).

[33] J.-P. Richard, *Phys. Rev. Lett.* **52**, 165 (1984); J. C. Price, *Phys. Rev. D* **36**, 3555 (1987).

[34] R. W. P. Drever, J. Hough, R. Bland, and G. W. Lessnoff, *Nature* **246**, 340 (1973).

[35] H. Heffner, *Proc. IRE* **50**, 1604 (1962).

[36] R. P. Giffard, *Phys. Rev. D* **14**, 2478 (1976).

[37] J. Weber, *Phys. Rev. Lett.* **24**, 276 (1970).

[38] J. Weber, *Phys. Rev. Lett.* **25**, 180 (1970).

[39] W. H. Press and K. S. Thorne, *Ann. Rev. Astron. Astrophys.* **10**, 334 (1972).

[40] P. S. Aplin, *Gen. Rel. Grav.* **3**, 111 (1972).

[41] J. Weber, V. Hughes, P. Kafka, R. W. P. Drever, and C. W. Misner, *Ann. N.Y. Acad. Sci.* **224**, 101 (1973).

[42] D. Bramanti and K. Maischberger, *Lett. Nuovo Cimento* **4** , 1007 (1972); D. Bramanti, K. Maischberger, and D. Parkinson, *Lett. Nuovo Cimento* **7**, 665 (1973).

[43] H. Billing, P. Kafka, K. Maischberger, F. Meyer, and W. Winkler, *Lett. Nuovo Cimento* **12**, 111 (1975).

[44] J. A. Tyson, *Ann. N.Y. Acad. Sci.* **224**, 74 (1973).

[45] J. L. Levine and R. L. Garwin, *Phys. Rev. Lett.* **31**, 173 (1973); R. L. Garwin and J. L. Levine, *Phys. Rev. Lett.* **31**, 176 (1973); J. L. Levine and R. L. Garwin, *Phys. Rev. Lett.* **33**, 794 (1974).

[46] J. L. Levine and R. L. Garwin (1974), cited above.

[47] R. L. Garwin, *Physics Today* , December 1974, p. 9.

[48] D. Sciama, J. Weber, P. Kafka, R. W. P. Drever, J. A. Tyson, in *General Relativity and Gravitation*, Proceedings of GR7, Tel Aviv University, June 23-28, 1974, eds. G. Shaviv and J. Rosen (Wiley, New York 1974).

[49] H. M. Collins, *Changing Order* (Sage Publications, London, 1985). See also this exchange: A. Franklin, *Stud. Hist. Phil. Sci.* **25**, 463 (1994); H. M. Collins, *Stud. Hist. Phil. Sci.* **25**, 493 (1994).

[50] R. W. P. Drever, *Q. J. R. Astr. Soc.* **18**, 9 (1977).

[51] J. A. Tyson and R. P. Giffard, *Ann. Rev. Astron. Astrophys.* **16**, 521 (1978).

[52] P. Kafka and L. Schnupp, *Astron. Astrophys.* **70**, 97 (1978).

[53] S. P. Boughn, W. M. Fairbank, R. P. Giffard, J. N. Hollenhorst, E. R. Mapoles, M. S. McAshan, P. F. Michelson, H. J. Paik, and R. C. Taber, *Ap. J. Lett.* **261**, L19 (1982).

[54] P. Astone *et al.*, *Proceedings of the Xth Italian General Relativity Conference*, Bardonecchia, September 1-4, 1992, unpublished.

[55] M. Aglietta, A. Castellina, W. Fulgione, G. Trinchero, S. Vernetto, P. Astone, G. Badino, G. Bologna, M. Bassan, E. Coccia, I. Modena, P. Bonifazi, M. G. Castellano, M. Visco, C. Castagnoli, P. Galeotti, O. Saavedra, C. Cosmelli, S. Frasca, G. V. Pallottino, G. Pizzella, P. Rapagnani, F. Ricci, E. Majorana, D. Gretz, J. Weber, and G. Wilmot, *Nuovo Cimento* **106B**, 1257 (1991); G. Pizzella, in *Developments in General Relativity, Astrophysics and Quantum Theory: A Jubilee in Honour of Nathan Rosen*, eds. F. Cooperstock, L. P. Horwitz, and J. Rosen, *Annals of the Israel Physical Society* **9**, 195 (1990).

[56] C. A. Dickson and B. F. Schutz, *Phys. Rev. D* **51**, 2644 (1995).

[57] M. E. Gertsenshtein and V. I. Pustovoit, *Sov. Phys. JETP* **16**, 433 (1962).

[58] P. G. Roll, R. Krotkov, and R. H. Dicke, *Ann. Phys. (N.Y.)* **26**, 442 (1964).

[59] A. A. Michelson and E. W. Morley, *Am. J. Sci.* **34**, 333 (1887).

[60] R. H. Dicke, *Rev. Sci. Instrum.* **17**, 268 (1946).

[61] R. Buderi, *The Invention That Changed the World* (Simon and Schuster, New York, 1996).

[62] J. D. Kraus, *Radio Astronomy* (McGraw-Hill, New York, 1966).

[63] R. H. Dicke, *The Theoretical Significance of Experimental Relativity* (Gordon and Breach, New York, 1964).

[64] R. L. Forward, *Phys. Rev. D* **17**, 379 (1978).

[65] R. W. P. Drever, in *Gravitational Radiation*, eds. N. Deruelle and T. Piran (North Holland, Amsterdam, 1983), p. 321.

[66] R. W. Drever, J. L. Hall, F. V. Kowalski, J. Hough, G. M. Ford, A. J. Munley, and H. Ward, *Appl. Phys. B* **31**, 97 (1983); see also A. Schenzle, R. DeVoe, and G. Brewer, *Phys. Rev. A* **25**, 2606 (1982).

[67] R. V. Pound, *Rev. Sci. Instrum.* **17**, 490 (1946).

[68] B. J. Meers, *Phys. Rev. D* **38**, 2317 (1988); K. A. Strain and B. J. Meers, *Phys. Rev. Lett.* **66**, 1391 (1991).

[69] D. Shoemaker, R. Schilling, L. Schnupp, W. Winkler, K. Maischberger, and A. Rüdiger, *Phys. Rev. D* **38**, 423 (1988).

[70] P. F. Michelson, E.-K. Hu, and G. Pizzella, editors, *International Symposium on Experimental Gravitational Physics* (World Scientific, Singapore, 1988).

[71] D. Nicholson, C. A. Dickson, W. J. Watkins, B. F. Schutz, J. Shuttleworth, G. S. Jones, D. I. Robertson, N. L Mackenzie, K. A. Strain, B. J. Meers, G. P. Newton, H. Ward, C. A. Cantley, N. A. Robertson, J. Hough, K. Danzmann, T. M. Niebauer, A. Rüdiger, R. Schilling, L. Schnupp, and W. Winkler, *Phys. Lett. A* **218**, 175 (1996).

[72] A. Abramovici, W. E. Althouse, R. W. P. Drever, Y. Gürsel, S. Kawamura, F. J. Raab, D. Shoemaker, L. Sievers, R. E. Spero, K. S. Thorne, R. E. Vogt, R. Weiss, S. E. Whitcomb, and M. E. Zucker, *Science* **256**, 325 (1992).

[73] C. Bradaschia, R. Del Fabbro, A. Di Virgilio, A. Giazotto, H. Kautzky, V. Montelatici, D. Passuello, A. Brillet, O. Cregut, P. Hello, C. N. Man, P. T. Manh, A. Marraud, D. Shoemaker, J.-Y. Vinet, F. Barone, L. Di Fiore, L. Milano, G. Russo, S. Solimeno, J. M. Aguirregabiria, H. Bel, J.-P. Duruisseau, G. Le Denmat, P. Tourrenc, M. Capozzi, M. Longo, M. Lops, I. Pinto, G. Rotoli, T. Damour, S. Bonazzola, J. A. Marck, Y. Gourghoulon, L. E. Holloway, F. Fuligni, V. Iafolla, G. Natale, *Nucl. Instr. Meth. Phys. Res. A* **289**, 518 (1990).

[74] J. Hough, G. P. Newton, N. A. Robertson, H. Ward, A. M. Campbell, J. E. Logan, D. I. Robertson, K. A. Strain, K. Danzmann, H. Lück, A. Rüdiger, R. Schilling, M. Schrempel, W. Winkler, J. R. J. Bennett, V. Kose, M. Kühne, B. F. Schutz, D. Nicholson, J. Shuttleworth, H. Welling, P. Aufmuth, R. Rinkleff, A. Tünnermann, and B. Willke, in *Proceedings of the Seventh Marcel Grossmann Meeting on General Relativity*, eds. R. T. Jantzen and G. M. Keiser (World Scientific, Singapore, 1996), p. 1352.

[75] P. S. Linsay, P. R. Saulson, and R. Weiss, *A Study of a Long Baseline Gravitational Wave Antenna System* (MIT, Cambridge, Mass., 1983), unpublished.

[76] R. E. Vogt, R. W. P. Drever, K. S. Thorne, and R. Weiss, *Caltech/MIT Project for a Laser Interferometer Gravitational Wave Observatory*, proposal to the National Science Foundation (Caltech, Pasadena, 1987), unpublished.

[77] K.-X. Sun, M. M. Fejer, E. Gustafson, and R. L. Byer, *Phys. Rev. Lett.* **76**, 3053 (1996).

[78] R. E. Vogt, R. W. P. Drever, F. J. Raab, K. S. Thorne, and R. Weiss, *The Construction, Operation, and Supporting Research and Development of a Laser Interferometer Gravitational-Wave Observatory*, proposal to the National Science Foundation (Caltech, Pasadena, 1989), unpublished.

GRAVITATIONAL RADIATION SOURCES AND SIGNATURES

Lee Samuel Finn*
Center for Gravitational Physics and Geometry†, The Pennsylvania State University, University Park, Pennsylvania 16802

1 Introduction

The goal of these lecture notes is to introduce the developing research area of *gravitational-wave phenomenology.* In more concrete terms, they are meant to provide an overview of gravitational-wave sources and an introduction to the interpretation of real gravitational wave detector data. They are, of course, limited in both regards. Either topic could be the subject of one or more books, and certainly more than the few lectures possible in a summer school. Nevertheless, it is possible to talk about the problems of data analysis and give something of their flavor, and do the same for gravitational wave sources that might be observed in the upcoming generation of sensitive detectors. These notes are an attempt to do just that.

Despite an 83-year history, our best theory explaining the workings of gravity—Einstein's theory of general relativity—is relatively untested compared to other physical theories. This owes principally to the fundamental weakness of the gravitational force: the precision measurements required to test the theory were not possible when Einstein first described it, or for many years thereafter.

The direct detection of gravitational-waves is a central component of our first investigations into the dynamics of the weakest of the known fundamental forces: gravity. It is only in the last 35 years that general relativity has been put to significant test. Today, the first effects of static relativistic gravity beyond those described by Newton have been well-studied using precision measurements of the

© 1998 by Lee Samuel Finn.
*Supported by National Science Foundation awards PHY 98-00111 and PHY 95-03084 to The Pennsylvania State University, and PHY 93-08728 to Northwestern University.
†Also Department of Physics, and Department of Astronomy and Astrophysics.

motion of the planets, their satellites, and the principal asteroids. Dynamical gravity has also been tested through detailed and comprehensive observations of the slow, secular decay of the Hulse-Taylor binary pulsar system.[1,2] What has not heretofore been possible is the direct observation of the effects of dynamical gravity on a laboratory instrument, *i.e.,* the direct detection of gravitational radiation.

The scientific importance of the direct detection of gravitational-waves does not stop with its detection, however. Strong gravitational-waves are difficult to generate: so difficult, in fact, that there is no possibility of a gravitational-wave "Hertz"-type experiment, where both the source and receiver are under laboratory control. The strongest gravitational-waves incident on Earth, as measured by our ability to detect them in the sensitive detectors now under construction, arise from astronomical sources. These are also the only sources that we can hope to observe in our detectors. The strongest of these anticipated sources—inspiraling or colliding neutron stars or black holes—are, in fact, of cosmological origin.

Very little relevant detail is known about the gravitational-wave sources that we anticipate may be detectable in the instruments now under construction. Estimates of source strengths and event rates are difficult to make reliably. This is because, at a deep and fundamental level, our understanding of the cosmos is limited to what we can learn from photons. The mechanism by which gravitational-waves are generated, on the other hand, favors sources that either do not radiate electromagnetically (*e.g.,* black holes), are obscured from view (*e.g.,* the gravitational collapse of a stellar core), or are so distant and decoupled from the immediate origin of the corresponding electromagnetic radiation that we cannot reliably decipher the relevant source characteristics from the photons that reach us (*e.g.,* γ-ray bursts).[3]

Gravitational-wave observations thus add a new dimension to our ability to observe the Universe: the observations that we make will tell us things we don't already know through other means.

In order to describe sensibly the signature of gravitational-wave sources in real detectors we must first discuss in some detail how we characterize gravitational-waves, how we characterize gravitational-wave detectors, and how we give operational meaning to the word "detect." These three subjects are addressed in Secs. 2, 3, and 4, respectively. In the context of gravitational-wave detection, gravitational-wave signals fall fairly neatly into three categories: burst signals, periodic signals, and stochastic signals. Sources thought to be responsible for detectable signals in these categories are described in Secs. 5.1, 5.2, and 5.3, respectively.

1.1 Conventions

- The distance from detector to source will always be large compared to either a wavelength of the radiation field or the physical dimension of the detector; consequently, the incident radiation is effectively planar.

- We choose a sign convention for the line element of Minkowskii spacetime and recall the *Einstein summation convention,* wherein repeated Greek indices in a product are implicitly summed over their full range:

$$\begin{aligned} ds^2 &= \eta_{\mu\nu}\, dx^\mu\, dx^\nu, && (1)\\ &= \sum_{\mu,\nu=0}^{3} \eta_{\mu\nu}\, dx^\mu\, dx^\nu, && (2)\\ &= -c^2\left(dx^0\right)^2 + \left(dx^1\right)^2 + \left(dx^2\right)^2 + \left(dx^3\right)^2, && (3)\\ &= -c^2dt^2 + dx^2 + dy^2 + dz^2. && (4) \end{aligned}$$

- We will always treat the gravitational fields as weak and use coordinates on spacetime that are either "Cartesian"+time or "spherical-polar"+time.

- We will find it convenient to introduce the use of Latin indices to represent just the spatial components (*i.e.,* x, y, and z) of a tensor. We generalize the Einstein summation convention to apply to repeated Latin indices in a product expression, with the implicit sum running over just the spatial coordinates, *e.g.,*

$$x^i y_i = \sum_{j=1}^{3} x^j y_j. \qquad (5)$$

- Except in the first several sections of these lecture notes, we will always work in units where the Newtonian gravitational constant G and the speed of light c are numerically equal to unity and the appearance of these constants in various formulae will be suppressed. Dimensional analysis will always suffice to determine how the formulae should appear with the constants in place.

- Using c we can express time in units of length and frequency in units of inverse length; similarly, by exploiting G and c we can express mass and

energy in units of length. Power is then a dimensionless number. For CGS units, the conversion factors between mass, energy, and length, and the physical constant with units of power, are

$$G/c^2 = 7.42 \times 10^{-29}\,\mathrm{cm/gm}, \quad (6)$$

$$G/c^4 = 8.26 \times 10^{-50}\,\mathrm{cm/erg}, \quad (7)$$

$$c^5/G = 3.63 \times 10^{59}\,\mathrm{erg/s}. \quad (8)$$

2 Characterizing Gravitational Radiation

For our purpose here—recognizing gravitational waves incident on a detector—two different characterizations of gravitational radiation are useful. The first is the radiation waveform and the second is the signal "power spectrum." The waveform describes the radiation field's time dependence while the power spectrum describes its Fourier components. In Secs. 2.1 and 2.3 we describe these different characterizations of gravitational radiation. Several important physical insights regarding gravitational radiation sources can be gained by considering the instantaneous power radiated by a source: we discuss these insights in Sec. 2.2.

2.1 Radiation Waveform

In this subsection we review briefly the expression of the radiation incident on a detector. Much of this section is by way of review; for more details, see either the lectures by Bob Wagoner in this collection, one of the many textbooks on relativity,[4–8] or an excellent review article on the subject.[9]

Gravitation manifests itself as spacetime curvature and gravitational waves as ripples in the curvature that appear to us, moving through time, to be propagating. Detectors are generally not directly sensitive to curvature, but to the mechanical displacement of their components; so, we focus our attention on the spacetime metric, from which physical distances between points in spacetime are determined. (The curvature is a function of the metric's second derivatives.)

We assume that gravity is weak in and around our detector; correspondingly, we treat the spacetime metric as if it were the metric of Minkowskii spacetime, plus a small perturbation:

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}, \quad (9)$$

where $\eta_{\mu\nu}$ is the Minkowskii metric and $h_{\mu\nu}$ the metric perturbation. The corresponding line element, describing the proper distance between nearby spacetime events whose coordinate separation is the infinitesimal dx^μ, is

$$ds^2 = g_{\mu\nu}dx^\mu dx^\nu = \eta_{\mu\nu}dx^\mu dx^\nu + h_{\mu\nu}dx^\mu dx^\nu. \quad (10)$$

Detecting gravitational waves amounts to building instruments that are sensitive to the effects of the small perturbation $h_{\mu\nu}$; determining the signature of the gravitational waves in the detector thus requires determining $h_{\mu\nu}$ and its influence on the detector.

The metric $g_{\mu\nu}$ tells us how the proper distance between points in spacetime is associated with our choice of coordinate system. Since the gravitational fields near our detector are weak and the spacetime nearly Minkowskii, we can introduce coordinates that are, in the neighborhood of the detector, nearly the usual Minkowskii-Cartesian coordinates, with the deviations of the order of the perturbation.

Now, small changes in the coordinate system do not change the proper distance between events, only our labeling of them. If we make small changes in our coordinate system, of the order of the perturbation, then we will make corresponding changes in the perturbation $h_{\mu\nu}$. We can use this freedom to simplify the expression of $h_{\mu\nu}$. Coordinate changes do not change the physics or any observable constructed from $h_{\mu\nu}$, of course. For this reason, and in analogy with electromagnetism, coordinate choices like these are referred to as *gauge* choices.

With the separation of the metric into the Minkowskii metric $\eta_{\mu\nu}$ plus a small perturbation, the field equations of general relativity become (at first order in the perturbation) a set of second order, linear differential equations for the ten components of the symmetric $h_{\mu\nu}$. Consequently, fixing the coordinates allows us to impose eight conditions on the ten components of the symmetric $h_{\mu\nu}$, leaving just two dynamical degrees of freedom. These are identified as the two polarizations of the gravitational radiation field.

An important gauge choice, always possible for radiative perturbations about Minkowskii space, is the *Transverse-Traceless,* or TT, gauge. The Transverse-Traceless gauge is always associated with a particular observer of the radiation. Let the four-velocity of this observer have components U^μ. Without loss of generality let t mark the proper-time of this observer (so that U^μ is just the coordinate vector in the t direction) and x, y, and z be the usual Cartesian coordinates (to

$\mathcal{O}(h)$) in the neighborhood of the observer. In TT-gauge the field equations are

$$\left(-\frac{\partial^2}{\partial t^2}+\nabla^2\right)h_{\mu\nu}=-16\pi T_{\mu\nu}, \tag{11}$$

subject to the constraints

$$\begin{aligned} h_{\mu\nu}U^\mu &= 0, & (12)\\ \frac{\partial}{\partial x^k}\left(h_{j\mu}\eta^{\mu k}\right) &= h_j{}^k{}_{,k}=0, & (13)\\ h_{jk}\delta^{jk} &= 0. & (14)\end{aligned}$$

The metric perturbation satisfies a wave equation whose source is the *stress-energy density* $T_{\mu\nu}$ of the matter and (non-gravitational) fields. In more physical terms, the constraints are (in order):

- $h_{\mu\nu}$ is purely spatial;
- the (spatial) metric perturbation h_{ij} is purely transverse: *i.e.,* if the radiation wavevector is k^i (where the index i runs over just the spatial dimensions; see §1.1), then $h_{ij}k^i$ vanishes for all j; and
- the metric perturbation is *trace-free.*

When there might be confusion we denote a metric perturbation in TT-gauge with a superscript TT on $h_{\mu\nu}$; also, since the perturbation is purely spatial, we generally refer just to the spatial projection (in the coordinate system of the observer) of the perturbation, as in h^{TT}_{ij}.

Given a metric perturbation $h'_{\mu\nu}$, expressed in any gauge, corresponding to a plane wave propagating in the direction n^k, we can recover the corresponding metric perturbation in TT-gauge by applying the linear operator P_{lm} to the spatial h'_{ij}:

$$\begin{aligned} h^{TT}_{ij} &= P^l_i h'_{lm}P^m_j-\frac{1}{2}P_{ij}P^{lm}h'_{lm} & (15)\\ P_{lm} &= \delta_{lm}-n_l n_m. & (16)\end{aligned}$$

Here and henceforth we will always express the metric perturbation in TT-gauge.

As mentioned above, gravitational wave detectors work by sensing the relative motion of their components induced by a passing gravitational wave. Let's see how the TT-gauge metric perturbation is related to such relative motion.

Consider a single isolated test mass, initially at rest at coordinate position $\vec{x}_A$ in a TT-gauge coordinate system. No forces act on this test mass; so, it moves through spacetime in such a way that its four-velocity always remains tangent to itself. (Forces, of course, cause the four-velocity to change direction.) The corresponding equations of motion for the spatial coordinates of the test mass are

$$\frac{d}{d\tau}\left(\frac{dx^i_A}{d\tau}\right)+\Gamma^i_{\mu\nu}\frac{dx^\mu_A}{d\tau}\frac{dx^\nu_A}{d\tau}=0, \tag{17}$$

where τ is the proper time of the test mass (initially τ is equal to t since the test mass is at rest) and $\Gamma^\alpha_{\beta\gamma}$ is the metric connection

$$\Gamma^\alpha_{\beta\gamma}=\frac{1}{2}g^{\alpha\mu}\left(g_{\mu\beta,\gamma}+g_{\mu\gamma,\beta}-g_{\gamma\beta,\mu}\right). \tag{18}$$

(Recall that $_{,k}$ represents the derivative $\partial/\partial x^k$.) Since the test mass is initially at coordinate rest the $dx^i/d\tau$ vanish initially; so, the only connection component of interest is Γ^i_{tt}. In TT-gauge, however, this component of the connection is identically zero (recall that $h^{TT}_{\mu\nu}$ is purely spatial); so, the equations of motion reduce to

$$\frac{d}{d\tau}\frac{dx^i_A}{d\tau}=0; \tag{19}$$

i.e., a free test particle at (TT-gauge) coordinate rest remains at coordinate rest. This applies equally well for a second component of the detector, located at $\vec{x}_B$: it, too, remains at coordinate rest.

This may seem paradoxical: if the coordinate separation of any two components of a detector remain unchanged by the passage of a gravitational wave, what is there to show the wave's existence? The paradox vanishes when we realize that coordinate separation is not physical separation. To determine the physical separation of the detector's components, we must invoke the metric again. Let the coordinate separation between the two components of the detector at time t be the infinitesimal dx^i_{AB},

$$dx^i_{AB}=x^i_B-x^i_A. \tag{20}$$

(Of course, $dt=0$ since we are talking about separation at the same coordinate time.) The physical distance between these two neighboring points in spacetime is

$$\begin{aligned} ds^2 &= g_{\mu\nu}dx^\mu dx^\nu, & (21)\\ &= \eta_{\mu\nu}dx^\mu_{AB}dx^\nu_{AB}+h^{TT}_{\mu\nu}dx^\mu_{AB}dx^\nu_{AB}, & (22)\\ &= \eta_{jk}dx^j_{AB}dx^k_{AB}+h^{TT}_{jk}dx^j_{AB}dx^k_{AB}. & (23)\end{aligned}$$

The second term in equation 23 shows the effect of the gravitational wave on the separation between the two elements of the detector: as $h_{\mu\nu}$ oscillates, so does the distance. If the equilibrium separation between the components is L in the direction $\hat{n}^j$, to $\mathcal{O}(h)$ the net change δL in the separation is equal to

$$\delta L = \frac{1}{2} L h^{TT}_{jk} \hat{n}^j \hat{n}^k. \tag{24}$$

The physical distance between detector components does change, in an amount proportional to the undisturbed separation and the wave strength as projected on the separation. Gravitational wave detectors are designed to be sensitive to this displacement of their components.

As mentioned above, the TT-gauge conditions amount to eight constraints on the ten otherwise independent components of the (symmetric) $h_{\mu\nu}$. There are thus two components of $h_{\mu\nu}$ that are independent of the choice of coordinate system; correspondingly, in general relativity there are two dynamical degrees of freedom of the gravitational field. To see what these amount to, without loss of generality consider a plane wave propagating in the z direction. Then we can write

$$h^{TT}_{\mu\nu} dx^\mu dx^\nu = h_+(x^i, t)\left(dx^2 - dy^2\right) + 2h_\times(x^j, t) dx\, dy, \tag{25}$$

where h_+ and $h_\times$ are the two independent dynamical degrees of freedom, or polarizations, of the gravitational radiation field.

Solutions to the wave equation for h_{ij} (eq. 11) can be analyzed in a slow motion expansion in exactly the same way as solutions to the Maxwell equations.[9–11] The radiative h_{ij} in this expansion divide neatly into two classes of multipolar fields, which are (in analogy with electromagnetism) termed *electric* and *magnetic* multipoles. The electric multipolar radiative fields are generated by time-varying multipole moments of the source matter density in the same way that the analogous electric moments of the Maxwell field are generated by the time-varying moments of the electric charge density. Similarly, the magnetic radiative moments are generated by the time-varying multipole moments of the matter momentum density, which is the analog of the electric current density.

In electromagnetism, the first radiative moment of a charge distribution is a time-varying charge dipole moment. When electrical charge is replaced by gravitational charge—*i.e.*, mass—we see that the corresponding dipole is just the position of the system's center of mass, which (owing to momentum conservation) is unaccelerated. Consequently, in general relativity there is no gravitational dipole radiation. The first gravitationally radiative moment of a matter distribution arises from its "accelerating" quadrupole moment. Dotting the i's and crossing the t's, we find that, at leading order, the radiation field at a distant detector is related to the matter distribution of the source according to

$$h^{TT}_{ij}(t, \vec{x}) = \frac{2}{r}\frac{G}{c^5}\frac{d^2}{dt^2} Q^{TT}_{ij}(t - r) \tag{26}$$

$$Q^{TT}_{ij} = P_{ik}(\vec{x}) Q_{kl} P_{lj}(\vec{x}) - \frac{1}{2} P_{ij}(\vec{x}) Q_{lm} P_{lm}(\vec{x}) \tag{27}$$

$$Q_{ij}(t) = \int d^3x\, \rho(t, \vec{x}) \left(x_i x_j - \frac{1}{3}\delta_{ij}\right), \tag{28}$$

$$P_{jk}(\vec{x}) = \delta_{jk} - x_j x_k / x^2. \tag{29}$$

The expression for h^{TT}_{ij} given above is the famous "quadrupole formula" of general relativity, which relates the acceleration of a source's quadrupole moment to the gravitational radiation emitted. It is, for weak gravitational fields, the exact analog of the more famous "dipole formula" of electromagnetism.

2.2 Radiated Power or Energy

Gravitational radiation carries energy away from the radiating system. Important insights into gravitational radiation can be gained by considering the energetics of radiation sources, which we do in this section.

The instantaneous power carried by the radiation is, in the usual way, proportional to the square of the time derivative of the field integrated over a sphere surrounding the source:

$$L \propto \frac{c^5}{G} 4\pi r^2 \dot{h}^2. \tag{30}$$

The "exact"* expression for the power carried away in electric quadrupole radiation is

$$L = \frac{1}{5}\frac{G}{c^5}\left\langle \frac{d^3 Q_{ij}}{dt^3}\frac{d^3 Q^{ij}}{dt^3}\right\rangle, \tag{31}$$

where the <> indicates an average several periods of the radiation. Note that the power depends on Q_{ij} and *not* Q^{TT}_{ij}.

If we focus on the radiation emitted by a weak-field, dynamical source, we can use the multipolar expansions described above to replace the fields by the

*In the context of our approximation of everywhere weak gravitational fields.

multipole moments of the source. For an electric ℓ-pole field radiated by a matter source with mass M, typical dimension R, and internal velocity V,

$$L^{(\ell)}_{\text{electric}} \propto \frac{c^5}{G}\left[\frac{GM}{c^2R}\left(\frac{V}{c}\right)^{\ell+1}\right]^2; \tag{32}$$

similarly, for a magnetic ℓ-pole field

$$L^{(\ell)}_{\text{magnetic}} \propto \frac{c^5}{G}\left[\frac{GM}{c^2R}\left(\frac{V}{c}\right)^{\ell+2}\right]^2. \tag{33}$$

The total power radiated is the sum over the power radiated in each of the multipoles.

Aside from numerical factors and symmetries, power radiated in the electric ℓ-pole channel is suppressed relative to that in the electric quadrupole channel by a factor of $(V/C)^{2(\ell-2)}$. Similarly, the radiation in the magnetic ℓ-pole channel is suppressed from the electric quadrupole radiation by a factor of $(V/C)^{2(\ell-1)}$. Consequently, sources whose internal velocities are significantly less than the speed of light radiate principally in the electric quadrupole channel (again, unless suppressed by symmetries).

There is still another way of looking at the power radiated by a gravitational radiation source. For the gravitational wave detectors we can hope to build, all the radiation of interest is of astrophysical origin. Excepting only a stochastic gravitational wave background, the radiation sources are all distinct systems whose structure or dynamics are governed by gravity. For these systems, judicious application of the Virial Theorem[12] allows us to relate the internal velocities V to the depth of the gravitational potential GM/R,

$$V^2 \sim \frac{GM}{R}. \tag{34}$$

Thus, for astrophysical sources

$$L^{(\ell)}_{\text{electric}} \propto \frac{c^5}{G}\left(\frac{V}{c}\right)^{2(\ell+3)} \simeq \frac{c^5}{G}\left(\frac{GM}{c^2R}\right)^{\ell+3}, \tag{35}$$

$$L^{(\ell)}_{\text{magnetic}} \propto \frac{c^5}{G}\left(\frac{V}{c}\right)^{2(\ell+4)} \simeq \frac{c^5}{G}\left(\frac{GM}{c^2R}\right)^{\ell+4}. \tag{36}$$

Strong gravitational wave sources thus have strong internal gravitational fields.

Finally, dimensional analysis of equation 31 for the power radiated in the electric quadrupole leads to an important physical insight. Dimensionally, the system's quadrupole moment is proportional to MR^2. In a closed, radiating system there is a typical time scale T for motion within the source; consequently, the total radiated power can be written

$$L \propto \frac{MR^2/T^3}{c^5/G}\frac{MR^2}{T^3}. \tag{37}$$

The quantity $MR^2/T^2 \simeq MV^2$ can be interpreted as the kinetic energy of source matter *engaged in motion associated with a time-varying quadrupolar moment.* Similarly, we identify MR^2/T^3 as the instantaneous power *available to be radiated.* Not all this power is radiated, however. Equation 37 shows that the *fraction* of the available power actually radiated is equal to the available power divided by a "fundamental power" defined by the physical constants G and c:

$$\frac{c^5}{G} = 3.6 \times 10^{59}\,\text{erg/s}. \tag{38}$$

The magnitude of this fundamental power gives us a feeling for the weakness of the gravitational interaction. For a source to radiate even one part in 10^9 of the power available to be tapped by the radiation field, it must have internal motions where the kinetic energy involved in quadrupolar motion is greater than 3.6×10^{50} erg/s. For scale, this is 10^{27} times greater than the power liberated in all of the nuclear reactions occurring in the Sun!

2.3 Signal Power Spectrum

Observations of gravitational wave signals are always of finite duration: either the signal is a burst of duration less than the observation period or the signal duration is determined by the period between when the detector is turned on and when it is turned off. A useful characterization of this *observed signal* is its *spectrum:* the contribution to the overall mean square signal amplitude owing to its Fourier components at a given frequency.

For definiteness, focus attention upon some particular polarization $h(t)$ of a gravitational-wave signal that is observed over a period T beginning at $t = 0$. The Fourier transform of this signal is $\tilde{h}(f)$:†

$$\tilde{h}(f) = \int_0^T dt\, e^{-2\pi i f t} h(t). \tag{39}$$

†We use the engineering convention for the Fourier transform.

The signal spectrum is evaluated for positive frequencies and is twice the square modulus of its Fourier transform averaged over the observation, or

$$P_h(f) := \frac{2}{T}|\tilde{h}(f)|^2 \tag{40}$$

for non-negative f. Since $h(t)$ is real, we can use The Parseval Theorem to obtain

$$\int_0^\infty df\, P_h(f) = \frac{1}{T}\int_0^T dt\, |h(t)|^2 = \left\langle h^2 \right\rangle , \tag{41}$$

where $\langle\cdot\rangle$ denotes a time average. The signal spectral density is thus a measure of the contribution to the mean-square signal amplitude owing to its Fourier components in a unit bandwidth. (For non-burst—*i.e.,* stochastic or periodic—signals, we often take the limit as $T \to \infty$.)

As we have described it, the signal spectrum is derived from the signal waveform $h(t)$ by "throwing away" the phase information. There is clearly much less information in $P(f)$ than in the corresponding $h(t)$: why, then, is $P(f)$ an interesting characterization of a signal?

One reason is that real detectors are only sensitive to radiation in a limited bandwidth—*i.e.,* at certain frequencies. The integral of the signal power spectrum over the detector bandwidth is the contribution to the mean-square amplitude of h from power in the detector bandwidth.

A second reason is that it is not always possible to determine the waveform of a gravitational wave signal. For example, the waveform of a stochastic signal, arising from a primordial background or from the confusion limit of a large number of weak sources, is intrinsically unknowable. Nevertheless, the signal spectrum is straightforward to calculate. In this case, the spectrum embodies everything we can know about the gravitational wave signal.

Another example illustrates a different circumstance. Calculations of gravitational radiation waveforms $h_+(t)$ and $h_\times(t)$ from the kind of stellar core collapse that triggers type II supernovae are, even in their grossest details, extremely sensitive to the details of the stellar model and the physics included in the simulations. In the face of this variety of structure, however, the spectra all show a remarkable similarity.[13] It may be that this variety reflects our ignorance of the relevant physics and that with better understanding the waveforms would show much less variation and much greater predictability; it may also be that the details of the collapse waveform are in fact very sensitive to the initial conditions. Whether in practice or in principle, the waveform is today unknown; nevertheless, the spectrum does appear to characterize the signal quite well.

We close with a final reason that the spectrum is a useful characterization of a gravitational wave signal. The sensitivity of a gravitational wave detector is limited by the detector noise, which is an intrinsically stochastic process. In the best detectors, the noise is fully characterized by *its* spectrum (cf. 3.5). We expect intuitively that a signal is detectable only when its spectrum has greater magnitude than the detector noise spectrum over a sufficient range of frequencies. This qualitative notion finds quantitative expression in the *signal-to-noise* ratio, which we discuss in Sec. 3.5 below.

2.4 Conclusion

For the purposes of detection, gravitational waves are usefully characterized by their waveform or spectrum. There are important sources for which the explicit waveform is not known, either because it is intrinsically unknowable, our grasp of the underlying physics is not complete, or the calculations involved in determining it are beyond our capabilities. In these cases, it may still be possible to estimate the signal spectrum, which then serves to characterize it.

3 Characterizing The Detector

3.1 Introduction

Gravitational-wave detectors transform incident gravitational waves into, *e.g.,* electrical signals that we can more easily manipulate. In Sec. 3.2, we describe briefly and schematically two of the detector technologies currently being pursued to detect gravitational waves. For all detectors we might realistically imagine building, the detector response is linear in the incident radiation: *i.e.,* the time history of the detector output is linearly related to the time history of the incident radiation.

There are two aspects of this response that we must consider: differential sensitivity to the radiation incident from different directions, and differential sensitivity to incident radiation of different frequencies. The first of these is described by the detector's *antenna pattern,* which we discuss in Sec. 3.3, and the second of these

is described by the detector's *response function,* which we discuss briefly in Sec. 3.4.

The output of a gravitational wave detector *might* contain a particular gravitational wave signal; however, it *always* contains noise. Detection, discussed in Sec. 4 below, involves distinguishing gravitational wave signals from detector noise. To make this distinction we must have some characterization of the signal (e.g., by waveform or by spectral density) and detector noise. How we characterize detector noise is the subject of Sec. 3.5.

3.2 Gravitational Wave Detectors

3.2.1 Acoustic Detectors

The earliest and most mature detection technology is, conceptually, nothing more than a high quality tuning fork. Gravitational waves excite the tuning fork; gravitational waves at or near the tuning fork resonant frequency excite it into large amplitude oscillations. The tuning fork is instrumented so that its acoustic vibrations become electrical signals, which, when amplified, are the gravitational wave signal.

Physically, the tuning fork is realized as one or more normal modes of a large metal-alloy block: the fundamental longitudinal mode of a right cylinder for the currently operating detectors, the five quadrupole modes of a sphere or a truncated icosahedron[14] for the proposed next generation detectors. The choices made in the construction of the ALLEGRO[15,16] detector, built and operated at the Louisiana State University, are typical for the current generation of these right cylindrical "bars": diameter of 60 cm, length of 3 m, and cast of Al5056 alloy for a total mass of 2296 Kg.

The mechanical oscillations of the tuning fork are converted into electrical signals, which are then amplified, digitized, and otherwise manipulated to determine whether gravitational waves are present or absent. In all of the high-sensitivity bar detectors operating today, the transducer is not directly connected to the bar, but instruments a second mechanical oscillator, of lower mass and smaller physical dimension, that is itself coupled to the bar. Gravitational waves drive the bar, which in turn drives the second oscillator. In the process, the amplitude of the mechanical vibrations are amplified, and it is this mechanically amplified motion that is converted into electrical signals and further amplified, *etc.* The coupling of the two mechanical oscillators splits the fundamental longitudinal mode of the larger bar into two closely-spaced modes. For the ALLEGRO detector, the antenna's normal modes are at 896.8 Hz and 920.3 Hz.

At this writing there are five operating cryogenic acoustic gravitational wave detectors:

- ALLEGRO, at the Louisiana State University in the United States,[16]
- AURIGA, at the University of Padua in Italy,
- EXPLORER, operated by the Rome group and located at CERN,[17]
- NAUTILUS, operated by the Rome group and located at the Frascati INFN Laboratory,[17] and
- NIOBE, at the University of Western Australia.[18]

In addition to these classical "bar" detectors, several spherical or truncated icosahedral detectors have been proposed or are undergoing technical development: SFERA, TIGA,[14,19,20] GRAIL,[21] and OMNI.[22]

3.2.2 Interferometric Gravitational Wave Detectors

An alternative technology for the detection of gravitational waves involves the use of a right-angle Michelson interferometer with freely suspended mirrors. Gravitational waves which are incident normal to the plane of an interferometer will lead to differential changes in the distance between the corner and end mirrors. For frequencies much less than the reciprocal of the light storage time in an interferometer arm, the corresponding motion of the fringes is proportional to the incident radiation waveform.

There are currently two Km-scale interferometer projects under construction: the French/Italian VIRGO Project[23] and the US LIGO Project.[24] VIRGO will consist of a single interferometer with 3 Km long arms situated just outside of Pisa, Italy. LIGO will consist of two separate facilities, one at Hanford, Washington and one in Livingston, Louisiana. Each LIGO facility will house an interferometer with arms of length 4 Km; in addition, the Hanford facility will also hold an interferometer of 2 Km arm length in the same vacuum system.

In addition to these larger interferometer projects, there are three more interferometric detectors of somewhat smaller scale under construction: the Australian

ACIGA project, the German/UK GEO 600 project[25] and the Japanese TAMA 300 project. The ACIGA Project's ultimate goal is a multi-kilometer detector, to be located several hours outside of Perth; presently, they are beginning the construction of an approximately 80 m prototype at the same site. GEO 600, located in Hanover, Germany, is a folded Michelson interferometer with an optical arm length of 1.2 Km. The Japanese TAMA 300 is a 300 m Fabrey-Perot interferometer located just outside of Tokyo; it is hoped that the success of this project will lead to the construction of the proposed Laser Gravitational Radiation Telescope (LGRT), which would be located near the Super-K neutrino detector.

There are several ways to make an interferometer more sensitive at frequencies less than the reciprocal of the detector's light storage time. One is to increase its arm length (recall equation 24!). The Laser Interferometer Space Antenna—LISA—is an ambitious project to place in solar orbit a constellation of satellites that will act as an interferometric gravitational wave detector.[26,27] The arm length of this interferometer would be 5×10^6 Km. The LISA project has been approved by the European Space Agency as part of its Horizon 2000+ Program; additionally, the US National Aeronautics and Space Administration is actively considering joining ESA as a partner to accelerate the development and launch of this exciting project.

3.3 Antenna Pattern

Gravitational wave detectors respond linearly to the applied field. The interferometric and bar gravitational wave detectors now under construction or in operation have only a single "gravitational wave" output channel.[‡] When a plane wave is incident on such a detector, the time history of the output channel is linearly related to a superposition $h(t)$ of the + and × polarizations of the incident plane wave:

$$h = F_+ h_+ + F_\times h_\times. \tag{42}$$

The factors F_+ and $F_\times$ describe the detector's "antenna pattern," or differential sensitivity to radiation of different polarizations incident from different directions. (In fact, the antenna pattern may also be a function of radiation wavelength; however, when the wavelength is much larger than the detector this dependence is insignificant.) They depend on relative orientation of the plane-wave's propagation direction and the definition of the + and × polarizations.

[‡]Some proposed acoustic detectors are instrumented on several independent modes. In this case, each mode may be considered a separate detector and represented as a single gravitational wave channel.

If we fix the propagation direction and rotate the polarization of the incident radiation, then the detector response $h(t)$ will change. Define the polarization averaged root-mean-square (RMS) antenna pattern F,

$$F^2(\vec{k}) = \overline{F_+^2(\vec{k}) + F_\times^2(\vec{k})}, \tag{43}$$

where $\vec{k}$ is the wave-vector of the incident plane wave and the overline denotes an average over a rotation of the incident radiation's polarization plane. The result depends only on the wave-vector (or, alternatively, the wave's propagation direction and wavelength) and is proportional to the detector's root-mean-square response to plane-wave radiation incident from a fixed direction at fixed wavelength. For either the acoustic or interferometric detectors now operating or under construction, $F(\vec{k})$ is independent of the magnitude of $\vec{k}$ as long as the radiation wavelength is much larger than the detector.

A convenient pictorial representation of the detector's response results if we plot the surface defined by $\hat{n}F(\vec{k})$ for fixed $|\vec{k}|$, where $\hat{n}$ is the unit vector in the direction of the source relative to the detector (*i.e.*, $\hat{n} = -\vec{k}/k$). In such a figure, the response of the detector to a plane wave with wave-vector $\vec{k}$ (appropriately averaged over polarization) is proportional to the distance of the surface from the origin in the direction of the source ($\hat{n}$). In the remainder of this subsection we describe the antenna pattern of interferometric and acoustic bar detectors to incident gravitational plane waves.

3.3.1 Bar Detectors

In a classic bar detector, incident gravitational waves drive the fundamental longitudinal mode of a right cylindrical bar. The driving force—and thus the radiation—is determined by observing the motion of this mode. For definiteness, let the longitudinal axis of the bar be along the $\hat{z}$-direction, and consider a plane gravitational wave propagating in the $\hat{x}$-direction:

$$h_{ij}^{TT}(t) dx^i dx^j = h_+(t) \left(dy^2 - dz^2 \right) + 2h_\times(t) dy\, dz. \tag{44}$$

The + polarization mode changes the z-distance between the atoms in the bar. This change is resisted by inter-atomic forces in the bar; thus, the bar's longitu-

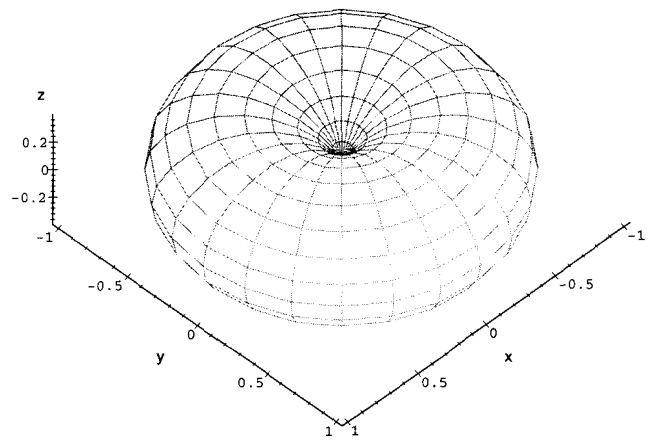


Fig. 1. The polarization-averaged RMS sensitivity of a bar detector to gravitational waves incident from any direction. The detector is at the origin of the figure and has its symmetry axis along the figure's z axis. The magnitude of the distance from the origin to the surface in a direction $\hat{n}$ is proportional to the relative response of the detector to radiation incident on the detector from that direction.

dinal normal mode is driven by this polarization component of the incident wave. The $\times$ polarization, on the other hand, does not excite the bar's mode in this way; so, the detector is insensitive to this component of the incident radiation.

If, on the other hand, the waves are incident along the *longitudinal* axis, then neither the $\times$ nor the $+$ polarization components cause any change in the longitudinal distance between the atoms in the bar; correspondingly, the bar is insensitive to waves incident along on the bar along its axis.

Finding the response to radiation of different polarizations incident from directions intermediate between these extremes is a relatively straightforward exercise in geometry. First define the bar's coordinate system. Let the bar's symmetry axis define the $\hat{z}$ direction and choose $\hat{x}$ and $\hat{y}$ such that $(\hat{x}, \hat{y}, \hat{z})$ defines a right-handed coordinate system.

Next consider a plane gravitational wave propagating in the $\hat{k}$ direction and define the polarizations of the gravitational wave. Introduce a plane orthogonal to $\hat{k}$. In this plane, spanned by the Cartesian coordinates $\hat{x}'$ and $\hat{y}'$, the TT-gauge metric perturbation can be written

$$h_{ij}dx^i dx^j = h_+\left(dx'^2 - dy'^2\right) + 2h_\times dx'dy'. \tag{45}$$

The choice of $\hat{x}'$ and $\hat{y}'$ is arbitrary: different choices correspond to either or both a rotation of one polarization state into another or a reflection that flips the sign of $h_\times$. For definiteness, we choose $\hat{x}'$ to be orthogonal to $\hat{z}$ and $\hat{y}'$ such that $(\hat{x}', \hat{y}', \hat{k})$ is a right-handed coordinate system. (In the degenerate case, where $\hat{k}$ is parallel to $\hat{z}$, we choose $\hat{x}'$ parallel to $\hat{x}$ and $\hat{y}'$ parallel to $\hat{y}$.) With these choices, it is straightforward to show that the response of the bar's longitudinal mode is proportional to h, where

$$h = F_+ h_+ + F_\times h_\times, \tag{46}$$

$$F_+ = \sin^2\theta, \tag{47}$$

$$F_\times = \sin^2\theta, \tag{48}$$

$$\cos^2\theta = \left(\hat{k}\cdot\hat{n}\right)^2. \tag{49}$$

Figure 1 shows the polarization-averaged RMS sensitivity of a right cylindrical acoustic detector to plane waves incident from a given direction. To interpret the figure, imagine a detector at the figure's origin with its symmetry axis coincident with the figure's z axis. A plane wave, arriving from direction $\hat{n}$, leads to a detector response proportional to *the distance in the direction $\hat{n}$ from the figure's origin to the surface.*

3.3.2 Interferometric Detectors

Interferometric gravitational wave detectors respond when incident gravitational waves cause a differential change in the length of the interferometer arms. Focus attention on interferometers whose arms meet in a right angle. To get a sense of the differential sensitivity of such a detector to radiation of different polarizations incident from different directions, define a right-handed interferometer coordinate system whose origin is the intersection of the arms and whose x and y coordinate directions are in the direction of the arms. Let a plane wave, described by the perturbation

$$h_{ij}dx^i dx^j = h_+\left(dx^2 - dy^2\right) + 2h_\times dx\,dy, \tag{50}$$

be incident on the detector from direction $\hat{z}$. There will be no detector output proportional to $h_\times$, since that component of the radiation does not lead to a differential change in the arm lengths; on the other hand, the polarization component proportional to h_+ does lead to a differential change in the arm lengths and, correspondingly, to detector output.

Similarly, consider radiation incident on the detector along the interferometer's x arm:

$$h_{ij}dx^i dx^j = h_+\left(dy^2 - dz^2\right) + 2h_\times dy\,dz. \tag{51}$$

Again, the $\times$ polarization mode does not lead to a differential change in the interferometer arm lengths (at first order in h); so, the detector is not sensitive to radiation with this polarization. On the other hand, radiation in the $+$ polarization mode, as we have defined it, leads to changes in the length of the y arm while leaving the x arm length unchanged; consequently, there is a *differential* change in the interferometer arm length and the detector is sensitive to radiation of this polarization incident from this direction.

To determine in general the coefficients F_+ and $F_\times$ that describe the response of an interferometric detector to incident plane waves, first describe the polarization modes of radiation incident on the detector relative to the detector coordinate system. In the usual (θ, ϕ) spherical coordinates associated with the interferometer coordinate system, the incident direction of a plane-wave propagating with wave-vector $\vec{k}$ is

$$\cos\theta \equiv -\vec{k}\cdot\vec{z}/|\vec{k}|, \tag{52}$$

$$\tan\phi \equiv \frac{\vec{k}\cdot\vec{y}}{\vec{k}\cdot\vec{x}}. \tag{53}$$

In the plane orthogonal to the radiation propagation direction $\hat{k}$, let the $\hat{x}'$ direction be parallel to the xy-plane and the $\hat{y}'$ direction be orthogonal to $\hat{x}'$ so that $(\hat{x}', \hat{y}', -\hat{k})$ forms a right-handed coordinate system. [In the degenerate case—radiation propagating parallel to the $\hat{z}$ direction—we take $\hat{x}'$ parallel to $\hat{x}$ and $\hat{y}'$ such that $(\hat{x}', \hat{y}', -\hat{k})$ is right-handed.] In terms of this coordinate system, define the $+$ and $\times$ polarizations of an incident gravitational wave by

$$h_{ij}dx^i dx^j = h_+\left(dx'^2 - dy'^2\right) + 2h_\times dx'dy'; \tag{54}$$

then, the antenna pattern factors F_+ and $F_\times$ are given by

$$F_+ \equiv \frac{1}{2}\left(1+\cos^2\theta\right)\cos 2\phi\cos 2\psi - \cos\theta\sin 2\phi\sin 2\psi, \tag{55}$$

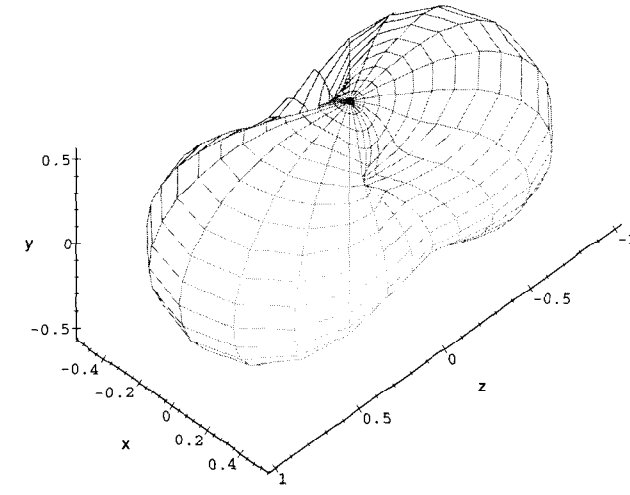


Fig. 2. The polarization-averaged RMS sensitivity of an interferometric gravitational wave detector to radiation incident from any direction. The detector is at the origin of the figure and has its arms aligned with the figure's x and y axes. The magnitude of the distance from the origin to the surface in a direction $\hat{n}$ is proportional to the relative response of the detector to radiation incident on the detector from that direction, averaged over all polarizations.

$$F_\times \equiv \frac{1}{2}\left(1+\cos^2\theta\right)\cos 2\phi\sin 2\psi + \cos\theta\sin 2\phi\cos 2\psi. \tag{56}$$

Figure 2 shows the polarization-averaged RMS sensitivity of a right-angle interferometric detector to plane waves incident from a given direction. The detector is at the origin of the figure, with its arms along the figure's $\hat{x}$ and $\hat{y}$ axes. The detector's sensitivity to radiation incident on the detector from direction $\hat{n}$ is proportional to the distance of the surface from the figure's origin in the direction $\hat{n}$.

3.4 Response Function

The output of a gravitational wave detector is a voltage, $v(t)$, that is linearly related to the incident radiation. Consider a gravitational plane wave, with polarizations h_+ and $h_\times$, incident on a detector with antenna pattern described by F_+ and $F_\times$. The detector response is given by

$$v(t) = \int_{-\infty}^{t} d\tau\, h(\tau) K(t-\tau), \qquad \text{where} \tag{57}$$

$$h(t) = F_+ h_+(t) + F_\times h_\times(t) \tag{58}$$

and K is the kernel of the linear transformation.

It is instructive to express this convolution in the frequency domain:

$$v(t) = \int_{-\infty}^{\infty} df\, \tilde{h}(f) \widetilde{K^*}(f) \exp(2\pi i f t), \tag{59}$$

where

$$\tilde{g}(f) \equiv \int_{-\infty}^{\infty} dt\, g(t) \exp(-2\pi i f t) \tag{60}$$

and we have assumed that $K(\tau)$ vanishes for negative τ.[§] From equation 59 we see that the response of the system—the output voltage—depends on the frequency of the incident radiation: depending on the character of the detector, the response may be relatively large for some frequencies and relatively small for others.

As an example, consider two equal masses M connected by a spring (spring constant $\omega_0^2 M$, quality factor Q). Denote the equilibrium separation by L. A passing gravitational wave of appropriate polarization disturbs the equilibrium separation of the system. The net result is that the passage of a gravitational wave acts as a driving force on the system's normal mode:

$$\ddot{x} + \frac{\omega_0}{Q}\dot{x} + \omega_0^2 x = \frac{1}{2} L \ddot{h}, \tag{61}$$

where x is the difference between the actual and equilibrium separation.

Suppose that we instrument this system with strain gauges to produce an output voltage $v(t)$ proportional to $x(t)$, the deviation from equilibrium separation. How is $v(t)$ [*i.e.*, $x(t)$] related to $h(t)$? In the frequency domain we see that

$$\tilde{v}(f) \propto \frac{f^2}{f_0^2 - f^2 + i f f_0/Q} \tilde{h}(f), \tag{62}$$

[§]Corresponding to a causal impulse response!

where

$$f_0 = \omega_0/2\pi. \tag{63}$$

The output voltage for excitations near the resonance can vary dramatically as a function of frequency.

The response function we have just described is equivalent conceptually to that of a modern acoustic detector: the radiation manifests itself as a driving force on the system's normal modes and the response is a strong function of the frequency in the neighborhood of the resonances.

The response function of an interferometric detector is quite different. For an interferometer, at frequencies much below the round-trip travel time (but greater than the pendulum frequency of the suspended mirrors and beam-splitter), the detector response is independent of frequency; only when the frequency becomes comparable to or larger than the round-trip light travel time in an interferometer arm does the response vary with frequency.[28¶]

The amplitude of the response determines those frequencies where an incident gravitational wave of unit amplitude gives relatively large amplitude output and where it gives relatively small amplitude output. It is not, however, the case that relatively large amplitude output corresponds to relatively large *sensitivity,* if by sensitivity we mean greater ability to detect. To address the question of sensitivity, we must turn to yet a different aspect of a detector's function: its noise.

3.5 Noise

The output channel of a gravitational wave detector is always alive with random fluctuations—*noise*—even in the absence of a gravitational wave signal. In a per-

¶It is commonly said that an interferometer responds to a passing gravitational wave proportionately with the differential change in the IFO arm length. This is not quite right. The response of an interferometric detector to a passing gravitational wave is proportional to the differential change in the round-trip light travel time in the arms. The round-trip light travel time involves the integrated change in the arm length over the past, as opposed to the instantaneous separation at the time of reflections. For frequencies small compared to the inverse round-trip light travel time, the difference is negligible.
It is also the case for interferometers that the frequency dependence of the response function varies with the incidence direction of the radiation though—again—this is only significant at frequencies comparable to or larger than the inverse round-trip light travel time in an arm.

fect world noise would arise exclusively from fundamental physical processes: *e.g.,* fluctuations owing to the finite temperature of the detector, counting statistics of individual photons on a photo-detector, *etc.* In the less than perfect world in which we live, there will be other contributions to the detector noise, beyond these fundamental processes, that arise from the imperfect construction of the detector (*e.g.,* bad electrical contacts), imperfections in the materials used to construct the detectors (*e.g.,* mechanical creep and strain release), and from the detector's interaction with the (non-gravitational wave) environment (*e.g.,* seismic vibrations, electromagnetic interactions, *etc.*).

Detection of gravitational waves requires that we be able to distinguish, in the detector output, between signal and noise. This requires that we have characterized the noise (and not only the signal). Since noise is intrinsically random in character, that characterization is in terms of its statistical properties. Some of these statistical properties we can predict, model, or anticipate *a priori,* based on the detector design; nevertheless, it is important to realize that an experimental apparatus is a real thing made in the real world and will never behave ideally. While a large part of the experimental craft involves building instruments that operate as close as possible to their theoretical limits or prior expectations, the final characterization of a detector will always be determined or verified empirically. In this section we describe something of how noise in gravitational wave detectors is characterized.

3.5.1 Correlations

Just as a probability distribution is fully characterized by its moments, so the random output of a gravitational wave detector can be fully characterized by its *correlations.* The N-point correlation function describes the mean value of the product of the detector output sampled at N different times. Mean, in this case, refers to an *ensemble* average, where the ensemble is an infinite number of identically constructed detectors. Denoting by $n(t)$ the noisy output of a gravitational wave detector in the absence of any signal, the N-point correlation function of the noise distribution is given by

$$C_N(\tau_0, \ldots, \tau_{N-1}) = \overline{n(\tau_0) \ldots n(\tau_{N-1})}, \tag{64}$$

where the over-bar signifies an ensemble average, which is also referred to as an average *across the process.*

As a practical matter ensemble averages are impossible to realize experimentally: one rarely has the opportunity of working with even two similar detectors, let alone an infinite number of identical ones. Thus, while a handy theoretical construct, the general set of correlation functions is not of great practical use in characterizing the behavior of a real detector.

3.5.2 Stationarity

If, however, the behavior of the detector noise does not depend significantly on time—*i.e.,* the noise is *stationary*—then the utility of the correlation function as a practical tool for characterizing detector noise increases dramatically. When the noise character is, figuratively, the same today as it was yesterday and as it will be tomorrow, then the detector yesterday (or an hour, or a minute, or a second ago) can be regarded as an identical copy of the detector we are looking at now, and both are identical copies of the detector tomorrow. Consequently, in the spirit of the ergodic theorem, we can replace the average across the process—the ensemble average—with an average *along* the process—a time average. The N-point correlation function is then a function of the difference in time between the N samples:

$$C_N(\tau_1, \ldots, \tau_{N-1}) = \lim_{T\to\infty} \frac{1}{T} \int_{t_0-T}^{t_0} n(t)n(t-\tau_1)\cdots n(t-\tau_{N-1})\, dt. \tag{65}$$

Of course, perfect stationarity is an impossible requirement. As a practical matter, what we require is that the noise process be stationary over a suitably long period. Let's try to make that concept more quantitative. To simplify the discussion, assume (without loss of generality) that the noise process has zero mean. Consider first the two-time correlation function of a stationary process:

$$C_2(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_{t_0-T}^{t_0} n(t)n(t-\tau)\, dt. \tag{66}$$

For sufficiently large τ we expect intuitively that $C_2(\tau)$ should vanish: the output now should be effectively uncorrelated with the output in either the distant past or the distant future. This will also be the case for the higher-order moments as well: for sufficiently large τ_k (any k), the correlation function C_N should vanish. Thus, we don't need to require perfect stationarity; rather, we require only that the statistical character be approximately stationary, varying significantly only over times long compared to the longest correlation time. In that case, we

can approximate the correlations C_N by averaging, as in equation 66, over *finite* periods.

3.5.3 Gaussian Noise

Noise from fundamental processes tends to be either Gaussian (*i.e.,* originating from contact with a heat bath or some dissipative process) or Poissonian (*e.g.,* originating in the counting statistics of identical and independently distributed—i.i.d.—events that occur at a fixed, average rate). For the gravitational wave detectors under construction, the intrinsically Poissonian processes (*e.g.,* photon counting statistics) have rates so high that they can be treated as Gaussian and we do so here and below.

One way to think about the detector noise is as a superposition of a Gaussian and approximately stationary component, a (hopefully lower amplitude) non-Gaussian, but still stationary component, superposed finally with a non-stationary component. General statements cannot be made about the non-Gaussian or non-stationary components: they differ from instrument to instrument and environment to environment and can only be characterized empirically. The characterization of the Gaussian-stationary component, however, is remarkably simple and has a useful physical interpretation, which we review in this and the next subsection. (For more information and detail, see Finn.[29])

Up to now we have considered the output of a detector as an analog process: *i.e.,* one that is continuous in time. In fact, the output we observe will have been sampled discretely at some sampling rate f_s, chosen to be something more than twice as great as the maximum frequency of interest for the detector output. So, instead of writing the noise at the detector output as $n(t)$ we write

$$n[k] \equiv n(t_k), \tag{67}$$

where

$$t_k = t_0 + k\Delta t \tag{68}$$

for constant Δt.

When the noise in the detector is Gaussian and stationary, any single sample $n[j]$ of the detector output is drawn from a normal distribution with a mean and variance that are independent of when the sample was taken. Without loss of generality we can assume that the mean $\overline{n}$ vanishes, in which case

$$P(n[j]) = \frac{\exp\left[-n[j]^2/2\sigma^2\right]}{\sqrt{2\pi\sigma^2}}. \tag{69}$$

We understand the variance σ^2 of the distribution to be the ensemble average of the square of the detector noise:

$$\sigma^2 = \overline{n[j]^2}. \tag{70}$$

Equation (69) holds true for each sample $n[j]$; consequently, the *joint probability* that the length N_T *sequence* of samples $n[j]$, j running from 1 to N_T, is a sample of detector noise and is given by the *multivariate* Gaussian distribution

$$P(n[1], n[2], \ldots, n[N_T]) = \frac{\exp\left[-\frac{1}{2}\sum_{j,k=0}^{N_T-1} n[j]\,||\mathbf{C}^{-1}||_{jk}\,n[k]\right]}{\sqrt{(2\pi)^{N_T}\det||\mathbf{C}||}}. \tag{71}$$

In place of the variance σ^2 that appears in the exponent of Eq. (69) is the *covariance* matrix $\mathbf{C}$. (The matrix $\mathbf{C}^{-1}$ that appears in Eq. (71) is, by construction, positive definite; consequently, it is non-singular and invertible.) Similarly, in place of the factor σ^2 that appears in the denominator of equation 69 is the determinant $\det||\mathbf{C}||$.

The mean over the product $n[j]n[k]$ is the value of the correlation function $C_2(t_j - t_k)$; it is also just the value of the jk element of the covariance matrix $||C||$:

$$C_2[j-k] = \overline{n[j]n[k]} = ||\mathbf{C}||_{jk}\,. \tag{72}$$

Since the detector noise is also assumed to be stationary, $||\mathbf{C}||_{jk}$ can depend only on the difference $j-k$; correspondingly, $\mathbf{C}$ is constant on its diagonals: *i.e.,* it is a *Toeplitz* matrix. Consequently, it is fully characterized by the sequence $c[k]$ of length $2N_T - 1$ whose elements are the first row and column of $\mathbf{C}$:

$$c[j-k] = ||\mathbf{C}||_{jk} = C_2(t_j - t_k). \tag{73}$$

The sequence c and the process mean (which we have assumed to vanish) fully characterize the random process. The sequence c, however, is just the two-time correlation function of the detector output! Thus, the two-time correlation function C_2 fully characterizes a Gaussian stationary process: all the higher order correlation functions C_N either vanish (for odd N) or are expressible as sums of products of C_2. Once we have determined C_2, then, we have completely determined the character of the Gaussian noise process.

3.5.4 Likelihood Function

In the last section we evaluated

$$P(v|0) \equiv \left(\begin{array}{l} \text{probability of observing} \\ \text{output sequence } v \text{ assuming} \\ \text{no signal is present} \end{array} \right) \tag{74}$$

for Gaussian-stationary detector noise. Since the detector is linear, the probability

$$P(v|h) \equiv \left(\begin{array}{l} \text{probability of observing} \\ \text{output sequence } v \text{ assuming} \\ \text{signal } h \text{ is present} \end{array} \right) \tag{75}$$

is just

$$P(v|h) = P(v - v_h|0), \tag{76}$$

where v_h is the detector response to the gravitational wave signal h. The ratio of these two probabilities,

$$\Lambda(v|h) \equiv \frac{P(v|h)}{P(v|0)}, \tag{77}$$

termed the *likelihood function,* is the *odds* that the data v is a combination of signal v_h and noise, as opposed to a noise alone. For a given observation v the likelihood can be viewed as a function of hypothesized signal h, in which case it has a convenient interpretation in terms of *plausibility:* in particular, $\Lambda(v|h)$ can be interpreted as the *plausibility* that the signal h is present given the particular observation v. (The likelihood is not, however, a probability.) This meaning of the likelihood is independent of the statistical character of the noise. The difficulty, if the noise is not Gaussian-stationary, is in evaluating Λ.

3.5.5 The Two-Time Correlation Function

The correlation function $C_2(\tau)$ describes the statistical relationship between pairs of samples drawn from the random process $n(t)$ at times separated by an interval τ. Given two samples separated in time by τ, a non-zero correlation $C_2(\tau)$ corresponds to an increased ability to predict the value of one member of the pair given the other.

The correlation function $C_2(\tau)$ is bounded by $\pm C_2(0)$, suggesting that we define the *correlation coefficient*

$$R_2(\tau) \equiv C_2(\tau)/C_2(0), \tag{78}$$

which is bounded by ± 1. If the correlation coefficient is zero for some τ, then samples taken an interval τ apart are entirely uncorrelated: knowledge of one does not lead to any increased ability to predict the other. A positive correlation coefficient tells us that the two samples are more likely close to each other in magnitude and sign than not, while a negative correlation coefficient tells us that the two samples are likely close to each other in magnitude but of opposite sign. The larger the coefficient magnitude the greater the tendency. When the correlation coefficient is unity then the correlation is perfect: *i.e.,* when it is $+1$ the two samples are always equal, and when it is -1 the two samples are always of equal magnitude but opposite sign.

3.5.6 Noise Power Spectral Density

Consider for a moment a simple harmonic oscillator—*e.g.,* a pendulum—coupled weakly to a heat bath. The heat bath excites the oscillator so that its mean energy is $k_B T$. Since the coupling to the heat bath is weak, the phase of the oscillator progresses nearly uniformly in time with rate ω_0 corresponding to the oscillator's natural angular frequency. Over long periods, however, the continual, random excitations of the oscillator cause the phase to drift in a random manner from constant rate.

Now suppose that we sample the position coordinate of the oscillator at intervals separated by exactly one period $2\pi/\omega_0$. Since the coupling to the heat bath is weak, the samples are very nearly identical: in fact, were it not for the contact with the heat bath, they would be exactly identical. Thus, we expect that the correlation coefficient corresponding to an interval equal to an oscillator period should be nearly unity. Continuing to focus on samples taken at intervals equal to exact multiples of the period, we expect that the correlation coefficients should remain large for small multiples, but should decrease as the interval increases since contact with the heat bath will lead, as time increases, to greater drift in the phase.

On the other hand, suppose that we sample the position coordinate of the oscillator at intervals separated by exactly odd integer multiples of a *half*-period π/ω_0. Now we expect the correlation coefficient to be nearly equal to -1 for small intervals, decreasing in magnitude to 0 as the interval increases.

Contact with a heat bath can take place in many ways, leading to subtly different correlation functions. Figure 3 shows the correlation function corresponding to two different kinds of heat bath contact: that which leads to velocity damping and that which leads to structural damping.[30] Note how, pictured in this way, there is apparently little difference between these two damping measures.

Since the correlation function is so oscillatory we are immediately led to consider its Fourier transform. In this case, since $C_2(\tau)$ is an even function of the lag τ, we consider the cosine transform, which we term the *one-sided power spectral density:*

$$S_v(f) = 4\int_0^\infty d\tau\, C_2(\tau)\cos(2\pi f\tau). \tag{79}$$

(*One-sided* refers to the fact that, in choosing a cosine transform, we have effectively folded the power in negative frequencies into the power at positive frequencies; so, the $S_v(f)$ includes the power at frequencies whose magnitude is $|f|$.) Figure 4 shows the power spectral densities corresponding to the correlation functions of figure 3. The strongly oscillatory nature of these functions shows up as a large peak at the oscillator resonant frequency (normalized to unity). In addition, however, the PSD shows clearly the very different off-resonance character of the noise. Noise from a structurally damped system rises in amplitude as the frequency falls below resonance, unlike the noise contribution from a viscously damped system; similarly, noise from a structurally damped system falls more steeply with frequency above resonance than does the noise from a viscously damped system. To see the same in the correlation function would require close inspection of the trends of the correlation function envelopes over very long lags.

Thus, even though it is completely equivalent to the correlation function, the power spectral density is often a more useful characterization of the noise character. In the case of Gaussian noise its equivalence to the correlation function guarantees it is also a full characterization of the detector noise. When the noise is not Gaussian, there are analogous spectra associated with the higher order correlations: for example, the bispectrum is the two-dimensional Fourier transform of $C_3(\tau_1, \tau_2)$,

$$\mathrm{Bi}(f_1, f_2) = \int_{-\infty}^{\infty} d\tau_1 e^{-2\pi i f_1 t_1}\int_{-\infty}^{\infty} d\tau_2\, e^{-2\pi i f_2 t_2}\, C_3(\tau_1, \tau_2), \tag{80}$$

and so on. These higher order spectra and their magnitudes play the same role for the higher-order correlation functions as the power spectral density plays for the auto-correlation function.

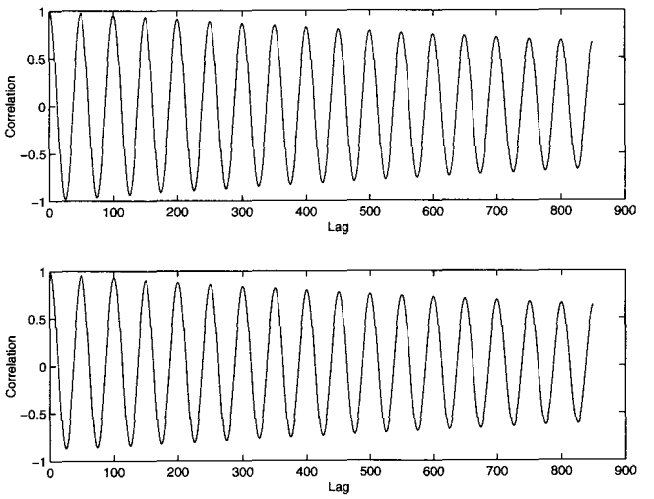


Fig. 3. The correlation functions for a harmonic oscillator in contact with a heat bath. Contact with a heat bath leads to damping; the nature and degree of the contact determine the character of the system's noise. In this figure we show the correlation function, over several periods, for two different kinds of heat bath contact with the same on-resonance damping. The upper panel corresponds to a viscous damped harmonic oscillator; the lower panel corresponds to a structurally damped oscillator. The difference between the two correlation functions is apparently very subtle.

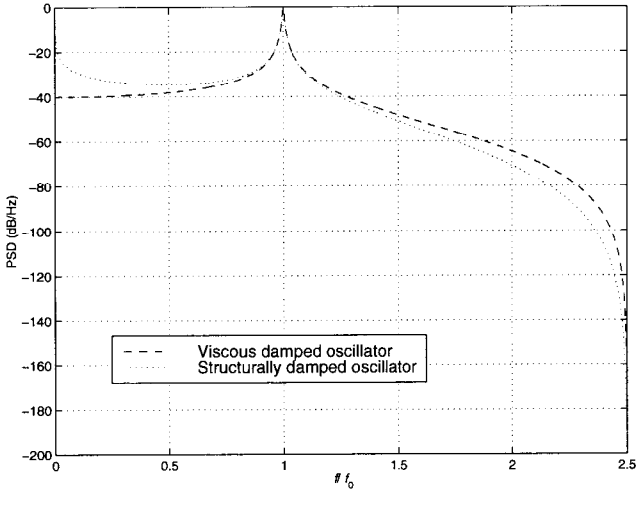


Fig. 4. The power spectral density of the two processes whose correlation functions are shown in figure 3. Note that, while the correlation functions appear very similar as functions of time, strong differences show up in the power spectral densities as functions of frequency.

3.6 Signal-to-Noise Ratio

When is a gravitational wave "detectable"? We haven't yet explored the meaning of "detection" qualitatively, let alone quantitatively; nevertheless, we have an intuitive feeling that a signal ought to be detectable if the detector's response to the signal is greater than the intrinsic noise amplitude. Let's develop that idea a bit.

Suppose that we have a detector with noise power spectral density $S_v(f)$ and particular output $v(t)$, which consists of a signal $v_h(t)$ superposed with detector noise $v_n(t)$. The variance of $v(t)$, over an interval $[0,T]$, is

$$\sigma_v^2 = \frac{1}{T}\int_0^T dt\, v(t)^2, \tag{81}$$

$$= \frac{2}{T}\int_0^\infty df\ |\tilde{v}(f)|^2\,. \tag{82}$$

The noise is a random process; so, then, is σ_v^2. Focus on the ensemble average of σ_v^2 and look in the frequency domain:

$$\frac{d\overline{\sigma_v^2}}{df} = \frac{2}{T}\overline{|\tilde{v}(f)|^2}, \tag{83}$$

$$= \frac{2}{T}\left(\overline{|\tilde{v}_n(f)|^2} + |\tilde{v}_h(f)|^2\,,\right) \tag{84}$$

where the final equality follows when we recognize that the noise is independent of the signal. The contribution to the mean signal variance thus consists of separate contributions from the signal and from the noise.

The ratio

$$\frac{|\tilde{v}_h(f)|^2}{\overline{|\tilde{v}_n(f)|^2}} \tag{85}$$

evidently tells us which—signal or noise—is expected (note ensemble average!) to contribute more to the amplitude of the detector output in a unit bandwidth about frequency f. We can compute a similar, dimensionless quantity over the full bandwidth

$$\int_{-\infty}^{\infty} df\, \frac{|\tilde{v}_h(f)|}{\overline{|\tilde{v}_n(f)|}/T} = 4\int_0^\infty df\, \frac{|\tilde{v}_h(f)|}{S_v(f)} \tag{86}$$

that tells us which of the signal v_h or the noise v_n is expected to contribute more to the variance of the output v.

Given a particular sample of detector output v, we don't know, *a priori*, what part is v_n and what part (if any) is v_h. Consider a quantity that we can calculate

directly from the detector output v:

$$\rho^2 \equiv 4\int_0^\infty df\,\frac{|\tilde{v}(f)|^2}{S_v(f)}. \tag{87}$$

The integrand is evidently the ratio of the actual contribution to the signal variance in a unit band about frequency f to the contribution that would be expected, in the same band, from noise alone. Not surprisingly, the ensemble mean $\overline{\rho^2}$ is

$$\overline{\rho^2} = 1 + 4\int_0^\infty df\,\frac{|\tilde{v}_h(f)|^2}{S_v(f)}\,. \tag{88}$$

We refer to ρ^2 as the *signal-to-noise ratio,* or SNR.‖

Our construction of ρ^2 has been physically motivated. It turns out, however, that exactly this same quantity arises from a consideration of the probability $P(v|0)$, which we explored in §3.5.3. In that section we found, for Gaussian-stationary noise,

$$P(v|0) = \frac{\exp\left[-\frac{1}{2}\sum_{j,k=0}^{N-1} v[j]\,||\mathbf{C}^{-1}||_{jk}\,v[k]\right]}{\sqrt{2\pi\,\det||\mathbf{C}||}}\,. \tag{89}$$

With just a little algebra, however, the argument of the exponential can be rewritten as[31]

$$\sum_{j,k=0}^{N-1} v[j]\left\|\mathbf{C}^{-1}\right\|_{jk} v[k] \;=\; \Re\left[\frac{1}{2N-1}\sum_{j=-(N-1)}^{N-1}\frac{\left|\tilde{V}[j]\right|^2}{\tilde{c}[j]}\right] \tag{90}$$

where the periodic sequence $\tilde{g}[j]$ is related to the discrete Fourier transform of the sequence $g[k]$:

$$\hat{g}[j] = \sum_{k=-(N-1)}^{N-1} e^{-2\pi ikj/(2N-1)} g[k] \tag{91}$$

and $V[k]$ is just $v[k]$ zero-padded for negative k:

$$V[k] = \begin{cases} v[k] & \text{for } k \geq 0 \\ 0 & \text{for } k < 0. \end{cases} \tag{92}$$

These summations can be regarded as approximations to integrals, in which case

$$\Re\left[\frac{1}{2N-1}\sum_{j=-(N-1)}^{N-1}\frac{\left|\tilde{V}[j]\right|^2}{\tilde{c}[j]}\right] \propto \int_0^\infty df\,\frac{\left|\tilde{V}\right|^2}{S_h(f)}. \tag{93}$$

‖Note that this definition of ρ^2 is different, by the additive factor of unity, than used elsewhere in the gravitational wave literature.

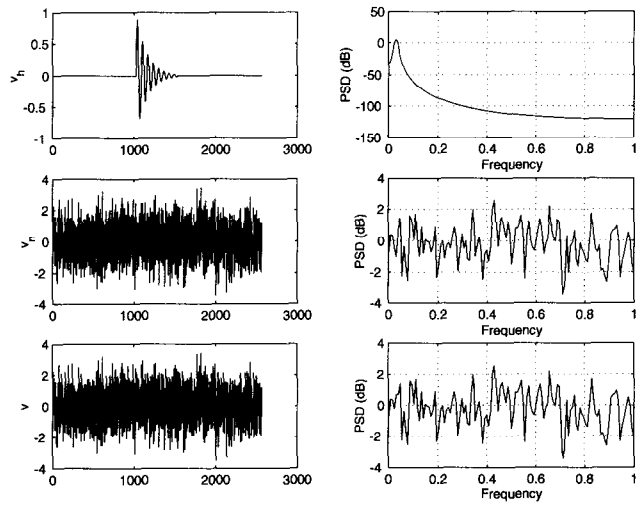


Fig. 5. An imagined gravitational wave signal (upper-left panel), detector noise (middle-left panel), and total detector output (signal+noise, lower-left panel). Note how the signal is not evident to the eye in the detector output. The right-hand panels show the power spectra of the corresponding left-hand panels; again, the signal power is not evident relative to the noise power.

Hence, the SNR associated with the observed detector output v is closely related to the probability that v is a sample of *just* detector noise, with no gravitational-wave signal present. The larger ρ^2, the smaller this probability. Should we observe detector output with large ρ^2, then, we are not too far wrong to be suspicious that we have seen evidence for gravitational waves.

To make this last judgment—which involves making quantitative the notion of "large" ρ^2—we need to know the probability distributions of the SNR in both the presence and absence of a signal: after all, since noise is a random process there is some non-zero probability that, in any given observation, ρ^2 will take on any particular value, large or small. We return to consider this point in Sec. 4.

3.6.1 Matched Filtering

Calculating ρ^2 defined by equation 87 does not require or make use of any information about the gravitational radiation source. Suppose that we know, *a priori*, the radiation waveform has the shape $V_h(t)$, and that the question is whether the corresponding signal $\alpha V_h(t-t_0)$, for some unknown constants α and t_0, is present in the detector observed output $v(t)$. Can we make use of this information—the signal shape $V_h(t)$—to boost our ability to observe the signal?

The answer is yes. To illustrate, figure 5 shows an imagined v_h, v_n, and v equal to $v_h + v_n$ in the left-hand panels, and the corresponding power spectra in the right-hand panels. For this illustration we have assumed that the noise is white across the detector bandwidth. The signal is not apparent to the eye in either v or its power spectrum $P_v(f)$. Figure 6 shows, in the top panel, the filter output when just v_h is passed through the filter K with impulse-response V_h set equal to v_h:

$$v'(t) = \int_{-\infty}^{t} d\tau\, v(\tau)K(t-\tau), \tag{94}$$

$$= v'_n(t) + v'_h(t), \tag{95}$$

where

$$K(\tau) = v_h(t), \tag{96}$$

$$v'_n(t) = \int_{-\infty}^{t} d\tau\, v_n(\tau)v_h(t+\tau), \tag{97}$$

$$v'_h(t) = \int_{-\infty}^{t} d\tau\, v_h(\tau)v_h(t+\tau). \tag{98}$$

Without loss of generality we assume v_h is non-zero only for positive t. The filtered detector output $v'(t)$ consists of a signal contribution $v'_h(t)$ and a noise contribution $v'_n(t)$. These are shown in the top and middle panels of figure 6, respectively. The bottom panel of figure 6 shows the filter output v' (equal to $v'_h + v'_n$). The presence of the "signal" v'_h is now much more evident.

The filter we have chosen has reduced the total power in the noise relative to that in the signal. How it does this is apparent by considering the power spectra in figures 5 and 6. In figure 5, the power in v_h is seen to be confined to a very narrow bandwidth about the frequency of the damped sinusoid. At its peak the signal power is about 5 dB greater than noise power. Nevertheless, the total noise power, integrated over the full bandwidth, is much greater than the signal power

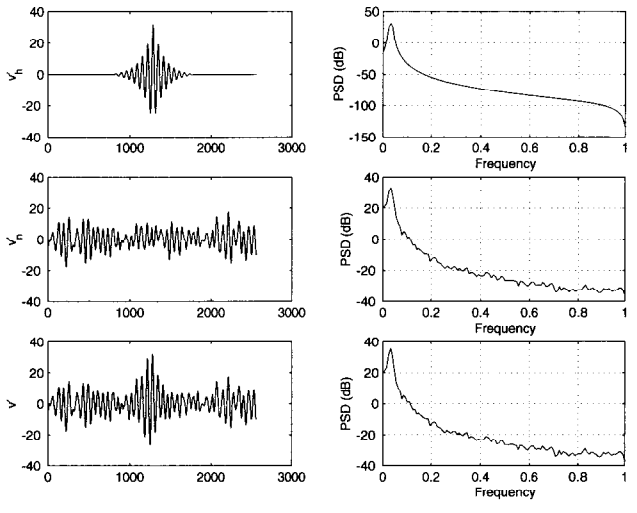


Fig. 6. The output of the filter described in equation 95 when just the signal v_h is filtered (upper panel) and when the detector output, consisting of signal and noise, is filtered (lower panel). In contrast to the lower-right panel of figure 5, the "signal" (*i.e.*, the upper panel) is quite evident even in the presence of noise.

and, consequently, the signal is overwhelmed by the noise (cf. the bottom panel of figure 5).

Now consider v'. The filter applied to the signal has the impulse response of the signal, or the squared magnitude frequency response given by the power spectrum in the top panel of figure 5. This is *matched* to the signal, in the sense that the power passed is in the band where the signal power is large and the power stopped is in the band where the signal power is small. Thus, what survives in v' is the signal power, together with only that noise power in the narrow band where the signal power is large. The signal to noise of the filtered detector output v' is correspondingly much higher in the presence of the signal than is the signal to noise ratio of v.

This example is illustrative. In fact, we can ask, for an arbitrary signal v_h embedded in noise with power spectrum $S_v(f)$, for the linear filter that maximizes

the ratio of the mean-square signal contribution to the mean-square noise contribution. That filter is referred to as the Wiener matched filter; in the frequency domain and for weak signals it is (up to an overall constant)

$$\widetilde{K}(f) = \frac{\widetilde{v}_h^*(f)}{S_h(f)}. \tag{99}$$

More generally, additional information is always useful for increasing our ability to detect a signal. This is true even if that information is not as complete as knowing the waveform. For example, consider the case where we know the signal spectrum, but not its waveform. In the frequency domain, we thus know the signal amplitude at each frequency, but not the corresponding phase. In the case where the waveform is known, we constructed the filter making full knowledge of both amplitude and phase information. We can also construct a filter that passes power in a given bandwidth, without regard to its phase. This filter will emphasize power in the bands where the ratio of signal power to noise power is relatively large over bands where the ratio is small; consequently, it will increase our ability to detect a signal whose spectrum is known in the same way that a matched filter increases our ability to detect a signal whose waveform is known.

3.7 The Effective Noise Power Spectral Density

How does one compare different detectors, with different response functions and different noise power spectral densities?

One possibility is the "performance benchmark": choose a prototypical source, evaluate the signal-to-noise that the source would give in the different detectors, and determine finally which detector is most likely to observe the source at a given level of confidence.[32]

This kind of judgment depends critically on the source: using different sources as your benchmark can lead to different conclusions. For example, sources whose power is concentrated at different frequencies focus attention on the detector noise at those frequencies. Thus, while benchmarking detectors against particular sources can be a powerful tool for comparing their relative performance, it is also a tool with a very narrow focus. We need some other way to compare the capability of detectors with a less specific emphasis on source.

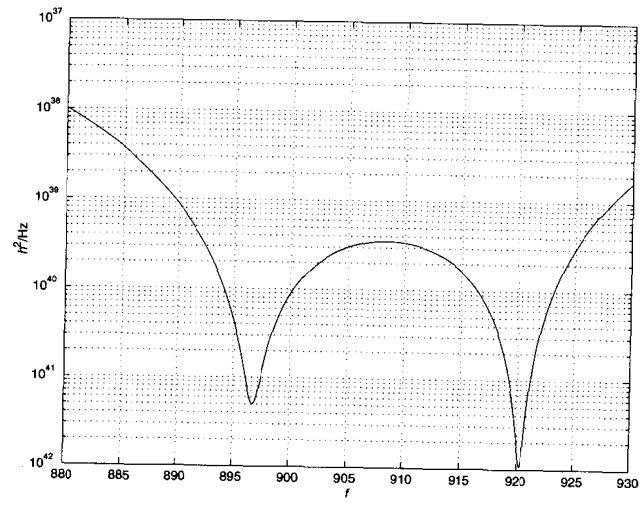


Fig. 7. The power spectral density of an effective stochastic gravitational wave signal that would mimic the noise in the output of a modern bar detector. Plotted is $\sqrt{S_h(f)}$ *vs* frequency f.

An important tool for making this more general comparison is the *effective power spectral density* $S_h(f)$,

$$S_h(f) \equiv \frac{S_v(f)}{|R(f)|^2}, \tag{100}$$

where $R(f)$ is the detector response function and $S_v(f)$ is the detector noise power spectral density. The quantity $S_h(f)$ describes an *effective detector noise:* it is the power spectral density of a stochastic gravitational wave signal that would have to be applied to a *noise-free* detector in order that the corresponding response have power spectral density $S_v(f)$. Over frequency bands where $S_h(f)$ is small, the detector is relatively sensitive; over frequency bands where it is large, the detector is relatively insensitive.

The effective power spectral density $S_h(f)$ has both a source and detector independent meaning, making it a particularly useful quantity for comparing gravitational wave detectors or for comparing a detector to a source. With it, one can

rank detectors according to their overall noise in a given bandwidth, *e.g.*,

$$\overline{h^2}_n(f_1, f_2) = \int_{f_1}^{f_2} df\, S_h(f), \tag{101}$$

or define an effective band $(f_0 - \Delta f/2, f_0 + \Delta f/2)$ over which the detector has greatest sensitivity, *e.g.*,

$$f_0 \equiv \frac{\int_0^\infty df\, f/S_h(f)}{\int_0^\infty df/S_h(f)}, \tag{102}$$

$$(\Delta f)^2 \equiv \frac{\int_0^\infty df\, (f - f_0)^2/S_h(f)}{\int_0^\infty df/S_h(f)}. \tag{103}$$

Finally, since the noise is referred directly to the amplitude of incident gravitational radiation, one can calculate the expected SNR of a given signal in the detector without reference to the detector's response function:

$$\begin{aligned} \overline{\rho^2} &= 1 + 4\int_0^\infty df\, \frac{\left|R(f)\tilde{h}(f)\right|^2}{S_v(f)} \\ &= 1 + 4\int_0^\infty df\, \frac{\left|\tilde{h}(f)\right|^2}{S_h(f)}. \end{aligned} \tag{104}$$

Figure 7 shows the modeled $S_h(f)$ for a modern bar detector, while figure 8 shows $S_h(f)$ for a model of the first-generation LIGO instrumentation. Note how the bar detector noise is particularly small in two narrow bands** about the resonant frequencies of the two-mode system consisting of the bar and its transducer, while the interferometer achieves its peak sensitivity over a much broader bandwidth.

3.7.1 An Aside: Noise in Bar Detectors

It is a common misconception that bar detectors are intrinsically narrow-band detectors. While the amplitude of a resonant detector's response is greatest for signal power in the neighborhood of the resonance, the thermal excitation of the bar is also concentrated in this band as well. The net result is that the contribution of the bar's thermal noise to the power spectral density expressed in units of h^2/Hz is effectively independent of frequency.

**Since the bar detector's "sensitivity" $1/S_h$ is multi-modal, it is more appropriate to define the effective band, as in eq. 102 and 103, separately about each peak.

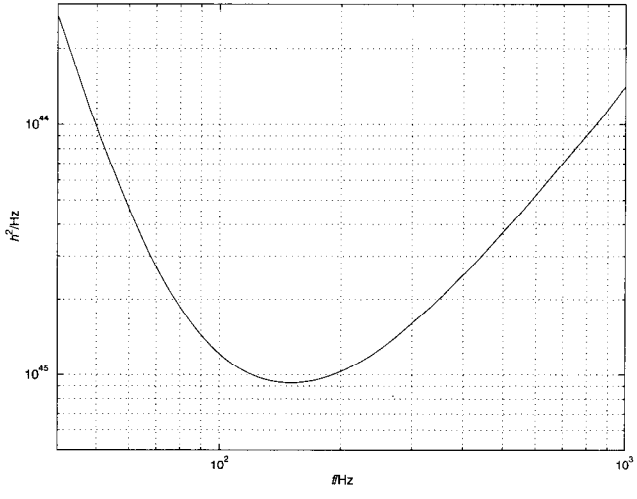


Fig. 8. The power spectral density of an effective stochastic gravitational wave signal that would mimic the noise in the output of first generation LIGO instrumentation. Plotted is $\sqrt{S_h(f)}$ *vs* frequency f.

To understand how resonant detectors become narrow band instruments, consider how the signal appears in the electronics that follow the transducer. The resonant character of the detector leads to large amplitude motion for signal power near the resonant frequency and small amplitude motion for signal power far from the resonance. Correspondingly, the amplified signal is large near to, and small far from, the resonance. *The amplifier contributes its own noise, however, which is approximately white at the amplifier output.* Thus, compared to the signal presented for amplification, the amplifier noise is *relatively* large far from resonance and *relatively* small near to resonance.

In present-day resonant cryogenic detectors, the bandwidth is limited by amplifier noise, referred back to h through the response function.

Since it is the amplifier noise, when referred back to h through the response function of the resonant bar, that limits the instrument bandwidth, why make the bar resonant at all? The purpose of making the detector resonant is to provide

mechanical amplification for the signal, so that, at least in a narrow bandwidth, it is much stronger than the limiting noise source: amplifier noise. Signal power at or near resonance leads to a large excitation of the bar, which translates into a large input to the amplifier; thus, the resonance of the bar amplifies the incident gravitational wave signal relative to all the noise sources that follow, including the limiting amplifier noise.

3.8 Conclusion

Gravitational wave detectors are characterized by their antenna patterns, which describe their differential sensitivity to radiation incident from different directions and with different polarizations, their response functions, which describe the differential amplitude of their response to signals of different frequency, and the character of their noise.

The distinction between the response function and the antenna pattern is sometimes an artificial one: the response function can (and for interferometric detectors, does) depend on the incident direction of the radiation.

Much of the experimental craft is devoted to making the detector noise approximately stationary and Gaussian (or in making the signal so large that the character of the noise is not significant for measurements of interesting precision). Stationary noise can be characterized by evaluating the correlations among samples taken at different relative times. For Gaussian noise, all the correlations are known once the pairwise correlation function is measured.

While the correlation functions are good conceptual tools for understanding the character of stationary detector noise, a more useful tool, fully equivalent, is the noise spectrum: the Fourier transform of the correlation function on its time arguments. The noise power spectral density is a particularly important and useful tool for characterizing the disposition of detector noise power.

4 Characterizing Detection

What does "detection" mean? Let's try to frame an answer by posing a specific question—*e.g.*, "with what confidence can we conclude that, in the last hour, gravitational waves from a new core collapse supernova in the Virgo cluster of galaxies passed through our gravitational wave detector?"—and exploring its meaning.

It turns out that, straightforward as it seems, there are two different ways of interpreting this question; correspondingly, there are two different meanings that "detection" can take. How we mean the question—or the kind of answer that we want to take away—determines the kind of analysis that we need to undertake with data collected at a gravitational wave detector.

4.1 Learning from Observation

"With what confidence can we conclude that, in the last hour, gravitational waves from a new core collapse supernova in the Virgo cluster of galaxies passed through our gravitational wave detector?"

As with most questions of detection, even before examining our observations we have some expectation of the answer. In this case, we know the rate of supernovae and this leads us to expect, on average, one such core collapse every four months; consequently, we believe the probability is approximately 3.4×10^{-4} that in any given hour—including the last—gravitational waves from a new Virgo cluster supernova were incident on our detector.

Probability, as we have used it here, means *degree of belief*. In this instance, our degree of belief coincided with the *expected frequency* of supernova events. This need not always be the case: we can assess degree of belief even when we can't assess relative frequency. For example, suppose that I have a coin that is known to be heavily biased toward either heads or tails. What is your degree of belief that, when I next flip the coin, it will land heads-up? Without telling you the direction or amount of the bias, you can't evaluate the expected relative frequency of heads or tails. You can, however, quantify your degree of belief: having no more reason to believe that the bias is toward heads than towards tails, you have no more reason to believe that the coin will, when next flipped, land heads-up than that it will land heads-down. Your *degree of belief* in either alternative, then, is 1/2. One does not have to search either long or hard to find examples from, *e.g.*, astrophysics, where probability as "degree of belief" exists and probability as "expected frequency" does not. For example, what is the probability that there exists a cosmological stochastic gravitational wave signal with a given amplitude and spectrum? In this case, "expected frequency" has no meaning: there is only one Universe, and it either does or does not have a stochastic gravitational wave background of given spectrum and amplitude.

After we examine the output of our gravitational-wave detector, our degree of belief in the supernova proposition may change: we may, on the basis of the observations, become more or less certain that radiation from a supernova passed through our detector. How do observations change our degree of belief in the different alternatives?

To explore how our degree of belief evolves with the examination of observations we need to introduce some notation:

$$H_0 = \left(\begin{array}{l} \text{proposition that gravitational waves from a} \\ \text{new supernova in the Virgo cluster } \textit{did not} \\ \text{pass through our detector in the last hour} \end{array} \right), \quad (105)$$

$$\mathcal{I} = \left(\begin{array}{l} \text{our prior knowledge of astrophysics, including} \\ \text{our best assessment of the supernova rate} \end{array} \right), \quad (106)$$

$$g = \left(\text{ observations from our gravitational wave detector } \right), \quad (107)$$

$$P(A|B) = \left(\text{ degree of belief in } A \text{ assuming that } B \text{ is true } \right), \quad (108)$$

$$\neg A = (\text{logical negation of proposition } A). \quad (109)$$

In this notation, $P(H_0|\mathcal{I})$ is the degree of belief we ascribe to the proposition that no gravitational waves from a core collapse supernova in the Virgo cluster passed through our detector in the last hour, given only our prior understanding of astrophysics; similarly, $P(H_0|g,\mathcal{I})$ is the degree of belief we ascribe to the same proposition, given *both* the observation g and our prior understanding of astrophysics.

To understand how $P(H_0|\mathcal{I})$ and $P(H_0|g,\mathcal{I})$ are related to each other, we need to recall two properties of probability. The first is unitarity: probability summed over all alternatives is equal to one. In our example, the two alternatives are that a supernova occurred or it did not:

$$P(H_0|g,\mathcal{I}) + P(\neg H_0|g,\mathcal{I}) = 1. \quad (110)$$

The second property we need to recall is Bayes Law, which describes how conditional probabilities "factor":

$$P(A|B,C)P(B|C) = P(A,B|C) = P(B|A,C)P(A|C). \quad (111)$$

Combining unitarity and Bayes Law, it is straightforward to show that

$$P(\neg H_0|g,\mathcal{I}) = \frac{\Lambda(g)}{\Lambda(g) + P(H_0|\mathcal{I})/P(\neg H_0|\mathcal{I})} \quad (112)$$

where

$$\Lambda(g) = P(g|\neg H_0,\mathcal{I})/P(g|H_0,\mathcal{I}), \quad (113)$$

$$P(g|H_0,\mathcal{I}) = \left(\begin{array}{l} \text{probability that } g \text{ is a sample of} \\ \text{detector output when } H_0 \text{ is true} \end{array} \right), \quad (114)$$

$$P(g|\neg H_0,\mathcal{I}) = \left(\begin{array}{l} \text{probability that } g \text{ is a sample of} \\ \text{detector output when } H_0 \text{ is false} \end{array} \right). \quad (115)$$

The two probabilities $P(g|H_0,\mathcal{I})$ and $P(g|\neg H_0,\mathcal{I})$ depend on the statistical properties of the detector noise and the detector response to the gravitational wave signal. In some cases they can be calculated analytically; in other circumstances it may be necessary to evaluate them using, *e.g.,* Monte Carlo numerical methods. Regardless of how one approaches data analysis, the detector must be sufficiently well-characterized that these or equivalent quantities are calculable.

Equation 112 describes how our degree of belief in the proposition $\neg H_0$ evolves as we review the observations. If Λ is large compared to the ratio $P(H_0|\mathcal{I})$ to $P(\neg H_0|\mathcal{I})$, then our confidence in $\neg H_0$ increases; alternatively, if it is small, then our confidence in $\neg H_0$ decreases. If Λ is equal to unity—*i.e.,* the observation g is equally likely given H_0 or $\neg H_0$—then the posterior probability $P(H_0|g,\mathcal{I})$ is equal to the prior probability $P(H_0|\mathcal{I})$ and our degree of belief in H_0 is unchanged: we learn nothing from the observation.

We can now answer the question that began this section. We understand confidence to mean *degree of belief* in the proposition that radiation originating from a new supernova in the Virgo cluster was incident on a particular detector during a particular hour. In response we make a quantitative assessment of our degree of belief in that proposition—the probability that the proposition is true.

4.2 Guessing Nature's State

Begin again: "With what confidence can we conclude that, in the last hour, gravitational waves from a new core collapse supernova in the Virgo cluster of galaxies passed through our gravitational wave detector?"

As before, we have the hypothesis H_0 and its logical negation, $\neg H_0$. The gravitational waves from a new Virgo cluster supernova either passed through our detector, or they did not. Our goal is to determine, as best we can, which of these two alternatives correctly describes what happened.

We decide which alternative is correct by consulting our observation g. Operationally, we adopt a rule or a procedure that, when applied to g, leads us to accept or reject H_0. The question that began this section asks us for our degree of confidence in the most reliable rule or procedure.

There are many procedures that we can choose from. Some are just plain silly: for example, always rejecting H_0 is a procedure. Similarly, accepting H_0 if a flipped coin lands heads is a procedure. Some procedures are more sensible: we can calculate a characteristic amplitude from the observation (*e.g.,* a signal-to-noise ratio) and reject H_0 if the amplitude exceeds a threshold. Nature doesn't always speak clearly; additionally, some crucial information is often hidden from us. Consequently, no procedure will, in the end, be perfect and every rule will, on unpredictable occasions, lead us to erroneous conclusions. Still, some procedures are clearly better than others: the question is, how do we distinguish between them quantitatively?

Better procedures are those that are less prone to error. Consequently, we focus on the frequency with which different procedures err. For our simple problem, where we want to decide only if we have or have not observed the radiation from a supernova (reject or accept H_0), there are two kinds of errors a decision procedure can make:

1. If no radiation is present (H_0 true), the rule may incorrectly lead us to conclude that radiation is present: a *false alarm,* or type I, error.

2. If radiation is present (H_0 false), the rule may incorrectly lead us to conclude that radiation is absent: a *false dismissal,* or type II, error.

The false alarm frequency is generally denoted α while the false dismissal frequency is denoted β.

If we have an ensemble of identical detectors, each observing simultaneously the same system for which H_0 (or $\neg H_0$) is true, and we apply our rule to each observation, then the fraction of observations in the ensemble that lead us to reject (accept) H_0 is just the false alarm (dismissal) frequency. False alarm and false dismissal frequencies can be interpreted as probabilities: in particular, the probability of our rule giving an incorrect result.

Even in the simple case at hand (a single hypothesis that we must accept or reject), there are at least two distinct kinds of errors that an inference procedure can make. Our measure of a rule's reliability thus involves at least two dimensions, and may involve more. How, then, do we order rules to settle upon a best, or optimal, rule?

To rank rules we must reduce the several error measures that describe a procedure's performance to a single figure of merit. How we choose to do this depends on the nature of our problem. In our case, rules that distinguish between H_0 and $\neg H_0$ are characterized by their false alarm and false dismissal frequencies; consequently, our criteria for ranking rules should depend on our relative intolerance to false alarms and false dismissals. For example, if we are testing for the presence of antibodies in an effort to diagnose and treat a serious illness, we might be very concerned to keep the false dismissal rate low, and not nearly as worried about a high false alarm rate: after all, a false dismissal might result in death, while a false alarm only in an unnecessary treatment with less serious repercussions. Judges or juries in criminal trials face different concerns: false dismissals let criminals go free, while false alarms send the innocent to prison—neither alternative being very palatable. Finally, in the case of gravitational wave detection, we may (at least initially) be very concerned to avoid false alarms, even at the risk of falsely dismissing many real signals.

Thus, in order to provide a relative ranking of different inference procedures for detection or parameter estimation, we must construct an *ad hoc* figure of merit that reflects our sensitivity to an incorrect decision. We term the best rule, under that *ad hoc* criteria, the "optimal" rule. "Optimality," however, is a relative concept: if the criteria change, the "optimal" rule changes also. In the three examples given above, the criteria might be

- *medical diagnosis:* fix a maximum acceptable false dismissal rate and choose the rule that, among all rules whose false dismissal rate is so constrained, has the minimum false alarm rate;

- *criminal justice:* choose a rule whose weighted total error $\alpha \cos\phi + \beta \sin\phi$ is minimized (ϕ being a matter of personal choice for an individual judge or juror);

- *gravitational wave detection:* fix a maximum acceptable false alarm rate and choose the rule that, among all rules whose false alarm rate is so constrained, has the minimum false dismissal rate.

False alarm and dismissal rates describe our confidence in the long-run behavior of the associated decision rule. To understand the implications of this measure of confidence, suppose that we have not one, but N independent and identical detectors all observing during the same hour. We use the same test, with false alarm rate α and false dismissal rate β, on the observations made at each detector, and find that, of these N observations, m lead us (through our inference rule) to reject H_0 and $N-m$ lead us to accept H_0. For a concrete example, suppose α is 1%, N is 10 and m is 3.

The probability of obtaining this outcome when the signal is absent (H_0 is true) is the probability of obtaining m false alarms in N trials, or

$$P(m|H_0, N) = \frac{N!}{(N-m)!m!}\alpha^m(1-\alpha)^{N-m}. \tag{116}$$

In our example, $P(m|H_0, N)$ evaluates to 1.1×10^{-4}. It is thus very unlikely that we would have made this observation if the signal were absent. Does this mean we should conclude the signal is present with, say, 99.99% confidence?

No! $P(m|H_0, N)$ describes the probability of observing m false alarms out of N observations. When the signal is *present,* however (*i.e.,* when H_0 is false), there are *no* false alarms and both α and $P(m|H_0, N)$ are irrelevant. There are, however, $N-m$ false dismissals; thus, the relevant quantity is $P(m|\neg H_0, N)$, the probability of observing $N-m$ false *dismissals:*

$$P(m|\neg H_0, N) = \frac{N!}{(N-m)!m!}(1-\beta)^m\beta^{N-m}. \tag{117}$$

If, in our example, the false dismissal rate β is 10%, then the probability of observing seven false dismissals out of ten trials is 8.7×10^{-5}.

The particular outcome of our example—three positive results out of ten trials—is, in the grand scheme of things, very unlikely; nevertheless, what is important to us is that it is *more* unlikely to have occurred when the signal is present than when it is absent. Despite the apparently overwhelming improbability of three false alarms in ten trials, it is nevertheless, slightly more likely than the alternative of seven false dismissals in ten trials.

We can now answer the question that began this section. We understand that question to ask for the error rate of the best general procedure for deciding between the alternative hypotheses. There is an implicit assumption regarding the decision criteria, which tells us what "best" means in this context. In the context of these criteria, we calculate the error rates for different inference rules, rank the different rules, and find the best rule and its corresponding error rates.

Contrast this with our understanding of the identically worded question as we understood it in the previous subsection. There, we understood confidence to mean the degree of belief that we should ascribe to alternative hypotheses; here, we understand confidence to refer to the overall reliability of our inference procedure. There we responded with a quantitative assessment of our degree of belief in the alternative hypotheses, *given a particular observation made in a particular detector over a particular period of time;* here we responded with an assessment of the relative frequency with which our rule errs given each alternative hypothesis.

There we did not make a choice between alternative hypotheses; rather, we rated them as more or less likely to be true in the face of a particular observation. Here, on the other hand, we do make choices and our concern is with the error rate of our procedure for choosing, averaged over many different observations and many different decisions.

Analyses like the ones in this section, where probability is interpreted as the limiting frequency of repeatable events and the focus is on false alarm and false dismissal frequencies, are termed Frequentist analyses. They have particular utility when it is possible to make repeated observations on identical systems: *e.g.,* particle collisions in an accelerator, where each interaction of particle bunches is a separate "experiment." Analyses like those in the previous section, where probability is interpreted as degree-of-belief and the focus is on the probability of different hypotheses conditioned on the observed data, are termed Bayesian analyses.

Bayesian analyses are particularly appropriate when the observations or experiments are non-repeatable: *e.g.,* when the sources are, like supernovae, non-identical and destroy themselves in the process of creating the signal. In this case we are interested in the properties of the individual systems and would prefer a measure of the relative degree of belief that we should ascribe to, for example, the proposition that the signal originated from a particular point in the sky.

That Bayesian and Frequentist analyses are different does not imply that one is right and the other wrong. Bayesian and Frequentist analyses do not address the same questions; so, they are not required to reach "identical" conclusions. On the other hand, it may well be that one analysis is more appropriate or responsive to

our concerns than the other. We can only make the choice of appropriate analysis tools when we understand the distinction between them.

5 Gravitational Radiation Sources

In this section we review briefly some of the different kinds of sources that are, at this writing, thought to be "important" for the generation of large interferometric detectors now under construction. "Important" is a term that requires definition in this context. Clearly, sources that we don't expect to detect, or to be able to detect, are unimportant. Detectable sources must radiate significant energy in the bandwidth where these detectors are most sensitive (ranging from the tens to hundreds of Hz) over a reasonable observation period. For periodic sources, this is the integrated power over a period of several months to perhaps as much as one to two years; for burst sources, this means that the expected rate of detectable bursts must be at least several per year.

The assessment of source strength and number or rate is difficult to make for most sources. Very often the radiation strength depends on physics and astrophysics that we don't know or understand in the requisite detail. For all proposed sources the rate, number, or even existence of sources whose signal strengths are large enough to be detectable is difficult to ascertain. This is not surprising: what we know of the heavens we know principally through electromagnetic observations; however, it is in the nature of gravitational wave sources that they leave little electromagnetic evidence of their existence.

Finally, since we are, with these instruments, looking at the universe in a fundamentally new way, we must keep an open mind to the possibility of sources unimagined: in this I side with John Haldane, who said (in a different context) "My own suspicion is that the universe is not only stranger than we suppose, but stranger than we can suppose."

One final note: until now, we have been careful to keep all factors G and c in our expressions for gravitational effects. Here and henceforth, we will write all expressions in units where G and c are unity: *e.g.,* units of length, with conversion factors from grams and ergs to centimeters as given in equations 6 and 7. These conversion factors can be invoked to find expressions in terms of quantities expressed in more conventional units.

5.1 Burst Sources

5.1.1 Compact Binary Inspiral

The source most-talked-about for the interferometric detectors now under construction are binary systems consisting of two compact, stellar mass objects—either neutron stars or stellar mass black holes. Like a rotating dumbbell, a binary star system has a large, accelerating quadrupole moment, which makes it (for its mass) a strong gravitational radiation source. The radiation carries away orbital binding energy and orbital angular momentum, which leads to a faster and more compact orbit. Kepler's Third Law relates the orbital frequency $f_{\rm orb}$, semi-major axis a, and total system mass M by

$$f_{\rm orb}^2 = \frac{M}{4\pi^2 a^3}. \tag{118}$$

Consequently, the radiated power is, in order of magnitude,

$$L \propto \left[M a^2 f_{\rm orb}^3\right]^2. \tag{119}$$

Note that radiated power *increases* as the orbit decays:

$$L \propto \left(\frac{M}{a}\right)^5. \tag{120}$$

For binaries that can become sufficiently compact, the power radiated gravitationally will, in the end, become large enough to dominate the system's evolution. Since the radiated power increases as the orbit decays, the system will then decay at an ever increasing rate, with ever increasing radiation amplitude, frequency, and power, until the components coalesce. It is the radiation from this *inspiral,* for binary systems of neutron stars or black holes, that is seen as an important source for the LIGO and VIRGO detectors.

Why compact components, like neutron stars or black holes, and why stellar mass, and not more or less massive? Recall that the proposed interferometric detectors have their greatest sensitivity at approximately 150 Hz. The quadrupole radiation from a binary system is at twice the system's orbital frequency; correspondingly, if the radiation is to be in the bandwidth of these detectors, the binary systems themselves must exist with orbital frequencies of at least 75 Hz. Kepler's Third Law places a lower bound for us on the matter density of the components, which must be much greater than the total system mass divided by the cube of

the orbital radius:

$$\rho_{1,2} \gg \frac{M}{a^3}, \tag{121}$$

$$\simeq \left(\pi f_{\rm gw}\right)^2, \tag{122}$$

$$\simeq 1.5 \times 10^{12} \left(\frac{f_{\rm gw}}{100\,{\rm Hz}}\right)^2 {\rm g/cm^3}. \tag{123}$$

Thus, irrespective of the total system mass, if a binary system is to radiate in a band where these detectors are sensitive, the central density of its components cannot be much less than nuclear density. With this we are forced, for astrophysical objects, to restrict attention to neutron stars or black holes.

The nuclear and super-nuclear equation of state place an upper limit on the neutron star mass, which does not apply for a black hole. The dynamics of the binary orbit, however, does place an upper limit on the mass of the black hole binaries that the ground-based interferometric detectors may observe. With every orbit the binary radiates away more of its binding energy, leading to a more compact orbit. Eventually the system coalesces: the two components merge, collide, or tidally disrupt. Even if we imagine that the components are point masses, so that there is no tidal disruption or collision that would terminate the inspiral signal at some finite orbital frequency, relativity appears to impose a maximum orbital frequency on binary systems. For approximately symmetric binary star systems (*i.e.,* those with equal mass components) this limit is[33]

$$f_{\rm max} \simeq 710 \frac{2.8\,{\rm M}_\odot}{M}\,{\rm Hz}, \tag{124}$$

where M is the system's total mass. Thus, the component black hole masses must be less than 15 $M_\odot$ if the inspiral signal is to survive into the bandwidth where the detector is most sensitive.

It is currently thought that, during the epoch when the radiation from the binary is in the bandwidth where the LIGO and VIRGO detector sensitivity is greatest, the binary components are well approximated as point masses for the purpose of computing the radiation and orbital evolution[34] (There is some small suggestion that resonant tidal interactions may complicate this picture[35]). During this epoch, the gravitational fields that determine the binary evolution are sufficiently strong that first order perturbation theory is not adequate to compute the orbits; nevertheless, the fields are not so strong that computing the orbits and the radiation via higher order perturbation theory is impractical.[36,37] For this overview, no additional insight is gained by considering anything higher the quadrupole formula radiation, in which case the excitation of the detector—an effective $h(t)$ that is a superposition of the radiation in the two polarization states of the wave—is[38,39]

$$h(t) = \frac{\mathcal{M}}{d_L} \Theta \left(\pi f \mathcal{M}\right)^{2/3} \cos\Phi(t), \tag{125}$$

where

$$\mathcal{M} \equiv \frac{\left(m_1 m_2\right)^{3/5}}{\left(m_1+m_2\right)^{1/5}}(1+z), \tag{126}$$

$$\Theta^2 \equiv 4\left[F_+^2(1+\cos^2\iota)^2 + 4F_\times^2\cos^2\iota\right], \tag{127}$$

$$f(t) \equiv \frac{1}{\pi\mathcal{M}}\left(\frac{5}{256}\frac{\mathcal{M}}{T_0-t}\right)^{3/8}, \tag{128}$$

$$\Phi(t) = \int^t 2\pi f(t')dt', \tag{129}$$

d_L is the cosmological luminosity distance to the source, m_1 and m_2 are the binary system's component masses, z is the source's cosmological redshift, ι is the angle between the binary's angular momentum axis and the line of sight to the detector, and T_0 is a constant of integration.

What can we determine through observation of the signal from such a system? The signal-to-noise, of course, which takes on a particularly simple form[38,39]:

$$\overline{\rho^2} \simeq 1 + 25\left(\frac{r_0}{d_L}\right)^2 \left(\frac{\mathcal{M}}{1.2\,{\rm M}_\odot}\right)^{5/3}, \tag{130}$$

where r_0 is a characteristic distance that depends only on the effective power spectral density of the source,

$$r_0^2 = \left(\frac{G{\rm M}_\odot}{c^2}\right)^2 \frac{5}{192\pi}\left(\frac{243}{7\times10^5}\right)^{1/3} \int_0^\infty \frac{df\ \left(\pi G{\rm M}_\odot/c^3\right)^2}{\left(\pi f G{\rm M}_\odot/c^3\right)^{7/3} S_h(f)} \tag{131}$$

and the average denoted by the over-bar is over both an ensemble of detectors and all relative orientations of the source and the detector. For the initial LIGO and VIRGO detectors, r_0 is about 13 Mpc. In order that we are confident that we have seen a source, the SNR ρ^2 should not be much less than about 65 in a single detector; so, we don't expect to see sources from distances beyond more than a few r_0.

How many of these sources can LIGO expect to see? Unfortunately, we know very little about the rate of compact binary coalescence, except that it is rare.

Black-hole/black-hole binaries are, of course, invisible to us except through gravitational waves. Binaries involving neutron star component(s) are observable to us only if one of the components is a pulsar. Pulsars are observable only if they are not too distant (in our galaxy or its satellite globular clusters) and if the pulsar beam intersects our line of sight.

There are only three known binary pulsar systems in or about our own galaxy that will coalesce in less than the age of the universe; of these three, one in particular drives the calculation of the rate density. If we attempt to project this meager observational data throughout the entire universe, accounting for observational biases that cause us to miss some fraction of the actual number of binary systems, we find that the rate density of coalescing binaries is

$$\dot{n} \simeq 10^{-7\pm2}\,\mathrm{Mpc}^{-3}\,\mathrm{yr}^{-1}. \tag{132}$$

The great uncertainty in how to project the observations throughout the universe is reflected in the factor of 10^4 uncertainty in this rate. (For a sense of the uncertainties and corresponding controversy surrounding the estimates of the rate density of inspiraling binary coalescence, see[40–46] and references therein.)

The quantity r_0, defined in equation 131, was constructed in such a way that, assuming that sources are distributed homogeneously and isotropically throughout space, the rate of sources observed above a signal-to-noise ρ of 8 (ρ^2 of 64) is equal to

$$\dot{N} = \frac{4\pi}{3} r_0^3 \dot{n}. \tag{133}$$

For the initial LIGO detectors, r_0 is only about 13 Mpc; so, the *anticipated* rate of binary inspiral is, even at its most optimistic, low for the first generation of detectors. Things get better for the more advanced instruments now on the drawing board: for these, r_0 climbs to over 100 Mpc and, by correlating the signal from the two LIGO 4 Km detectors and the LIGO 2 Km detector, the effective r_0 can be increased by another factor of 3/2.[39]

In addition to the SNR, we also observe the scaling of the radiation frequency with time. From equation 128 this gives us $\mathcal{M}$, the so-called chirp mass, which depends on the component masses and the source's cosmological redshift. Knowing both the SNR and the chirp mass raises the interesting possibility of measuring the Hubble constant: the rate of cosmological expansion. Measuring the SNR tells us, up to the complications of the noise and the unknown orientation angles, something about the luminosity distance to the source. Similarly, measuring the chirp mass tells us, up to the unknown component masses, something about the source redshift. There is thus a hidden redshift/luminosity-distance relationship in observations of binary inspiral. By statistical analysis of a large number of binary inspiral observations, that relationship can be extracted and with it the Hubble constant.[39]

5.1.2 Compact Binary Coalescence

Eventually the inspiraling orbit of a binary system with compact components must end: the neutron stars collide, the black holes merge, or the one black hole tidally disrupts its neutron star companion. The radiation arising from the last few inspiral orbits through the coalescence of the compact objects that compose the binary may also be a significant source of radiation.

Unfortunately, very little is really known about the gravitational waves that result from the late stages of inspiral or the coalescence of either neutron star or black hole binary systems. In both cases the gravitational fields are quite strong and dynamical, which would appear to rule-out a perturbative approach and require a (numerical) solution to the fully non-linear Einstein field equations.[47] (In the neutron star case, the problem is further complicated by the need to model the dynamics of the fluid, which cannot be ignored in a coalescence.)

Numerical simulations of the coalescence of two black holes in a head-on collision have been calculated and provide some guidance: these tend to show that the total energy is disappointingly small: on order 10^{-4} of the system's total mass-energy.[48–52] There are also some recent calculations of off-axis collisions, which suggest strongly that the maximum energy radiated in such a collision will be less than 1% of the total mass energy.[53] In particular, it is difficult to justify the additional, *ad hoc* factor of 10% Mc^2 assumed by some authors in estimating the detectability of these sources.[54,55]

Surprisingly, perturbation calculations of the radiation arising from colliding black hole spacetimes give results that are in close accord with the limited number of fully relativistic numerical simulations that have been performed.[51,56–61] This accord is difficult to explain and may signal that we have something new to learn about the nature of the solutions to the full field equations.

5.1.3 Black Hole Formation

Black holes form in the collision of neutron stars at the end-point of neutron star binary inspirals; they also form in the core collapse of sufficiently massive stars. Unless the formation mechanism is especially symmetric, the new black holes that form will be initially quite deformed and will need to radiate away their deformations before they can settle down into a quiescent state, which is axisymmetric.

Quiescent black holes are characterized only by their mass M and angular momentum J. (And electric charge, too; however, astrophysical black holes are unlikely to carry any significant electric charge.) Correspondingly, while the initial radiation from the formation of a black hole depends on the details of the formation, the final radiation depends principally on M and J. In fact, the late-time waveform from a perturbed black hole is a superposition of exponentially damped sinusoids, whose frequencies and damping times depend only on M and J, the overtone number n, and the harmonic order ℓ and m of the perturbation.

Almost all of the modes of a black hole are very strongly damped. The most weakly damped modes are associated with the fundamental quadrupole-order excitations. Even these are strongly damped unless the black hole is very rapidly rotating. For this reason, we focus attention on the fundamental quadrupole modes, which are the most likely to be detectable. Setting aside the start-up transient associated with the details of the initial excitation, a good model for the "ring-down" of a newly-formed or perturbed black hole is thus[62,29]

$$h_{\rm RMS}(t) = 2\sqrt{\frac{2\epsilon}{Q(a)F(a)}}\frac{M}{r}e^{-\pi ft/Q}\sin(2\pi ft) \qquad (t>0), \tag{134}$$

where the amplitude is averaged (in a root-mean-square sense) over all orientation angles,

$$f \simeq 12.\,{\rm KHz}\left(\frac{{\rm M}_\odot}{M}\right)\left(\frac{F(a)}{37/100}\right), \tag{135}$$

$$Q \simeq 2(1-a)^{-9/20}, \tag{136}$$

$$a \equiv \frac{J}{M^2}, \tag{137}$$

$$F(a) \simeq 1-\frac{63}{100}(1-a)^{3/10}, \tag{138}$$

r is the distance from the black hole to the detector, ϵ is the fraction of the total mass of the black hole carried away in radiation, and we have assumed that all five of the fundamental tone quadrupole modes are excited equally. Corresponding to this radiation is an estimated signal-to-noise ratio of

$$\overline{\rho^2} = 1+34\,G(a)^2\frac{\epsilon}{10^{-4}}\left(\frac{20\,{\rm Mpc}}{r}\right)^2\left(\frac{M}{13\,{\rm M}_\odot}\right)^3\frac{10^{-46}\,{\rm Hz}^{-1}}{S_h}, \tag{139}$$

where we have assumed

- an efficiency ϵ for fraction of the rest mass of the system radiated gravitationally that is equivalent to what is found in black hole collisions, and
- the effective noise power spectral density is approximately constant over the signal bandwidth (which is broad for strongly damped oscillations).

The rate of black hole formation is entirely uncertain; however, most astrophysicists see no reason why the same mechanisms that make neutron stars cannot also make black holes at approximately the same rate.[45] By our present understanding of formation mechanisms, this rate is not high even at the distance of the Virgo cluster (~ 20 Mpc): perhaps as many as a few per year, but likely much less. Consequently, $\overline{\rho^2}$ should be at least on order 30–35 for a confident detection in ideal circumstances.[63] The caveat of "ideal circumstances" is an important one, however: the character of the waveform for this source—an exponentially damped sinusoid—is exactly the kind of technical noise one might expect in a real interferometer owing to transient disturbances that affect, for example, the suspension of the interferometer mirrors. Thus, without strong assurance that what is observed is not a weak disturbance intrinsic to the detector, prospects are not good for observing radiation from this source.

5.1.4 Stellar Core Collapse

Theoretical models of stellar core collapse, and the corresponding gravitational wave luminosity, have a long and checkered history: estimates of the gravitational wave luminosity have, over the last 30 years, ranged over more than four orders of magnitude.[63–67] It is not simply the luminosity that is unknown: the waveforms themselves are also entirely uncertain, leading to a further difficulty in estimating the detectability of this source. (Examples in the literature can be found in the citations.[63,65–70,13] Nevertheless, it is still possible to evaluate what is required of stellar core collapse in order that it be observable in a given detector.[71]

Suppose that the waveform from supernovae is given by

$$h_+ = \frac{2M_\odot}{r}\alpha f_+ m(t), \tag{140}$$

$$h_\times = \frac{2M_\odot}{r}\beta f_\times m(t), \tag{141}$$

α and β are constants, f_+ and $f_\times$ are functions of the relative orientation of the source with respect to the detector, and $m(t)$ is some function of time which we leave undetermined for now. The power radiated into each polarization mode is given by

$$\dot{E}_+ = \alpha^2 \left\langle f_+^2 \right\rangle M_\odot^2 |\dot{m}|^2, \tag{142}$$

$$\dot{E}_\times = \beta^2 \left\langle f_\times^2 \right\rangle M_\odot^2 |\dot{m}|^2, \tag{143}$$

where <> signifies an *average* over a sphere centered on the source.

Now assume that equal power is radiated into the two polarization modes. Then we can write α and β in terms of a single parameter ϵ as

$$\alpha^2 = \frac{\epsilon}{2M_\odot \left\langle f_+^2 \right\rangle \int dt\, |\dot{m}|^2}, \tag{144}$$

$$\beta^2 = \frac{\epsilon}{2M_\odot \left\langle f_\times^2 \right\rangle \int dt\, |\dot{m}|^2}. \tag{145}$$

In terms of ϵ the power radiated into the + and × polarization states is thus

$$\dot{E}_+ = \dot{E}_\times = \frac{\epsilon M_\odot |\dot{m}|^2}{2 \int dt |\dot{m}|^2}. \tag{146}$$

Finally, return to consider the time dependence of the waveform $m(t)$. Note that, by the Parseval Theorem,

$$\int dt\, |\dot{m}(t)|^2 = \int df\, (2\pi f)^2 |\widetilde{m}(f)|^2, \tag{147}$$

where $\widetilde{m}$ is the Fourier transform of m. Assume, as suggested by numerical simulations, that the (real) power radiated per unit bandwidth is approximately constant for frequencies in the interval $(f_{\rm min}, f_{\rm max})$ and falls off rapidly outside that band, with $f_{\rm min}$ on order 100 Hz and $f_{\rm max}$ approximately 1 KHz. In this approximation,

$$(2\pi f |\widetilde{m}(f)|)^2 \simeq \frac{\epsilon M_\odot}{f_{\rm max} - f_{\rm min}}. \tag{148}$$

We can now evaluate the SNR we expect from core collapse supernova gravitational waves incident on the detector. Current calculations suggest ϵ in the range 10^{-9}–$10^{-8} M_\odot$, with peak power in the 200–300 Hz band.[63,67,72] For convenience here, suppose that the detector noise power spectral density S_h is approximately constant over the bandwidth of the signal (*i.e.*, from $f_{\rm min}$ to $f_{\rm max}$) and (optimistically) equal to its approximate value at 100 Hz, and that $f_{\rm min}$ is very much less than $f_{\rm max}$. The mean-square signal-to-noise ratio is thus

$$\begin{aligned}\overline{\rho^2} &\simeq 1 + \frac{\epsilon M_\odot}{r^2} \frac{1}{2\pi^2 S_h} \frac{1}{f_{\rm max} f_{\rm min}}, \\ &\simeq 1 + 2.3 \frac{\epsilon}{10^{-8}} \left(\frac{15\,{\rm Mpc}}{r}\right)^2 \frac{10^{-46}\,{\rm Hz}^{-1}}{S_h} \frac{100\,{\rm Hz}}{f_{\rm min}} \frac{1\,{\rm KHz}}{f_{\rm max}}.\end{aligned} \tag{149}$$

The distance of 15 Mpc is the range to the center of the Virgo Cluster of galaxies: if supernovae can be reliably observed to distance, we can expect a rate of approximately three per year. Reliable observation of millisecond bursts at a rate of three per year, however, requires an SNR $\overline{\rho^2}$ of somewhat more than 30. Thus, without a very optimistic efficiency ϵ, we can't expect to be able to observe supernovae much beyond our own galaxy, and certainly not out to the Virgo cluster.

5.2 Periodic Sources of Gravitational Radiation

All the anticipated sources of periodic gravitational waves for the ground-based detectors now under construction involve tapping the stored rotational energy of rapidly rotating neutron stars. This means that the neutron star must be in some way non-asymmetric. The strength of the radiation depends on the degree of asymmetry. All the uncertainty associated with periodic sources of gravitational radiation arises either with the mechanism for producing the asymmetry or the degree of asymmetry.

At this writing four different kinds of asymmetries, or mechanisms for producing asymmetries, are discussed as possibly leading to detectable gravitational waves. We discuss these in subsections 5.2.1, 5.2.4, 5.2.5, and 5.2.6 below. Additionally, we discuss observational constraints on the radiation from isolated pulsars in subsection 5.2.2, and some issues related to the detection of periodic radiation in subsection 5.2.3.

5.2.1 Non-Axisymmetric Rotators

All but the youngest neutron stars have a solid, crystalline crust, covered by a fluid surface and with a fluid interior. The fluid surface and interior adjust

themselves to the star's rotation, remaining always axisymmetric and, therefore, not contributing to any gravitational radiation. As the star cools or spins-down (owing to, *e.g.*, magnetic multipolar radiation if it is a pulsar), the shape of its crust cannot adjust continuously to its new conditions. The stresses in the crust build until the crust fractures, relieving the stress. The final crust shape is likely to be non-axisymmetric and responsible for gravitational radiation as the star rotates.

Suppose that the star is rotating about a principal axis of its moment of inertia tensor with rotational rate f. Let I_3 be the moment of inertia along the axis or rotation, I_1 and I_2 be the other two principal moments of inertia, and define ϵ to be the difference between I_1 and I_2 relative to I_3:

$$\epsilon = (I_2 - I_1)/I_3. \tag{150}$$

Setting aside the very slow spin-down of the system as its angular velocity changes and averaging over the angles that describe the relative orientation of the pulsar with respect to the detector, the characteristic radiation from this system is given by

$$h(t) \simeq h_0 \cos(4\pi f t + \phi_0), \tag{151}$$

$$\overline{h}_0 = \frac{32\pi^2\sqrt{2}}{5}\frac{\epsilon f_0^2 I_3}{r}, \tag{152}$$

$$= 4.8 \times 10^{-26} \frac{I}{10^{45}\,\mathrm{g\,cm^{-3}}} \frac{\epsilon}{10^{-6}} \frac{10\,\mathrm{Kpc}}{r} \left(\frac{f}{300\,\mathrm{Hz}}\right)^3. \tag{153}$$

The power radiated gravitationally through this mechanism depends, through ϵ, on the degree of asymmetry that can be supported by the neutron star crust. Alpar and Pines[73] have looked at the structure of the crust and the likely strain that it can support. For our purposes it is instructive to look at the most extreme possibility they considered: that the crust is well approximated as a pure Coulomb-lattice crust. Such a lattice could sustain a strain some 10^3 to 10^4 times as much as is typical of terrestrial material. When the maximum allowable strain is supposed to be supported by the solid part of the neutron star (which is only a small fraction of the entire star), one arises with a maximum ϵ of approximately 10^{-6}.

This is an extreme value: it depends on the crust being a pure Coloumb lattice, assumes that some mechanism has led it to be stressed to its fracture point, and that the corresponding strain is principally quadrupolar. For young neutron stars one might imagine this conspiracy of circumstance possible; however, for older neutron stars plastic flow of the crust would lead to relaxation over the age of the most rapidly rotating neutron stars—the so-called millisecond pulsars—reducing the maximum ϵ for these systems to no greater than 4×10^{-10}.

5.2.2 Observational Constraints

There is observational evidence that, at least for the older, millisecond pulsars, ϵ cannot be much greater than this limit. The power L radiated gravitationally by a spinning neutron star comes directly from the star's rotation; consequently, radiation back-reaction must slow the star in such a way that energy is conserved. This leads to a slow spin-down of the star: if P is the spin period, then its rate of change $\dot{P}$, assuming that gravitational radiation reaction is the only source of spin-down, is

$$\dot{P} = \frac{LP^3}{(2\pi)^2 I}. \tag{154}$$

Since the radiated power L is proportional to $\epsilon^2 I/P^6$ (cf. 31), the measured period and period derivative place a strict upper limit on ϵ for isolated pulsars.

The spin-down rate ($\dot{P}$) of most pulsars has been measured. If we take the most extreme view and ascribe all of the spin-down to angular momentum carried off by gravitational waves, the oblacity of millisecond pulsars still cannot exceed 10^{-9} for most millisecond pulsars.[73]

For a few young, isolated pulsars, timing observations are so good that we can place still stronger limits on ϵ: limits that exclude the possibility that a significant part of the spin-down is owing to radiation reaction. For these pulsars, not only the rate of the spin $\dot{P}$ but also its second derivative $\ddot{P}$ has been measured. Using only that the rotational energy of the star is proportional to I/P^2 and that the radiated power is proportional to P^{-1-n}, one can quickly show that

$$\frac{P\ddot{P}}{\dot{P}^2} = 2 - n. \tag{155}$$

If the spin-down is due to quadrupole gravitational radiation reaction, n is equal to 5 and higher-order radiative moments would lead to larger n. On the other hand, if the spin-down is due to, say, magnetic dipole radiation (from the rotation of the pulsar's magnetic dipole moment), n is equal to 3. There are no isolated pulsars for

which the measured $P\ddot{P}/P^2$ yields an n approaching 5. In particular, for the Crab pulsar, which is the prototypical young pulsar, the measured n is approximately 2.5. This is strongly inconsistent with the suggestion that gravitational radiation damping plays an important—let alone dominant—role in the spin-down of the Crab, or any other known, pulsar. Thus, we must conclude that effective oblacity ϵ of pulsars is, in fact, quite small and the radiation quite weak compared to the detector noise.

5.2.3 Detecting Periodic Sources

That the signal amplitude is small is, by itself, not of overwhelming concern. Consider for a moment the signal-to-noise associated with an observation of a periodic signal, as is given in equation 151, over a period T long compared to the periodicity of the radiation $1/f_{\rm gw}$. The SNR is given by

$$\rho^2 = 1 + 4\int_0^\infty df \frac{\left|\tilde{h}(f)\right|^2}{S_h(f)}. \tag{156}$$

Since the signal is monochromatic over the observation period, $\tilde{h}$ is appreciable only at frequencies near $f_{\rm gw}$; consequently, we can set $S_h(f)$ equal to $S_h(f_{\rm gw})$ in equation 156. Then, invoking The Parseval Theorem, we can rewrite equation 156 in the time domain:

$$\begin{aligned}
\rho^2 &= 1 + 4\int_0^\infty df \frac{\left|\tilde{h}(f)\right|^2}{S_h(f)}, &(157)\\
&= 1 + 2\int_{-\infty}^\infty df \frac{\tilde{h}(f)^2}{S_h(f_{\rm gw})},\\
&= 1 + \frac{2}{S_h(f_{\rm gw})}\int_0^T dt\, h(t)^2,\\
&= 1 + h_0^2 \frac{T}{S_h(f_{\rm gw})}; &(158)
\end{aligned}$$

i.e., for periodic sources the SNR grows with the observation period. As a result, a signal with a small peak amplitude can be detected much more readily if it is periodic than if it is a burst signal of finite duration.

In reality, there are practical limits to how large the SNR can be made. An observation can only be so long: for LIGO an observation lasting a full year would be quite long. Additionally, the signal is not quite as simple as we have assumed. The detector's motion with respect to the source is non-uniform, owing both to Earth's rotation about its axis and its orbit about the sun. The time-dependent Doppler shift of the signal leads to a frequency modulation, which depends on the position of the source in the detector's sky. In order to obtain the growth of ρ^2 with time that we found above that frequency, modulation must be removed.

Neither of these issues is significant if we know the position of the source on the sky and its frequency. If, on the other hand, we are contemplating a "blind" search, across the sky or over a wide bandwidth in frequency, the story changes. If we are only interested in one frequency, say f_0, then by folding the data we can find $\tilde{h}(f_0)$. If we are interested in a wide range of frequencies, however, the most effective way to find $\tilde{h}(f)$ in that band involves a Fourier transform over the interesting bandwidth. For a significant bandwidth (say, several hundred Hz) and an observation lasting several months, this is a multi-gigapoint FFT. If, in addition, we are interested in searching over the sky, then we must demodulate differently for different points in the sky. For a several month observation, the number of independent patches in the sky is quite large and the computational resources required exceed any that might be available now or in the foreseeable future.[74] Thus, different ways of searching for periodic sources must be developed, which will not have the same growth rate of SNR with time.

Finally—and this is true even if we are interested only in one point on the sky and one frequency—as the observation gets longer, we must become concerned about the stability of our instrument and the characterization of its noise. One cannot expect the noise level to remain stationary over indefinitely long periods; additionally, as we observe for longer periods, we are, in an important sense, exploring in greater and greater detail the character of the noise in very narrow bandwidths. As the observation period increases we must ask, with an increasing degree of concern, how confident we are that there are no technical noise sources, such as weak, drifting oscillators, that may be masquerading as signal over the period of our observation.

5.2.4 Precession

An axisymmetric neutron star, rotating about its symmetry axis, does not radiate gravitationally. On the other hand, if the angular momentum is not coincident with the symmetry axis—*e.g.,* if the star is precessing—then it will radiate grav-

itationally. Misalignment of an axisymmetric neutron star's angular momentum and body axes could arise as the result of crustal fractures associated with a neutron star quake.

The same observational constraints that apply to gravitational radiation arising from the rotation of a non-axisymmetric neutron star about a principal axis also apply to radiation arising from precession of an axisymmetric neutron star (cf. Sec. 5.2.2).

While neutron star precession is—in principle—possible, if the neutron star is also a pulsar, the precession should also manifest itself as periodic variations in the electromagnetic pulse shape. At present there is no observational evidence for pulse shape variations induced by free precession. This may be because the misalignment is too small to be observed in the pulse shape or because the stresses associated with misalignment quickly bring the star back in to alignment.

Gravitational radiation associated with precession of an axisymmetric star occurs at both the rotational frequency and twice the rotational frequency[75]; consequently, it can be distinguished from the radiation associated with a fully non-axisymmetric star rotating about a principle axis. Interestingly, observing the amplitude of the radiation at both the rotation frequency and twice the rotation frequency allows one to determine all the angles that characterize the orientation of the star relative to the line-of-sight (LOS): the angle between the angular momentum and the LOS as well as the angle between the body axis and the angular momentum. If the star is also observable as a pulsar, then one can test models of pulsar beaming, since, together with the observed pulse shape, these make predictions about the angle between the LOS and the magnetic axis.

5.2.5 Thermally Driven Non-Axisymmetry

Timing of the x-ray emission from several accreting neutron stars has revealed quasi-periodic variability that can be explained as arising from the rapid rotation of the underlying neutron star. An intriguing coincidence in these observations is that the rotation rate of all these systems appears to be close to equal. This suggests that there is some underlying mechanism that ensures that accretion spins these stars up to—but not beyond—this limiting angular velocity. One possibility is that the rotation rate is limited by gravitational radiation reaction.

How might gravitational radiation limit the rotation rate of an accreting system? If the accretion leads to a non-axisymmetry in the neutron star then, as the star spins-up, the angular momentum radiated by this rotating non-axisymmetry increases until it balances the angular momentum accreted, limiting the star's rotation rate. The angular momentum radiated is, like the radiated power, a strong function of angular velocity ($\dot{J}$ is proportional to Ω^5, where Ω is the angular rotation rate); so, it is not surprising that the limiting angular velocity should be similar for these systems.

Proposals like this are characteristically made for systems where there appears to be some upper limit to the rotation rate. To be plausible, there must be some universal mechanism whereby the same process that spins the star up also leads to a non-axisymmetry that can cause a radiative loss of angular momentum. Recently Bildsten[76] offered some promising ideas for a mechanism like this that would operate in rapidly accreting, low magnetic field neutron stars like Sco X-1. At the core of Bildsten's proposal is the observation that localized heating of the neutron star owing to non-isotropic accretion leads to differential electron capture rates in the neutron star fluid. These lead, in turn, to density gradients as nuclear reactions in the neutron star adjust its composition. Bildsten suggested that, if the rotation axis is not aligned with the accretion axis and if some other mechanism (in Bildsten's original suggestion, a magnetic field) can break the symmetry still further, these density gradients may form in a non-axisymmetric fashion.

There are two big "ifs" in this proposal. Even if the accretion disk is misaligned with the star's rotation axis, the density perturbations will be distributed symmetrically about the star's rotation axis unless some other mechanism can be shown to break the symmetry further. Additionally, though not recognized in the original proposal, buoyancy forces will lead the density perturbations to re-distribute themselves in the star symmetrically about its rotation axis, significantly suppressing the gravitational radiation. For these two reasons the initial excitement over the Bildsten proposal has dampened. It should not be extinguished, however: the recognition that accretion can lead, through pyro-nuclear reactions, to density perturbations that may radiate gravitationally is certainly sound and new. With time will come greater understanding of where and how this effect may arise in nature, and that greater understanding may yet include a robust mechanism for producing significant gravitational radiation from accreting neutron stars.

5.2.6 R-Mode Instability

Also in the last year, Andersson[77,78] discovered a new class of unstable perturbative modes of rotating relativistic stars. In the absence of gravitational radiation these modes are all stable; however, gravitational radiation back-reaction on the modes causes them to undergo exponential growth, feeding off of the rotational energy of the star.

These particular modes are unusual in two different respects:

1. In the absence of viscosity they are unstable for *all* angular velocities.

2. At leading order the radiation is entirely "magnetic" in character: *i.e.,* the radiation couples to the momentum density distribution and its time evolution, not the mass density distribution and its time evolution.

Figure 9 gives a schematic view of the character of the fluid velocity perturbation, relative to the star's rotation, for the lowest order radiating mode (magnetic quadrupole) and for low rotational velocity. The lines indicate fluid flow-lines; the arrow indicates the relative direction of fluid flow on adjacent flow-lines. At low rotation rates there is very little radial component to the fluid motion; so, we show only the angular components.

In the absence of rotation, the fluid in the star can circulate in the pattern shown with constant velocity: *i.e.,* there are no restoring forces acting on the fluid's inertia and the eigenfrequency of the mode is zero. Correspondingly, the magnetic quadrupole moment of this fluid mode has constant amplitude. Rotation introduces Coriolis forces, which act back on the fluid's inertia and cause the circulation in these cells to be periodic. The quadrupole moment of the fluid momentum now has a second time derivative; correspondingly, it radiates gravitationally.

The power radiated gravitationally comes, ultimately, from the star's rotational energy. The radiation thus carries off positive angular momentum from the star. The greatest angular momentum can be carried by the modes with azimuthal quantum number m equal to $\pm\ell$; so, focus attention on these modes. *Relative to the star's rotation,* one of these modes is co-rotating and carries positive angular momentum and one is counter-rotating and carries negative angular momentum. It turns out, however, that both of these modes are dragged, by the star's overall rotation, so that *viewed from an inertial frame, both carry positive angular mo-*

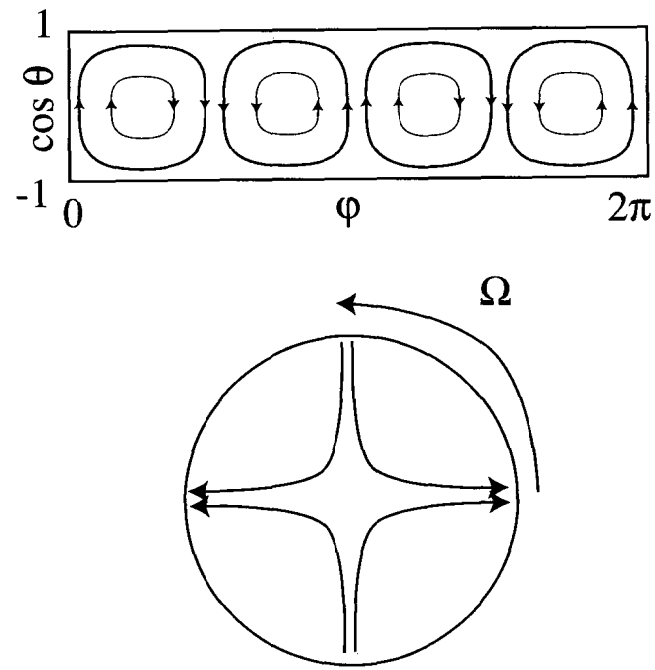


Fig. 9. Schematic illustration of the fluid flow-lines for the lowest order r-mode perturbation relative to the star's rotation. The arrow indicates the relative direction of fluid flow on adjacent flow-lines. In the upper panel the star's surface (co-latitude θ, longitude ϕ) is mapped to the plane; in the lower panel, the view is looking down along the star's rotation axis. In both cases, the flow pattern is dragged along by the star's rotation. At low rotation there is essentially no radial component to the fluid's motion; so, we show only the angular components.

mentum. The back-reaction of the radiation adds negative angular momentum to each of these modes; correspondingly, the co-rotating mode is damped and the counter-rotating mode is anti-damped: *i.e.,* it grows. As it grows, of course, it radiates more strongly, leading to greater anti-damping: the mode undergoes exponential growth.

This general mechanism by which gravitational radiation reaction can lead to amplification of a mode that is counter-rotating in the body frame but co-rotating in the inertial frame was first discussed by Zel'dovich in the context of rotating black holes. His concern was that, through this mechanism, a rotating black hole should radiate—a result that anticipated the more general result of Hawking. The first application to stars came from Chandrasekhar[79] and Friedman and Schutz[80] whose focus, however, was on a different set of modes.

Viscosity damps the R-modes. If the viscosity is large enough, then the viscous damping exceeds the anti-damping caused by radiation reaction and the mode is stabilized. Present understanding of the viscosity of neutron star fluid suggests that there is a short period in the life of a new-born neutron star, lasting perhaps one year, when the mode may be unstable and a rapidly rotating star may be a strong, nearly periodic source of gravitational radiation. The radiation is "nearly" periodic because its amplitude is so great that, over the course of the year, the star's angular rotation rate may evolve from 10^3 to 10^2 Hz, simply due to the angular momentum carried away in the radiation.[81,82]

5.3 Stochastic Gravitational Radiation

In the previous subsections we discussed burst and periodic sources of gravitational radiation. In both cases the discussion focused on the source, and the radiation was characterized typically in terms of a waveform $h(t)$, which depends on the details of a source and its orientation with respect to the detector.

The situation for a stochastic gravitational wave signal is different. A stochastic gravitational wave signal is intrinsically random in character. In particular, it is not generated by an isolated source, it is not incident on the detector from a single direction, and it has no characteristic waveform.

In fact, a stochastic signal can be treated as just another source of detector noise. The stochastic radiation has an $h(t)$ that is characterized solely by its correlation functions or associated spectra; correspondingly, the detector output owing to the stochastic signal is characterized in terms of correlation functions or associated spectra.

Detection of any signal hinges on observing some characteristic that distinguishes signal from noise. If a stochastic signal appears in a detector no different from intrinsic detector noise, how do we make the critical distinction that allows us to say we have detected a signal?

The essential difference between the action of a stochastic signal and intrinsic detector noise is that stochastic radiation incident on two detectors is correlated, and the correlation depends in a completely predictable way on the relative orientation and separation of the detectors. Any gravitational wave signal—stochastic or otherwise—can be resolved into a superposition of plane waves of different frequencies and propagation directions bathing the detectors. A component of the radiation of given frequency and incident direction drives two or more detectors coherently, with a phase delay that depends on the incidence direction, detector separation, and radiation wavelength. For components whose wavelength is much larger than the separation between the detectors, the phase difference is only weakly dependent on the radiation wavelength or incident direction; so, summed over incident direction there is a strong correlation between the output of the detectors. On the other hand, for radiation components with wavelengths much smaller than the separation between the detectors, the phase difference depends strongly on the incident direction and the wavelength; so, summed over incident directions, the correlation between the output of the detectors is weak. Thus, in the presence of a stochastic gravitational wave background, the output of two or more detectors should show predictable correlations that depend on their relative separations, relative orientations, and the stochastic signal's spectrum.[83]

When considering sources of detectable stochastic gravitational radiation for ground-based detectors, it is conventional to enumerate the contributions of primordial origin: *e.g.,* radiation arising during an inflationary epoch in the early universe,[84] from the decay of a cosmic string network,[85] or from a phase transition in the early universe.[86–88] Less frequently discussed—perhaps because it is so mundane—are the contributions arising from the *confusion limit* of discrete but unresolved sources: *e.g.,* core-collapse supernovae[18] or binary inspiral. In fact, the contribution at low-frequencies (periods of hours to minutes) to the stochastic signal from unresolved galactic binary systems is expected to be many times greater than the intrinsic detector noise of the proposed space-based interferomet-

Interval	# events	"rate"	
0–5 s	3	0.6/s	number in
5–10 s	4	0.8/s	any interval
10–15 s	11	2.2/s	varies from
15–20 s	2	0.4/s	long-run rate
20–25 s	5	1.0/s	
0–25 s	25	1.0/s	long-run rate

Table 1. Number of Poisson distributed events, with mean rate 1/s, occurring in consecutive five second intervals. Note that the number of events in any given interval varies: the rate is only the mean of the distribution of "rates" calculated over many intervals.

ric detector LISA.[89,27]

Let's talk a bit more about how discrete, well-defined sources—whose waveforms may be well known—can superpose to form a stochastic signal of predictable spectrum. (We focus on burst sources here; however, a similar story can be told for periodic ones.[89])

Astrophysical sources of gravitational-wave bursts strong enough to be observed by conceivable detectors are certainly independent events. Setting aside the slow evolution of the source population on cosmological timescales, the number of events whose radiation impinges upon the receiver during any finite length observation is Poisson distributed: if the event rate is $\dot{N}$ and the observation period is τ, then

$$P(n|\tau\dot{N}) = \frac{\left(\tau\dot{N}\right)^n}{n!} e^{-\tau\dot{N}}. \tag{159}$$

It is critical to recognize that the number of events exciting the detector at any given moment is not the constant equal to the product of the event rate and the signal duration. The actual number of events in the observation period varies. Averaged over many intervals it has a mean value, which is the product of the event rate and the signal duration. For example, table 1 shows the number of Poisson distributed events over consecutive five second intervals, the rate calculated separately in each interval, and the overall average rate.

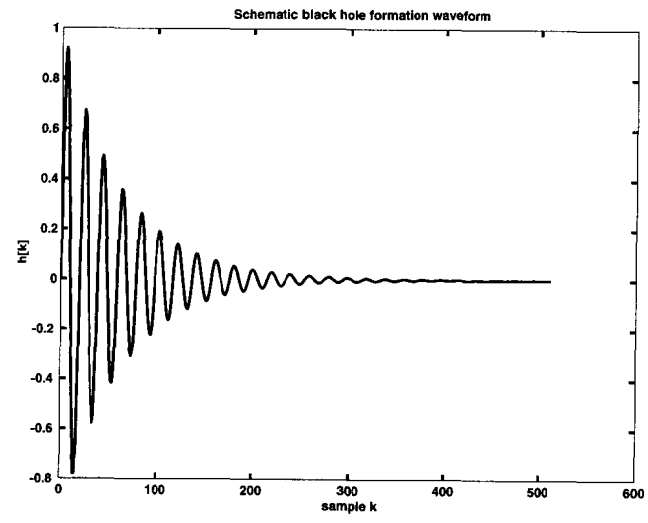


Fig. 10. The prototypical waveform associated with an impulsive perturbation of a black hole: an exponentially damped sinusoid.

Now let's suppose that the "events" we are discussing are the arrival at our detector of the initial wavefront associated with an impulsive perturbation of a black hole. That waveform is a damped sinusoid:

$$h(t) \propto e^{-2\pi f(t-t_0)/Q} \sin 2\pi f(t-t_0) \qquad \text{for } t > t_0. \tag{160}$$

The signal from a single, prototypical event is shown in figure 10. The left-hand panels of figure 11, on the other hand, shows $h(t)$ at the detector when signals exactly like these arrive at the detector with different rates. In the top panel the signals are, for the most part, clearly distinguishable. In the middle panel the rate is higher and separate signals can be distinguished only in exceptional cases. In the bottom panel the rate is higher still and the identity of the individual signals is completely obscured.

Especially in the bottom panel of figure 11, the signal $h(t)$ at any given t is the superposition of a large, but random, number of signals. Similarly, the amplitude of each contributing signal is itself random (corresponding to the signal amplitude

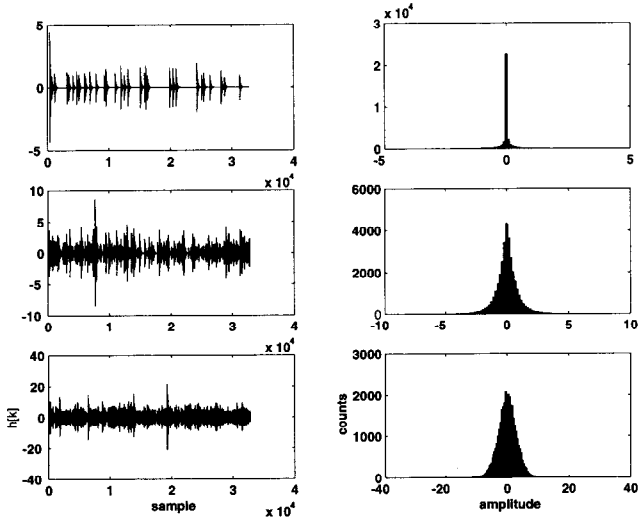


Fig. 11. The left-hand panes show waveforms corresponding to the superposition of many discrete and idealized black hole formation events, with the number of events in a fixed interval Poisson distributed. The rate increases by an order of magnitude from the top to the middle panel, and again from the middle to the bottom panel. The right-hand panes show the distribution of wave amplitude derived from the corresponding left-hand panel.

at a random moment relative to the signal start time). In circumstances like these we expect The Central Limit Theorem[90] to apply, leading to a normal distribution for $h(t)$. On the other hand, at the lower rate pictured in the top panel, there are not enough events superposed at any given moment for us to expect $h(t)$ to exhibit a normal distribution. In the right-hand panels of figure 11 we show the distribution of $h(t)$ taken from the left-hand panel and find that these intuitions are borne out.

Note that none of these conclusions depend in any way on the details of the signal from an individual source: instead of the superposition of damped sinusoids we could just as well have constructed $h(t)$ from the superposition of binary

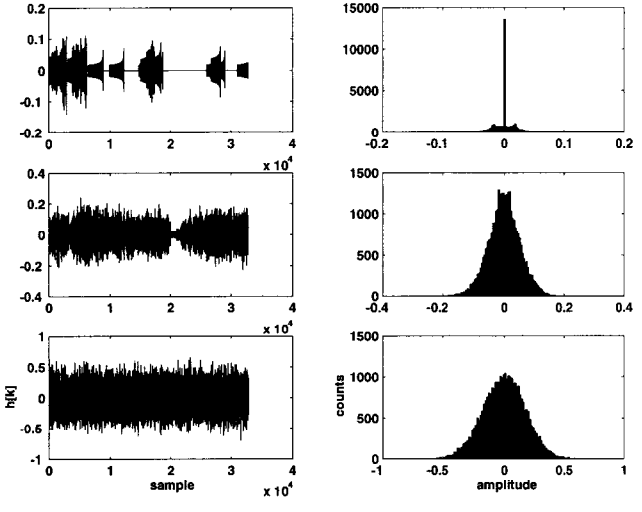


Fig. 12. As in figure 11, but with band-limited binary inspiral waveforms.

inspiral signals in a fixed bandwidth. We do exactly that in figure 12. Note how the distributions of $h(t)$ at high rate are, for superpositions of large numbers of signals, identical (*i.e.,* they are normal distributions).

Despite the fact that the distribution of $h(t)$ arising from the superposition of a large number of sources are identical, visual inspection suggests that there are still differences. These differences are associated with the correlations: the distribution of the products $h(t)h(t+\tau)$ as a function of τ. These are very different for the damped sinusoids, which are characteristic of black hole perturbations, and the "chirps," which are characteristic of binary inspirals. In figure 13 we show the power spectral densities of the superpositions in figures 11 and 12. Note how the power spectral densities of the stochastic signals formed from the random superposition of events of a given character preserve the spectral shape of the underlying signal, with the overall amplitude proportional to the event rate. It is this property of the superposition that permits us to predict the character of the stochastic signal arising from the confusion limit of a large number of sources; conversely, observation of a stochastic signal provides us, through its amplitude

and spectrum, with information about the underlying source, its rate, and spatial distribution.

This last point gives us the prospect of using a detected stochastic background, arising from the superposition of unresolved sources, to perform a source "census": a determination of the density of sources in space. This information, in turn, can shed light on astrophysical processes that are otherwise unobservable. An excellent example of this comes from LISA observations of the background from close white dwarf binary systems (CWDBs).[91] These binary systems, which have orbital periods ranging from days to hours, are so close that they are optically unobservable as binaries. We know they exist, nevertheless, because we can see their progenitor systems. CWDBs arise as one of the end-points of binary star evolution, following a particularly difficult to understand and model evolutionary stage where the two stars are orbiting each other within a single envelope of gas. Once they emerge from this final stage of *common envelope evolution,* they orbit each other as if point masses. Gravitational radiation reaction leads to a slow, secular decay in the orbit of these systems until they become so close that mutual tidal interactions lead to rapid orbital decay or stellar disruption.

Let dn/df be the number density of CWDBs with orbital frequency f. The amplitude of the stochastic gravitational wave signal from this population of objects is, at frequency $2f$, proportional to dn/df. CWDBs are introduced into the population at a rate $d\dot{n}_+/df$ as they emerge from the final stage of common envelope evolution at orbital frequency f; similarly, they leave the population at a rate $d\dot{n}_-/df$ as they disrupt at orbital frequency f. Knowing the rate $d\dot{n}_+/df$ would provide us with valuable information about common envelope evolution, which we cannot obtain through optical observations; similarly, the rate $d\dot{n}_-/df$ depends on the white dwarf mass spectrum (we know the white dwarf equation of state quite well) and knowing it would reveal the mass spectrum of binaries emerging from the common envelope evolution stage.

We can determine both $d\dot{n}_+/df$ and $d\dot{n}_-/df$ from LISA observation of the CWDB stochastic signal spectrum. Observation of this spectrum determines, as we have seen, dn/df. Once injected into the population, a CWDB's orbit evolves owing exclusively to gravitational radiation reaction, which proceeds at the rate df/dt, until it is removed from the population through coalescence or disruption. In steady state we thus have a *continuity equation* governing the

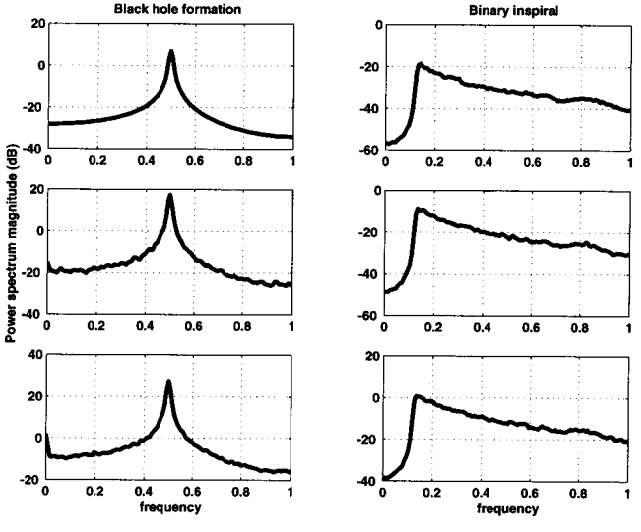


Fig. 13. Power spectra corresponding to the superpositions of black hole ringdown signals in figure 11 and binary inspiral signals in figure 12. The power spectra reflects the underlying source of the stochastic signal (black hole ringdown or binary inspiral) while the amplitude is proportional to the rate of individual sources contributing to the signal.

CWDB population:

$$\frac{d\dot{n}_+}{df} - \frac{d\dot{n}_-}{df} = \frac{d}{df}\left(\frac{dn}{df}\frac{df}{dt}\right). \quad (161)$$

The orbital frequency evolution rate df/dt is known (it is proportional to $f^{11/3}$) and LISA observations will determine the number density dn/df; consequently, from LISA observations we can determine $d\dot{n}_+/df$ and $d\dot{n}_-/df$ and learn about the end of common envelope evolution and the mass spectrum of white dwarfs in CWDBs.

6 Conclusions

In these lectures I've tried to give a brief overview of how we think about gravitational waves when we set out to detect them, and provide a snapshot of current thinking on the anticipated wave sources. Along the way, I've tried to describe some of the science we can hope to do once we can reliably detect gravitational wave sources.

The principal difficulty in discussing the sources that we hope to observe is our real lack of knowledge of their character. As is often the case, however, this difficulty is really a disguised opportunity: when the detectors come on-line and we begin to detect gravitational radiation sources, we will not simply be confirming what we already know, but learning things entirely new about the cosmos!

References

[1] J. H. Taylor and J. M. Weisberg. A new test of general relativity: gravitational radiation and the binary pulsar PSR 1913+16. *Astrophys. J.*, 253:908–920, 1982.

[2] J. H. Taylor and J. M. Weisberg. Further experimental tests of relativistic gravity using the binary pulsar PSR 1913+16. *Astrophys. J.*, 345:434–450, October 1989.

[3] S. D. Mohanty, Lee Samuel Finn, and Joseph Romano. Detecting and association between gamma-ray and gravitational wave bursts. *Phys. Rev.*, D60:121101, 1999.

[4] Steven Weinberg. *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity.* Wiley, New York, 1972.

[5] C. W. Misner, K. S. Thorne, and J. A. Wheeler. *Gravitation.* Freeman, San Francisco, 1973.

[6] Bernard F. Schutz. *A first course in general relativity.* Cambridge University Press, 1990.

[7] Hans Stephani. *General Relativity.* Cambridge University Press, Cambridge, second edition, 1993.

[8] Hans C. Ohanian and Remo Ruffini. *Gravitational and Spacetime.* Norton, New York, second edition, 1994.

[9] Kip S. Thorne. Multipole expansions of gravitational radiation. *Rev. Mod. Phys.*, 52(2):299–339, April 1980.

[10] John David Jackson. *Classical Electrodynamics.* Wiley, New York, second edition, 1975.

[11] Lee Samuel Finn. Gravitational waves from solar oscillations: Proposal for a transition-zone test of general relativity. *Class. Quant. Grav.*, 2:381–402, 1985.

[12] Stuart L. Shapiro and Saul A. Teukolsky. *Black Holes, White Dwarfs, and Neutron Stars.* Wiley, New York, 1983.

[13] Thomas Zwerger and Ewald Müller. Gravitational radiation from rotational core collapse. http://www.mpa-garching.mpg.de/~ewald/GRAV/grav.html, 1998.

[14] Warren W. Johnson and Stephen M. Merkowitz. Truncated icosahedral gravitational wave antenna. *Phys. Rev. Lett.*, 70(16):2367–2370, 19 April 1993.

[15] W. O. Hamilton, Ziniu K. Geng, Warren W. Johnson, Evan Mauceli, Stephen M. Merkowitz, Andrew Morse, and Norert Solomonson. Performance of the ALLEGRO detector—and what our experience tells us about spherical detectors. In Velloso et al.,[22] pages 19–26.

[16] E. Mauceli, Z. K. Geng, W. O. Hamilton, W. W. Johnson, S. Merkowitz, A. Morse, B. Price, and N. Solomonson. The ALLEGRO gravitational wave detector: Data acquistion and analysis. *Phys. Rev. D*, 54:1264, 1996.

[17] P. Astone, M. Bassan, P. Bonifazi, P. Carelli, E. Coccia, C. Cosmelli, V. Fafone, S. Frasca, A. Marini, G. Mazzitelli, Y. Minekov, I. Modena, G. Modestino, A. Moleti, G. V. Pallotino, M. A. Papa, G. Pizzella, P. Rapagani, F. Ricci, F. Ronga, R. Terenzi, M. Visco, and L. Votano. The gravitational wave detectors EXPLORER and NAUTILUS. In Velloso et al.,[22] pages 39–50.

[18] D. Blair and L. Ju. A cosmological background of gravitaitonal waves produced by supernovae in the early Universe. *Mon. Not. R. Astron. Soc.*, 283:648–650, 1996.

[19] Stephen M. Merkowitz and Warren W. Johnson. Spherical gravitational wave antennas and the truncated icosahedral arrangements. *Phys. Rev. D*, 51(6):2546–2558, 15 March 1995.

[20] Stephen M. Merkowitz and Warren W. Johnson. First tests of a truncated icosahedral gravitational wave antenna. *Phys. Rev. D*, 53(10):5377–5381, 15 May 1996.

[21] G. Frossati. A 100 TON 10mK spherical gravitational wave detector. In Velloso et al.,[22] pages 163–168.

[22] W. F. Velloso, Jr., O. D. Aguiar, and N. S. Magalhaes, editors. *Omnidirectional Gravitational Radiation Observatory*, Singapore, 1997. World Scientific.

[23] C. Bradaschia, R. del Fabbro, A. di Virgilio, A. Giazotto, H. Kautzky, V. Montelatici, D. Passuello, A. Brillet, O. Cregut, P. Hello, C. N. Man, P. T. Manh, A. Marraud, D. Shoemaker, J. Y. Vinet, F. Barone, L. Di Fiore, L. Milano, G. Russo, J. M. Aguirregabiria, H. Bel, J. P. Duruisseau, G. Le Denmat, Ph. Tourrenc, M. Capozzi, M. Longo, M. Lops, I. Pinto, G. Rotoli, T. Damour, S. Bonazzola, J. A. Marck, Y. Gourghoulon, L. E. Holloway, F. Fuligni, V. Iafolla, and G. Natale. The VIRGO project—a wide band antenna for gravitational-wave detection. *Nucl. Instrum. Methods Phys. Research*, A289:518–525, 1990.

[24] Alex Abramovici, William E. Althouse, Ronald W. P. Drever, Yekta Gürsel, Seiji Kawamura, Fredrick J. Raab, David Shoemaker, Lisa Sievers, Robert E. Spero, Kip S. Thorne, Rochus E. Vogt, Rainer Weiss, Stanley E. Whitcomb, and Michael E. Zucker. LIGO: The laser interferometer gravitational-wave observatory. *Science*, 256:325–333, April 1992.

[25] Karsten Danzmann. The GEO project: A long baseline laser interferometer for the detection of gravitational waves. *Lecture Notes in Physics*, 410:184–209, 1992.

[26] P. Bender, I. Ciufolini, K. Danzmann, W. Folkner, J. Hough, D. Robertson, A. Rüdiger, M. Sanford, R. Schilling, B. Schutz, R. Stebbins, T. Sumner, P. Touboul, S. Vitale, H. Ward, W. Winkler, J. Cornelisse, F. Hechler, Y. Jafry, and R. Reinhanrd. LISA. Laser Interferometer Space Antenna for the detection and observation of gravitational waves. Max-Planck-Institut für Quantenoptik, Hans-Kopfermann-Str. 10, D-85748 Garching, February 1996. MPQ 208.

[27] William M. Folkner, editor. *Laser Interferometer Space Antenna*, number 456 in *AIP Conference Proceedings*, Woodbury, New York, July 1998. American Institute of Physics.

[28] P. R. Saulson. How an interferometer extracts and amplifies power from a gravitational wave. *Class. Quant. Grav*, 14:2435–2454, 1997.

[29] Lee Samuel Finn. Detection, measurement, and gravitational radiation. *Phys. Rev. D*, 46(12):5236–5249, 1992.

[30] Peter R. Saulson. Thermal noise in mechanical experiments. *Phys. Rev. D*, 42(8):2437–2445, 1990.

[31] Lee Samuel Finn. Gravitational-wave data analysis with multiple detectors: The gravitational-wave receiver. I. Deterministic signals. In preparation.

[32] Lee Samuel Finn. Science benchmarks for LIGO. Technical Report T970167, California Institute of Technology, LIGO Laboratory, California Institute of Technology, Pasadena, CA 91125, September 1997.

[33] Lawrence E. Kidder, Clifford M. Will, and Alan G. Wiseman. Innermost stable orbits for coalescing binary systems of compact objects. *Class. Quant. Grav.*, 9:L125–L131, 1992.

[34] Lars Bildsten and Curt Cutler. Tidal interactions of inspiraling compact binaries. *Astrophys. J.*, 400:175–180, November 1992.

[35] Wynn C. G. Ho and Dong Lai. Resonant tidal excitations of rotating neutron stars in coalescing binaries. *Mon. Nat. R. Astron. Soc.*, 308:153–166, 1999.

[36] Luc Blanchet, Guillaume Faye, and Bénédicte Posnot. Gravitational field and equations of motion of compact binaries to 5/2 post-Newtonian order. *Phys. Rev. D*, 58, 1998.

[37] Luc Blanchet, Thibault Damour, Bala R. Iyer, Clifford M. Will, and Alan G. Wiseman. Gravitational-radiation damping of compact binary systems to second post-Newtonian order. *Phys. Rev. Lett.*, 74(18):3515 3518, 1 May 1995.

[38] Lee Samuel Finn and David F. Chernoff. Observing binary inspiral in gravitational radiation: One interferometer. *Phys. Rev. D*, 47(6):2198–2219, 1993.

[39] Lee Samuel Finn. Binary inspiral, gravitational radiation, and cosmology. *Phys. Rev. D*, 53(6):2878–2894, 15 March 1996.

[40] I. H. Stairs, Z. Arzoumanian, F. Camilo, A. G. Lyne, D. J. Nice, J. H. Taylor, S. E. Thorsett, and A. Wolszczan. Measurement of relativistic orbital decay in the PSR B1534+12 binary system. *Astrophys. J.*, 505:352, 1998.

[41] Hans A. Bethe and G. E. Brown. Evolution of binary compact objects which merge. *Astrophys. J.*, 506:780, 1998.

[42] Hans A. Bethe and G. E. Brown. Contribution of high-mass black holes to mergers of compact binaries. astro-ph/9805355, May 1998.

[43] Philippe Bagot, Simon F. Portegies Zwart, and Lev R. Yungelson. Gamma-ray bursts and density evolution of neutron star binary mergers. astro-ph/9802094, February 1998.

[44] Lev Yungelson and Simon F. Portegies Zwart. Evolution of close binaries: Formation and merger of neutron star binaries. astro-ph/9801127, January 1998.

[45] Ramesh Narayan, Tsvi Piran, and A. Shemi. Neutron-star and black hole binaries in the galaxy. *Astrophys. J. Lett.*, 379:L17–L20, 1991.

[46] E. Sterl Phinney. The rate of neutron star binary mergers in the universe: Minimal predictions for gravity wave detectors. *Astrophys. J.*, 380(1):L17–L21, 1991.

[47] Lee Samuel Finn. A numerical approach to binary black hole coalescence. In M. Francavigli, G. Longhi, L. Lusanna, and E. Sorace, editors, *Proceedings of the 14th International Conference on General Relativity and Gravitation*, pages 147–166, Singapore, 1997. International Society on General Relativity and Gravitation, World Scientific.

[48] L. Smarr. Gauge conditions, radiation formulae, and the two black hole collision. In Larry L. Smarr, editor, *Sources of Gravitational Radiation*, pages 245–274. Cambridge University Press, Cambridge, 1979.

[49] Peter Anninos, David Hobill, Edward Seidel, Larry Smarr, and Wai-Mo Suen. Collision of two black holes. *Phys. Rev. Lett.*, 71(18):2851–2854, November 1993.

[50] Peter Anninos, David Hobill, Edward Seidel, Larry Smarr, and Wai-Mo Suen. The head-on collision of two equal mass black holes. *Phys. Rev. D*, 52:2044–2058, 1995.

[51] John Baker, Andrew Abrahams, Peter Anninos, Steve Brandt, Richard Price, Jorge Pullin, and Edward Seidel. The collision of boosted black holes. *Phys. Rev. D*, 55:829, 1997.

[52] Peter Anninos and Steven Brandt. Head-on collision of two unequal mass black holes. *Phys. Rev. Lett.*, 81:508–511, 1998.

[53] Gaurav Khanna, John Baker, Reinaldo J. Gleiser, Pablo Laguna, Carlos O. Nicasio, Hans-Peter Nollert, Richard Price, and Jorge Pullin. Inspiralling black holes: the close limit, January 1999. Submitted to *Science*; Center for Gravitational Physics and Geometry Preprint CGPG-99/3-1.

[54] Éanna É. Flanagan and Scott A. Hughes. Measuring gravitational waves from binary black hole coalescences. I. Signal to noise for inspiral, merger, and ringdown. *Phys. Rev. D*, 57:4535, 15 April 1998.

[55] Éanna É. Flanagan and Scott A. Hughes. Measuring gravitational waves from binary black hole coalescences. II. The waves' information and its extraction, with and without templates. *Phys. Rev. D*, 57:4566, 15 April 1998.

[56] Richard H. Price and Jorge Pullin. Colliding black holes: the close limit. *Phys. Rev. Lett.*, 72(21):3297–3300, 23 May 1994.

[57] P. Anninos, R. H. Price, J. Pullin, E. Seidel, and W.-M. Suen. Head-on collision of two black holes: Comparison of different approaches. *Phys. Rev. D*, 52:4462–4480, 1995.

[58] A. M. Abrahams and R. H. Price. Black-hole collisions from Brill-Lindquist initial data: Predictions of perturbation theory. *Phys. Rev. D*, 53:1972–1976, 1996.

[59] A. M. Abrahams and R. H. Price. Applying black hole perturbation theory to numerically generated spacetime. *Phys. Rev. D*, 53:1963–1971, 1996.

[60] Zeferino Andradea and Richard H. Price. Head-on collisions of unequal mass black holes: close limit predictions. *Phys. Rev. D*, 56:6336, 1997.

[61] Reinaldo Gleiser, Oscar Nicasio, Richard Price, and Jorge Pullin. The collision of boosted black holes: second order close limit calculations. *Class. Quant. Grav.*, 13:L117, 1999.

[62] Fernando Echeverria. *Phys. Rev. D*, 40:3194, 1989.

[63] Lee Samuel Finn. Detectability of gravitational radiation from supernovae. In J. R. Buchler, S. L. Detweiler, and J. Ipser, editors, *Nonlinear Problems in Relativity and Cosmology*, pages 156–172. New York Academy of Sciences, New York, 1991.

[64] Ewald Müller. Gravitational radiation from collapsing rotating stellar cores. *Astron. Astrophys.*, 114:53, 1982.

[65] Lee Samuel Finn. Gravitational radiation from rotating stellar core collapse. In B. F. Schutz, editor, *Gravitational Radiation Data Analysis*, pages 33–54. Kluwer, Dordrecht, 1989.

[66] Lee Samuel Finn and Charles R. Evans. Determining gravitational radiation from Newtonian self-gravitating systems. *Astrophys. J.*, 351:588–600, 1990.

[67] R. Mönchmeyer, G. Schäfer, E. Müller, and R. E. Kates. *AA*, 256:417, 1991.

[68] R. A. Saenz and S. L. Shapiro. Gravitational radiation from stellar collapse – ellipsoidal models. *Astrophys. J.*, 221:286, April 1978.

[69] R. A. Saenz and S. L. Shapiro. Gravitational and neutrino radiation from stellar core collapse. Improved ellipsoidal model calculations. *Astrophys. J.*, 229:286, May 1979.

[70] R. A. Saenz and S. L. Shapiro. Gravitational radiation from stellar core collapse III. Damped ellipsoidal oscillations. *Astrophys. J.*, 244:1033, March 1981.

[71] Lee Samuel Finn. Gravitational radiation sources for acoustic detectors. In Velloso *et al.*,[22] pages 3–12.

[72] Adam Burrows and John Hayes. Pulsar recoil and gravitational radiation due to asymmetrical stellar collapse and explosion. *Phys. Rev. Lett.*, 76(3):352–355, 15 January 1996.

[73] A. M. Alpar and David Pines. Gravitational radiation from a solid-crust neutron star. *Nature (London)*, 314:334, 1985.

[74] Patrick R. Brady, Teviet Creighton, Curt Cutler, and Bernard F. Schutz. Searching for periodic sources with LIGO. *Phys. Rev. D*, D57, 2101-2116 1998.

[75] Mark Zimmerman and Eugene Szednits. *Phys. Rev. D*, 20:351, 1979.

[76] Lars Bildsten. Gravitational radiation and rotation of accreting neutron stars. *Astrophys. J. Lett.*, 501:89, July 1998.

[77] Nils Andersson. A new class of unstable modes for rotating relativistic stars. *Astrophys. J.*, 502:708, 1998.

[78] Nils Andersson, Kostas D. Kokkotas, and Nikolaos Stergioulas. On the relevance of the R-mode instability for accreting neutron stars and white dwarfs. *Astrophys. J.*, S16: 307–314, 1999..

[79] S. Chandrasekhar. *Phys. Rev. Lett.*, 24:611, 1970.

[80] J. Friedman and B. F. Schutz. Secular instability of rotating Newtonian stars. *Astrophys. J.*, 222:281–296, 15 May 1978.

[81] Lee Lindblom, Benjamin J. Owen, and Sharon M. Morsink. Gravitational radiation instability in hot young neutron stars. *Phys. Rev. Lett.*, 80(22), 1 June 1998.

[82] Benjamin J. Owen, Lee Lindblom, Curt Cutler, Bernard F. Schutz, Alberto Vecchio, and Nils Andersson. Gravitational waves from hot young rapidly rotating neutron stars. *Phys. Rev. D*, 58, 1998.

[83] Éanna É. Flanagan. Sensitivity of the Laser Interferometer Gravitational Wave Observatory to a stochastic background, and its dependence on the detector orientations. *Phys. Rev. D*, 48(6):2389–2407, 15 September 1993.

[84] Edward W. Kolb and Michael S. Turner. *The Early Universe*, volume 69 of *Frontiers in Physics*. Addison-Wesley, Reading, Massachusetts, paperback edition, 1994.

[85] A. Vilenkin and S. Shellard. *Cosmic Strings and Other Topological Defects*. Cambridge University Press, Cambridge, 1995.

[86] Marc Kamionkowski, Arthur Kosowsky, and Michael S. Turner. Gravitational radiation from first-order phase transitions. *Phys. Rev. D*, 49:2837–2851, 15 March 1994.

[87] Arthur Kosowsky, Michael S. Turner, and Richard Watkins. Gravitational radiation from colliding vacuum bubbles. *Phys. Rev. D*, 45(12):4514–4535, 15 June 1992.

[88] Arthur Kosowsky and Michael S. Turner. Gravitational radiation from colliding vacuum bubbles: Envelope approximation to many-bubble collisions. *Phys. Rev. D*, 47(10):4372–4391, 15 May 1993.

[89] D. Hils, P. L. Bender, and R. F. Webbink. *Astrophys. J.*, 360:75, 1990.

[90] Jon Mathews and R. L. Walker. *Mathematical Methods of Physics*. Benjamin/Cummings, Menlo Park, California, second edition, 1970.

[91] Lee Samuel Finn. Gravitational Wave Astronomy. In S. Dhurandhar and T. Padmanabhan, editors, *Gravitation and Cosmology*, pages 95–110, Netherlands, 1997. Kluwer.

QUANTUM GRAVITY AT THE PLANCK LENGTH

Joseph Polchinski*
Institute for Theoretical Physics
University of California
Santa Barbara, CA 93106-4030

ABSTRACT

I describe our understanding of physics near the Planck length, in particular the great progress of the last four years in string theory.

*Supported by NSF Grants PHY94-07194 and PHY97-22022.

© 1998 Joseph Polchinski.

1 Introduction

For obvious reasons, the SLAC Summer Institute is usually concerned with the three particle interactions. It is very appropriate, though, that the subject of the 1998 SSI is gravity, because the next step in understanding the weak, strong, and electromagnetic interactions will probably require the inclusion of gravity as well. There are many reasons for making this statement, but I will focus on two, one based on supersymmetry and one based on the unification of the couplings.

What is supersymmetry? I will answer this in more detail later, but for now let me give two short answers:

A. A lot of new particles.

B. A new *spacetime* symmetry.

Answer A is the pragmatic one for a particle experimentalist or phenomenologist. In answer B, I am distinguishing internal symmetries like flavor and electric charge, which act on the fields at each point of spacetime, from symmetries like Lorentz invariance that move the fields from one point to another. Supersymmetry is of the second type. If the widely anticipated discovery of supersymmetry actually takes place in the next few years, it not only means a lot more particles to discover. It also will be the first new spacetime symmetry since the discovery of relativity, bringing the structure of the particle interactions closer to that of gravity; in a sense, supersymmetry is a partial unification of particle physics and gravity.

The unification of the couplings is depicted in figure 1. This is usually drawn with a rather different vertical scale. Here the scale is compressed so that the three gauge couplings can hardly be distinguished, but this makes room for the fourth coupling, the gravitational coupling. Newton's constant is dimensionful, so what is actually drawn is the dimensionless coupling $G_N E^2$ with E the energy scale and $\hbar = c = 1$. This dimensionless gravitational coupling depends strongly on energy, in contrast to the slow running of the gauge couplings.

It is well-known that the three gauge couplings unify to good accuracy (in supersymmetric theories) at an energy around 2×10^{16} GeV. Note however that the fourth coupling does not miss by much, a factor of 20 or 30 in energy scale. This is another way of saying that the grand unification scale is near the Planck scale. In fact, the Planck scale $M_P = 2 \times 10^{19}$ GeV is deceptively high because of various factors like 4π that must be included. Figure 1 suggests that the grand

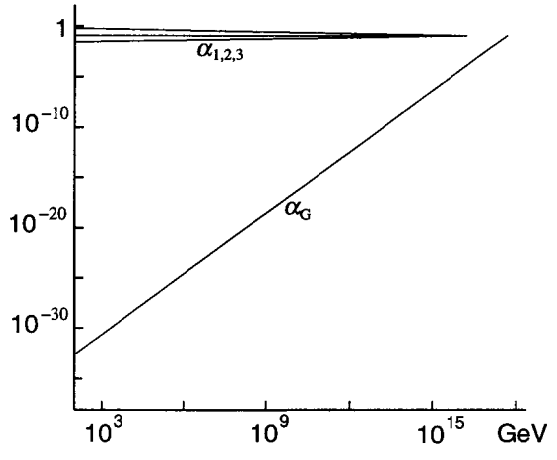


Fig. 1. The three gauge couplings and the dimensionless gravitational coupling as functions of the energy. Here $\alpha_G = G_N E^2/8\pi$.

unification of the three gauge interactions will actually be a very grand unification including gravity as well. The failure of the four couplings to meet exactly could be due to any of several small effects, which I will discuss briefly later.*

Figure 1 also shows why the phenomenologies of the gauge interactions and gravity are so different: at accessible energies the coupling strengths are very different. For the same reason, the energy scale where the couplings meet is far removed from experiment. Nevertheless, we believe that we can deduce much of what happens at this scale, and this is the subject of my lectures. At the end I will briefly discuss experimental signatures, and Michael Peskin and Nima Arkani-Hamed will discuss some of these in more detail.

In section 2 I discuss the idea that spacetime has more than four dimensions: first why this is not such a radical idea, and then why it is actually a good idea. In section 3 I review string theory as it stood a few years ago: the motivations from the short distance problem of gravity, from earlier unifying ideas, and from the search for new mathematical ideas, as well as the main problem, vacuum

*I will also discuss briefly the idea of low energy string theory, in which figure 1 is drastically changed.

selection. In section 4 I introduce the idea of duality, including weak–strong and electric–magnetic. I explain how supersymmetry gives information about strongly coupled systems. I then describe the consequences for string theory, including string duality, the eleventh dimension, D-branes, and M-theory. In section 5 I develop an alternative theory of quantum gravity, only to find that "all roads lead to string theory." In section 6 I explain how the new methods have solved some of the puzzles of black hole quantum mechanics. This in turn leads to the Maldacena dualities, which give detailed new information about supersymmetric gauge field theories. In section 7 I discuss some of the ways that the new ideas might affect particle physics, through the unification of the couplings and the possibility of low energy string theory and large new dimensions. In section 8 I summarize and present the outlook.

2 Beyond Four Dimensions

Gravity is the dynamics of spacetime. It is very likely that at lengths near the Planck scale ($L_P = 10^{-33}$ cm) it becomes evident that spacetime has more than the four dimensions that are visible to us. That is, spacetime is as shown in figure 2a, with four large dimensions (including time) and some additional number of small and highly curved spatial dimensions. A physicist who probes this spacetime with wavelengths long compared to the size of the small dimensions sees only the large ones, as in figure 2b. I will first give two reasons why this is a natural possibility to consider, and then explain why it is a good idea.

The first argument is cosmological. The universe is expanding, so the dimensions that we see were once smaller and highly curved. It may have been that initially there were more than four small dimensions, and that only the four that are evident to us began to expand. That is, we know of no reason that the initial expansion had to be isotropic.

The second argument is based on symmetry breaking. Most of the symmetry in nature is spontaneously broken or otherwise hidden from us. For example, of the $SU(3) \times SU(2) \times U(1)$ gauge symmetries, only a $U(1)$ is visible. Similarly the flavor symmetry is partly broken, as are the symmetries in many condensed matter systems. This symmetry breaking is part of what makes physics so rich: if all of the symmetry of the underlying theory were unbroken, it would be much easier to figure out what that theory is!

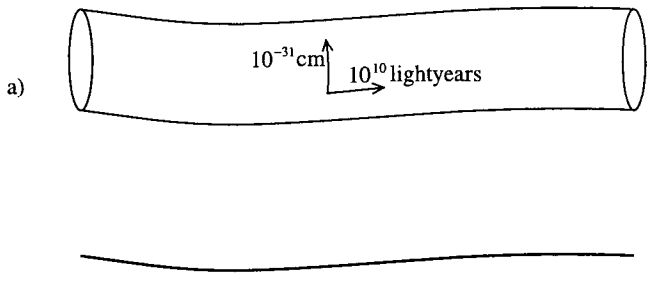


Fig. 2. a) A spacetime with one large dimension and one small one. We assume here that the small dimensions are nearly Planck sized; the possibility of larger dimensions will be considered later. b) The same spacetime as seen by a low energy observer.

Suppose that this same symmetry breaking principle holds for the spacetime symmetries. The visible spacetime symmetry is $SO(3,1)$, the Lorentz invariance of special relativity consisting of the boosts and rotations. A larger symmetry would be $SO(d,1)$ for $d > 3$, the Lorentz invariance of $d+1$ spacetime dimensions. Figure 2 shows how this symmetry would be broken by the geometry of spacetime.

So extra dimensions are cosmologically plausible, and are a natural extension of the familiar phenomenon of spontaneous symmetry breaking. In addition, they may be responsible for some of the physics that we see in nature. To see why this is so, consider first the following cartoon version of grand unification. The traceless 3×3 and 2×2 matrices for the strong and weak gauge interactions fit into a 5×5 matrix, with room for an extra $U(1)$ down the diagonal:

$$\left[\begin{array}{c|c} 3\times 3 & X,Y \\ \hline X,Y & 2\times 2 \end{array}\right]. \qquad (1)$$

Now let us try to do something similar, but for gravity and electromagnetism. Gravity is described by a metric $g_{\mu\nu}$, which is a 4×4 matrix, and electromagnetism

by a 4-vector A_μ. These fit into a 5×5 matrix:

$$\left[\begin{array}{c|c} g_{\mu\nu} & A_\mu \\ \hline A_\nu & \phi \end{array}\right]. \tag{2}$$

In fact, if one takes Einstein's equations in five dimensions, and writes them out in terms of the components (2), they become Einstein's equations for the four-dimensional metric $g_{\mu\nu}$ plus Maxwell's equation for the vector potential A_μ. This elegant unification of gravity and electromagnetism is known as *Kaluza–Klein theory.*

If one looks at the Dirac equation in the higher-dimensional space, one finds a possible explanation for another of the striking patterns in nature, the existence of quark and lepton generations. That is, a single spinor field in the higher-dimensional space generally reduces to several four-dimensional spinor fields, with repeated copies of the same gauge quantum numbers.

Unification is accompanied by new physics. In the case of grand unification this includes the X and Y bosons, which mediate proton decay. In Kaluza–Klein theory it includes the dilaton ϕ, which is the last element in the matrix (2). I will discuss the dilaton further later, but for now let me note that it is likely not to have observable effects. Of course, in Kaluza–Klein theory there is more new physics: the extra dimension(s)!

Finally, let me consider the threshold behavior as one passes from figure 2b to figure 2a. At energies greater than the inverse size of the small dimensions, one can excite particles moving in those directions. The states are quantized because of the finite size, and each state of motion looks, from the lower-dimensional point of view, like a different kind of particle. Thus the signature of passing such a threshold is a whole tower of new particles, with a spectrum characteristic of the shape of the extra dimensions.

3 String Theory

3.1 The UV Problem

To motivate string theory, I will start with the UV problem of quantum gravity. A very similar problem arose in the early days of the weak interaction. The original

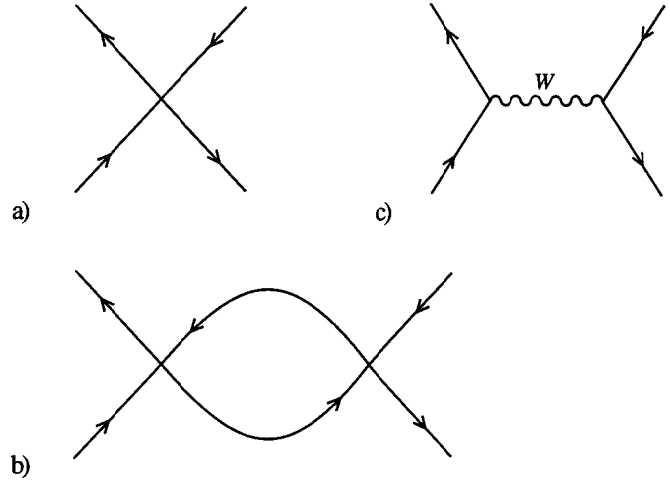


Fig. 3. a) A leptonic weak interaction in four-fermi theory. b) A divergent second order amplitude. c) The weak interaction in Weinberg-Salam theory. At short distance the contact interaction is resolved into the exchange of a W boson.

Fermi theory was based on an interaction of four fermionic fields at a spacetime point as depicted in figure 3a. The Fermi coupling constant $G_{\rm F}$ has units of length-squared, or inverse energy-squared. In a process with a characteristic energy E the effective dimensionless coupling is then $G_{\rm F}E^2$. It follows that at sufficiently high energy the coupling becomes arbitrarily strong, and this also implies divergences in the perturbation theory. The second order weak amplitude of figure 3b is dimensionally of the form

$$G_{\rm F}^2 \int^\infty E' dE', \tag{3}$$

where E' is the energy of the virtual state in the second order process, and this diverges at high energy. In position space the divergence comes when the two weak interactions occur at the same spacetime point (high energy = short distance). The divergences become worse at each higher order of perturbation theory and so cannot be controlled even with renormalization.

Such a divergence suggests that the theory one is working with is only valid

up to some energy scale, beyond which new physics appears. The new physics should have the effect of smearing out the interaction in spacetime and so softening the high energy behavior. One might imagine that this could be done in many ways, but in fact it is quite difficult to do without spoiling Lorentz invariance or causality; this is because Lorentz invariance requires that if the interaction is spread out in space it is also spread out in time. The solution to the short-distance problem of the weak interaction is not quite unique, but combined with two of the broad features of the weak interaction — its $V - A$ structure and its universal coupling to different quarks and leptons — a unique solution emerges. This is depicted in figure 3c, where the four-fermi interaction is resolved into the exchange of a vector boson. Moreover, this vector boson must be of a very specific kind, coming from a spontaneously broken gauge invariance. And indeed, this is the way that nature works.†

For gravity the discussion is much the same. The gravitational interaction is depicted in figure 4a. As we have already noted in discussing figure 1, the gravitational coupling G_N has units of length-squared and so the dimensionless coupling is $G_N E^2$. This grows large at high energy and gives again a nonrenormalizable perturbation theory.‡ Again the natural suspicion is that new short-distance physics smears out the interaction, and again there is only one known way to do this. It involves a bigger step than in the case of the weak interaction: it requires that at the Planck length the graviton and other particles turn out to be not points but one-dimensional objects, loops of "string," figure 5a. Their spacetime histories are then two-dimensional surfaces as shown in figure 4b.

At first sight this is an odd idea. It is not obvious why it should work and not other possibilities. It may simply be that we have not been imaginative enough, but because UV problems are so hard to solve we should consider carefully this one solution that we have found. And in this case the idea becomes increasingly attractive as we consider it.

†It could also have been that the divergences are an artifact of perturbation theory but do not appear in the exact amplitudes. This is a logical possibility, a "nontrivial UV fixed point." Although possible, it seems unlikely, and it is not what happens in the case of the weak interaction.

‡Note that the bad gravitational interaction of figure 4a is the same graph as the smeared-out weak interaction of figure 3c. However, its high energy behavior is worse because gravity couples to energy rather than charge.

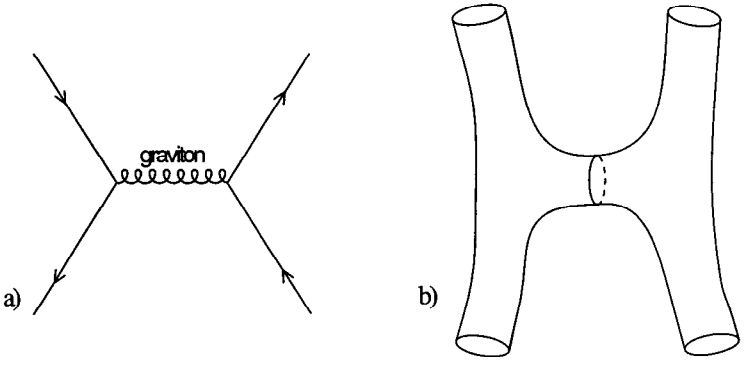


Fig. 4. a) Exchange of a graviton between two elementary particles. b) The same interaction in string theory. The amplitude is given by the sum over histories, over all embeddings of the string world-sheet in spacetime. The world-sheet is smooth: there is no distinguished point at which the interaction occurs (the cross section on the intermediate line is only for illustration).

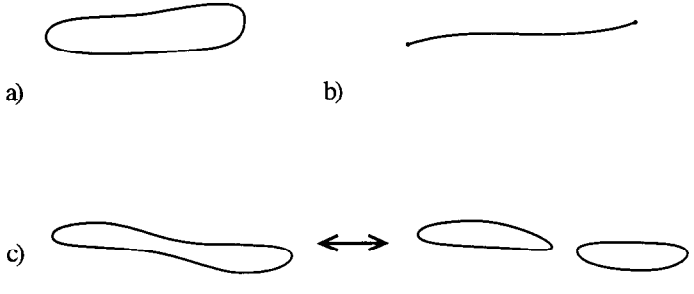


Fig. 5. a) A closed loop of string. b) An open string, which appears in some theories. c) The basic splitting–joining interaction.

3.2 All Roads Lead to String Theory

The basic idea is that the string has different states with the properties of different particles. Its internal vibrations are quantized, and depending on which oscillators are excited it can look like a scalar, a gauge boson, a graviton, or a fermion. Thus the full Standard Model plus gravity can be obtained from this one building block. The basic string interaction is as in figure 5c, one string splitting in two or the reverse. This one interaction, depending on the states of the strings involved, can look like any of the interactions in nature: gauge, gravitational, Yukawa.

A promising fact is that string theory is unique: we have known for some time that there are only a small number of string theories, and now have learned that these are actually all the same. (For now, this does not lead to predictive power because the theory has many vacuum states, with different physics.)

Further, string theory dovetails very nicely with previous ideas for extending the Standard Model. First, string theory automatically incorporates supersymmetry: it turns out that in order for the theory to be consistent the strings must move in a "superspace" which has "fermionic" dimensions in addition to the ordinary ones. Second, the spacetime symmetry of string theory is $SO(9,1)$, meaning that the strings move in ten dimensions. As I have already explained, this is a likely way to explain some of the features of nature, and it is incorporated in string theory. Third, string theory can incorporate ordinary grand unification: some of the simplest string vacua have the same gauge groups and matter that one finds in unifying the Standard Model.

From another point of view, if one searches for higher symmetries or new mathematical structures that might be useful in physics, one again finds many connections to string theory. It is worthwhile to note that these three kinds of motivation — solving the divergence problem, explaining the broad patterns in the Standard Model, and the connection with mathematics, were also present in the weak interaction. Weinberg emphasized the divergence problem as I have done. Salam was more guided by the idea that non-Abelian gauge theory was a beautiful mathematical structure that should be incorporated in physics. Experiment gave no direct indication that the weak interaction was anything but the pointlike interaction of figure 3a, and no direct clue as to the new physics that smears it out, just as today it gives no direct indication of what lies beyond the Standard Model. But it did show certain broad patterns — universality and the $V-A$ structure — that were telltale signs that the weak interaction is due to exchange of a gauge boson. It appears that nature is kind to us, in providing many trails to a correct theory.

3.3 Vacuum Selection and Dynamics

So how do we go from explaining broad patterns to making precise predictions? The main problem is that string theory has many approximately stable vacua, corresponding to different shapes and sizes for the rolled-up dimensions. The physics that we see depends on which of these vacua we are in. Thus we need to understand the dynamics of the theory in great detail, so as to determine which vacua are truly stable, and how cosmology selects one among the stable vacua.

Until recently our understanding of string theory was based entirely on perturbation theory, the analog of the Feynman graph expansion, describing small numbers of strings interacting weakly. However, we know from quantum field theory that there are many important dynamical effects that arise when we have large numbers of degrees of freedom and/or strong couplings. Some of these effects, such as confinement, the Higgs mechanism, and dynamical symmetry breaking, play an essential role in the Standard Model. If one did not know about them, one could not understand how the Standard Model Hamiltonian actually gives rise to the physics that we see.

String theory is seemingly much more complicated than field theory, and so undoubtedly has new dynamical effects of its own. I am sure that all the experimentalists would like to know, "How do I falsify string theory? How do I make it go away and not come back?" Well, you can't. Not yet. To understand why, remember that in the '50s Wolfgang Pauli thought that he had falsified Yang–Mills theory, because it seems to predict long range forces not seen in nature. The field equations for the weak and strong forces are closely parallel to those for electromagnetism, and so apparently of infinite range. It is the dynamical effects, symmetry breaking and confinement, that make these short range forces. Just as one couldn't falsify Yang–Mills theory in the '50s, one cannot falsify string theory today. In particular, because we cannot reach the analog of the parton regime where the stringy physics is directly visible, the physics that we see is filtered through a great deal of complicated dynamics.

There is a deeper problem as well. The Feynman graph expansion does not

converge, in field theory or string theory. Thus it does not define the theory at finite nonzero coupling. One needs more, the analog of the path integral and renormalization group of field theory.

Happily, since 1994 we have many new methods for understanding both field theories and string theory at strong coupling. These have led to steady progress on the questions that we need to answer, and to many new results and many surprises. This progress is the subject of the rest of my lectures.

4 Duality in Field and String Theory

4.1 Dualities

One important idea in the recent developments is *duality*. This refers to the equivalence between seemingly distinct physical systems. One starts with different Hamiltonians, and even with different fields, but after solving the theory one finds that the spectra and the transition amplitudes are identical. Often this occurs because a quantum system has more than one classical limit, so that one gets back to the same quantum theory by "quantizing" either classical theory.

This phenomenon is common in quantum field theories in two spacetime dimensions. The duality of the Sine-Gordon and Thirring models is one example; the high-temperature–low-temperature duality of the Ising model is another. The great surprise of the recent developments is that it is also common in quantum field theories in four dimensions, and in string theory.

A particularly important phenomenon is *weak–strong duality*. I have emphasized that perturbation theory does not converge. It gives the asymptotics as the coupling g goes to zero, but it misses important physics at finite coupling, and at large coupling it becomes more and more useless. In some cases, though, when g becomes very large there is a simple alternate description, a weakly coupled dual theory with $g' = 1/g$. In one sense, as $g \to \infty$ the quantum fluctuations of the original fields become very large (non-Gaussian), but one can find a dual set of fields which become more and more classical.

Another important idea is *electric–magnetic duality*. A striking feature of Maxwell's equations is the symmetry of the left-hand side under $\mathbf{E} \to \mathbf{B}$ and $\mathbf{B} \to -\mathbf{E}$. This symmetry suggests that there should be magnetic as well as electric charges. This idea became more interesting with Dirac's discovery of the quantization condition

$$q_e q_m = 2\pi n\hbar \ , \tag{4}$$

which relates the quantization of the electric charge (its equal magnitude for protons and electrons) to the existence of magnetic monopoles. A further key step was the discovery by 't Hooft and Polyakov that grand unified theories predict magnetic monopoles. These monopoles are solitons, smooth classical field configurations. Thus they look rather different from the electric charges, which are the basic quanta: the latter are light, pointlike, and weakly coupled while monopoles are heavy, "fuzzy," and (as a consequence of the Dirac quantization) strongly coupled.

In 1977 Montonen and Olive proposed that in certain supersymmetric unified theories the situation at strong coupling would be reversed: the electric objects would be big, heavy, and strongly coupled and the magnetic objects small, light, and weakly coupled. The symmetry of the sourceless Maxwell's equations would then be extended to the interacting theory, with an inversion of the coupling constant. Thus electric–magnetic duality would be a special case of weak–strong duality, with the magnetically charged fields being the dual variables for the strongly coupled theory.

The evidence for this conjecture was circumstantial: no one could actually find the dual magnetic variables. For this reason the reaction to this conjecture was skeptical for many years. In fact the evidence remains circumstantial, but in recent years it has become so much stronger that the existence of this duality is in little doubt.

4.2 Supersymmetry and Strong Coupling

The key that makes it possible to discuss the strongly coupled theory is *supersymmetry*. One way to think about supersymmetry is in terms of extra dimensions — but unlike the dimensions that we see, and unlike the small dimensions discussed earlier, these dimensions are "fermionic." In other words, the coordinates for ordinary dimensions are real numbers and so commute with each other: they are "bosonic;" the fermionic coordinates instead satisfy

$$\theta_i \theta_j = -\theta_j \theta_i \ . \tag{5}$$

For $i = j$ this implies that $\theta_i^2 = 0$, so in some sense these dimensions have zero size. This may sound rather mysterious but in practice the effect is the same as having just the bosonic dimensions but with an extra symmetry that relates the masses and couplings of fermions to those of bosons.

To understand how supersymmetry gives new information about strong coupling, let us recall the distinction between *symmetry* and *dynamics*. Symmetry tells us that some quantities (masses or amplitudes) vanish, and others are equal to one another. To actually determine the values of the masses or amplitudes is a dynamical question. In fact, supersymmetry gives some information that one would normally consider dynamical. To see this, let us consider in quantum theory the Hamiltonian operator H, the charge operator G associated with an ordinary symmetry like electric charge or baryon number, and the operator Q associated with a supersymmetry. The statement that G is a symmetry means that it commutes with the Hamiltonian,

$$[H, G] = 0 \ . \tag{6}$$

For supersymmetry one has the same,

$$[H, Q] = 0 \ , \tag{7}$$

but there is an additional relation

$$Q^2 = H + G \ , \tag{8}$$

in which the Hamiltonian and ordinary symmetries appear on the right. There are usually several Gs and several Qs, so that there should be additional indices and constants in these equations, but this schematic form is enough to explain the point. It is this second equation that gives the extra information. To see one example of this, consider a state $|\psi\rangle$ having the special property that it is neutral under supersymmetry:

$$Q|\psi\rangle = 0 \ . \tag{9}$$

To be precise, since we have said that there are usually several Qs, we are interested in states that are neutral under *at least one* Q but usually not all of them. These are known as *BPS (Bogomolnyi–Prasad–Sommerfield) states.* Now take the expectation value of the second relation (8) in this state:

$$\langle\psi|Q^2|\psi\rangle = \langle\psi|H|\psi\rangle + \langle\psi|G|\psi\rangle \ . \tag{10}$$

The left side vanishes by the BPS property, while the two terms on the right are the energy E of the state $|\psi\rangle$ and its charge q under the operator G. Thus

$$E = -q \ , \tag{11}$$

and so the energy of the state is determined in terms of its charge. But the energy is a dynamical quantity: even in quantum mechanics we must solve Schrödinger's equation to obtain it. Here, it is determined entirely by symmetry information. (There is a constant of proportionality missing in (11), because we omitted it from (8) for simplicity, but it is determined by the symmetry.)

Since the calculation of E uses only symmetry information, it does not depend on any coupling being weak: it an exact property of the theory. Thus we know something about the spectrum at strong coupling. Actually, this argument only gives the allowed values of E, not the ones that actually appear in the spectrum. The latter requires an extra step: we first calculate the spectrum of BPS states at weak coupling, and then adiabatically continue the spectrum: the BPS property enables us to follow the spectrum to strong coupling.

The BPS states are only a small part of the spectrum, but by using this and similar types of information from supersymmetry, together with general properties of quantum systems, one can usually recognize a distinctive pattern in the strongly coupled theory and so deduce the dual theory. Actually, this argument was already made by Montonen and Olive in 1977, but only in 1994, after this kind of reasoning was applied in a systematic way in many examples starting with Seiberg, did it become clear that it works and that electric–magnetic duality is a real property of supersymmetric gauge theories.

4.3 String Duality, D-Branes, M-Theory

Thus far the discussion of duality has focused on quantum field theory, but the same ideas apply to string theory. Prior to 1994 there were various conjectures about duality in string theory, but after the developments described above, Hull, Townsend, and Witten considered the issue in a systematic way. They found that for each strongly coupled string theory (with enough supersymmetry) there was a unique candidate for a weakly coupled dual. These conjectures fit together in an intricate and consistent way as dimensions are compactified, and evidence for them rapidly mounted. Thus weak–strong duality seems to be a general property in string theory.

Weak–strong duality in field theory interchanged the pointlike quanta of the original fields with smooth solitons constructed from those fields. In string theory, the duality mixes up various kinds of objects: the basic quanta (which are now strings), smooth solitons, black holes (which are like solitons, but with horizons and singularities), and new stringy objects known as *D-branes*.

The D-branes play a major role, so I will describe them in more detail. In string theory strings usually move freely. However, some string theories also predict localized objects, sort of like defects in a crystal, where strings can break open and their endpoints get stuck. These are known as D-branes, short for Dirichlet (a kind of boundary condition—see Jackson, *Classical Electrodynamics*) membranes. Depicted in figure 6, they can be points (D0-branes), curves (D1-branes), sheets (D2-branes), or higher-dimensional objects. They are dynamical objects—they can move, and bend—and their properties, at weak coupling, can be determined with the same machinery used elsewhere in string theory.

Even before string duality it was found that one could make D-branes starting with just ordinary strings (for string theorists, I am talking about T-duality). Now we know that they are needed to fill out the duality multiplets. They have many interesting properties. One is that they are smaller than strings; one cannot really see this pictorially, because it includes the quantum fluctuations, but it follows from calculations of the relevant form factors. Since we are used to thinking that smaller means more fundamental, this is intriguing, and we will return to it.

Returning to string duality, figure 7 gives a schematic picture of what was learned in 1995. Before that time there were five known string theories. These differed primarily in the way that supersymmetry acts on the string, and also the type I theory in that it includes open strings. We now know that starting with any one of these theories and going to strong coupling, we can reach any of the others. Again, the idea is that one follows the BPS states and recognizes distinctive patterns in the limits. The parameter space in the figure can be thought of as two coupling constants, or as the radii of two compact dimensions.

In figure 7 there is a sixth limit, labeled *M-theory*. We have emphasized that the underlying spacetime symmetry of string theory is $SO(9,1)$. However, the M-theory point in the figure is in fact a point of $SO(10,1)$ symmetry: the spacetime symmetry of string theory is larger than had been suspected. The extra piece is badly spontaneously broken, at weak coupling, and not visible in the perturbation theory, but it is a property of the exact theory. It is interesting that $SO(10,1)$ is

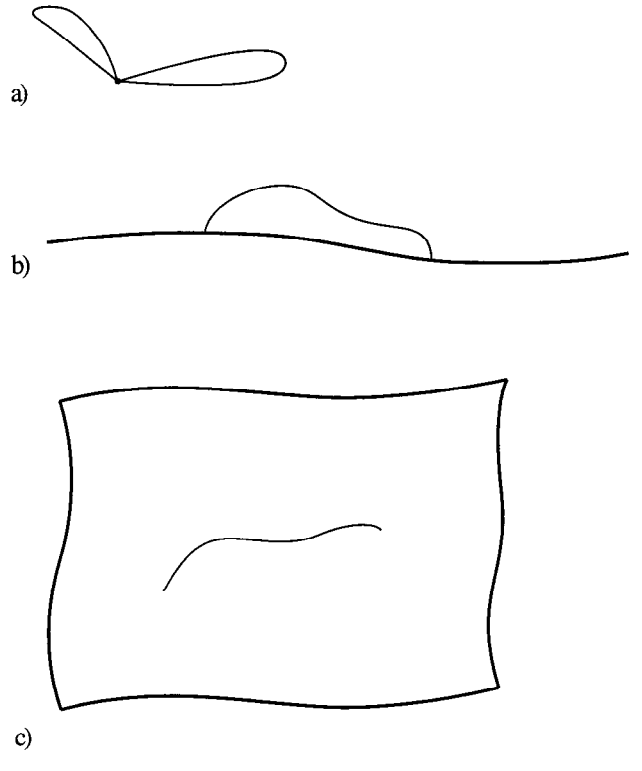


Fig. 6. (a) A D0-brane with two attached strings. (b) A D1-brane **(bold)** with attached string. (c) A D2-brane with attached string.

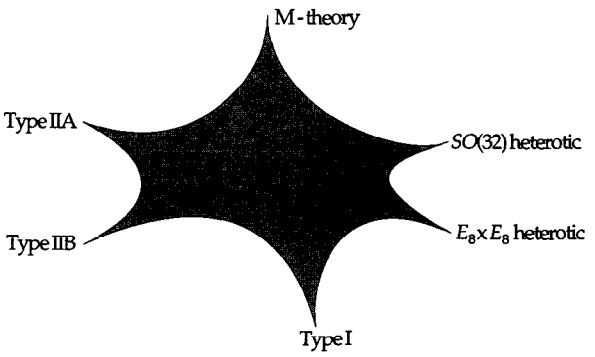


Fig. 7. The five string theories, and M-theory, as limits of a single theory.

known to be the largest spacetime symmetry compatible with supersymmetry.

Another way to describe this is that in the M-theory limit the theory lives in eleven spacetime dimensions: a new dimension has appeared. This is one of the surprising discoveries of the past few years. How does one discover a new dimension? It is worthwhile explaining this in some more detail. The D0-brane mass is related to the characteristic string mass scale m_s and the dimensionless string coupling g_s, by

$$m_{D0} = \frac{m_s}{g_s} . \tag{12}$$

When g_s is small this is heavier than the string scale, but when g_s is large it is lighter. Further, the D0-brane is a BPS state and so this result is exact. If one considers now a state with N D0-branes, the mass is bounded below by Nm_{D0}, and in fact this bound is saturated: there is a BPS bound state with

$$m_{ND0} = N\frac{m_s}{g_s} \tag{13}$$

exactly. Now observe that for g_s large, all of these masses become small. What can the physics be? In fact, this is the spectrum associated with passing a threshold where a new spacetime dimension becomes visible. The radius of this dimension is

$$R = \frac{g_s}{m_s} . \tag{14}$$

That is, small g_s is small R and large g_s is large R. In particular, perturbation theory in g_s is an expansion around $R = 0$: this is why this dimension has always been invisible!

5 An Alternative to String Theory?

On Lance Dixon's tentative outline for my lectures, one of the items was "Alternatives to String Theory." My first reaction was that this was silly, there are no alternatives, but on reflection I realized that there was an interesting alternative to discuss. So let us try to construct a quantum theory of gravity based on a new principle, not string theory. We will fail, of course, but we will fail in an interesting way.

Let us start as follows. In quantum mechanics we have the usual position-momentum uncertainty relation

$$\delta x \delta p \geq \hbar^2 . \tag{15}$$

Quantum gravity seems to imply a breakdown in spacetime at the Planck length, so perhaps there is also a position-position uncertainty relation

$$\delta x \delta x \geq L_P^2 . \tag{16}$$

This has been discussed many times, and there are many ways that one might try to implement it. We will do this as follows. Suppose that we have N nonrelativistic particles. In normal quantum mechanics the state would be defined by N position vectors

$$\mathbf{X}_i , \quad i = 1, \ldots, N . \tag{17}$$

Let us instead make these into Hermitean matrices in the particle-number index

$$\mathbf{X}_{ij} , \quad i, j = 1, \ldots, N . \tag{18}$$

It is not obvious what this means, but we will see that it leads to an interesting result. For the Hamiltonian we take

$$H = \frac{1}{2M} \sum_{m=1}^{D-1} \sum_{i,j=1}^{N} (p_{ij}^m)^2 + M'^5 \sum_{m,n=1}^{D-1} \sum_{i,j=1}^{N} |[X^m, X^n]_{ij}|^2 . \tag{19}$$

The first term is just an ordinary nonrelativistic kinetic term, except that we now have N^2 coordinate vectors rather than N so there is a momentum for each, and

we sum the squares of all of them. The indices m and n run over the $D-1$ spatial directions, and M and M' are large masses, of order of the Planck scale. The potential term is chosen as follows. We want to recover ordinary quantum mechanics at low energy. The potential is the sum of the squares of all of the components of all of the commutators of the matrices $\mathbf{X}_{ij}$, with a large coefficient. It is therefore large unless all of these matrices commute. In states with energies below the Planck scale, the matrices will then commute to good approximation, so we do not see the new uncertainty (16) and we recover the usual quantum mechanics. In particular, we can find a basis which diagonalizes all the commuting X^m_{ij}. Thus the effective coordinates are just the N diagonal elements X^m_{ii} of each matrix in this basis, which is the right count for N particles in ordinary quantum mechanics: the X^m_{ii} behave like ordinary coordinates.

The Hamiltonian (19) has interesting connections with other parts of physics. First, the commutator-squared term has the exact same structure as the four-gluon interaction in Yang–Mills theory. This is no accident, as we will see later on. Second, there is a close connection to supersymmetry. In supersymmetric quantum mechanics, one has operators satisfying the algebra (7,8). Again in general there are several supersymmetry charges, and the number $\mathcal{N}$ of these Qs is significant. For small values of $\mathcal{N}$, like 1, 2, or 4, there are many Hamiltonians with the symmetry. As $\mathcal{N}$ increases the symmetry becomes more constraining, and $\mathcal{N}=16$ is the maximum number. For $\mathcal{N}=16$ there is only one invariant Hamiltonian, and it is none other than our model (19). To be precise, supersymmetry requires that the particles have spin, that the Hamiltonian also has a spin-dependent piece, and that the spacetime dimension D be 10. In fact, supersymmetry is necessary for this idea to work. The vanishing of the potential for commuting configurations was needed, but we only considered the classical potential, not the quantum corrections. The latter vanish only if the theory is supersymmetric.

So this model has interesting connections, but let us return to the idea that we want a theory of gravity. The interactions among low energy particles come about as follows. We have argued that the potential forces the $\mathbf{X}_{ij}$ to be diagonal: the off-diagonal pieces are very massive. Still, virtual off-diagonal excitations induce interactions among the low-energy states. In fact, the leading effect, from one loop of the massive states, produces precisely the (super)gravity interaction among the low energy particles.

So this simple idea seems to be working quite well, but we said that we were going to fail in our attempt to find an alternative to string theory. In fact we have failed because this is not an alternative: it *is* string theory. It is actually one piece of string theory, namely the Hamiltonian describing the low energy dynamics of N D0-branes. This illustrates the following principle: that all good ideas are part of string theory. That sounds arrogant, but with all the recent progress in string theory, and a fuller understanding of the dualities and dynamical possibilities, string theory has extended its reach into more areas of mathematics and has absorbed previous ideas for unification (including $D=11$ supergravity).

We have discussed this model not just to introduce this principle, but because the model is important for a number of other reasons. In fact, it is conjectured that it is not just a piece of string theory, but is actually a *complete* description. The idea is that if we view any state in string theory from a very highly boosted frame, it will be described by the Hamiltonian (19) with N large. Particle physicists are familiar with the idea that systems look different as one boosts them: the parton distributions evolve. The idea here is that the D0-branes are the partons for string theory; in effect the string is a necklace of partons. This is the *matrix theory* idea of Banks, Fischler, Shenker, and Susskind (based on earlier ideas of Thorn), and at this point it seems very likely to be correct or at least a step in the correct direction.

To put this in context, let us return to the illustration in figure 7 of the space of string vacua, and to the point made earlier that the perturbation theory does not define the theory for finite g. In fact, every indication is that the string description is useful only near the five cusps of the figure in which the string coupling becomes weak. In the center of the parameter space, not only do we not know the Hamiltonian but we do not know what degrees of freedom are supposed to appear in it. It is likely that they are not the one-dimensional objects that one usually thinks of in string theory; it is more likely that they are the coordinate matrices of the D-branes.

6 Black Hole Quantum Mechanics

6.1 Black Hole Thermodynamics

In the '70s it was found that there is a close analogy between the laws of black hole mechanics and the laws of thermodynamics. In particular, the event horizon area (in Planck units) is like the entropy. It is nondecreasing in classical gravitational processes, and the sum of this *Bekenstein–Hawking entropy* and the entropy of radiation is nondecreasing when Hawking radiation is included. For more than 20 years it has been a goal to find the statistical mechanical picture from which this thermodynamics derives — that is, to count the quantum states of a black hole. There have been suggestive ideas over the years, but no systematic framework for addressing the question.

I have described the new ideas we have for understanding strongly interacting strings. A black hole certainly has strong gravitational interactions, so we might hope that the new tools would be useful here. Pursuing this line of thought, Strominger and Vafa were able in early 1996 to count the quantum states of a black hole for the first time. They did this with the following thought experiment. Start with a black hole and imagine adiabatically reducing the gravitational coupling G_N. At some point the gravitational binding becomes weak enough that the black hole can no longer stay black, but must turn into ordinary matter. A complete theory of quantum gravity must predict what the final state will look like. The answer depends on what kind of black hole we begin with — in other words, what are its electric, magnetic, and other charges (the no-hair theorem says that this is all that identifies a black hole). For the state counting we want to take a supersymmetric black hole, one that corresponds to a BPS state in the quantum theory. For the simplest such black holes, the charges that they carry determine that at weak coupling they will turn into a gas of weakly coupled D-branes, as depicted in figure 8. For these we know the Hamiltonian, so we can count the states and continue back to strong coupling where the system is a black hole. Indeed, the answer is found to agree precisely with the Bekenstein–Hawking entropy. Our initial motivation was one problem of quantum gravity, the UV divergences. Now, many years later, string theory has solved another, very different, long-standing problem in the subject.

This result led to much further study. It was found that in addition to the agreement of the entropy of BPS states with the Bekenstein–Hawking entropy, agreement was also found between the general relativity calculation and the D-brane calculation for the entropies of near-BPS states, and for various dynamical quantities such as absorption and decay amplitudes. This goes beyond the adiabatic continuation argument used to justify the entropy calculation, and in late 1997 these results were understood as consequences of a new duality, the *Maldacena* duality. This states that, not only are the weakly coupled D-branes the adiabatic continuation of the black hole, but that *the D-brane system at all couplings is dual (physically equivalent) to the black hole.* In effect, D-branes are the atoms from which certain black holes are made, but for large black holes they are in a highly quantum state while the dual gravitational field is in a very classical state. A precise statement of the Maldacena duality requires a low energy limit in the D-brane system, while on the gravitational side one takes the limit of the geometry near the horizon.

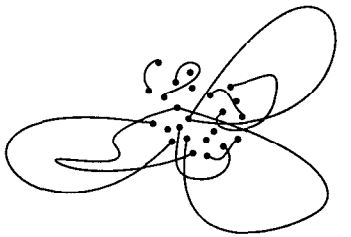


Fig. 8. A state of many D0-branes, with attached strings.

6.2 The Information Paradox

This new duality has two important consequences. The first is for another of the nagging problems of quantum gravity, the black hole information paradox. A black hole emits thermal Hawking radiation, and will eventually decay completely. The final state is independent of what went into the black hole, and incoherent. In other words, an initially pure state evolves into a mixed state; this is inconsistent with the usual rules of quantum mechanics. Hawking argued that in quantum gravity the evolution of states must be generalized in this way.

This has been a source of great controversy. While most physicists would be pleased to see quantum mechanics replaced by something less weird, the particular modification proposed by Hawking simply makes it uglier, and quite possibly inconsistent. But 20 years of people trying to find Hawking's "mistake," to identify the mechanism that preserves the purity of the quantum state, has only served to sharpen the paradox: because the quantum correlations are lost behind the horizon, either quantum mechanics is modified in Hawking's way, or the locality of physics must break down in a way that is subtle enough not to infect most of physics, yet act over long distances.

The duality conjecture above states that the black hole is equivalent to an ordinary quantum system, so that the laws of quantum evolution are unmodified. However, to resolve fully the paradox one must identify the associated nonlocality in the spacetime physics. This is hard to do because the local properties of spacetime are difficult to extract from the highly quantum D-brane system: this is related to the *holographic principle.* This term refers to the property of a hologram, that the full picture is contained in any one piece. It also has the further connotation that the quantum state of any system can be encoded in variables living on the *boundary* of that system, an idea that is suggested by the entropy–area connection of the black hole. This is a key point where our ideas are in still in flux.

6.3 Black Holes and Gauge Theory

Dualities between two systems give information in each direction: for each system there are some things that can be calculated much more easily in the dual description. In the previous subsection we used the Maldacena duality to make statements about black holes. We can also use it in the other direction, to calculate properties of the D-brane theory.

To take full advantage of this we must first make a generalization. We have said that D-branes can be points, strings, sheets, and so on: they can be extended in p directions, where here $p = 0, 1, 2$. Thus we refer to Dp-branes. The same is true of black holes: the usual ones are local objects, but we can also have black strings — strings with event horizons — and so on. A black p-brane is extended in p directions and has a black hole geometry in the orthogonal directions. The full Maldacena duality is between the low energy physics of Dp-branes and strings in the near-horizon geometry of a black p-brane. Further, for $p \leq 3$ the low energy physics of N Dp-branes is described by $U(N)$ Yang–Mills theory with $\mathcal{N} = 16$ supersymmetries. That is, the gauge fields live on the D-branes, so that they constitute a field theory in $p + 1$ "spacetime" dimensions, where here spacetime is just the world-volume of the brane. For $p = 0$, this is the connection of matrix quantum mechanics to Yang–Mills theory that we have already mentioned below (19).

The Maldacena duality then implies that various quantities in the gauge theory can be calculated more easily in the dual black p-brane geometry. This method is only useful for large N, because this is necessary to get a black hole which is larger than string scale and so described by ordinary general relativity. Of course we have a particular interest in gauge theories in $3 + 1$ dimensions, so let us focus on $p = 3$. The Maldacena duality for $p = 3$ partly solves an old problem in the strong interaction. In the mid-'70s 't Hooft observed that Yang Mills theory simplifies when the number of colors is large. This simplification was not enough to allow analytic calculation, but its form led 't Hooft to conjecture a duality between large-N gauge theory and some unknown string theory. The Maldacena duality is a precise realization of this idea, for supersymmetric gauge theories.[§] For the strong interaction we need of course to understand nonsupersymmetric gauge theories. One can obtain a rough picture of these from the Maldacena duality, but a precise description seems still far off. It is notable, however, that string theory, which began as an attempt to describe the strong interaction, has now returned to its roots, but only by means of an excursion through black hole physics and other strange paths.

6.4 Spacetime Topology Change

This subsection is not directly related to black holes, but deals with another exotic question in quantum gravity. Gravity is due to the bending of spacetime. It is an old question, whether spacetime could not only bend but break: does its topology as well as its geometry evolve in time?

Again, string theory provides the tools to answer this question. The answer

[§]For $p = 3$ the near-horizon geometry is the product of an anti-de Sitter space and a sphere, while the supersymmetric gauge theory is conformally invariant (a conformal field theory), so this is also known as the *AdS–CFT correspondence.*

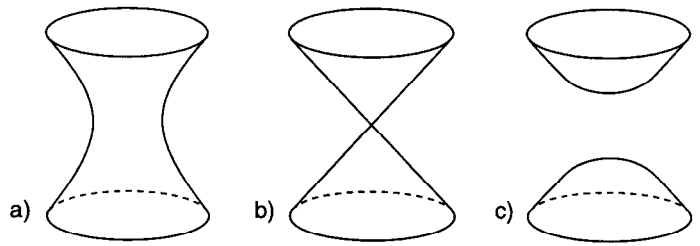


Fig. 9. An example of spacetime topology change.

is "yes" — under certain controlled circumstances the geometry can evolve as shown schematically in figure 9. It is interesting to focus on the case that the topology change is taking place in the compactified dimensions, and to contrast the situation as seen by the short-distance and long-distance observers of figures 2a and 2b. The short distance observer sees the actual process of figure 9. The long distance observer cannot see this. Rather, this observer sees a *phase transition.* At the point where the topology changes, some additional particles become massless and the symmetry breaking pattern changes. Thus the transition can be analyzed with the ordinary methods of field theory; it is this that makes the quantitative analysis of the topology change possible.

Incidentally, topology change has often been discussed in the context of spacetime foam, the idea that the topology of spacetime is constantly fluctuating at Planckian distance scales. It is likely that the truth is even more strange, as in matrix theory where spacetime becomes "non-Abelian."

7 Unification and Large Dimensions

My talks have been unapologetically theoretical. The Planck length is far removed from experiment, yet we believe we have a great deal of understanding of the very exotic physics that lies there. In this final section I would like to discuss some ways in which the discoveries of the last few years might affect the physics that we see.

Let me return to the unification of the couplings in figure 10a, and to the failure of the gravitational coupling to meet the other three exactly. There are

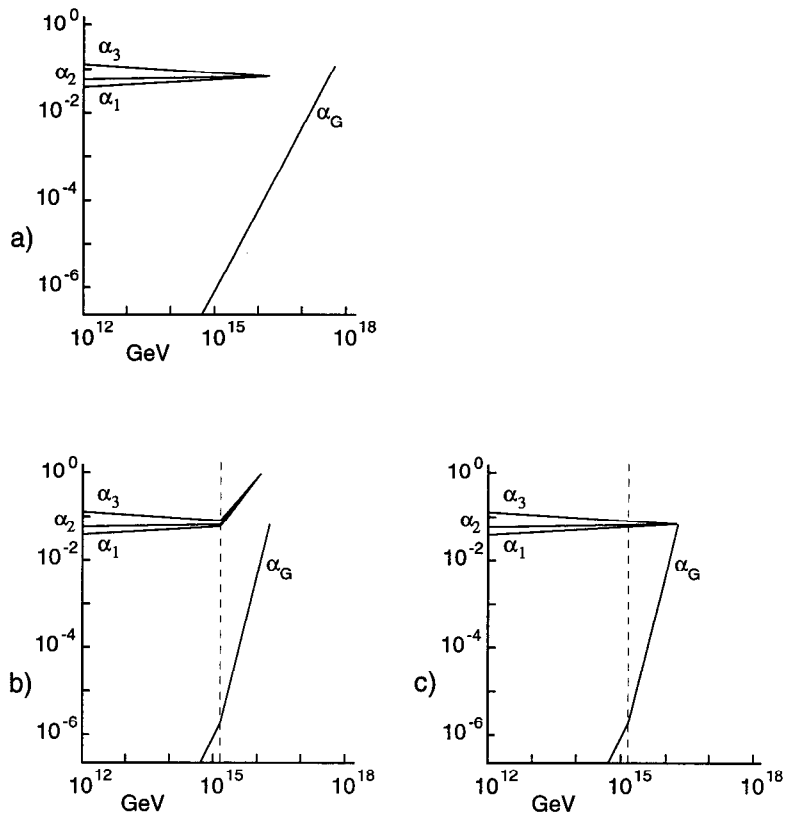


Fig. 10. a) Running of the three gauge couplings and the dimensionless gravitational coupling with energy. b) Effect of a fifth dimension below the unification scale. c) Effect of a fifth dimension of the Horava-Witten type.

many ideas to explain this. There could be additional particles at the weak, intermediate, or unified scales, which change the running of the gauge couplings so as to raise the unification point. Or it may be that the gauge couplings actually do unify first, so that there is a normal grand unified theory over a small range of scales before the gravitational coupling unifies. These ideas focus on changing the behavior of the gauge couplings. Since these already unify to good approximation, it would be simpler to change the behavior of the gravitational coupling so that it meets the other three at a lower scale — to lower the Planck scale. Unfortunately this is not so easy. The energy-dependence of the gravitational coupling is just dimensional analysis, which is not so easy to change.¶

There is a way to change the dimensional analysis — that is, to change the dimension! We have discussed the possibility that at some scale we pass the threshold to a new dimension. Suppose that this occurred below the unification scale. For both the gauge and the gravitational couplings the units change, so that both turn upward as in figure 10b. This does not help; the couplings meet no sooner.

There is a more interesting possibility, which was first noticed in the strong coupling limit of the $E_8 \times E_8$ heterotic string. Of the five string theories, this is the one whose weakly coupled physics looks most promising for unification. Its strong-coupling behavior, shown in figure 11, is interesting. A new dimension appears, but it is not simply a circle. Rather, it is bounded by two walls. Moreover, all the gauge fields and the particles that carry gauge charges move only in the walls, while gravity moves in the bulk. Consider now the unification of the couplings. The dynamics of the gauge couplings, and their running, remains as in $3+1$ dimensions; however, the gravitational coupling has a kink at the threshold, so the net effect can be as in figure 10c. If the threshold is at the correct scale, the four couplings meet at a point.

As it stands this has no more predictive power than any of the other proposed solutions. There is one more unknown parameter, the new threshold scale, and one more prediction. However, it does illustrate that the new understanding of string theory will lead to some very new ideas about the nature of unification. Figure 11 is only one example of a much more general idea now under study, that

¶It was asked whether the gravitational coupling has additional β-function type running. Although this could occur in principle, it does not do so because of a combination of dimensional analysis and symmetry arguments.

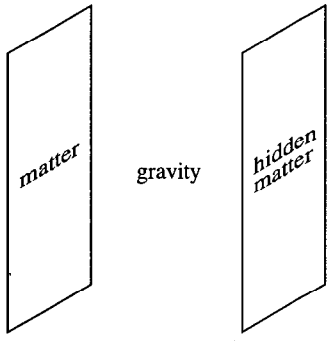


Fig. 11. A Horava–Witten spacetime. The two planes represent $3+1$ dimensional walls, in which all the Standard Model particles live, while gravity moves in the $4+1$ dimensional bulk between the walls. In string theory there are six additional dimensions, which could be much smaller and are not shown. The wall is then $9+1$ dimensional in all, and the spacetime $10+1$ dimensional.

the Standard Model lives in a brane and does not move in the full space of the compact dimensions, while gravity does do so.

This new idea leads in turn to the possibility of radically changing the scales of new physics in string theory. To see this, imagine lowering the threshold energy (the kink) in figure 10c; this also lowers the string scale, which is where the gravitational coupling meets the other three. From a completely model-independent point of view, how low can we go? The string scale must be at least a TeV, or else we would already have seen string physics. The five-dimensional threshold must correspond to a radius of no more than a millimeter, or else Cavendish experiments would already have shown the four-dimensional inverse square law turning into a five-dimensional inverse cube. Remarkably, it is difficult to improve on these extreme model-independent bounds. The large dimension in particular might seem to imply a whole tower of new states at energies above 10^{-4} eV, but these are very weakly coupled (gravitational strength) and so would not be seen. It may be that construction of a full model, with a sensible cosmology, will raise these scales, but that they will still lie lower than we used to imagine.

I had been somewhat skeptical about this idea, for a reason that is evident in

figure 10c. If the threshold is lowered further, the gravitational coupling meets the other three before they unify and one loses the successful prediction of $\sin^2\theta_w$. However, it is wrong to pin so much on this one number; the correct prediction might come out in the end in a more complicated way. One should certainly explore the many new possibilities that arise, to see what other consequences there are and to broaden our perspective on the possible nature of unification.

8 Outlook

I will start with the more theoretical problems.

1. The black hole information problem. It seems that the necessary ingredients to solve this are at hand, and that we will soon assemble them correctly. However, it has seemed this way before, and the clock on this problem is at 22 years and counting. Still, our understanding is clearly deeper than it has ever been.

2. The cosmological constant problem. In any quantum theory the vacuum is a busy place, and should gravitate. Why is the cosmological constant, even if nonzero, so much smaller than particle or Planck energies? This is another hard problem, not just in string theory but in any theory of gravity. It has resisted solution for a long time, and seems to require radical new ideas.

The new ideas that I have described have not led to a solution, but they have suggested new possibilities. One important ingredient may be supersymmetry. Throughout the discussion of duality this plays a central role in canceling quantum fluctuations, suggesting that it also does so in the vacuum energy. The problem is that supersymmetry is broken in nature; we need a phase with some properties of the broken theory and some of the unbroken. We have learned about many new phases of string theory, but not yet one with just the right properties.

Another ingredient may be nonlocality. The cosmological constant affects physics on cosmic scales but is determined by dynamics at short distance: this suggests the need for some nonlocal feedback mechanism. Recall that the black hole information problem also seems to need nonlocality; perhaps these are related.

3. Precise predictions from string theory? Our understanding of string dynamics is much improved, but still very insufficient for solving the vacuum selection/stability problem, especially with nonsupersymmetric vacua. It is hard to see how one could begin to address this before solving the cosmological constant problem, since this tells us that we are missing something important about the vacuum.

An optimistic projection is that we soon solve the information problem, that this gives us the needed idea to solve the cosmological constant problem, and then we can address vacuum selection. When will this happen? Well, if I didn't have to write these d**n notes, ... More likely, we still are missing some key concepts.

4. What is string theory? We are closer to a nonperturbative formulation than ever before: the things that we have learned in the past few years have completely changed our point of view. It may be that again the ingredients are in place, in that both matrix theory and the Maldacena duality give nonperturbative definitions, and we simply need to extract the essence.

5. Distinct signatures of string theory? Is there any distinctively stringy experimental signature? All of the new physics may lie far beyond accessible energies, but we might be lucky instead. I have discussed the possibility of low energy string theory and large dimensions. I am still inclined to expect the standard picture to hold, but the new ideas are and will remain a serious alternative. Another possibility is a fifth force from the dilaton or other moduli (scalars that are common in string theory). These are massless to first approximation, but quantum effects almost invariably induce masses for all scalars. The resulting mass is likely in the range

$$\frac{m_{\rm weak}^2}{m_{\rm P}} < m_{\rm scalar} < m_{\rm weak} \ . \tag{20}$$

The lower limit is interesting for a fifth force, while the whole range is interesting for dark matter.

The most interesting hope is for something unexpected, perhaps cosmological and associated with the holographic principle, or perhaps a distinctive form of supersymmetry breaking.

6. Supersymmetry. Supersymmetry has played a role throughout these talks. In string theory it is a symmetry at least at the Planck scale, but is broken somewhere between the Planck and weak scales. The main arguments for breaking at the weak scale are independent of string theory: the Higgs hierarchy problem, the unification of the couplings, the heavy top quark. In addition, the ubiquitous role that supersymmetry plays in suppressing quantum fluctuations in our discussion of strongly coupled physics supports the idea that it suppresses the quantum corrections to the Higgs mass. The one cautionary note is that the cos-

mological constant suggests a new phase of supersymmetry, whose phenomenology at this point is completely unknown. Still, the discovery and precision study of supersymmetry remains the best bet for testing all of these ideas.

In conclusion, the last few years have seen remarkable progress, and there is a real prospect of answering difficult and long-standing problems in the near future.

References

Two texts on string theory:

- M. B. Green, J. H. Schwarz, and E. Witten, *Superstring Theory, Vols. 1 and 2* (Cambridge University Press, Cambridge, 1987).
- J. Polchinski, *String Theory, Vols. 1 and 2* (Cambridge University Press, Cambridge, 1998).

An earlier version of these lectures, with extensive references:

- J. Polchinski, "String Duality," Rev. Mod. Phys. **68**, 1245 (1996), hep-th/9607050.

The talk by Michael Peskin at the SSI Topical Conference gives more detail on some of the subjects in my lectures. Other popular accounts:

- G. P. Collins, "Quantum Black Holes Are Tied to D-Branes and Strings," Physics Today **50**, 19 (March 1997).
- E. Witten, "Duality, Spacetime, and Quantum Mechanics," Physics Today **50**, 28 (May 1997).
- B. G. Levi, "Strings May Tie Quantum Gravity to Quantum Chromodynamics," Physics Today **51**, 20 (August 1998).

A summer school covering many of the developments up to 1996:

- C. Efthimiou and B. Greene, editors, *Fields, Strings, and Duality, TASI 1996* (Singapore: World Scientific, 1997).

Lectures on matrix theory:

- T. Banks, "Matrix Theory," Nucl. Phys. Proc. Suppl. **67**, 180 (1998), hep-th/9710231.
- D. Bigatti and L. Susskind, "Review of Matrix Theory," hep-th/9712072.

Lectures on developments in black hole quantum mechanics:

- G. T. Horowitz, "Quantum States of Black Holes," gr-qc/9704072.
- A. W. Peet, "The Bekenstein Formula and String Theory," Classical and Quantum Gravity **15**, 3291 (1998), hep-th/9712253.

Two very recent reviews, including the Maldacena conjecture:

- A. Sen, "Developments in Superstring Theory," hep-ph/9810356.
- J. Schwarz, "Beyond Gauge Theories," hep-th/9807195.

For more on low energy string theory and millimeter dimensions, see the talk by Nima Arkani-Hamed at the SSI Topical Conference, and references therein.

PRECISION ELECTROWEAK RESULTS AT LEP-I AND LEP-II

Manel Martinez
Institut de Fisica d'Altes Energies, Edifici Cn, U. A. B.
E-08193 Bellaterra (Barcelona), Spain

ABSTRACT

The most relevant LEP electroweak measurements obtained during the LEP-I and LEP-II operation phases are briefly reviewed, and their internal consistency, as well as their agreement with the Minimal Standard Model (MSM) predictions, are discussed.

© 1998 by Manel Martinez.

1 Introduction

Since its start-up in 1989, LEP has proven to be the most powerful machine to probe the theory of the electroweak interactions, the MSM. A full bunch of high precision measurements performed by the four LEP experiments allow now the extraction of many important parameters to an unprecedented level of accuracy.

With the start of the LEP-II program in fall 1995, the LEP-I program arrived to its end although, as of now, part of the LEP-I data has not yet reached final publication. The reason for that is that several measurements have been reanalyzed in the light of a better theoretical and experimental understanding of the systematic effects. Nevertheless, altogether the LEP-I measurements have been already the subject of many good and extensive reviews (see, for instance Ref. 1) and therefore, in spite of their great importance, in this note we are going to omit the details about their experimental determination and discuss just the results, their meaning, and their interpretation. On the contrary, we want to describe in some more detail the LEP-II results which, apart from being more recent, have now just started reaching maturity.

In this note, the most relevant LEP Electroweak measurements performed so far are briefly presented, paying special attention to their systematic limitations, and their impact in the Standard Model (SM). The results presented here come, for most observables, from all data taken between 1989 and 1997 by the four LEP experiments: about 15 million hadronic Z decays and about 1.7 million charged-lepton Z decays coming from the Z resonance scan, together with about 4000 W pair events coming from the WW production at and above threshold. Most of the results discussed are still preliminary although they have been already presented in public.[2]

The Z observables are classified into three groups: Z lineshape and lepton forward-backward asymmetry measurements, tau polarization measurements, and finally, electroweak results with quarks, where the present results, analyzed together with the ones obtained at SLAC, have originated some discussion about possible systematic effects in the quark couplings to the Z, as we shall see later. The WW measurements are classified into three groups as well: WW cross section and branching ratios, the W mass, and the triple gauge couplings.

They are all discussed in turn in the following sections.

2 The LEP-I Measurements

2.1 Z Lineshape and Lepton F-B Asymmetry Measurements

The cross section for the production of a fermion pair $f\bar{f}$ can be written as

$$\sigma_{f\bar{f}}(s) = \int_{4m_f^2}^{s} ds' H(s,s') \hat{\sigma}_{f\bar{f}}(s') \,, \tag{1}$$

where $H(s,s')$ is the so-called radiator function which takes care of initial state radiation corrections, and the reduced cross section $\hat{\sigma}$ is written as

$$\hat{\sigma}_{f\bar{f}}(s) = \sigma^0_{f\bar{f}} \cdot \frac{s\Gamma_Z^2}{\left(s - M_Z^2\right)^2 + \left(s\Gamma_Z/M_Z\right)^2} + (\gamma - Z) + |\gamma|^2 \,, \tag{2}$$

where $\sigma^0_{f\bar{f}}$ is called the peak cross section, $\Gamma_Z = \Gamma_h + \Gamma_e + \Gamma_\mu + \Gamma_\tau + \Gamma_{inv}$ stands for the total Z width, and M_Z is the Z mass. This parameterization assumes the validity of QED for the photon exchange part and also takes from the MSM the interference between the photon- and Z-mediated amplitudes. * This interference is very small around the Z pole. In the case of Bhabha scattering, $f = e$, one can either subtract from the data the t-channel contributions, also taken from the theory, or add them to the previous expression. The cross section at the peak can be written in turn in terms of the Z mass and width, and the Z partial widths to the initial state Γ_e and the final state Γ_f:

$$\sigma^0_{f\bar{f}} = \frac{12\pi}{M_Z^2} \cdot \frac{\Gamma_e \Gamma_f}{\Gamma_Z^2} \,, \tag{3}$$

where the partial widths can be written in terms of effective vector and axial couplings of the fermions to the Z:

$$\Gamma_l = \frac{G_F M_Z^3}{6\pi\sqrt{2}} \left({g_{V_l}}^2 + {g_{A_l}}^2\right) \left(1 + \frac{3\alpha}{4\pi}\right) \,. \tag{4}$$

Assuming lepton universality, only four parameters are needed to describe the s-dependence of the hadronic and leptonic cross sections: M_Z, Γ_Z, σ^0_h, and the ratio of hadronic to leptonic partial widths ($R_l = \Gamma_h/\Gamma_l = \sigma^0_h/\sigma^0_l$), where the lepton is taken to be massless. If lepton universality is not assumed, R_l is substituted by three analogous quantities, R_e, R_μ, and R_τ. This choice of parameters, given the fact that they are directly related to geometrical characteristics of the lineshape, has two advantages: on the one hand, their correlations are small, and on the other hand they simplify the task of disentangling common uncertainties:

*For some experiments, this is only the case for hadronic final states.

- M_Z and Γ_Z are the only measurements sensitive to the energy scale and they are strongly affected by common LEP energy uncertainties which, thanks to the use of the resonant depolarization technique for the determination of the LEP energy, have been reduced to 1.7 MeV and 1.3 MeV respectively.[3] The present LEP averages are $M_Z = 91.1867 \pm 0.0021$ GeV and $\Gamma_Z = 2.4939 \pm 0.0024$ GeV.
- σ_h^0 is the only lineshape observable in which the overall normalization, and hence the luminosity measurement uncertainties, enter. Presently this measurement is already dominated by the common systematical error due to the theoretical uncertainty in the low-angle Bhabha cross section, which presently is estimated to be of 0.11% (Ref. 4). Since R_l is a ratio of cross sections, the luminosity measurement cancels and there is, so far, no relevant source of common systematics affecting it sizably. The present LEP averages are $\sigma_h^0 = 41.491 \pm 0.058$ nb and $R_l = 20.765 \pm 0.026$.

The forward-backward asymmetry is defined as

$$A_{FB} = \frac{\sigma_F - \sigma_B}{\sigma_F + \sigma_B}, \tag{5}$$

where F and B indicate the forward or backward hemisphere. Normally it is obtained by fitting the measured angular distribution to the formula

$$\frac{d\sigma^l}{d\cos\theta}(s) = \frac{3}{8}\sigma^l(s) \cdot \left(1 + \cos^2\theta + \frac{8}{3}A_{FB}^l(s)\cos\theta\right) . \tag{6}$$

In the case of the e^+e^- final state, the t-channel contribution is either subtracted from the observed asymmetry or added to the previous expression.

Once the different $A_{FB}^l(s_i)$ are obtained, they are fitted together with the lineshape data to get the lineshape parameters mentioned above and the peak asymmetry, $A_{FB}^{0,l}$:

$$\begin{aligned} A_{FB}^{0,l} &= \frac{3}{4}A_e A_l \\ A_f &= \frac{2g_{V_f}/g_{A_f}}{1 + \left(g_{V_f}/g_{A_f}\right)^2}. \end{aligned} \tag{7}$$

from which the effective weak mixing angle is measured. The main error in this measurement is still statistical. Experimental systematics can only come from simultaneous charge and forward-backward asymmetries in the detector, which are bound to be very small. The present knowledge of the beam energy contributes a non-negligible 0.0005 to $\Delta A_{FB}^{0,l}$. The LEP average for the peak asymmetry is $A_{FB}^{0,l} = 0.01683 \pm 0.00096$.

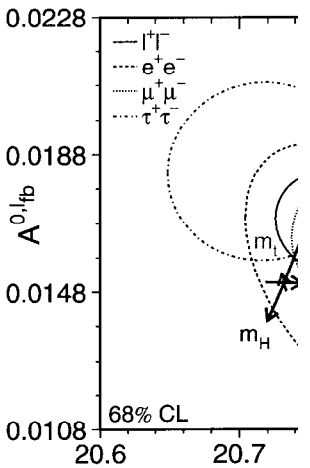


Fig. 1. Contours of 68% probability in the R_l- to correspond to a massless lepton. Unless note the arrows showing the Standard Model $m_t = 173.8 \pm 5.0$ GeV, $M_H = 300\ ^{+700}_{-210}$ GeV, in the direction of increasing values.

Figure 1 summarizes the tests for lepton uni' forward-backward lepton asymmetry analyses

The use of the sketched procedure does in rameterization of the lineshape and asymmetr contribution to the hadronic cross section, bec dividually measured for each flavor. This tern general approach, which is to parameterize th the Z exchange amplitude, has been implem so-called S-matrix ansatz.[5] Fitting the LEP-I

center-of-mass energies between 130 and 172 GeV in that approach, the Z mass as defined above turns out to be $M_Z = 91.1882 \pm 0.0031$ and the parameter which measures the hadronic γZ interference turns out to be $j^{tot}_{had} = 0.14 \pm 0.14$ in good agreement with the MSM expectation $j^{tot}_{had} = 0.22$ (Ref. 2).

2.2 Tau Polarization

The measurement of the final state longitudinal polarization asymmetry for tau leptons:

$$< P_\tau > = \frac{\sigma_R - \sigma_L}{\sigma_R + \sigma_L} = -A_\tau, \tag{8}$$

where $\sigma_{R(L)}$ is the integrated cross section for right (left) handed taus, provides a means to measure the tau couplings to the Z directly. Since the taus decay inside the detectors, their helicity information can be obtained by using their parity-violating weak decays as a spin analyzer. The main decays used are: electron, muon, pion, rho, and a_1. Maximal sensitivity is obtained for the semileptonic modes (pion and rho) since then just one neutrino escapes detection. In this case, the decay angle of the hadron in the tau rest frame (or equivalently its scaled energy) is measured and the integrated tau polarization is extracted from

$$\frac{1}{N}\frac{dN}{dx} = 1 + \alpha < P_\tau > \cos\theta^* , \tag{9}$$

where α is a sensitivity coefficient linked to the spin of the hadron. For pions ($s = 0$), $\alpha = 1$, and for ρ and a_1 ($s = 1$), $\alpha < 1$, but α can be improved by studying the hadron helicity through the analysis of its decay products. The systematics are in this case comparable to the statistical errors and their eventual reduction would require a lot of effort in the further understanding of the calorimeters.

By measuring the tau polarization as a function of the tau production angle, θ one can write

$$P_\tau(\cos\theta) = -\frac{A_\tau + A_e \cdot \frac{2\cos\theta}{1+\cos^2\theta}}{1 + A_e A_\tau \cdot \frac{2\cos\theta}{1+\cos^2\theta}} . \tag{10}$$

From this expression it is apparent that, while the integrated polarization measures A_τ, its forward-backward asymmetry measures the electron coupling, A_e.

The present LEP average results for these two measurements are $A_\tau = 0.1431 \pm 0.0045$ and $A_e = 0.1479 \pm 0.0051$.

This last measurement can be directly compared with the left-right initial state polarization asymmetry from SLD[6] since, in a rather model independent way, both determine the same quantity: $A^0_{LR} = A_e$. The SLD measurement $A_e = 0.1504 \pm 0.0023$ turns out to be in perfect agreement with the LEP average.

2.3 Electroweak Results with Quarks

Given the lack of techniques to tag the light quark flavors efficiently, most of the LEP experiments compute inclusively the forward-backward asymmetry in all hadronic events by estimating the average quark charge via a momentum-weighted mean of charges of the hadrons belonging to each quark's jet. This way, the inclusive forward-backward asymmetry for the actual Z decay mixture of quark flavors, or Quark Charge Asymmetry $< Q_{FB} >$ can be extracted leading to a LEP average of $\sin^2\theta^{lept}_{eff} = 0.2321 \pm 0.0010$. This uncertainty is dominated by the Monte Carlo modeling for light quarks.

The measurement of the Z decay width into b hadrons is especially important because, within the MSM, it receives a vertex correction involving the top quark, which is absent from any other final state. By taking the ratio $R_b = \Gamma_b/\Gamma_h$, most of the vacuum polarization corrections depending on the top quark and the Higgs mass cancel out, and one is left with the following approximate expression:

$$R_b \simeq R_d \cdot \left[1 - \frac{20}{13}\frac{\alpha}{\pi}\left(\frac{m_t^2}{M_Z^2} + \frac{13}{6}\log\frac{m_t^2}{M_Z^2}\right)\right] . \tag{11}$$

Therefore, R_b has a singular role since its accurate measurement should provide an indirect determination of m_t independent of M_H, which is something that none of the rest of precision measurements can do. Anyway, the effect of the top quark vertex corrections is only of order 2% for a top mass of 150 GeV. Therefore, only a precise measurement, to better than 1%, is useful to get information on the top mass.

The relatively large b mass ($\sim$ 4.7 GeV) and lifetime ($\sim 1.5 \times 10^{-12}$ s), makes possible the use of its decay kinematics to have the largest identification efficiency and purities among all the quarks. Several methods have been used to tag b events at LEP:

- Lepton tag: It uses high p and high p_t leptons from b decays. High purity can be achieved but one has to pay for the small semileptonic b branching ratio.
- Event shape tag: High mass, high momentum b mesons, or baryons give rise to particular event shapes which have been used to tag b events with high efficiency although rather modest purity. Recognition has been optimized using Neural Network techniques.
- Lifetime tag: The long lifetime of the b quark can be used, using silicon microvertex detectors, to tag b events by looking for tracks not coming from the Z production vertex. This is currently the best performing method with both high purity and efficiency.

The main systematic errors come from the evaluation of the efficiency and the background of the selection. The best option is to try to use data to estimate both. In practice, the techniques mentioned above can be used to tag only one hemisphere and look at the other one to measure the tag efficiency and the background. This is the basic idea of the "double tag" methods, in which these quantities are obtained from the measurement of the number of the tagged hemispheres N_t and double-tagged events N_{tt} through the following equations:

$$\begin{aligned} N_t &= 2N_{had}(R_b\varepsilon_b + R_c\varepsilon_c + (1 - R_b - R_c)\varepsilon_{uds}) \\ N_{tt} &= N_{had}(R_b\varepsilon_b^2 C_b + R_c\varepsilon_c^2 C_c \\ &\quad +(1 - R_b - R_c)\varepsilon_{uds}^2 C_{uds}) \,, \end{aligned} \tag{12}$$

where ε_i stand for the efficiencies [ε_b is of $\mathcal{O}(25\%)$ with lifetime tag, ε_c typically $\mathcal{O}(0.5\%)$ so that the typical b purity is of $\mathcal{O}(95\%)$ and ε_{uds} are negligible] and C_i are the tagging efficiency hemisphere correlations [typically $\mathcal{O}(1)$]. To solve these equations, R_b and ε_b are taken to be the unknowns while the rest of the quantities are computed using Monte Carlo simulation. Given the high purity of the lifetime tags,

$$\begin{aligned} R_b &\approx N_t^2/(4N_{tt}N_{had}) \\ \varepsilon_b &\approx 2N_{tt}/N_t \end{aligned} \tag{13}$$

and therefore, the uncertainty in the additional quantities determined with Monte Carlo, enters R_b in the following way:

$$\begin{aligned} \Delta R_b/R_b &\approx -1.5 \;\; \Delta\varepsilon_c/\varepsilon_b \\ \Delta R_b/R_b &\approx \Delta C_b \end{aligned} \tag{14}$$

so that, for instance, an absolute error of 1% on C_b translates directly into a 1% relative error on R_b. Given the fact that disentangling in the Monte Carlo the different correlation sources and their actual size is a rather difficult task, the estimation of their actual systematic uncertainty is being reanalyzed in some detail and, its final understanding might produce still sizable changes in the actual measurements. The present results from the four collaborations are shown in Fig. 2.

The c quarks can be tagged in two different ways:

- using b tag techniques (lepton tag, event shape), extending them to the lower p and p_t regions, and then fitting simultaneously the b and c information,
- through the reconstruction of charmed meson decays. The cleanest one is $D^{*+} \rightarrow$

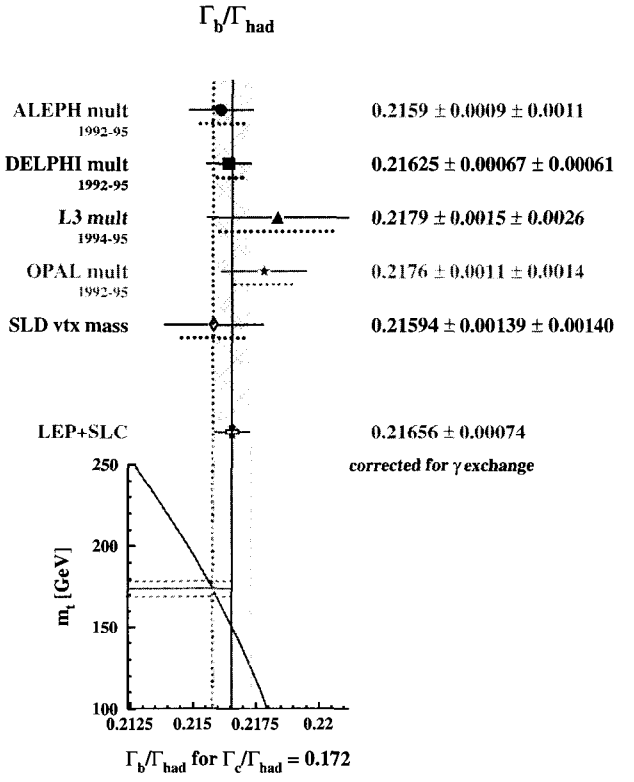


Fig. 2. The ratio of the Z partial width to b hadrons to the one to all hadrons measured by the four experiments, together with the mean and the MSM prediction as a function of the top quark mass. The error bars reflect the full uncertainty.

$D^0\pi^+$ and then $D^0 \to K^-\pi^+$ and its charge conjugate, because the low $D^* - D^0$ mass difference produces the signature of a soft π accompanying a D^0 with opposite charge to the K.

In either case, the efficiencies and purities are much lower than for b quarks and the dependence on external input for the production and decay branching ratios are very important. The results from the four collaborations are shown in Fig. 3, where one can see that the agreement between the different measurements is excellent. Usually the selection cuts are chosen to optimize the ratio between statistical and systematical uncertainties and, in this sense, these measurements are not systematically limited. At any rate, the weight of correlated systematical uncertainties is very important in the final errors and therefore, the proper averaging of the measurements from the different experiments and using different techniques is non-trivial. A detailed discussion, which also applies to the heavy flavor asymmetries to be presented later, can be found in Ref. 7. The improvement in the tag purities in the final analyses of the data have reduced the contamination of b events into the c sample and vice versa, leading to a correlation between the final averages of R_b and R_c of about –0.17.

The heavy quark forward-backward asymmetries provide presently the most precise determination of the effective weak mixing angle at LEP. What is measured is

$$A_{FB}^{0,q} = \frac{3}{4} A_e A_q \,, \tag{15}$$

where A_q is $\sim 0.66, 0.93$ for c- and b-type quarks, respectively, and depends only mildly on $\sin^2\theta_{eff}^{lept}$. Therefore, the asymmetry is quite large and mainly sensitive to the $\sin^2\theta_{eff}^{lept}$ dependence of A_e. The main difficulties measuring the forward-backward asymmetry for quark final states are the flavor identification and the charge assignment.

For b quarks, the charge assignment is done in two different ways: (i) either the charge of a high p, high p_t lepton from semileptonic decays, which identifies the b events, is also used to extract the charge of the parent quark, or (ii) the lifetime information in one hemisphere is used to tag the event while the weighted mean charge (jet charge) in the other hemisphere measures the quark charge. The two methods lead to samples which are almost completely statistically independent. The systematical errors differ as well: in the first method, the knowledge of the lepton purities and of the semi-leptonic branching ratios is crucial; in the second, the charm background in the b sample is the main worry, though, as in the case of the measurement of R_b, a lot of effort has been put by the LEP collaborations in redoing it as much as possible for the final analyses. The results obtained by the four LEP c methods explained can be seen in Fig. 4.

For c quarks, the charge is extracted either fom th global fits using the low p_t leptons from semiletonic

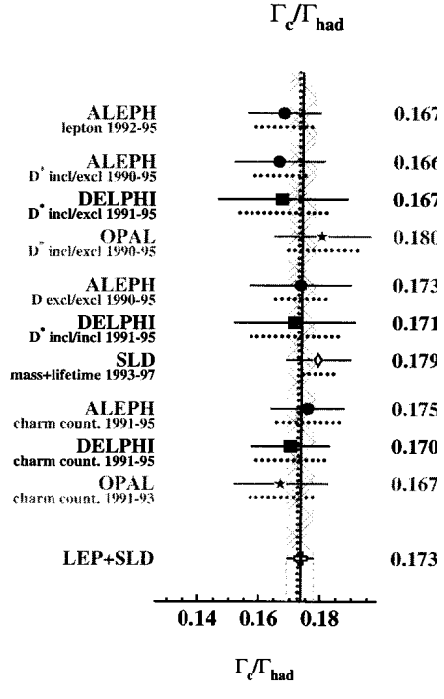


Fig. 3. The ratio of the Z partial width to c hadros to th by the four experiments, together with the mean ɛd the of the top quark mass. The error bars reflect the fil unc

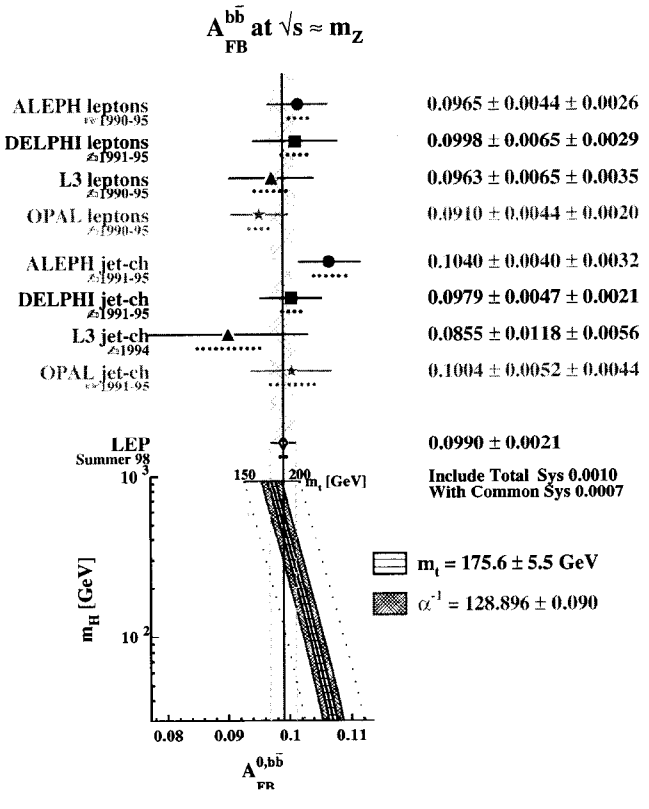


Fig. 4. The b-quark forward-backward charge asymmetry measured by the four experiments, together with the mean and the MSM prediction as a function of the top quark mass. The error bars reflect the full uncertainty.

the $D^{*\pm}$ meson. The results obtained by the four LEP collaborations with the different methods explained can be seen in Fig. 5.

The LEP results are combined at the "raw" asymmetry level, as measured at the average LEP energy and therefore, they have to be corrected for QED, QCD, and energy effects to obtain $A_{FB}^{0,q}$ as appearing in Eq. (15). After the corrections, and including the

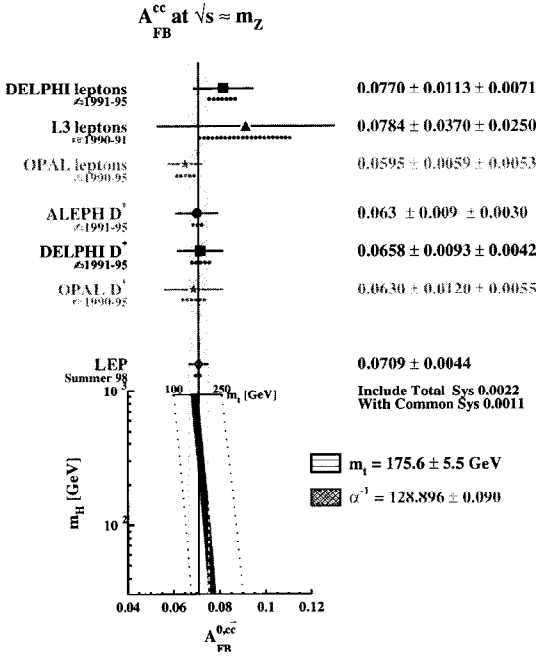


Fig. 5. The c-quark forward-backward charge asymmetry measured by the four experiments, together with the mean and the MSM prediction as a function of the top quark mass. The error bars reflect the full uncertainty.

relevant data from SLD as well, the results are

$$\begin{aligned} A_{FB}^{0,b} &= 0.0990 \pm 0.0021 \\ A_{FB}^{0,c} &= 0.0709 \pm 0.0044 \,. \end{aligned} \tag{16}$$

3 Interpretation of the Z Measurements

3.1 Direct Results

3.1.1 Tests of Lepton Universality

Lepton universality in the Z couplings can be tested by analyzing simultaneously the information coming from the lepton partial widths and the leptonic forward-backward and tau polarization asymmetries. Then, following Eqs. (4) and (7), the vector and axial couplings of the Z to each lepton species can be disentangled. The results (see Fig. 6) show perfect agreement with the hypothesis of lepton universality for both, the vector and axial couplings:

$$\begin{aligned} g_{A_\mu}/g_{A_e} &= 0.9996 \pm 0.0014 \\ g_{A_\tau}/g_{A_e} &= 1.0015 \pm 0.0015 \end{aligned}$$

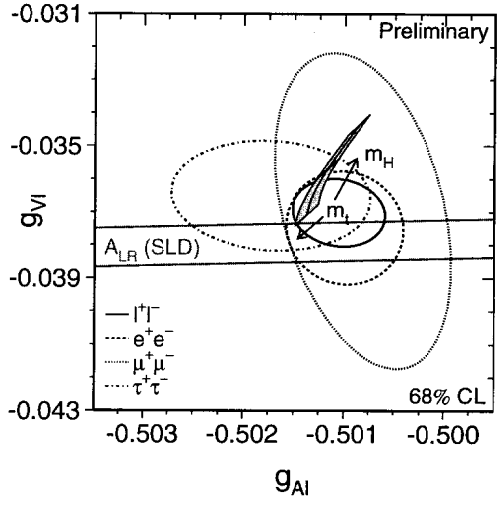


Fig. 6. The 68% probability contours in the g_{V_l}-g_{A_l} plane compared to the MSM predictions.

$$\begin{aligned} g_{V_\mu}/g_{V_e} &= 0.984 \pm 0.098 \\ g_{V_\tau}/g_{V_e} &= 0.974 \pm 0.043 \,. \end{aligned} \quad (17)$$

3.1.2 Tests of Z Couplings to Leptons and Quarks

Assuming lepton universality and using all the above asymmetry measurements, leptonic and hadronic, one can test whether the measured Z couplings to leptons and quarks do agree with their SM expectations or not. The combination of all the LEP leptonic asymmetries leads to a determination of the Z coupling to the leptons of

$$A_l = 0.1469 \pm 0.0027 \,.$$

As already explained, the quark asymmetries determined by LEP do not measure directly the quark couplings to the Z but instead their product with the leptonic ones. This is illustrated in Figs. 7 and 8 where the direct determination of A_b, A_c, and A_l performed at SLD by using initial state polarization is also shown. From these figures it is clear that no significant discrepancy with respect to the SM prediction can be claimed from the LEP measurements. Also, it is clear that the leptonic measurements lead to a value of A_l which in the SM requires a rather light Higgs, as we will see in more detail later.

3.1.3 The Effective Weak Mixing Angle

Figure 9 shows the compilation of all the values of $\sin^2\theta^{lept}_{eff}$ obtained from the asymmetry measurements at LEP presented in the previous sections. The SLD determination, coming from the measurement of the left-right polarization asymmetry,[6] is also shown. Its discrepancy with the average of the LEP measurements alone is about 2.1 σ. The overall agreement of all the measurements is reasonably good and the mean value provides a very precise determination of the effective weak mixing angle, which is very sensitive to the top quark and the Higgs masses.

Figure 10 summarizes the information coming from the measurement of the LEP asymmetries and the LEP lineshapes, and shows clearly the need for the existence of genuine electroweak radiative corrections beyond the running of the electromagnetic coupling constant to explain the measured values. It also shows that the uncertainty in $\Delta[1/\alpha(M_Z^2)] = 0.090$ that we'll assume throughout this note, has an impact in the interpretation of $\sin^2\theta^{lept}_{eff}$ comparable to the experimental error.

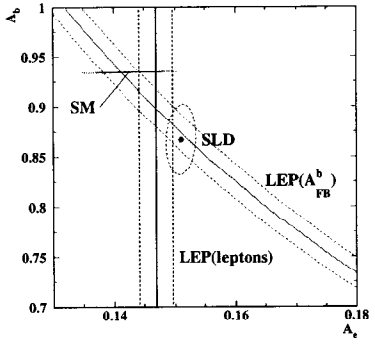


Fig. 7. One-sigma constraints in the A_b versus A_l plane coming from the leptonic and hadronic asymmetry measurements.

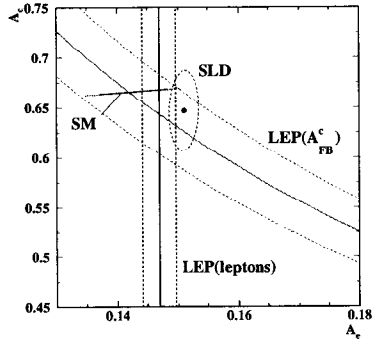


Fig. 8. One-sigma constraints in the A_c versus A_l plane coming from the leptonic and hadronic asymmetry measurements.

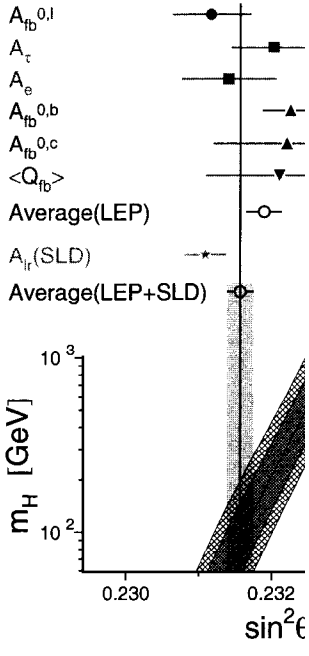


Fig. 9. The LEP averages for $\sin^2\theta^{lept}_{eff}$ obtain
ables discussed in the text, together with the S

3.1.4 Quantities Derived from Ratios

The ratios of partial widths, due to the cancela
numerator and denominator, allow the direct
the theory without relying, in principle, too
tensions of the MSM which would manifest
polarization would produce very similar predi

R_l depends little on m_t and M_H whereas
coupling constant, $R_l \sim R^0_l(1+\alpha_s(M_Z)/\pi)$. T
of $\alpha_s(M_Z)$ with minimum theoretical uncerta

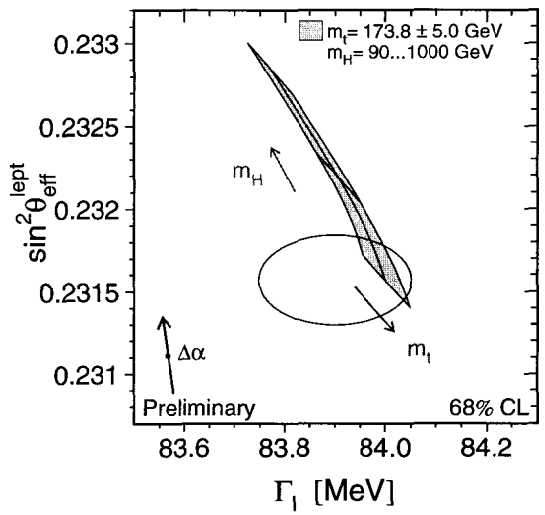


Fig. 10. The average for $\sin^2\theta^{lept}_{eff}$ and for Γ_l compared with the SM predictions. The arrow shows these predictions when the only electroweak radiative correction included is the running of the electromagnetic coupling constant.

using the formulae suggested in Ref. 8, which relate R_l to the QCD prediction, known to $\mathcal{O}(\alpha_s^3)$, one gets

$$\alpha_s(M_Z^2) = 0.122 \pm 0.004 \pm 0.001\ ,$$

where the second error accounts for varying M_H in the range, $M_H = 76^{+85}_{-47}$ GeV. This result is in good agreement with the world average, $\alpha_s(M_Z^2) = 0.119 \pm 0.002$.

The invisible width, Γ_{inv}, can be derived from the direct lineshape measurements by subtracting from the total width the one due to visible channels. If one assumes that all the invisible width is due to neutrino final states, then one can derive the number of light neutrino species N_ν by writing

$$\frac{\Gamma_{inv}}{\Gamma_l} = N_\nu \cdot \frac{\Gamma_\nu}{\Gamma_l} \qquad (18)$$

and taking the ratio Γ_ν over Γ_l from the MSM: $\Gamma_\nu/\Gamma_l = 1.991 \pm 0.001$. A small error should be noted in the MSM prediction for this ratio, which does not depend on α_s and in which the top and Higgs mass dependences largely cancel. Using the LEP average for $\Gamma_{inv}/\Gamma_l = 5.961 \pm 0.023$ one obtains

$$N_\nu = 2.994 \pm 0.011\ .$$

Since the result favors three species without any doubt, this measurement can then be interpreted as the determination of the width for additional invisible decays of the Z, which yields:

$$\Delta\Gamma_{inv} = -1.6 \pm 1.9\ \text{MeV}$$

or, equivalently

$$\Delta\Gamma_{inv} < 2.8\ \text{MeV}$$

at the 95 % confidence level.

4 The LEP-II Measurements

The LEP-I phase finished in 1995 and Table 1 shows the evolution of the center-of-mass energy in LEP after the LEP-II phase started, and the accumulated luminosity in each energy regime. Altogether, the data already analyzed amounts to less than 4000 W pairs. The data taken in 1998 is still being analyzed and therefore will not be covered in this note. The plans for the two forthcoming years are increasing the LEP energy as much as possible (up to 200 GeV) and taking as much luminosity as possible (about 100 pb^{-1}/exp./year).

Basically three different regimes can be distinguished in what concerns the LEP operation after the LEP-I phase:

- The so-called LEP 1.5 phase in which the energy was not sufficient for the production of W pairs and, hence the only relevant electroweak studies were those of the measurement of the γZ interference terms in the S-matrix language.
- In 1996 LEP raised the center-of-mass energy up to 161 GeV, reaching the threshold for the W-pair production and delivering about 40 pb^{-1} in total, which meant about 120 WW events. That "threshold regime" allowed, as we shall see, the indirect determination of the W mass through its effect on the observable cross section.

Table 1. The LEP center-of-mass energy and luminosity after the LEP-I phase.

Year	$\sqrt{s}$ (GeV)	$L(pb^{-1})$/exp.	W pairs
Nov. 1995	130/136 (LEP 1.5)	6	-
June-Aug. 1996	161 (WW threshold)	10	~ 30
Oct.-Nov. 1996	172	10	~ 100
July-Nov. 1997	183	55	~ 850
Nov. 1997	130/136	7	-
1998	189	~ 200	?

- After that, the increase in the LEP energy allowed the study of the WW cross section and asymmetries, independent of the direct measurement of their mass, by fitting the invariant mass distribution of their decay products.

These last two regimes are the relevant ones for the study of the W properties, which is the main electroweak subject of the LEP-II run. The other electroweak analyses with that data (mainly related to the Z events) are not going to be covered in this note.

The Ws are pair produced at LEP and, given the large W width, only their decay products can be detected. Ws are expected to decay into a lepton and a neutrino 33% of the time, and hadronically into jets 67% of the time. Therefore, the events observed at LEP can be classified into three categories:

- Fully Leptonic: ($BR = 11\%$), in which both Ws decay leptonically, leaving just two visible leptons. The events are characterized by a low multiplicity, high missing p_T, and large acoplanarity. Selection efficiencies are typically of about 70% for purities of about 90%. For this channel the backgrounds are not a big issue and come mainly from two-photon events and from radiative return Z leptonic decays (especially into taus) and at energies above the ZZ threshold from $ZZ \to l^+l^-\nu\bar{\nu}$ events.
- Semileptonic: ($BR = 44\%$), in which one W decays leptonically and the other decays hadronically. These events have large multiplicity, large missing p_T, and an isolated energetic lepton which should be rather antiparallel to the missing momentum direction. The selection efficiency is typically of the order of 85% for the lepton being an electron, 90% if it is a muon, and about 60% when it is a tau lepton, for purities well above 90%. The backgrounds for this channel are very low and come mainly from $q\bar{q}$ events, from Zee, and from other four-fermion processes.
- Fully hadronic: ($BR = 45\%$), in which both Ws decay hadronically, leading basically to four-jet events. The signature in this case is the high multiplicity and the nonexistence of missing mass. The background for this channel is rather severe, especially the one due to the process $q\bar{q} \to 4jets$ and therefore, the selection requires the use of the most sophisticated multivariable techniques (linear discriminant analyses, neural networks, etc.). The use of all sorts of kinematical variables, together with the use of variables related to lifetime tagging within these techniques (unlike the Zs, the Ws do not decay into b quarks), allow efficiencies above 80% for purities above 80%.

In the following sections we are going to describe how these different samples are used to determine the W boson parameters.

4.1 WW Cross Section and Branching Ratios

As already said, due to the extremely short W lifetime, only the four fermions coming from the W pair decay are observed in our detector. Therefore, in practice, all the possible ways in which these identified four fermions could come from the e^+e^- interaction should be considered as interfering channels to produce our detected final state. This interference which in principle could be large, can be drastically reduced by requiring the four fermions to be well inside the detector and well separated. Nevertheless, one needs a precise prescription to define what the WW cross section is and how can it be obtained from the measured four fermion cross section. The definition agreed among the four LEP collaborations corresponds to the diagrams depicted in Fig. 11 with the subsequent decay of the produced Ws:

$$\sigma_{WW} \equiv \sigma(e^+e^- \to W^+W^- \to 4f) \ ,$$

that is, the cross section through only the double W resonating diagrams (technically called CC03 diagram subset after the LEP-II workshop).[9]

The actual cross section measured by the LEP collaborations is corrected to this definition either by applying a correction factor or just by subtracting the remaining contributions estimated with a Monte Carlo simulation.

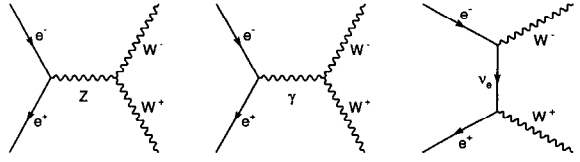


Fig. 11. The Feynman diagrams for the process, $e^+e^- \rightarrow W^+W^-$.

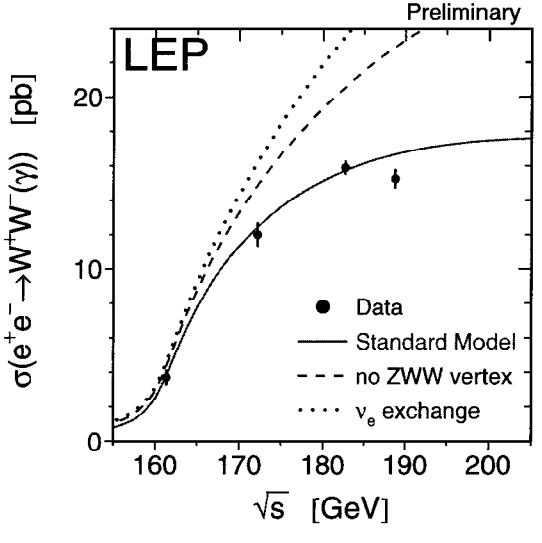


Fig. 12. The LEP average of the W pair cross section as a function of the center-of-mass energy compared with the SM prediction. The (very preliminary) measurement at 189 GeV is also shown. For illustration purposes the predictions neglecting the triple gauge coupling contributions are also given.

Figure 12 shows the LEP averaged total CC03 cross section as a function of the center-of-mass energy. The agreement with the SM expectation is excellent and excludes clearly a description in which the triple gauge coupling is neglected.

The analysis of the above cross sections for each individual channel allows the precise determination of the W branching ratios as can be seen from Figs. 13 and 14. In this second case, within the Standard Model, the branching ratio of the W bosons depends on the six elements of the Cabibbo-Kobayashi-Maskawa quark-mixing matrix (V_{CKM}) not involving the top quark and on α_s. Using the current world-averages and errors of other matrix elements, not assuming the unitarity of V_{CKM}, and from the world average of $\alpha_s(M_Z^2) = 0.119 \pm 0.002$ evolved to M_W^2, a constraint on the least well measured CKM matrix element is obtained:

$$|V_{cs}| = 1.04 \pm 0.04 \, .$$

4.2 W Mass

The W mass has been measured so far at LEP with the following two rather different techniques.

4.2.1 The Threshold Measurement

At threshold energies, the influence of the mass in the cross section is fundamentally a kinematic effect in the measured cross section which gives the highest sensitivity at about 161 GeV. The procedure to extract the W mass consists of the measurement of the cross section and its subsequent analysis in terms of the W mass using the most accurate theoretical predictions.[10] Although at these energies, no observable effects can be expected from the eventual existence of anomalous W couplings, it is true that one has to implicitly rely on the validity of the SM prediction of the cross section to extract the W mass. In this sense it is not a "model independent" measurement of M_W although it is free from some of the problems that affect other techniques.

4.2.2 The Direct Mass Reconstruction

Above threshold, the W mass is mainly obtained by the analysis of the invariant-mass distribution of the W decay products. This technique is only possible for semileptonic

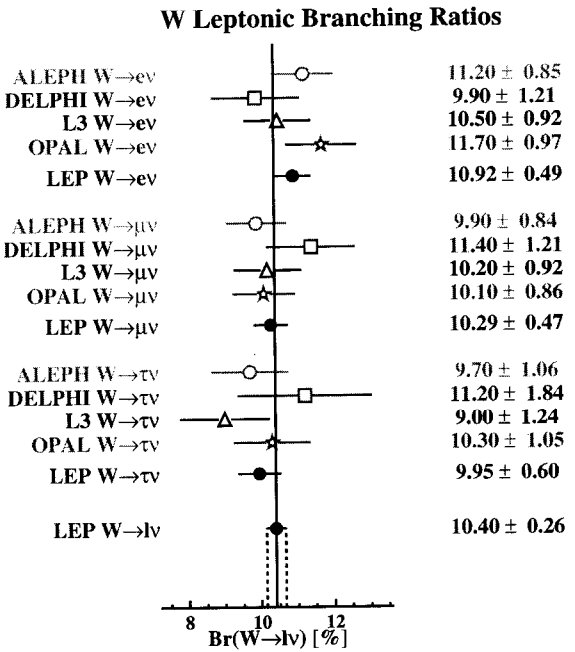


Fig. 13. The LEP measurements of the W leptonic branching ratios.

and fully hadronic events. In the first case, the events are forced to a two-jet configuration after excluding the lepton candidate. In the second case they are forced to a four (or more)-jet configuration using the jet algorithm giving the best invariant mass resolution.

In both cases, the invariant mass resolution is improved by applying a kinematical fit to reconstruct four objects using the energy and momentum constraints. The invariant mass resolution is further improved by constraining the two masses of the event to be (exactly or approximately) equal. This is called two-C fit in the semileptonic channel and five-C fit in the fully hadronic one.

In the fully hadronic case, the impossibility of unequivocally identifying the quark generating each jet introduces additional complications. Apart from the non-negligible

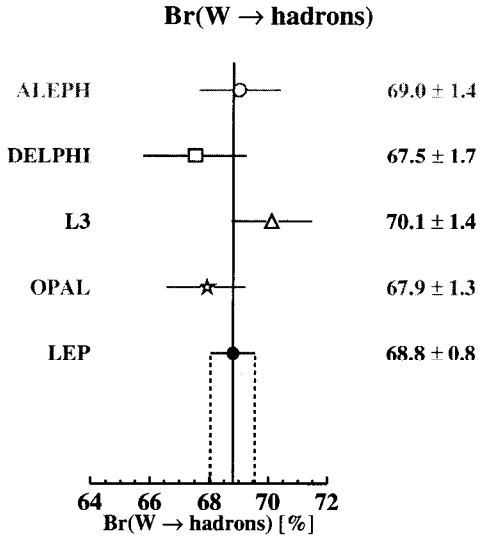


Fig. 14. The LEP measurements of the W hadronic branching ratios.

physics background already mentioned, the need for pairing the jets to determine the two W invariant masses introduces a non negligible combinatorial background which is reduced by the use of dedicated pairing algorithms.

After these steps, the invariant mass distributions have a Breit-Wigner like shape as can be seen in Figs. 15 and 16, but which is sizably distorted by several effects such as initial state radiation, phase space boundary, detector resolution, misassignment of particles to Ws, backgrounds, event selection, etc. Therefore, to get a precise determination of the W mass, elaborate fitting techniques should be applied to the reconstructed events. Out of the different techniques tried by the LEP experiments, one of the most popular is the use of Monte Carlo reconstructed mass distributions to compare directly with the data mass distributions. The idea is to have many large Monte Carlo template distributions generated with different W parameters to see which one fits the data best. Technically, instead of generating large MC samples at different W parameters, only a large sample is generated at a reasonable value of the W parameters (M_W^{MC}, Γ_W^{MC}) and

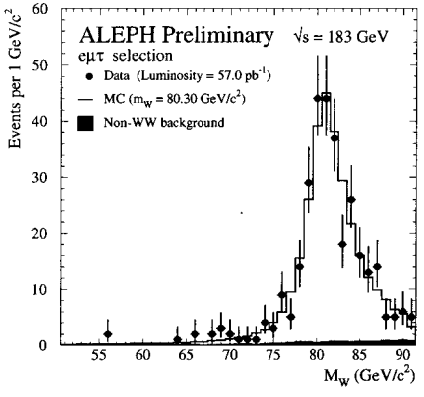


Fig. 15. The W reconstructed invariant mass distribution in semileptonic events.

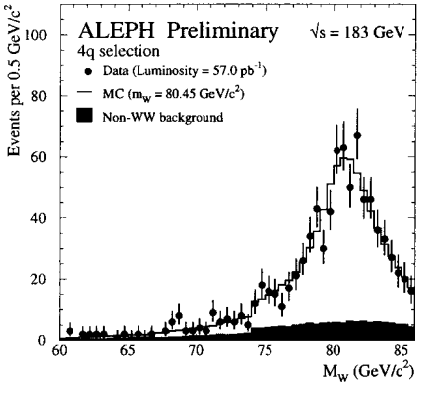


Fig. 16. The W reconstructed invariant mass distribution in fully hadronic events.

then the events are reweighted using the ratio of the matrix elements squared,

$$\omega_i(M_W, \Gamma_W) = \frac{|M(M_W, \Gamma_W; p_i^1, p_i^2, p_i^3, p_i^4)|}{|M(M_W^{MC}, \Gamma_W^{MC}; p_i^1, p_i^2, p_i^3, p_i^4)|}$$

to reproduce in a continuous manner the predictions for any other value of the W parameters with the same MC events.

In practice, from the fit to the invariant mass distributions, all the experiments extract M_W and some of them also Γ_W, though for this last parameter the present accuracy is still very poor.

4.2.3 Final State Interactions

An additional complication in the case of the fully hadronic channel is the fact that the expected separation of the W decay vertices ($1/\Gamma_W \sim 0.1$ fm) is smaller than the decay product hadronization scale [$\mathcal{O}(1\text{fm})$] so that the final state particle production cannot be expected to be dictated by the separate decays of the W bosons and therefore one may expect the identities of the individual W decay products not to be well defined any more. Two different kind of effects have been suggested as possible sources of W "crosstalking":

- The "color reconnection" between the quarks from different Ws. Since the Ws are color singlets, this effect is expected to be very small (a factor $\sim 1/N_c^2$ suppression) in the perturbative regime. Nevertheless it could, in principle, be sizable in the non-perturbative regime although, since in that regime there are no precise predictions, all we have are just phenomenological models which has just now started to be contrasted with the data. The models considered so far can be grouped into two families:
  - Color Singlet Models,[11] which predict mass shift effects ranging from zero to 50 MeV depending on the model parameters and the experimental procedure used.
  - Non-singlet Models,[12] which predict larger shifts but are already disfavored by the Z data and by the multiplicity distributions.
- The Bose-Einstein correlations between identical bosons (pions) from different Ws. As before, the expected mass shift depends on the model parameters[13] and on the experimental procedure used, and ranges from zero to about 50 MeV.

The detailed analysis of all the event properties should allow to discriminate between these models and to adjust their parameters independently from the W mass determination. Unfortunately, so far the statistical power of the data has been too limited to make any conclusive statement, although the preliminary results are encouraging and show the potential of the methods with the forthcoming statistics.

4.2.4 Results

The LEP results for the W mass determination, using the techniques explained above, are shown in Fig. 17. As can be seen, the most precise determination comes from the direct reconstruction and is in perfect agreement with the cross-section determination. Figure 18 shows that the LEP measurement is in good agreement with all the direct and indirect determinations of the W mass performed so far and that its accuracy is already comparable to the one obtained in the proton colliders, which so far were giving the best direct measurement.

In the direct reconstruction result, the uncertainty can be split as follows:

$$M_W = 80.32 \pm 0.08 \pm 0.05\,(FSI) \pm 0.02\,(LEP)\ \text{GeV}\,,$$

where the first term is the experimental uncertainty, FSI stands for the theoretical uncertainty due to Final State Interactions, and LEP stands for the uncertainty in the LEP center-of-mass energy. The direct reconstruction results for semileptonic and fully hadronic channels separately are:

$$\begin{aligned} M_W^{semileptonic} &= 80.31 \pm 0.11 \pm 0.02\,(LEP)\ \text{GeV} \\ M_W^{hadronic} &= 80.39 \pm 0.10 \pm 0.09\,(FSI) \pm 0.02\,(LEP)\ \text{GeV} \end{aligned}$$

so that no discrepancy pointing to an eventually large FSI effect is observed.

4.3 Triple Gauge Couplings

Since the couplings ZW^+W^- and γW^+W^- do enter at tree level in the prediction of the $e^+e^- \to W^+W^-$ cross section (see Fig. 11), LEP-II allows for the first time the direct study of the non-abelian structure of the electroweak SM in e^+e^- collisions.

The most general structure of the Lagrangian describing each one of these triple gauge couplings (TGC) which would preserve Lorentz invariance would have seven different terms, while in the Standard Model at tree level, just two terms for each coupling are predicted. Nevertheless, out of these seven terms, just three conserve C and P,

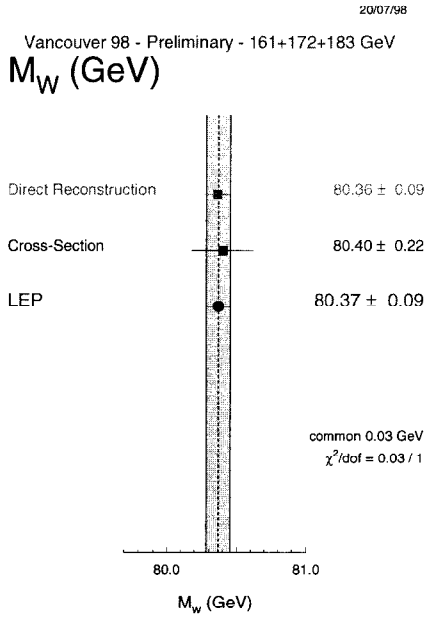


Fig. 17. The LEP determinations of the W mass.

while the other three violate CP and the last one just conserves CP. For several reasons to follow we will consider only those terms that conserve C and P, namely g_1, κ, and λ, which are related to the static electromagnetic moments for the W^+ in the following way:

$$\begin{aligned} Q_W &= eg_1^\gamma && \text{Electric charge} \\ \mu_W &= \tfrac{e}{2M_W}(g_1^\gamma + \kappa_\gamma + \lambda_\gamma) && \text{Magnetic dipole moment} \\ q_W &= -\tfrac{e}{M_W^2}(\kappa_\gamma - \lambda_\gamma) && \text{Electric quadrupole moment} \end{aligned}$$

and the same for the weak charge and moments by just substituting γ by Z. Any theory

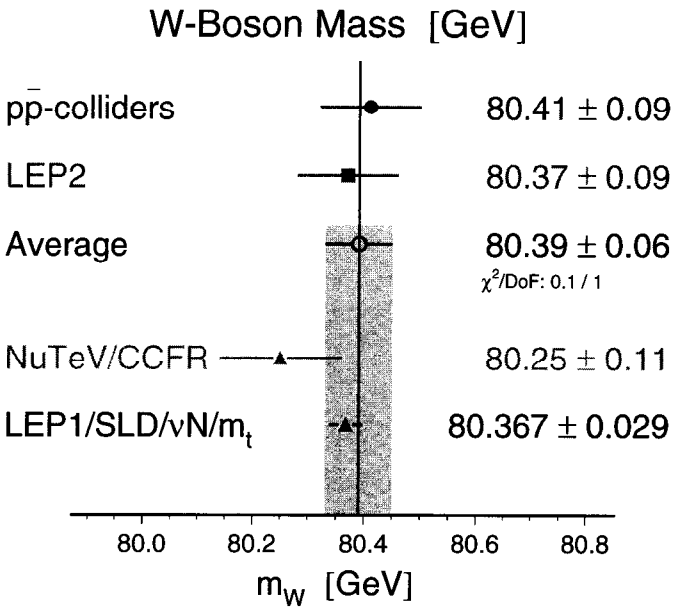


Fig. 18. Summary of all the determinations of the W mass.

incorporating new physics beyond the SM and including the electroweak theory as an effective low-energy limit may introduce deviations of these parameters from their SM values, which are:

$$\begin{aligned} g_1^\gamma &= g_1^Z = 1 \\ \kappa_\gamma &= \kappa_Z = 1 \\ \lambda_\gamma &= \lambda_Z = 0 \end{aligned}$$

so that the deviations can be quantified using $\Delta g_1 = g_1 - 1$, $\Delta\kappa = \kappa - 1$, and λ directly.

The assumption of electromagnetic gauge invariance reduces further the number of free parameters to five since it implies $g_1^\gamma = 1$ and, if in addition, global $SU(2)$ invariance is required, then the following relationship among these five parameters should hold:

$$\begin{aligned} \Delta\kappa_Z &= \Delta g_1^Z - \Delta\kappa_\gamma \tan^2\theta_W \\ \lambda_Z &= \lambda_\gamma \,. \end{aligned}$$

Therefore, we are left with just three free parameters which are Δg_1^Z, $\Delta\kappa_\gamma$, and λ_γ. Deviations of any of these three parameters from zero imply changes in the WW cross section, W production angles, W decay angles, and their correlations, which one can try to detect in the data.

For this purpose, so far only the semileptonic and fully hadronic channels are used. The selection and jet definition proceed typically in a similar way as for the mass determination. In the semileptonic events, the lepton charge is used to sign the W $\cos\theta$ direction and only a two-fold ambiguity remains to compute the decay angles in the hadronic W decay. In the fully hadronic case, the jet charge Q_{JET} is used to sign the W $\cos\theta$ direction and, on top of the combinatorial background already mentioned while discussing the mass determination, a four-fold ambiguity in the decay angles remains due to the two hadronic W decays. Due to these problems, the semileptonic channel is the most powerful one to study the triple gauge couplings.

A remaining problem is how to fit the mentioned three free parameters from the five-dimensional kinematical information, in addition to the total cross section. Several techniques have been developed in the different LEP collaborations which can be basically grouped as follows:

- Multidimensional Maximum Likelihood: It should allow to exploit the full angular information but in practice it is technically rather complex. Two approaches have been tried:
  - Unbinned using a semianalytical expression: It allows a relatively fast parameter fitting but requires afterwards, the application of rather involved multidimensional corrections in order to account for the biases and sensitivity losses introduced by backgrounds and experimental distortions.
  - Binned using Monte Carlo events: It incorporates properly all these corrections but requires a very large number of Monte Carlo events to be sure that all the bins in five-D do have significant population.
- Optimal Observables: The idea is to project the relevant kinematical information for a certain TGC observable onto a one-D distribution. Since triple gauge cou-

plings contribute at tree level only linearly to the process amplitude, they show up in the differential cross section as a quadratic effect:

$$\frac{d\sigma(\Omega,\bar{\alpha})}{d\Omega} = S_0(\Omega) + \sum_i S_{1;i}(\Omega)\,\alpha_i + \sum_{i,j} S_{2;i,j}(\Omega)\,\alpha_i\alpha_j\ ,$$

where α_i stands for the couplings and Ω stands for the observed kinematical variables. From this simple equation it can be shown that the optimal observable for coupling "i" can be defined as:

$$OO_i \equiv \frac{\frac{d}{d\alpha_i}\left(\frac{d\sigma(\Omega,\bar{\alpha})}{d\Omega}\right)|_{\alpha_i=0}}{\frac{d\sigma(\Omega,\bar{\alpha})}{d\Omega}|_{\alpha_i=0}} = \frac{S_{1;i}(\Omega)}{S_0(\Omega)}\ .$$

In practice, the Monte Carlo $d\sigma/d\Omega$ is used to account for the backgrounds and experimental effects, and the parameters are extracted by computing the value of the probability of the observed value of the OO_i for each event and then maximizing their combined likelihood function.

By now, given the limited statistical power of the measurements, the results from the different LEP experiments are provided as likelihood curves for a single parameter fit while the other parameters are zeroed. These likelihood curves are then combined, as can be seen in Fig. 19. From the results collected in that figure, no significant discrepancy with the SM is observed.

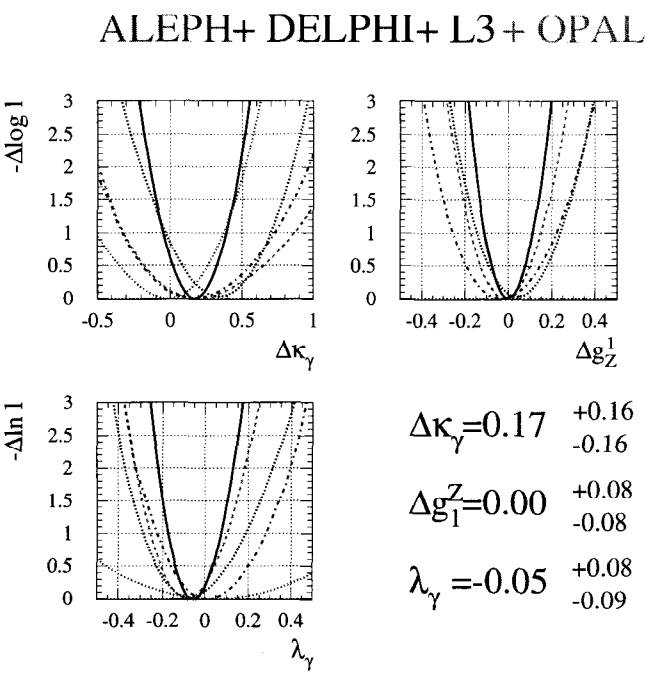


Fig. 19. Summary of the LEP constraints on the triple gauge couplings.

5 Standard Model Fits

The results presented in the previous sections can be interpreted in the context of the MSM, allowing a check of its validity in describing all the observations, and enabling the inference of some of the fundamental MSM parameters.

The comparison is made through a fit of the measurements shown in Table 2 to their MSM predictions in terms of the top quark mass m_t, the Higgs boson mass M_H, and the value of $\alpha_s(M_Z)$.

The most up-to-date MSM calculations have been used and their estimated theoretical uncertainties[14] have also been propagated in the fitting procedure. The uncertainty in $1/\alpha(M_Z^2) = 128.878 \pm 0.090$,[15] due to the contribution of light quarks in the photon vacuum-polarization, is the one that dominates by far.

The results of the fit with different data sets are presented in Table 3. It is clear from these results that the data prefers a rather light Higgs boson mass as can be seen also in Fig. 20. In that figure, the impact of taking $1/\alpha(M_Z^2) = 128.905 \pm 0.036$, as advocated in Ref. 16, is also shown.

In addition, from Table 3 one sees that there is a nice agreement between the direct measurement of m_t and M_W and their indirect determination through their effect in the precision electroweak measurements. This fact constitutes a successful deep test of the Standard Model predictions beyond the tree level structure. This agreement can also be seen in Fig. 21 which also illustrates the fact mentioned above: in a completely independent manner, the precision electroweak measurements and the direct determinations of m_t and M_W point to a rather light Higgs boson mass.

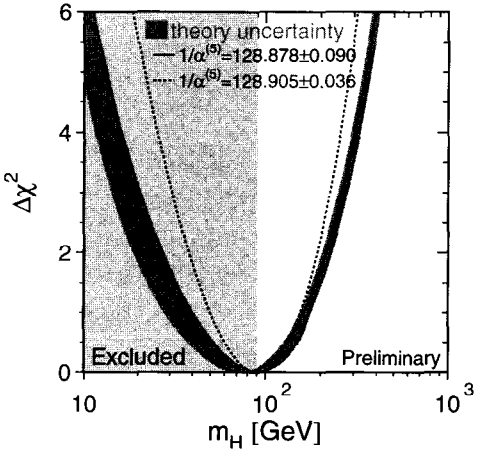


Fig. 20. $\Delta\chi^2$ vs $M_{\rm H}$ using all data.

	Measurement	Correlation Matrix	Standard Model Fit	Pull
$\alpha(M_Z^2)^{-1}$	$128.878 \pm 0.090(83)$		128.878	0.0
a) LEP				
line-shape and				
lepton asymm.:				
M_Z [GeV]	$91.1867 \pm 0.0021(17)$	1.00	91.1865	0.1
Γ_Z [GeV]	$2.4939 \pm 0.0024(13)$	0.00 1.00	2.4958	−0.8
σ_h^0 [nb]	$41.491 \pm 0.058(57)$	-0.04 -0.18 1.00	41.473	0.3
R_l	$20.765 \pm 0.026(20)$	-0.01 0.00 0.12 1.00	20.748	0.7
$A_{FB}^{0,l}$	$0.01683 \pm 0.00096(60)$	0.06 0.00 0.01 -0.07 1.00	0.01613	0.7
τ polarization:				
A_τ	$0.1431 \pm 0.0045(27)$		0.1467	−0.8
A_e	$0.1479 \pm 0.0051(10)$		0.1467	0.2
b and c quark:				
R_b^0	$0.21656 \pm 0.00074(57)$	1.00	0.21590	0.9
R_c^0	$0.1735 \pm 0.0044(34)$	-0.17 1.00	0.1722	0.3
$A_{FB}^{0,b}$	$0.0990 \pm 0.0021(10)$	-0.06 0.05 1.00	0.1028	−1.8
$A_{FB}^{0,c}$	$0.0709 \pm 0.0044(22)$	0.02 -0.04 0.13 1.00	0.0734	−0.6
$q\bar{q}$ charge asymm.:				
$\sin^2\theta_{eff}^{lept}(Q_{FB})$	$0.2321 \pm 0.0010(8)$		0.23157	0.5
M_W [GeV]	$80.37 \pm 0.09(6)$		80.37	0.0
b) SLD				
$\sin^2\theta_{eff}^{lept}(A_{LR})$	$0.23109 \pm 0.00029(18)$		0.23157	−1.7
A_b	$0.867 \pm 0.035(25)$		0.935	−1.9
A_c	$0.647 \pm 0.040(23)$		0.668	−0.5
c) $p\bar{p}$ and νN				
M_W [GeV] ($p\bar{p}$)	$80.41 \pm 0.09(07)$		80.37	0.4
$1 - M_W^2/M_Z^2(\nu N)$	$0.2254 \pm 0.0021(10)$		0.2232	1.1
m_t [GeV] ($p\bar{p}$)	$173.8 \pm 5.0(3.9)$		171.1	0.5

Table 2. Summary of measurements included in the combined analysis of SM parameters. Section a) summarizes LEP averages (which in the case of R_b and R_c, also include SLD); section b) summarizes the relevant additional SLD results; and section c) summarizes the electroweak results from hadron colliders and νN-scattering. The size of the estimated systematic component of the quoted uncertainty is indicated in parentheses. The MSM fit result in column three and the pulls in column four are derived from the fit including all data with the Higgs mass as a free parameter.

Table 3. Results of fits to LEP and other data for m_t, M_H, and $\alpha_s(M_Z^2)$. No external constraint on $\alpha_s(M_Z^2)$ has been imposed.

	LEP	No M_W,no m_t	No M_W	All data
m_t (GeV)	$160\ ^{+13}_{-9}$	$158\ ^{+9}_{-8}$	171.0 ± 4.9	151.1 ± 4.9
M_H (GeV)	$60\ ^{+127}_{-35}$	$32\ ^{+41}_{-15}$	$82\ ^{+95}_{-51}$	$76\ ^{+85}_{-47}$
$\alpha_s(M_Z^2)$	0.121 ± 0.003	0.120 ± 0.003	0.120 ± 0.003	0.119 ± 0.003
$\chi^2/(d.o.f.)$	4/9	13/12	15/13	15/15
$\sin^2\theta^{lept}_{eff}$	0.23182 ± 0.00023	0.23157 ± 0.00018	0.23159 ± 0.00020	0.23157 ± 0.00019
M_W (GeV)	80.314 ± 0.038	80.332 ± 0.037	80.367 ± 0.029	80.371 ± 0.026

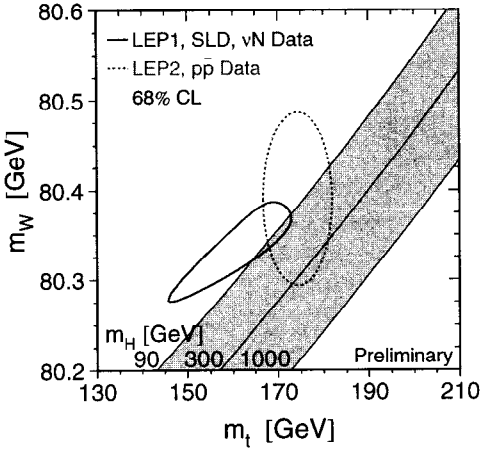


Fig. 21. The direct and indirect determinations of m_t and M_W compared with the SM predictions.

6 Summary

The precision electroweak measurements at LEP-I have been a big success, even beyond the most optimistic expectations. Thanks to these measurements, the Z boson properties are known today with very high precision.

The LEP-II electroweak precision program started just a few years ago and is now reaching maturity. The most precise measurement, M_W, has already reached the same precision as in collider experiments. The statistics accumulated during 1998 and the ones foreseen in the coming years should also allow precision measurements of the rest of the W boson properties.

Altogether, these measurements allow an unprecedented test of the good health of the MSM in describing the electroweak data and, together with the relevant electroweak measurements performed around the world, open a significant window for the inference of the last missing piece of the puzzle: M_H.

References

[1] M. Martinez *et al.*, "Precision Tests of the Electroweak Interaction and the Z Pole," Rev. Mod. Phys. **71**, 575 (1999).

[2] The LEP Electroweak Working Group, "A Combination of Preliminary Electroweak Measurements and Constraints on the Standard Model," CERN-EP/99-015, February 1999. Prepared from Contributions of the LEP and SLD Experiments to the 1998 summer conferences.

[3] The LEP Energy Working Group, "Calibration of the Center-of-Mass Energies at LEP1 for Precise Measurements of Z Properties," Eur. Phys. J. C **6**, 187 (1999).

[4] S. Jadach *et al.*, "New Results on the Theoretical Precision of the LEP/SLC Luminosity," Phys. Lett. B **450**, 262 (1999).

[5] A. Leike, T. Riemann, and J. Rose, Phys. Lett. B **273**, 513 (1991); T. Riemann, Phys. Lett. B **293**, 451 (1992); S. Kirsh and T. Riemann, Comput. Phys. Commun. **88**, 89 (1995).

[6] K. Baird for the SLD Collaboration, "Measurements of A_{LR} and A_l from SLD," talk presented at ICHEP 98, Vancouver, B.C., Canada, 23–29 July 1998.

[7] The LEP Heavy Flavor Group, "Input Parameters for the LEP/SLD Electroweak Heavy Flavor Results for Summer 1998 Conferences," Internal note LEPHF/98-01, "http://www.cern.ch/LEPEWWG/heavy/lephf9801.ps.gz."

[8] T. Hebbeker *et al.*, Phys. Lett. B **331**, 165 (1994); P. A. Raczka and A. Szymacha, Phys. Rev. D **54**, 3073 (1996); D. E. Soper and L. R. Surguladze, Phys. Rev. D **54**, 4566 (1996).

[9] "Physics at LEP2," edited by G. Altarelli, T. Sjöstrand, and F. Zwirner, CERN 96-01 Yellow Report, Vol. 1, February 1996.

[10] D. Bardin, *et al.*, Nucl. Phys. B (Proc. Suppl.) **37**, 148 (1994); S. Jadach *et al.*, Comput. Phys. Commun. **94**, 216 (1996).

[11] T. Sjöstrand and V. A. Khoze, Z. Phys. C **62**, 281 (1994); L. Lönnblad and T. Sjöstrand, Phys. Lett. B **351**, 293 (1995); V. A. Khoze and T. Sjöstrand, Eur. Phys. J. C **6**, 271 (1999) and references therein.

[12] J. Ellis and K. Geiger, Phys. Lett. B **504**, 230 (1997).

[13] S. Jadach and K. Zalewski, CERN-TH/97-29, Acta Phys. Polon. B **28**, 1363 (1997).

[14] D. Bardin *et al.*,"Precision Calculation Working Group" (PCWG), CERN 95-03 Yellow Report, 31 March 1995 and references therein; D. Bardin and G. Passarino, "Upgrading of Precision Calculations for Electroweak Observables," CERN-TH/98-92 [hep-ph/9803425].

[15] S. Eidelmann and F. Jegerlehner, Z. Phys. C **67**, 585 (1995); H. Burkhardt and B. Pietrzyk, Phys. Lett. B **356**, 398 (1995).

[16] M. Davier and A. Höcker, Phys. Lett. B **419**, 419 (1998).

ELECTROWEAK RESULTS FROM NUTEV

Robert B. Drucker
University of Oregon
Eugene, OR 97403

Representing the NuTeV Collaboration

ABSTRACT

We report our measurement of the weak mixing angle, $\sin^2\theta_W$, in neutrino–nucleon deep inelastic scattering in the NuTeV experiment at Fermilab. Using separate neutrino and antineutrino beams, we are able to use the Paschos-Wolfenstein relations to extract $\sin^2\theta_W$ with low systematic errors. Our preliminary result is $\sin^2\theta_W{}^{(on-shell)} \equiv 1 - m_W^2/m_Z^2 = 0.2253 \pm 0.0019(stat.) \pm 0.0010(syst.)$, which in the Standard Model is equivalent to a W mass measurement of 80.26 ± 0.11 GeV/c^2.

© 1998 by Robert B. Drucker.

1 Introduction

Neutrino scattering experiments have played an important role in the development of the electroweak standard model. Precision measurements of neutrino-nucleon (νN) neutral current (NC) to charged current (CC) ratios contributed to predictions of the W, Z, and top quark masses before these particles were directly observed. Today, in conjunction with the large samples of direct Z and W production from collider experiments, precision measurements from νN scattering can help constrain the mass of the Higgs boson. Neutrino–nucleon scattering results are also sensitive to a variety of non-Standard Model physics, including extra Z bosons, leptoquarks, quark compositeness, and neutrino oscillations.[1–3]

In the past, neutrino experiments have had a large common systematic uncertainty due to charm quark production.[4–6] Production of charm quarks from low-x sea quarks has large systematic uncertainties associated with the effective mass suppression, see Fig. 1. The NuTeV experiment ran in separate neutrino and antineutrino modes, which greatly reduces the sea quark uncertainties by using the quantity suggested by Paschos and Wolfenstein[7] for measuring the weak mixing angle:

$$R^- \equiv \frac{\sigma^{\nu N}_{NC} - \sigma^{\bar{\nu} N}_{NC}}{\sigma^{\nu N}_{CC} - \sigma^{\bar{\nu} N}_{CC}} = \frac{1}{1-r}\left(\frac{\sigma^{\nu N}_{NC}}{\sigma^{\nu N}_{CC}} - r\frac{\sigma^{\bar{\nu} N}_{NC}}{\sigma^{\bar{\nu} N}_{CC}}\right) \equiv \frac{R^\nu - rR^{\bar{\nu}}}{1-r}, \qquad (1)$$

where $r \equiv \sigma^{\bar{\nu} N}_{CC}/\sigma^{\nu N}_{CC} \approx 1/2$. Since R^- is a function of the difference between neutrino and antineutrino cross sections, effects from sea quarks tend to cancel.

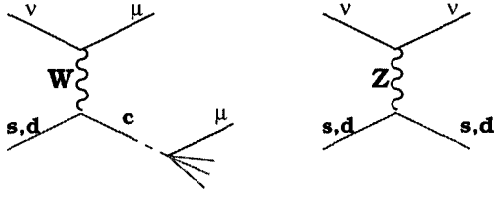


Figure 1: Feynman diagrams showing how a charm quark is produced in νN scattering. Note that this process only occurs in the charged current events; therefore, it alters the NC/CC ratio.

The $\sin^2\theta_W$ measured in νN scattering is approximately the on-shell mixing angle, $\sin^2\theta_W^{on-shell} \equiv 1 - m_W^2/m_Z^2$. Our result, therefore, can be compared with direct measurements of the mass of the W boson by collider experiments. This comparison is relatively insensitive to uncertainties in the top quark mass. A discrepancy between the NuTeV measurement of the W mass and other direct measurements would arise in the various Standard Model extensions mentioned above.

2 The NuTeV Experiment

The NuTeV detector consists of a 690 ton iron–scintillator calorimeter, followed by an iron toroidal spectrometer, see Fig. 2. The calorimeter consists of 168 steel plates, each 3 m × 3 m × 5.15 cm, instrumented with liquid scintillator counters placed after every two steel plates and drift chambers after every four plates. The calorimeter has a total length of 18 m and an average density of 4.2 g/cm^3. The muon spectrometer consists of three toroidal magnets with magnetic fields varying from 1.7 T near the center to 1.5 T near the edge. Drift chambers are positioned downstream of each of the toroids for tracking the muon trajectory and determining the muon momentum. The toroid spectrometer is not used directly in this analysis, but it is used to measure the neutrino flux from the beamline.

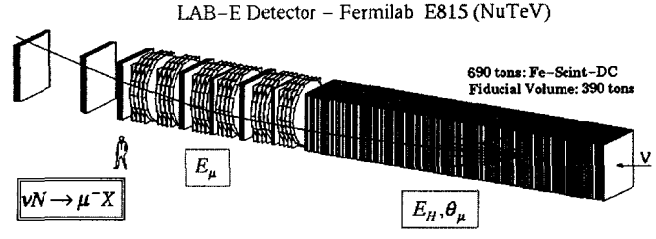


Figure 2: A drawing of the NuTeV detector, showing the calorimeter and the toroidal spectrometer. The figure includes a drawing of a charged current event showing the incoming neutrino, the hadronic shower in the calorimeter, and the muon going through the spectrometer.

The NuTeV beamline, known as the Sign Selected Quadrupole Triplet (SSQT), received 3×10^{18} protons at 800 GeV from the Fermilab Tevatron during the 1996–97 fixed-target run. Figure 3 shows a diagram of the SSQT elements. Pions and kaons of one charge are bent by a dipole immediately downstream of the BeO target. Neutral and wrong-sign particles were stopped in beam dumps. A quadrupole magnet triplet then focuses the beam into the 0.5 km decay region. The center of the decay region is 1.3 km from the NuTeV detector. Figure 4 shows

the predicted flux for the NuTeV neutrino beam. Measured $\bar{\nu}_\mu$ contamination in the ν_μ beam is less than 1/1000, and ν_μ contamination in the $\bar{\nu}_\mu$ beam is less than 1/500. Electron neutrinos in the beam account for 1.3% of the observed interactions in neutrino mode and 1.1% in antineutrino mode.

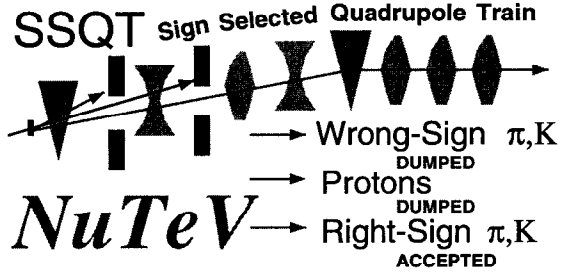


Figure 3: This diagram shows the elements of the NuTeV beamline SSQT elements. The protons enter from the left and hit the BeO target shown as a small rectangle in the picture. The wrong-sign and neutral secondaries are dumped in the beam dumps, and the right-sign particles are focused into the decay space.

Event length in the detector, defined as the number of scintillation counters with signal in them, is used to separate NC and CC events. Figure 5 shows examples of NC and CC candidate events in the NuTeV detector. CC events appear long in the detector because they contain a muon in the final state. NC events only have a short range hadronic shower. Experimentally, the ratio of NC to CC events is measured by the ratio of the number of short to long events.

Contamination in the beam from ν_e interactions alters the ratio of short to long events, because CC ν_e events do not contain a final state muon and so appear short. Figure 6 shows the length distributions separately for CC, NC, and electron neutrino events. It is critical to have a precise understanding of the number of electron neutrino events in the beam. Most of the ν_e contamination comes from $K^\pm_{e3}$ decay (93% in the ν_μ beam and 70% in the $\bar{\nu}_\mu$ beam). We use a Monte Carlo to estimate the number of electron neutrinos from $K^\pm_{e3}$ decay. The Monte Carlo was tuned to agree with the number of observed $\nu_\mu/\bar{\nu}_\mu$ events. Figure 7 shows the ν_μ energy distribution and a comparison with the Monte Carlo. The SSQT was very precisely aligned, which leads to a very well understood prediction of the number of $K^\pm_{e3}$ decays. The estimate for the number of $\nu_e/\bar{\nu}_e$ from $K^\pm_{e3}$ decays is accurate to approximately 1.5%, and is dominated by the

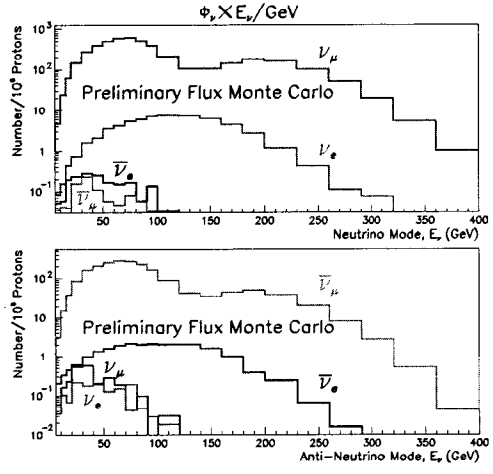


Figure 4: This plot shows the predicted number of νN events from the NuTeV beam as a function of neutrino energy. The upper plot is for neutrino mode and the lower plot is anti-neutrino mode. Note the very small contamination of "wrong-sign" neutrinos.

uncertainty in the measured $K^\pm_{e3}$ branching ratio. Uncertainties on neutral kaon production and decay are much larger, but the SSQT design greatly reduces the acceptance for neutrinos from neutral particle decays.

NuTeV had a separate calibration beamline that delivered beam during each 60 second accelerator cycle. Calibration data with muons, pions, and electrons were recorded throughout the entire run. This provided a large sample of calibration data as well as a constant monitor of detector performance and calibration.

3 Event Selection

Events used for this analysis are required to deposit at least 20 GeV in the calorimeter to ensure efficient triggering and good vertex reconstruction. The event vertex is required to be in the fiducial region of the detector. In the transverse dimensions, we require the vertex to be in the central two-thirds of the calorimeter, in order to reduce losses of final state particles out the sides of the detector. The longitudinal position of the vertex is required to be at least 0.4 m

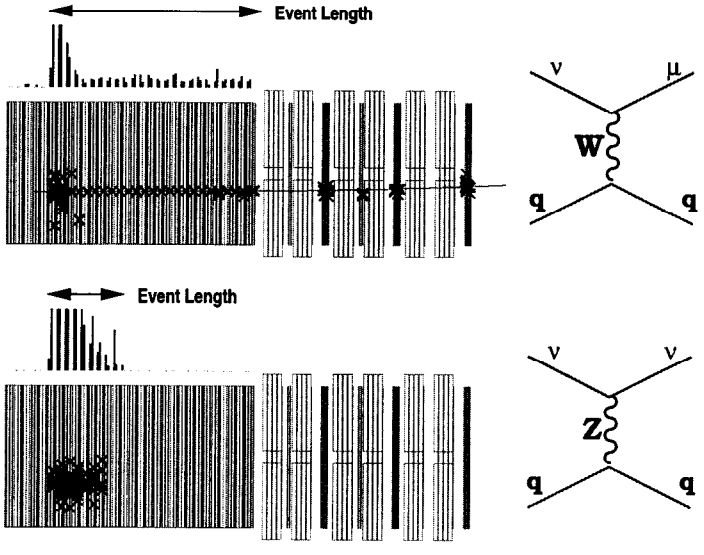


Figure 5: Event displays of a NuTeV charged current candidate and a neutral current candidate. The X's are hits in the tracking chambers and the vertical bars above the calorimeter show the energy recorded in each scintillation counter. This figure illustrates the definition of "long" and "short" events.

steel-equivalent from the upstream end of the calorimeter to ensure that the event was neutrino induced, and at least 2.4 m steel-equivalent from the downstream end to ensure that a minimal event length can be measured. Cosmic-ray events are a small background, and they are statistically subtracted using events taken during beam-off gates. Muon-induced events are also statistically subtracted using events recorded while the beam was on. After all cuts, 1.3 million neutrino and 0.3 million antineutrino events are observed.

4 Extraction of $\sin^2\theta_W$

Figure 8 shows the measured event length distributions for the neutrino and antineutrino samples. Long events, considered *CC* candidates, are those that penetrate more than 20 counters, or 2 m steel-equivalent. The ratio of short to long events in neutrino mode is $R^{\nu}_{meas.} = 0.4198 \pm 0.0008$, and in antineutrino mode $R^{\bar{\nu}}_{meas.} = 0.4215 \pm 0.0017$.

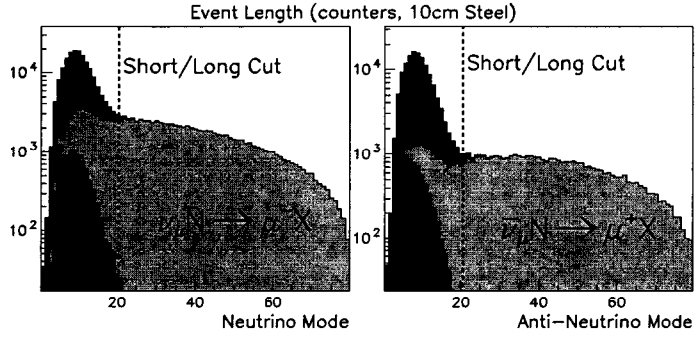


Figure 6: This figure shows the number of neutrino events as a function of event length. The separate contributions from charged current, neutral current, and electron neutrino events are shown. Note that the ν_e events only contribute to the short event sample.

$\sin^2\theta_W$ is extracted using a detailed detector Monte Carlo simulation. The Monte Carlo must carefully simulate all detector and cross-section effects that cause cross-talk between the long and short event samples.

Several detector effects must be considered to properly simulate the expected event length distribution. For *CC* events the event length is affected by the scintillation counter efficiency, noise, and active area. These properties were measured using neutrino data and muons from the calibration beam. Determining the event vertex position is critical to the length measurement. Vertex finding resolutions and biases were studied using a detailed GEANT simulation of the detector. Vertex position bias was also checked using the CCFR two-muon data sample, which allows a very precise independent vertex position measurement from the two muon tracks. It is also important to carefully model the length distribution of hadronic showers, since they determine the event length in the NC sample. The hadronic shower length was studied using calibration beam pions. As a check, shower lengths from a sample of calibration beam kaons were also studied, because neutrino-induced events contain a mixture of pion and kaon secondaries. No significant differences were observed between the pion and kaon shower lengths. Figure 9 shows the hadronic shower lengths from the calibration beam pions compared with the tuned Monte Carlo simulation. The Monte Carlo does an excellent job of reproducing the hadronic shower length.

The Monte Carlo used a leading-order cross-section model for neutrino-nucleon

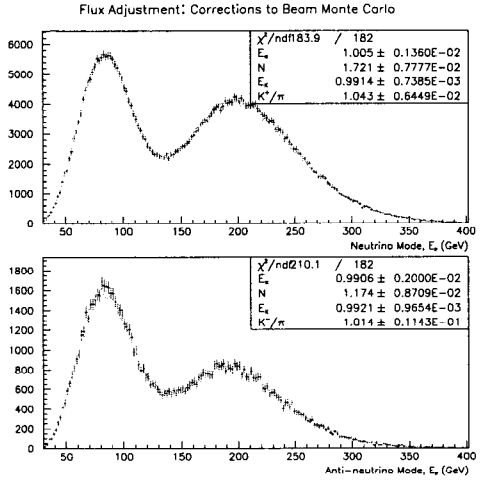


Figure 7: The NuTeV neutrino flux measured using a sample of charged current events. The plots show the number of neutrino events as a function of neutrino energy; the upper plot is neutrino mode and the lower plot is antineutrino mode. The peak near 90 GeV is due to pion decay and the peak near 200 GeV is from kaon decay. The solid lines show the Monte Carlo flux used in the analysis. The Monte Carlo was tuned slightly (within errors) to match the data.

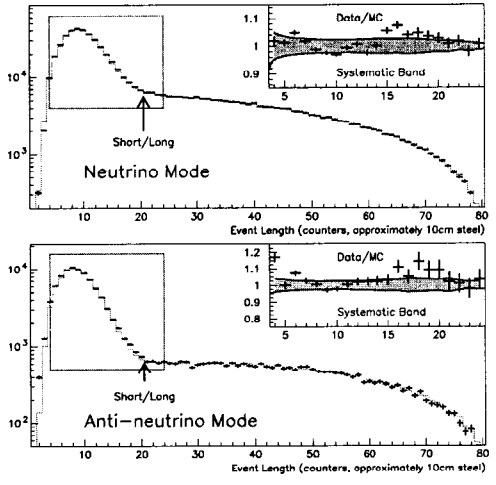


Figure 8: These plots show the number of events as a function of the event length. The solid curve is Monte Carlo, and shows excellent agreement with the data.

deep inelastic scattering, and also included neutrino-electron and quasi-elastic scattering. Leading-order parton distribution functions, as measured by the CCFR experiment[8] and fit with the Buras-Gaemers parameterization,[9] were used. CCFR was the predecessor to NuTeV and used the same calorimeter as NuTeV. These parton distribution functions were modified to include measured u/d valence asymmetries and sea quark asymmetries.[10,11] The strange sea distribution used was that measured by CCFR from its two-muon sample.[12] Mass suppression effects for heavy quark production are modeled using slow rescaling, with parameters taken from the CCFR two-muon data.

Electroweak and QED radiative corrections to the scattering cross sections are applied using a computer program supplied by Bardin.[13] The charm sea was included using CTEQ4L parton distribution functions,[14] and was assigned a 100% systematic. The longitudinal structure function, R_L, used was based on QCD predictions and data.[15] Higher-twist effects[16] (effects from the nuclear environment of the interaction) were included with a 100% systematic.

$\sin^2\theta_W$ was extracted from the data by varying the value of $\sin^2\theta_W$ used in the Monte Carlo until the predicted NC/CC ratios agreed with the data. The comparison with data was based on the quantity

$$\hat{R}^-_{meas.} \equiv R^\nu_{meas.} - \alpha R^\nu_{meas.},$$

where $\alpha = 0.5136$ was chosen, using the Monte Carlo, to minimize the systematic uncertainty due to charm quark production. Note that $\hat{R}^-_{meas.}$ is quite close to being R^- as defined in Eq. (1). Using $\hat{R}^-_{meas.}$, which is possible because of NuTeV's ability to run separate neutrino and antineutrino beams, greatly reduces many systematic uncertainties from sea quark effects and other correlated systematics. Figure 10 shows the $R^\nu_{meas.}$ and $R^{\bar{\nu}}_{meas.}$ distributions as a function of energy deposited in the calorimeter. Figure 11 shows the $\hat{R}^-$ distribution. The Monte Carlo agrees well with the data. Note that there appear to be slight systematic wiggles in the $R^\nu_{meas.}$ and $R^{\bar{\nu}}_{meas.}$ distributions, but they are correlated and cancel out in the $\hat{R}^-_{meas.}$ plot. This is exactly the reason for using $\hat{R}^-$.

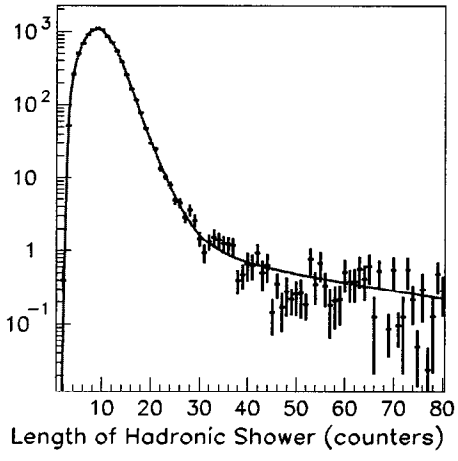


Figure 9: Length of hadronic showers in the NuTeV calorimeter. The points show the number of events as a function of event length for calibration beam pions. The solid curve is the Monte Carlo modeling of the event length. The agreement is excellent.

Table 1 shows a summary of the systematic uncertainties in the $\sin^2\theta_W$ measurement. The total error is dominated by statistical uncertainty. Note how small the charm mass effect error is with this R^- analysis.

5 Results

The preliminary result for $\sin^2\theta_W$ from NuTeV is

$$\begin{aligned}\sin^2\theta_W{}^{(\text{on-shell})} &= 0.2253 \pm 0.0019(stat.) \pm 0.0010(syst.) \\ &\quad -0.00142 \times \left(\frac{{M_{top}}^2 - (175\ \text{GeV})^2}{(100\ \text{GeV})^2}\right) \\ &\quad +0.00048 \times \log_e\left(\frac{M_{Higgs}}{150\ \text{GeV}}\right). \qquad (2)\end{aligned}$$

This result represents a factor of three improvement in systematic uncertainty over the previous best νN result. The small residual dependence of our result on M_{top} and M_{Higgs} comes from the leading terms in the electroweak radiative corrections.[13] Figure 12 shows the NuTeV result in the M_W–M_{top} plane. Since

Table 1: This table shows a summary of all the NuTeV systematic uncertainties. Note that the uncertainty in the measurement is dominated by the statistical error. Note also the very small uncertainty associated with the charm mass.

Source of Uncertainty		$\delta \sin^2\theta_W$
Statistics:	Data	0.00188
	Monte Carlo	0.00028
Total statistics		0.00190
$\nu_e/\bar{\nu}_e$ *Flux:*	$K^\pm$ (1.1%)	0.00024
	Other sources of ν_e's	0.00048
Energy measurement:	Calibrations (0.5%)	0.00043
	Muon energy deposition (3%)	0.00004
	Energy resolution	0.00004
Event length:	Hadron shower	0.00015
	Longitudinal vertex determination	0.00015
	Counter edge location	0.00010
	Counter efficiency and noise	0.00016
Total experimental systematics		0.00075
Sea quarks:	Strange Sea	0.00034
	V_{cd}	0.00004
	Charm sea	0.00009
	Charm mass	0.00009
Other $\nu/\bar{\nu}$ cross-section differences: $\sigma^\nu/\sigma^{\bar{\nu}}$		0.00023
	Nonisoscalar target	0.00013
	Radiative corrections	0.00051
Non-QPM cross section:	Higher twist	0.00013
	Longitudinal structure function	0.00007
Total physics model		0.00070
Total uncertainty		0.0022

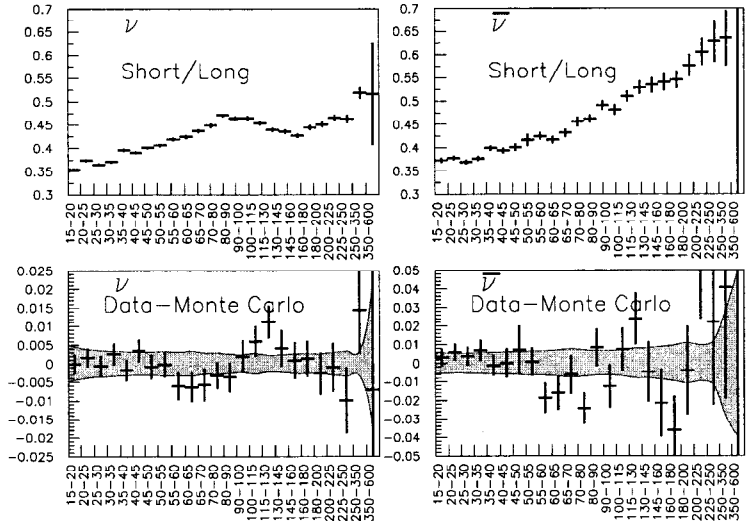


Figure 10: The upper plots show the $R^{\nu}_{meas.}$ and $R^{\bar{\nu}}_{meas.}$ distributions from the NuTeV data as a function of hadronic shower energy. The lower plots show the difference between the data and the Monte Carlo predictions. The band shows the systematic uncertainty. The Monte Carlo is in good agreement with the data. There are slight correlated wiggles in the data, within the errors. The advantage of using R^- is that these correlated effects subtract out.

$\sin^2\theta_W{}^{(\text{on-shell})} \equiv 1 - M_W^2/M_Z^2$, this result is equivalent to

$$\begin{aligned} M_W &= 80.26 \pm 0.10(stat.) \pm 0.05(syst.) \\ &\quad +0.073 \times \left(\frac{M_{top}{}^2 - (175\ \text{GeV})^2}{(100\ \text{GeV})^2} \right) \\ &\quad -0.025 \times \log_e \left(\frac{M_{Higgs}}{150\ \text{GeV}} \right). \end{aligned} \tag{3}$$

A comparison of this result with direct measurements of M_W is shown in Fig. 13. The NuTeV result is comparable in precision to the current world direct measurements of the W boson mass.

5.1 Model Independent Couplings

The NuTeV result can be expressed in a more model independent way. The above result has some model dependence in correcting from our measured $\sin^2\theta_W$ to

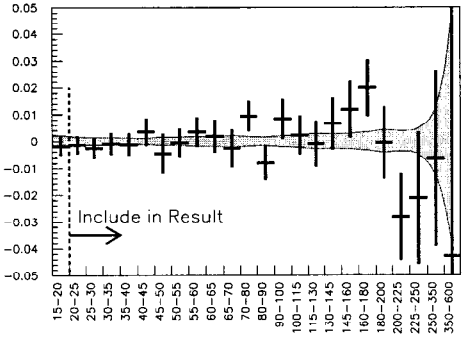


Figure 11: This plot shows the difference between data and Monte Carlo for $\hat{R}^-_{meas.}$ as a function of hadronic shower energy. The band shows the systematic uncertainty. The data and Monte Carlo show excellent agreement. Note that the slight wiggles seen in the $R^{\nu}_{meas.}$ distributions have no effect on $\hat{R}^-_{meas.}$.

the on-shell value. R^- can be expressed in terms of right- and left-handed quark couplings:

$$\begin{aligned} R^- &= \frac{\sigma^{\nu}_{NC} - \sigma^{\bar{\nu}}_{NC}}{\sigma^{\nu}_{CC} - \sigma^{\bar{\nu}}_{CC}} \\ &= \left(\frac{1}{2} - \sin^2\theta_W \right) \\ &= u_L^2 + d_L^2 - u_R^2 - d_R^2. \end{aligned}$$

The variable used in the NuTeV Monte Carlo extraction, $\hat{R}^-_{meas.}$, can be expressed in terms of the quark couplings

$$\begin{aligned} 0.8587u_L^2 + 0.8828d_L^2 - 1.1657u_R^2 - 1.2288d_R^2 &= 0.2277 \pm 0.0022 \\ &= 0.4530 - \sin^2\theta_W. \end{aligned}$$

Figure 14 shows the regions in the g_L^2–g_R^2 plane that the NuTeV measurement allows, where $g^2_{L,R} \equiv u^2_{L,R} + d^2_{L,R}$. In order to present the result on the two-dimensional plane, it is assumed that $\delta^2_{L,R} \equiv u^2_{L,R} - d^2_{L,R}$ take on the Standard Model values. Also shown in the figure is the region measured by CCFR using R^{ν}. The CCFR measurement is less precise, but when combined with NuTeV it gives a small allowed region. In the future, the NuTeV data will be used for both

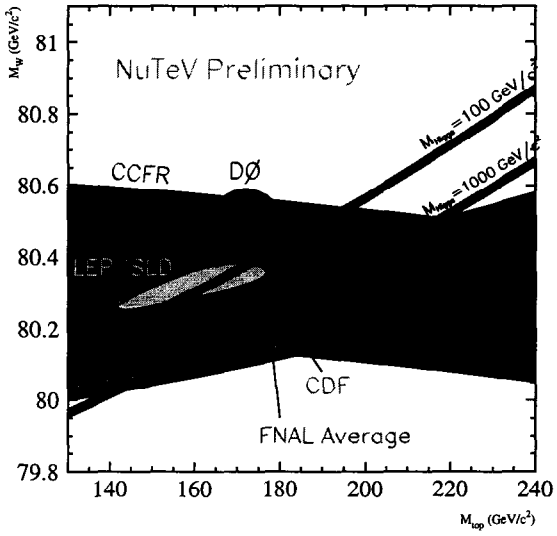


Figure 12: This plot shows the NuTeV result in the M_W–M_{top} plane compared with other world measurements. The light band is the NuTeV result. The dark band is the CCFR result. The center ellipse shows the combined Fermilab results from NuTeV, CCFR, DØ, and CDF. The Standard Model predictions for $M_{\rm Higgs}$ = 100 GeV/c^2 and 1000 GeV/c^2 are shown as narrow bands. The Fermilab and LEP/SLD results tend to slightly favor a light Higgs mass. Note that the NuTeV result depends on the Higgs mass. This dependence is small, and is shown by the side-bands drawn on the NuTeV result which show how much the band would shift in the 100–1000 GeV/c^2 Higgs mass range.

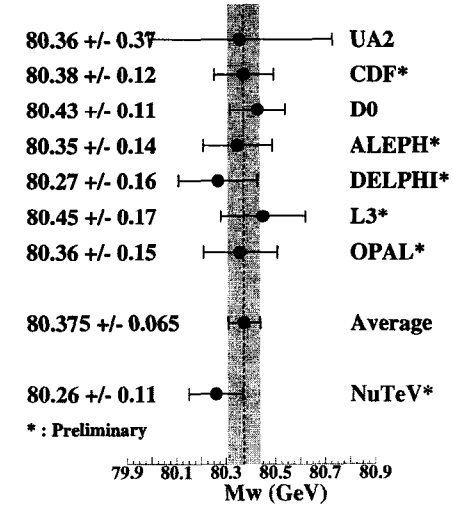


Figure 13: The NuTeV result for the W boson mass, compared with other direct measurements. The NuTeV result is currently the most precise direct measurement of the W mass. The NuTeV result is consistent with the other results (unfortunately).

parameters and will reduce the allowed region even further. The plot also shows the Standard Model prediction, which is in good agreement with the data, and therefore, shows no evidence of non-Standard Model physics.

6 Conclusions

NuTeV ran in the Fermilab 1996-1997 fixed-target run and took 3×10^{18} protons on target. By running in separate neutrino and antineutrino modes, we were able to make use of the Paschos-Wolfenstein R^- variable to measure the on-shell weak mixing angle, $\sin^2\theta_W^{on-shell}$. We have presented a preliminary result that is a factor of two more precise than previous νN scattering experiments. Our result

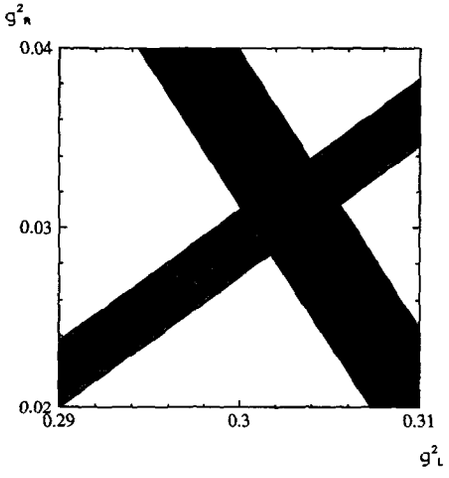


Figure 14: Areas in g_L^2-g_R^2 space allowed by the NuTeV and CCFR model-independent measurements of the quark couplings. Since NuTeV measures R^- and CCFR measures R^ν, the allowed regions for the two experiments are nearly perpendicular. The Standard Model prediction is shown as a point.

is equivalent to a measurement of the W mass and is consistent with other direct measurements of the W mass.

Acknowledgments

This research was supported by the U.S. Department of Energy and the National Science Foundation. We would also like to thank the staff of Fermilab for their substantial contributions to the construction and support of this experiment during the 1996–97 fixed-target run.

References

[1] P. Langacker *et al.*, Rev. Mod. Phys. **64**, 87 (1991).

[2] K.S. McFarland *et al.*, Phys. Rev. Lett **75**, 3993 (1995).

[3] D. Naples *et al.*, Phys. Rev. D **59**, 31101 (1999); E-Print Archive: hep-ex/9809023.

[4] K.S. McFarland, *et al.*, Eur. Phys. J. C **1**, 509 (1998).

[5] A. Blondel *et al.*, Z. Phys. C **45**, 361 (1990).

[6] J. Allaby *et al.*, Z. Phys. C **36**, 611 (1985).

[7] E. A. Paschos and L. Wolfenstein, Phys. Rev. D **7**, 91 (1973).

[8] W. G. Seligman *et al.*, Phys. Rev. Lett. **79**, 1213 (1997).

[9] A. J. Buras and K. J. F. Gaemers, Nucl. Phys. B **132**, 249 (1978).

[10] M. Arneodo *et al.*, Nucl. Phys. B **487**, 3 (1997).

[11] E. A. Hawker *et al.*, Phys. Rev. Lett. 80, 3715 (1998).

[12] S. A. Rabinowitz *et al.*, Phys. Rev. Lett. 70, 134 (1993).

[13] D. Yu. Bardin and V. A. Dokuchaeva, JINR-E2-86-260 (1986); and private communication.

[14] CTEQ Collaboration, Phys. Rev. D **55**, 1280 (1997).

[15] L. W. Whitlow, SLAC-Report-357, 109 (1990).

[16] J. Pumplin, Phys. Rev. D **64**, 2751 (1990). $S_0 \leq 2$ GeV^2 is allowed by data summarized in M. Virchaux and A. Milsztajn, Phys. Lett. B **274**, 221 (1992).

ELECTROWEAK PHYSICS RESULTS FROM THE SLD

Frank E. Taylor*
Laboratory for Nuclear Science
Massachusetts Institute of Technology
Cambridge, MA 02139

Representing the SLD Collaboration

ABSTRACT

Data taken by the SLD Collaboration at the Stanford Linear Collider (SLC) on $e^+e^- \rightarrow Z^0$, obtained with a polarized electron beam, have enabled many incisive tests of the electroweak sector of the Standard Model to be performed. We discuss our recent determinations of $\sin^2\theta_W$, derived from the total cross-section asymmetry, and the quark final state asymmetries, A_s, A_c, A_b, and branching ratios, R_c and R_b. Aspects of the precision tests of the standard electroweak theory, involving radiative corrections, are described. Limits on the mass of the Higgs particle are given.

* Work supported in part by the Department of Energy.

© 1998 by Frank E. Taylor.

1 Introduction

The theoretical description of electron-positron annihilation into the Z^0 vector boson is rich with experimentally testable predictions. Since the first data taken at LEP and SLC in 1990, experimental tests have been made of the Standard Model (SM) with ever increasing statistical and systematic precision. For some time, the data have been of sufficient quality to probe the radiative corrections related to the top quark and Higgs particle. To date, no experimental test of the SM has shown a persistent and confirmed deviation from theory. Rather than spectacular anomalies discovered, the experimental program at LEP-I and SLC is characterized as one of ever increasing refinement and verification of theory. At this time, the Z^0 pole data in conjunction with measurements at the FNAL collider of the W^+ boson and top quark masses are beginning to approach the untested Higgs sector of the theory.

The SLD-SLC facility[1] brings a number of unique features to the exploration of the SM at the Z^0 pole. The SLC produces Z^0s with polarized electrons colliding with unpolarized positrons in an exquisitely small intersection region that is roughly $1.6\ \mu m^2$ (700 μm) transverse (longitudinal) to the beam. The electron polarization provided by the SLC allows a quite precise measurement of $\sin^2\theta_W$ to be made with a statistical "boost" proportional to $|P_e|^2/A_e^2$, where P_e is the electron polarization and A_e is the axial coupling of the electron to the Z^0, by using the total cross section left-right asymmetry, A_{LR}. Tests of the SM can be performed on measurements of the properties of the final state, such as a study of the forward-backward left-right cross section asymmetry for b-quark final states. The SLD is equipped with a pixel vertex detector, based on the CCD technology, which is used to identify b- and c-quark final states and a Cherenkov ring imaging detector (CRID) for particle ID. The vertex detector has been recently upgraded, permitting better vertex resolution and coverage.[2]

Table 1 is a summary of the data taken at the SLD. Noteworthy is the outstanding accelerator operation during 1997-98, when the SLC delivered up to 20K polarized Z^0 events in a week. The total number of polarized Z^0 events is 560K with an "event-weighted" polarization of ~ 72%.

Table 1: History of SLD data taking.

Year	Purpose	Events	e^- Polarization
1992	physics	10K	22%
1993	physics	50K	63%
1994-5	physics	100K	77%
1996	physics & VXD3 commissioning	50K	77%
1997-8	physics	350K	73%

2 Basics of the Standard Model

For leptons at lowest order, the $SU(2)_L$ X U(1) standard model of Glashow-Weinberg-Salam[3] has the following Feynman rules for the neutral current (NC) and charged current (CC) fermion-gauge boson interactions, written in terms of the members of the first generation:

(a) $\gamma - e^+ e^-$ interactions (NC_γ):

$$-iQ_e \bar{e}\,\gamma_\lambda e, \qquad (1a)$$

(b) $Z^0 - f\bar{f}$ (f = e, ν) interactions (NC_Z):

$$\frac{-i}{\sqrt{2}}\left(\frac{G_F M_Z^2}{\sqrt{2}}\right)^{1/2} \bar{f}\gamma_\lambda(g_{V_f} - g_{A_f}\gamma_5)f, \tag{1b}$$

(c) $W^{\pm} - e\nu$ interactions (CC):

$$-i\left(\frac{G_F M_w^2}{\sqrt{2}}\right)^{1/2} \bar{\nu}\gamma_\lambda(1-\gamma_s)\,e, \tag{1c}$$

where G_F is the Fermi coupling constant, M_z the Z-boson mass, and g_v, g_A are NC vector and axial vector coupling constants, respectively. The two NC interactions, involving the exchange of γ or Z, are characterized by coupling strengths e and g', respectively, whereas the charged current (exchange of $W^{\pm}$) has a coupling strength $g \sim \sqrt{G_F}\, M_W$. The coupling constants involving the exchange of massive vector bosons ($W^{\pm}$ and Z^0) are related by the Weinberg mixing angle, θ_w, with $\tan\theta_w = g'/g$ and the electromagnetic neutral current is "unified" with the $W^{\pm}$ and Z interactions by $e = g'g/\sqrt{g'^2 + g^2}$. The NC interaction for a massive fermion pair coupled to the Z^0 has generally unequal right and left terms couplings g_{Re} and g_{Le}, respectively. The CC interaction is pure "V-A."

By the Higgs mechanism the vector boson masses, M_w and M_z, acquire mass of magnitude determined by the vacuum expectation value of the Higgs field, given by $\langle v \rangle = (\sqrt{2}G_F)^{-1/2} = 246$ GeV, and the coupling constants g and g'. The Fermi coupling constant, G_F, is the only "dimensioned number" in the theory. It is conventional to express the boson masses in terms of the experimentally determined constants α_{em}, $\sin^2\theta_w$ and G_μ, where α_{em} is the fine structure constant and is determined by the quantum hall effect and Thomson scattering at low energies, $\sin^2\theta_w$ is measured by a number of experiments, and $G_\mu = G_F$ by precision muon decay data. In this convention we have:

$$M_z^2 = \frac{\pi\alpha_{em}}{\sqrt{2}G_\mu \sin^2\theta_w \cos^2\theta_w \xi_Z}, \tag{2a}$$

$$M_w^2 = \frac{M_z^2}{2}\left(1 + \sqrt{1 - \frac{4\pi\pi_{em}}{\sqrt{2G_F}\xi_w M_z^2}}\right), \tag{2b}$$

where ζ_z and ζ_w are radiative correction terms, of order one, to be described later. The NC vector and axial vector couplings (or right- and left-handed couplings) for fermion f are given by:

$$g_{Vf} = (g_{Lf} + g_{Rf}) = I^3_{Lf} - 2Q_f \sin^2\theta_w, \tag{3a}$$

$$g_{Af} = (g_{Lf} - g_{Rf}) = I^3_{Lf}, \tag{3b}$$

respectively, where I^3_{Lf} is the third component of the weak isospin of fermion f, and Q_f is the corresponding electric charge. Since most of the precision electroweak data are at the Z pole, it is conventional to take $\sin^2\theta_w^{eff}$ to be the *effective value* at the pole,[4] thereby absorbing the radiation corrections in the gauge boson propagators. As such,

$$\sin^2\theta_w^{effective} = (1 - g_{Vf}/g_{Af})/4|Q_f|, \tag{4}$$

where g_{Vf} and g_{Af} are the *effective* vector and axial vector couplings of fermion f at the Z^0 pole.

Two important measurables at the Z^0-pole dependent on the coupling of $Z^0 \to f\bar{f}$ are the partial decay width given by

$$\Gamma_{f\bar{f}} = \frac{\sqrt{2}\, G_\mu\, M_Z^3}{12\pi} N_c \left(g_{Vf}^2 + g_{Af}^2\right) \zeta_{ff}, \tag{5a}$$

where N_c is the number of color degrees of freedom (N_c = 1 for leptons and 3 for quarks) and ζ_{ff} is a radiative correction term, of order one, and the coupling constant asymmetry, defined as:

$$A_f = \frac{g_{Lf}^2 - g_{Rf}^2}{g_{Lf}^2 + g_{Rf}^2} = \frac{2\, g_{Vf}\, g_{Af}}{g_{Vf}^2 + g_{Af}^2} = \frac{2\, g_{Vf} / g_{Af}}{1 + (g_{Vf} / g_{Af})^2} \quad . \tag{5b}$$

The coupling asymmetry for charged leptons is especially sensitive to the effective mixing angle, θ_w^{eff}. Table 2 shows the couplings, asymmetry, and sensitivity of the asymmetry to the effective mixing angle for the first generation quarks and leptons.

Table 2: First generation coupling constants and sensitivity of A_f to $\sin^2\theta_w^{eff}$ for $\sin^2\theta_w^{eff} = 0.23$.

f	I_{Lf}^3	Q_f	g_{Af}	G_{Vf}	G_{Lf}	G_{Rf}	A_f	$\frac{\delta A_f}{\delta \sin^2\theta_w}$
ν	1/2	0	1/2	1/2	1/2	0	1	0
e	-1/2	-1	-1/2	-0.04	-0.27	0.23	0.16	-7.9
u	1/2	2/3	1/2	0.19	0.35	-0.16	0.69	-3.5
d	-1/2	-1/3	-1/2	-0.35	-0.43	0.08	0.94	-0.6

Equations 1, 2, and 3, with radiative corrections, form the basis of precision tests of the electroweak sector of the SM. Equations 4, 5 and Table 2 furnish the tools and predictions of the SM. Note that the charged lepton couplings (e, μ, τ) are dominated by the axial term and have the greatest sensitivity for the determination of $\sin^2\theta_w$ and that the down quark couplings (d, s, b) are similar, but opposite sign, to the neutrino couplings and have a small sensitivity to $\sin^2\theta_w$. As a consequence, the down-like quark couplings lead to their being strongly polarized left-handed in Z^0 decay.

Some of the questions we have addressed to the data are:

- Is the Lorentz structure of the electroweak interaction described correctly by Eq. 1?
- Are the couplings for all generations and weak isospin states given by the theory isospin assignments and a *universal value* of $\sin^2\theta_w^{eff}$?
- Are the physical gauge boson masses related by Eq. 2?
- Do the higher order radiative corrections result in better consistency of the data?

2.1 Measurement of Weak Couplings

One of the fundamental tests of the SM afforded by data at the Z^0 pole is the check of the $SU(2)_L$ x U(1) gauge group structure of the theory. It is of interest to see if Eqs. 3, 4, and 5 and the weak isospin and charge assignments of Table 2 are operative for all generations with the *same value* of $\sin^2\theta_w^{eff}$. This test is sensitive to putative Z^0s, indicating a new gauge group, or, as yet unobserved, new particles which could lead to non-universal radiative corrections for different generations.

In the context of the SM the differential cross section for $e^+e^- \to Z^0 \to f\bar{f}$ in the case, for example, of a left-handed electron producing a left-handed fermion in the final state, has the form

$$\sigma(\cos\theta) \sim g_{Le}^2\, g_{Lf}^2 \left(1 + \cos\theta\right)^2, \tag{8}$$

where θ is the angle of the final fermion with respect to the electron beam direction and g_{Le} and g_{Lf} are the left-handed electron and final fermion couplings to Z^0, respectively. Averaging over final fermion helicities for an electron beam with a polarization of P_e, the differential cross section has the form

$$\frac{d\sigma}{d\cos\theta} \sim (1-P_e A_e)\left(1+\cos\theta^2\right)+2\cos\theta\,(A_e - P_e)\,A_f \qquad (9)$$

Here, the electron and final fermion coupling asymmetry parameters A_e, A_f are given by Eq. 5b.

Figure 1 shows the differential cross section for the production of $(b\bar{b})$ for three electron beam polarization cases. Note that for left-handed incident electrons the b-quark tends to go in the forward hemisphere. This large production asymmetry is exploited by the SLD in b-quark studies in tests of the SM through electroweak coupling asymmetries and b-quark flavor mixing.

The basis of the measured asymmetries at LEP and SLC is given by Eqs. 10 (a-d). At LEP the forward-backward asymmetry ($P_e = 0$), given by

$$A_{FB}^f = \frac{\sigma_F^f - \sigma_B^f}{\sigma_F^f + \sigma_B^f} = \frac{3}{4} A_e A_f, \qquad (10a)$$

is a measure of the product of coupling asymmetries A_e, A_f. The forward-backward asymmetry is exploited at LEP to obtain lepton and quark couplings.

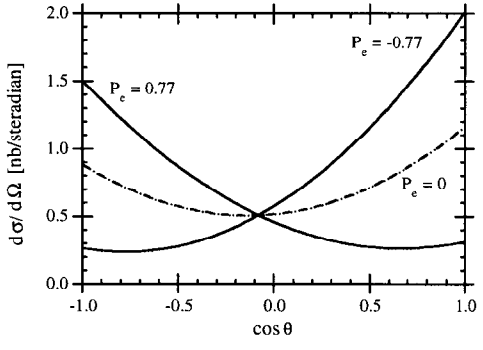


Figure 1: The differential cross section for $e^+e^- \rightarrow Z^0 \rightarrow b\bar{b}$ is shown for $P_e = \pm 0.77$, corresponding to polarized electron beams at SLC and $P_e = 0$, the operating condition for LEP.

The left-right total cross section asymmetry, exploited by the SLD Collaboration,[5] is given by

$$A_{LR} = \frac{\sigma_L - \sigma_R}{\sigma_L + \sigma_R} = |P_e| A_e, \qquad (10b)$$

where σ_{LR} are the $e^+e^- \rightarrow Z^0$ total cross sections for incident L, R electron polarizations and is the *centerpiece* of our precision electroweak tests at the SLC. Knowledge of the electron polarization $|P_e|$ allows the electron Z^0 coupling asymmetry, A_e, to be directly and quite accurately measured. The polarized forward-backward asymmetry, given by

$$A_{FB_LR}^f = \frac{\left(\sigma_{FL}^f - \sigma_{BL}^f\right) - \left(\sigma_{FR}^f - \sigma_{BR}^f\right)}{\sigma_{FL}^f + \sigma_{BL}^f + \sigma_{FR}^f + \sigma_{BR}^f} = \frac{3}{4} |P_e| A_f, \qquad (10c)$$

allows a *direct determination* of the final state fermion coupling asymmetry parameters, A_f. This asymmetry is used at the SLC to measure A_b and A_c.

The τ-lepton polarization is measured both at LEP and SLC by, for example, the energy distribution of the π in $\tau^- \to \pi^- \nu_\tau$ decay. In the SM the τ-polarization, $P_\tau(\cos\theta)$, is dependent on both A_τ and A_e by

$$P_\tau(z) = -\frac{(1+z^2)(1+A_eP_e)A_\tau + 2(A_e+P_e)z}{(1+z^2)(1+A_eP_e) + 2(A_e+P_e)A_\tau z}, \qquad (10d)$$

where $z = \cos\theta$ and θ is the angle between the τ-lepton and the electron beam direction and P_e is the incident electron beam polarization ($P_e = 0$ at LEP). Both LEP and SLD have determined A_e and A_τ by this method. The SLD collaboration exploits the polarized incident electron beam[6] to measure an asymmetry enhanced by a large factor (~3.4 at large $|\cos\theta|$).

3 Lepton Asymmetries

Lepton asymmetries derived from pure leptonic final states $Z^0 \to l^+l^-$ are the easiest to interpret as precision tests of the SM. Nevertheless, the b-quark asymmetry, exploited by the LEP experiments, offers a high statistical precision method within the context of the SM for determining the electron asymmetry, A_e.

3.1 Pure Lepton Modes

The single most precise determination of the electron asymmetry and thus of $\sin^2\theta_w^{eff}$ comes from the measurement at the SLD of the total e^+e^- cross-section asymmetry A_{LR}, defined by Eq. 10b. The measurement directly determines A_e, the electron coupling asymmetry, with only small corrections and is only possible with incident polarized beams. Events for the A_{LR} analysis are selected to be "hadronic" and not a beam background, 2γ, or Bhabha event. Small experimental corrections of order $\leq 0.3\%$ are made for backgrounds, experimental asymmetries and direct channel γ exchange, and Z^0 γ interference. The systematic error of the measurement of A_{LR} is dominated by the electron beam polarization uncertainty. The electron beam polarization is directly measured by means of a Compton polarimeter[1] throughout data taking, with running averages updated at least as frequently as every hour, giving rise to an estimated systematic error presently at the level of $\Delta P_e/P_e \sim 0.7\%$ for the '94-'95 period and ~ 1.1% for '97-'98 data.

The lepton coupling asymmetry parameters A_L, for L = e, μ, and τ have been directly measured[7] using the corresponding *lepton final states* by the FB-LR asymmetry of Eq. 10c. The resulting cross sections are shown in Fig. 2. Although this test is of limited statistics (e, μ, τ: 9.4K, 7.6K, 7.1K events), is a direct measurement of A_μ and A_τ, a confirmation of A_e using the electron final state, and a probe of lepton universality. A maximum likelihood method is used to determine the coupling asymmetries which takes into account all known cross section terms (ten terms for the e^+e^- final state), backgrounds, and detection efficiencies.

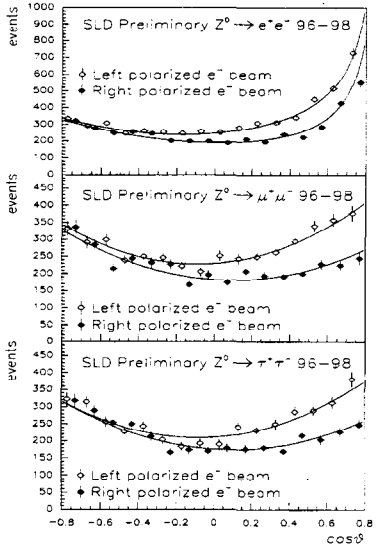


Figure 2: The differential cross sections measured at the SLD for $e^+e^- \rightarrow Z^0 \rightarrow e^+e^-$, $\mu^+\mu^-$, and $\tau^+\tau^-$ are shown.

At the end of the '97-'98 run, several important checks of the SLC parameters were performed which are important to certify the small corrections of γ exchange and γ–Z^0 interference and to make certain that the SLC positrons are not polarized. In order to check the energy calibration of the electron and positron spectrometers, which follow the beam interaction point, the Z^0 cross section peak was scanned and matched to the Z^0 mass determined at LEP-I. It was determined that the energy scale is known to about 40 MeV, which results in a negligible correction to $\sin^2\theta_w^{eff}$ (0.00011 ± 0.00009). In addition, a putative positron polarization was sought by directly measuring the positron beam polarization by means of a Møller polarimeter located in End Station A and found to be consistent with zero ($P_{e+} = -0.02 \pm 0.07\%$). Finally, the Compton polarimeter, which measures the electron polarization, was questioned but found to be in good agreement with measurements of the energy asymmetry of back-scattered Compton gammas detected by a threshold Cherenkov gas counter and a quartz fiber- tungsten radiator calorimeter.

In Table 3 we summarize the world's data on lepton coupling asymmetries[7,8] (both SLD[7] and LEP[8] data are preliminary).

Table 3: Lepton Coupling Asymmetries.

Quantity	Asymmetry
SLD: A_{LR}	0.1510 ± 0.0025
SLD: A_e	0.1504 ± 0.0072
SLD: A_μ	0.120 ± 0.019
SLD: A_τ	0.142 ± 0.019
SLD: $\langle A_{e,\mu,\tau} \rangle$	0.1459 ± 0.0063
SLD: A_{LR}, $\langle A_{e,\mu,\tau} \rangle$	0.1503 ± 0.0023
LEP: A_{FB} (l), $A_{e,\mu,\tau}$	0.1469 ± 0.0027
SLD+LEP	0.1489 ± 0.0018

We note that the LEP-I values and those from the SLC-SLD are in good agreement with the resulting value of leptonic asymmetries differing by only ~ 1 σ.

3.2 A_e from b-Quark Asymmetry

The experiments at LEP have exploited $Z^0 \to b\bar{b}$ data to determine A_{FB}^{0b}, the b-quark asymmetry at the pole, and hence A_e within the context of the SM (see Eq. 10a). The method rests on the assumption that A_b, the asymmetry parameter of the b-quark coupling, agrees with the SM prediction since what is actually measured in A_{FB}^{0b} is the product of A_b and A_e. Even within the context of the SM, important corrections are needed before an accurate value of A_e can be extracted from A_{FB}^{0b}. The c-quark background can dilute the b-asymmetry since this background enters with a smaller asymmetry and opposite electric charge. The jet charge method is especially sensitive to this correction. QCD effects, which tend to smear the direction of the final state quarks and make the measured asymmetry smaller, and $B-\bar{B}$ flavor mixing are two other corrections, which must be included. The resulting A_{FB}^{0b} measurement from LEP is 0.0991 ± 0.0021. Using the SM value of $A_b = 0.935$, the corresponding value of $A_e = 0.14145 \pm 0.0030$, which is 2.1 σ different from the pure lepton-based determinations above.[7]

4 Determination of $\sin^2\theta_w^{eff}$

The value of $\sin^2\theta_w^{eff}$ can be determined through Eqs. 3 and 5b, and the results are given in Table 4. The right-most column indicates the χ^2/DoF of the LEP measurements alone and of the LEP and SLC preliminary measurements combined. We take these data to be in generally good agreement, although we note that the pure leptonic-derived values of $\sin^2\theta_w^{eff}$ measured at the SLC and LEP are smaller than the value derived from A_{FB}^{0b} at LEP.

Table 4: Effective mixing angle at the Z^0 pole.

Measurement	$\sin^2\theta_w^{eff}$	χ^2/DoF
SLC:A_{LR},<A_e,μ,τ>	0.23110 ± 0.00029	
LEP: A_{FB}^{0l}	0.23117± 0.00054	
LEP: A_τ	0.23202 ± 0.00057	
LEP: A_e	0.23141± 0.00065	
LEP leptonic	0.2316 ± 0.0004	1.2/2
SLC+LEP leptonic	0.23127 ± 0.00023	
LEP: A_{FB}^{0b}	0.23223 ± 0.00038	
LEP: A_{FB}^{0c}	0.2320 ± 0.0010	
LEP: <Q_{FB}>	0.2321± 0.0010	
LEP average	0.23187 ± 0.00024	3.2/5
LEP+SLC average	0.23156 ± 0.00018	7.3/6 all

The χ^2/DoF of all data is 7.3/6 (~ 32 % CL), whereas for only lepton-based measurements χ^2/DoF = 3.7/3 (~ 35 % CL)—both in reasonable agreement. Later, we make an observation on the nature of the $\sin^2\theta_w^{eff}$ value derived from A_{FB}^{0b} at LEP and the experimental value of A_b.

5 Standard Model Tests with s, c, and b Quarks

It is interesting to see if the coupling constants g_V and g_A for the s, c, and b quarks follow the predictions of the SM. Two variables are accessible for this test of the SM: the coupling constant asymmetry and the branching ratio. The coupling constant asymmetry measures the combination $A_f = 2g_{Vf}/g_{Af}/(1+(g_{Vf}/g_{Af})^2)$ and the branching

ratio for Z^0 to decay into a quark favor, f, determines the combination $R_f = \Gamma_f / \Gamma_{had} \sim (g_{Vf}^2 + g_{Af}^2)$. To check theory, the vector and axial vector couplings are taken from the average value of $\sin^2\theta_w^{eff}$ derived from global fits, where the overall consistency of the data is checked. Or more strictly, the value of $\sin^2\theta_w^{eff}$ is taken from purely lepton measurements, thereby probing the universality of $\sin^2\theta_w^{eff}$ across lepton-quark families. In the case of the $Z^0 \to b\bar{b}$ vertex (also $Z^0 \to s\bar{s}$), since $g_L^2 \sim 30\ g_R^2$, R_b has a large sensitivity to the *left-handed* coupling, such as vertex corrections involving $W^{\pm}$, $t^{\pm}$ exchange, whereas A_b has a greater sensitivity to the *right-handed* components. We find the following sensitivities for $\sin^2\theta_w^{eff} \approx 0.23$

$$\frac{\delta R_b}{R_b} \sim -3.6\ \delta g_{Lb} + 0.7\ \delta g_{Rb}\ , \qquad (11a)$$

$$\frac{\delta A_b}{A_b} \sim -0.3\ \delta g_{Lb} + 1.7\ \delta g_{Rb}\ . \qquad (11b)$$

Thus, R_b and A_b are complementary measurements and probe different sectors of the theory.

The SLD brings a number of experimental tools to the study of the electroweak couplings of s, c, and b quarks. The pixel vertex detector can identify displaced vertices from the IP as a signature for b and c quark final states, and the Cherenkov Ring Imaging Detector (CRID) can identify $\pi^{\pm}$, $K^{\pm}$, $p^{\pm}$ particles in the final state. Various discriminants are used to isolate s-, c-, and b-quark events, which result for example in a hemisphere b-tag efficiency and purity in the range of 45% and 99%, respectively.

The most powerful test of the SM for quark final states have come from a study of the b-system. Several factors conspire to allow this.

(1) The asymmetry parameter A_b is expected to be only weakly dependent on $\sin^2\theta_w^{eff}$ (see the entry for d-quarks in Table 2), thus the measurement of A_{FB}^{0b}, given by Eq. 10a, is mostly a determination of A_e.

(2) The b-quark has a large branching ratio (~22%) and thus has a small statistical error for reasonable detection efficiency.

(3) b-quark events can be separated from other hadronic events by detection of displaced secondary vertices and various event attributes, such as high transverse mass.

(4) The b-quark charge can be determined by a variety of experimental signatures, such as a high P_t charged lepton, jet charge, kaon charge and, in the SLD, initial state tagging through the incident electron beam polarization.

(5) A_{FB} varies with center-of-mass energy, and when measured below and above the Z^0 peak at LEP, has a larger asymmetry, thereby adding information to the small asymmetry at the pole.

Similar arguments hold for s-quark and c-quark final states (no flavor mixing), although the c-quark final state is more sensitive to $\sin^2\theta_w$, thereby making the test of the SM more involved. Further, c-quarks are more difficult to separate from background and their detection frequently exploits "charm counting" where all charm hadrons which eventually decay via D^0, D^+, D_s, and Λ_c are summed, or by the reconstruction of D*s, yielding a clean sample of events. A lepton fit is also used to separate c (and b) events from background.

The coupling constant asymmetry measurements are not sensitive to the detection efficiency, although the background dilution in the asymmetry is. For the branching ratio analysis the detection efficiency critically enters and it is highly beneficial to achieve high detection efficiency since, for example, $\delta R_b \sim 1/\varepsilon_b$.

5.1 Measurements of A_S, A_C, and A_b

The SLD has measured A_s, A_c, and A_b directly through Eq. 10c, which depends critically on a precise knowledge of the electron beam polarization, P_e, but is a direct and powerful method largely independent of the Z^0-leptonic coupling physics.[9,10,11] Figure 3 shows the left and right asymmetries as determined by jet-charge for b-quarks as observed by the SLD.[11] (See Fig. 1 for the relevant physics at the tree-level.) LEP derives the asymmetry parameters from the forward-backward composite asymmetry, A_{FB}, given by Eq. 10a. For the extraction of A_b and A_c using A_{FB}, a value of A_e derived from other measurements must be used.

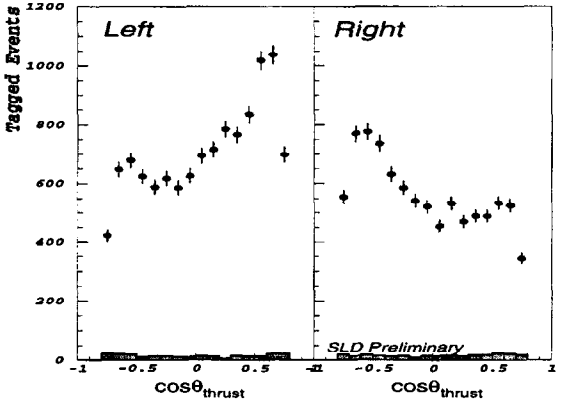


Figure 3: The θ-angle distribution for b-quarks produced from left and right electron beam polarizations at the SLD, indicating the forward-backward left-right asymmetry in b-quark production as measured by the jet-charge asymmetry.

The resulting asymmetries determined at LEP and SLC are given in Tables 5, 6, and 7 for A_s, A_c, and A_b, respectively. The LEP A_{FB} asymmetries have been corrected by the factor $4/(3\,A_e)$ with $A_e = 0.1491 \pm 0.0018$ (average of A_{LR} and LEP leptonic). In the tables the combined averages have taken into account correlations between measurements.

The SLD determination of A_S is derived from a sample of approximately 300K hadronic decays of Z^0 bosons produced with a polarized electron beam. The $s\bar{s}$ sample was selected by suppressing heavy quark decays (c and b) by requiring well-reconstructed tracks to have a transverse impact parameter with the IP of less than three times the estimated resolution. The impact parameter resolution obtained from our upgraded CCD vertex detector, VXD3, is given by $\sigma_d = 9 \oplus 29/(p_\perp \sin^{3/2}\theta)\,\mu m$, where θ is the polar angle of the track with respect to the beam. Further refinements of the $s\bar{s}$ sample were obtained by using the Cherenkov Ring Imaging Detector (CRID) to identify $K^\pm$. In addition, K^0_S and $\Lambda^0/\bar{\Lambda}^0$ were reconstructed and employed as tags of the $s\bar{s}$ final state.

Table 5: $A_{d,s}$ measurements from LEP and SLC (Summer 1998).

Experiment	Method	Asymmetry
SLD	A_S	$0.82 \pm 0.10 \pm 0.07$
DELPHI: (1995)	A_S	$1.165 \pm 0.311 \pm 0.116$
DELPHI: (1995)	$A_{d,s}$	$0.996 \pm 0.276 \pm 0.480$
OPAL: (1997)	$A_{d,s}$	$0.605 \pm 0.311 \pm 0.098$
LEP average		0.91 ± 0.21
LEP+SLD average		0.84 ± 0.10

Each of the two hemispheres, defined by the perpendicular plane bisecting the thrust axis, was required to have at least one identified strange particle. Overall, the purity of the $s\bar{s}$ final state was estimated to be 69% with 10% $u\bar{u}$, 9% $d\bar{d}$, 11% $c\bar{c}$, and 1% $b\bar{b}$ contamination. A binned maximum likelihood fit was performed to determine A_s. The procedure included particle detection efficiencies and acceptances.

Taking $\sin^2\theta_w^{eff} = 0.23155$, the SM predicts $A_S = 0.935$, a value which is slowly varying with $\sin^2\theta_w^{eff}$. The table indicates that the preliminary SLC value of A_S is 1.0 σ low with respect to the SM and the LEP value is 0.1 σ low (with A_e correction made). The SLD+LEP average is dominated by the SLC value and is 1.0 σ low.

The SLD brings excellent vertex detector and particle identification to the isolation of $c\bar{c}$ events. From the ensemble of hadronic events, secondary vertices possibly signifying $c\bar{c}$ events are identified by means of a topological algorithm,[12] which searches for space points in three dimensions where the reconstructed tracks overlap. Tracks, in this algorithm, are considered probability "tubes." The method finds secondary vertices in 84% of b, 38% of c, and 2% of light quark events under the requirement that the secondary vertex be displaced from the IP by at least 1 mm. Further refinement of the charm sample is obtained by reconstructing the mass of the tracks associated with the secondary vertex—a calculation aided by the well-defined IP of the SLC and our high resolution pixel vertex detector. The mass is defined as $M = \sqrt{M_{ch}^2 + P_t^2} + |P_t|$, where M_{ch}^2 is the mass of the ensemble of charged tracks associated with the secondary vertex and P_t is the net transverse momentum with respect to the line connecting the secondary vertex and the IP. Charm candidates were required to have a vertex mass in the range $0.55 < M_c < 2$ GeV/c^2. By vetoing on b-events and imposing momentum cuts, charm events are isolated with 81% purity. Several methods are employed to determine the direction of the charm quark needed for the $A_{FB\text{-}LR}$ evaluation, such as vertex charge, $K^{\pm}$ tag, lepton tag and D^*, D^+ identification. These tags have different efficiencies in the range of ~ 30%. A maximum likelihood fit to all tagged events is used to determine A_c.

Referring to Table 6, the A_c LEP-SLC average is in good agreement (only 0.6 σ) with the SM value of $A_c = 0.668$ for $\sin^2\theta_w^{eff} = 0.23155$ with the preliminary SLD value lower than the SM by about 0.5 σ and the LEP value low by 0.9 σ. The LEP values are derived from measurements which average to $\langle A_{FB}^{0,c} \rangle = 0.0714 \pm 0.0044$, which is then corrected by $4/(3\ A_e)$.

Table 6: A_c measurements from LEP and SLC (Summer 1998).

Experiment	Method	A_c
SLD	K & vtx-Q	$0.65 \pm 0.04 \pm 0.03$
SLD	lepton	$0.70 \pm 0.09 \pm 0.07$
SLD	D^*, D^+	$0.63 \pm 0.06 \pm 0.04$
SLD average		0.649 ± 0.040
DELPHI	lepton	$0.73 \pm 0.10 \pm 0.11$
L3	lepton	$0.82 \pm 0.29 \pm 0.19$
ALEPH	D^*	$0.63 \pm 0.08 \pm 0.02$
DELPHI	D^*	$0.64 \pm 0.08 \pm 0.04$
OPAL	D^*	$0.62 \pm 0.11 \pm 0.05$
LEP average		0.634 ± 0.040
LEP+SLD average		0.642 ± 0.028

A summary of measurements of A_b is shown in Table 7, where correlations were taken into account in the computation of averages. The SM predicts $A_b = 0.935$ for $\sin^2\theta_w^{eff} = 0.23155$, which is 3.2 σ larger than the world average for the measurement. We note that the LEP and preliminary SLC values are consistent; the SLC value is 1.9 σ lower than the SM, and LEP is 2.3 σ low.

Table 7: A_b measurements from LEP and SLC (Summer 1998).

Experiment	Method	A_b
SLD	jet charge	0.849 ± 0.026 ± 0.031
SLD	lepton	0.932 ± 0.058 ± 0.038
SLD	$K^{\pm}$ tag	0.854 ± 0.088 ± 0.106
SLD average		0.866 ± 0.036
ALEPH	lepton	0.908 ± 0.041 ± 0.020
DELPHI	lepton	0.904 ± 0.057 ± 0.026
L3	lepton	0.869 ± 0.055 ± 0.030
OPAL	lepton	0.851 ± 0.038 ± 0.021
ALEPH	jet charge	0.953 ± 0.037 ± 0.029
DELPHI	jet charge	0.898 ± 0.042 ± 0.021
L3	jet charge	0.806 ± 0.106 ± 0.051
OPAL	jet charge	0.898 ± 0.047 ± 0.037
LEP average		0.885 ± 0.022
LEP+SLD average		0.880 ± 0.017

The analysis of A_b involves many of the same considerations as discussed above for the extraction of A_c. A vertex mass tag is used, as in the case of A_c, but for the $b\bar{b}$ events it becomes more effective because the b-quark is heavier than any other visible quark, and hence has a large vertex mass which can be well separated from background. By requiring $M_b > 2$ GeV/c^2, the b-quark purity is 97% with an efficiency of 76%. Several strategies are used to determine the direction of the b-quark, such as jet charge, lepton tag, or a $K^{\pm}$ tag, which exploits the $b \rightarrow c \rightarrow s$ decay cascade. All methods have to rely on Monte Carlo simulations to a greater or lesser extent. The jet charge method has the advantage in that it is self-calibrating. Again, the LEP values for A_b are derived from measurements of A_{FB}^{0b}, which have an average value of $\langle A_{FB}^{0b} \rangle = 0.0991 \pm 0.0021$, and are corrected by the factor $4/(3A_e)$.

5.2 Measurements of R_c and R_b

The Z^0 branching fractions R_c and R_b are interesting places to search for violations of the SM. The largest known radiative corrections to R_b in the SM are charged-current vertex diagrams, which are dominated by the top quark.[13] In the past, R_b has stirred considerable interest in that the measurements were in significant disagreement with the SM. In the summer of 1995, for example, the value of R_b was almost four standard deviations different from the SM. However, a new analysis presented at ICHEP '96 by ALEPH,[14] which, among several improvements, reduced hemisphere correlations by reconstructing the primary vertex in each hemisphere for each event, resulted in a precise value of R_b in close agreement with the SM.

Tables 8 and 9 summarize the R_c and R_b measurements, respectively, from LEP[8] and SLC.[15,16] A double tag method is used in both the R_c and R_b analyses, where the heavy quark tagging procedure outlined above is applied independently to each thrust hemisphere of all hadronic events. Determining the single tag, double tag, and mixed tag rates, enables the tagging probability to be estimated by the data itself. A set of equations is developed in which either R_c and R_b can be determined, which avoids

large uncertainties of a direct calculation of the single tag efficiency. The b-tagging efficiency is typically $\varepsilon_b \sim 35\%$ with a hemisphere correlation of $\lambda_b \sim 0.6\%$.

The R_c world average is within 0.2 σ of the SM value of $R_c = 0.1723$. The SLC values are considered preliminary. The R_b world average of Table 9 is within 1.4 σ of the SM value of $R_b = 0.2155$. Hence the interesting 4 σ discrepancy with the SM of the summer 1995 has dissolved into fairly good agreement.

Table 8: Measurements of R_c (Winter 1998).

Experiment	Method	R_c
SLD	VTX-mass	0.1794 ± 0.0085 ± 0.0061
ALEPH	lepton	0.168 ± 0.006 ± 0.010
DELPHI	lepton	0.164 ± 0.009 ± 0.020
ALEPH	c-counting	0.176 ± 0.005 ± 0.011
DELPHI	c-counting	0.168 ± 0.011 ± 0.013
OPAL	c-counting	0.167 ± 0.011 ± 0.011
ALEPH	D* incl/excl	0.166 ± 0.012 ± 0.009
DELPHI	D* incl/excl	0.176 ± 0.015 ± 0.015
OPAL	D* incl/excl	0.180 ± 0.011 ± 0.013
ALEPH	D* excl/excl	0.173 ± 0.014 ± 0.009
DELPHI	D* incl/incl	0.171 ± 0.013 ± 0.015
SLD+LEP average		0.1731 ± 0.0044

Table 9: Measurements of R_b (Summer 1998).

Experiment	Method	R_b
SLD (prelim.)	VTX+mass	0.2159 ± 0.0014 ± 0.0014
ALEPH	multi-var	0.2159 ± 0.0009 ± 0.0011
DELPHI	multi-var	0.2163 ± 0.0007 ± 0.0006
L3	impact+lept	0.2176 ± 0.0015 ± 0.0026
OPAL	VTX+lept	0.2176 ± 0.0011 ± 0.0014
LEP+SLD average		0.21656 ± 0.00074

5.3 Comment on A_b and $\sin^2\theta_w^{eff}$

A_b itself shows the same difference as we observed in the $\sin^2\theta_w^{eff}$ value derived from $A_{FB}^{0,b}$ above—namely $A_{FB}^{0,b}$ measured is smaller than the prediction of the SM, resulting in a larger value of $\sin^2\theta_w^{eff}$ from the LEP measurements. Remember that we have used the pure leptonic determination of A_e to correct the LEP forward-backward asymmetries in order to obtain the value of A_b by the factor $4/(3A_e)$. Also note that the SLD measures A_b directly by means of the forward-backward left-right asymmetry.

Examining the factors involved, we conclude that the discrepancy in $\sin^2\theta_w^{eff}$ between LEP and the SLD is consistent with a discrepancy of the value of A_b from the SM. In fact, if we take the direct measurement of A_b from the SLD to derive the

value of A_e from the LEP value of $A_{FB}^{0,b}$ we obtain a value of $\sin^2\theta_w^{eff}$ which is *consistent with the pure leptonic value*. The equation below shows this calculation.

$$A_e = \frac{4}{3}\frac{A_{FB}^{0,b}(LEP)}{A_b(SLD)} = \frac{4}{3}\frac{0.0991\pm0.0021}{0.866\pm0.036} = 0.1526\pm0.0071 \qquad (12)$$

which results in $\sin^2\theta_w^{eff} = 0.2308 \pm 0.0009$—in good agreement (0.5 σ) with the pure leptonic value of $\sin^2\theta_w^{eff} = 0.23127 \pm 0.00023$. If the value $A_b(LEP) = 0.885 \pm 0.022$ (completely correlated with $A_{FB}^{0,b}$) is used instead, then $\sin^2\theta_w^{eff} = 0.23112 \pm 0.00038$ is obtained.

In a more sophisticated context, the analysis of Takeuchi, Grant, and Rosner[17] shows that the data of LEP and SLD are consistent in the pure leptonic sector and that the (current) major discrepancy (3.2 σ too low) is in the value of A_b. The left-handed coupling of the b-quark is tightly constrained by R_b and is in agreement with the SM. A_b is sensitive primarily to the right-handed coupling and is not in agreement with the SM (see Eqs. 11a and 11b). Figure 4 shows the analysis of the current experimental situation.

The TGR[17] variable ζ_b is sensitive to the parity violation of the b-quark, as measured by A_b. Its value is insensitive to the value of $\sin^2\theta_w^{eff}$—hence the horizontal band in Fig. 4. A_{LR} from SLC and A_e from LEP are most strongly dependent on $\sin^2\theta_w^{eff}$—hence the vertical bands.

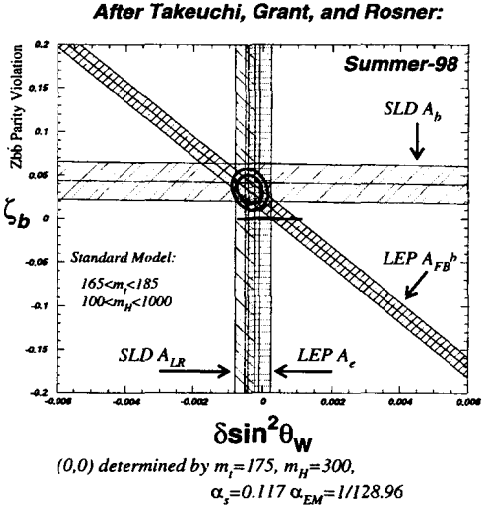


Figure 4: The consistency of the world's data is shown by the overlap of the measurement bands. The inconsistency with the SM is indicated by the offset of the overlap of measurements with the expectation of the SM given by the (0,0) point of the figure.

6 Radiative Corrections

We have observed that the data of LEP and SLD generally agree with the predictions of the Standard Model, although there is some concern over the discrepancy of A_b, and consequently the value of $\sin^2\theta_w^{eff}$ derived from the b-quark sector. So far, we have checked that Eq. 3 is generally valid, namely that the vector and axial vector couplings of a fermion to the Z^0 boson follow the weak isospin assignments and depend on the electric charge and effective weak mixing angle in a universal way. This is a test of the $SU(2)_L$ X U(1) gauge structure of the theory.

In order to check the theory beyond the "tree level" radiative corrections must be included. This process is usually performed by fitting all the measured quantities at the Z^0 pole with radiative corrections to derive the top quark mass and the Higgs mass, or constraining the top quark mass by the measured value at FNAL and deriving the Higgs mass. The consistency of the theory is measured by the quality of the fit. Sensitivities to M_t and M_h arise from different measured quantities having different dependencies on the parameters of the radiative corrections. We have already absorbed the vertex corrections for leptons (but not for quarks) by defining an effective mixing angle determined by A_e, the electron coupling constant asymmetry.

There have been a number of theoretical treatments which make the comparison of theory to data less model-dependent.[18] Here we adopt a more pedestrian approach and use "experimental" parameters, such as M_t and M_h to obtain an estimate of the magnitude of various terms. The need for radiative corrections is easily demonstrated in a naive evaluation of Eq. 2 using the value of $\sin^2\theta_w^{eff}$ determined above, α_{em} given by atomic physics experiments, and G_F from muon decay. This naive exercise predicts M_Z = 88.38 ± 0.03 GeV and M_W = 77.48 ± 0.04 GeV, values which are embarrassingly discrepant with the measured values of 91.1867 ± 0.0021 GeV and 80.37 ± 0.090 GeV, respectively [Vancouver 1998].[8]

Radiative corrections are grouped into two general types: (1) electromagnetic corrections, which include initial and final state radiation, vertex diagrams, etc., and the running of α_{em} for the atomic energy scale $q^2 \approx 0$ to the $q^2 = M_Z^2$; and (2) electroweak corrections where some of the corrections are absorbed in the definition of $\sin^2\theta_w^{eff}$, isospin-breaking loop terms in W and Z propagators, running of the Z self-energy, corrections to the $Z \rightarrow b\bar{b}$ vertex, and corrections to the W mass.

7.1 Hadronic Vacuum Polarization

Much of the discrepancy in the vector boson mass relations is corrected by the running of $\alpha_{em}(Q^2)$. The QED coupling at the M_Z scale is related to its value at low energy, given precisely in atomic physics experiments by

$$\alpha_{em}(M_Z^2) = \frac{\alpha_{em}(0)}{1 - \Delta\alpha_{em}(M_Z^2)}, \qquad (13)$$

where $\Delta\alpha_{em}(M_Z^2) = -\Pi_{\gamma\gamma}$ is the photon self-energy. The photon self-energy is evaluated quite accurately for leptons, but not for quarks. In order to evaluate the quark contribution, the measured e^+e^- hadronic cross section is employed to determine the value of a dispersion integral. Ironically, the data in the 1.05 to 5 GeV region, well beneath the scale of the Z^0 pole, contribute a large fraction of the integral as well as a large part of the uncertainty. Much of the data in this energy region are old and have large normalization errors. Further, above the charm threshold, many channels with different properties exist, complicating the evaluation of the integral. A number of evaluations have been performed.[19] The result adopted by the LEP EW-WG is $\alpha_{em}(M_Z^2) = 1/(128.896 \pm 0.090)$, which is an evolution of $\Delta\alpha/\alpha \sim 0.063$. In this formulation G_F does not run.[20]

7.2 Electroweak Radiative Corrections

The radiative corrections are furnished by the terms ζ in Eqs. 2 and 5 above. At the one-loop level the correction terms for the M_Z mass relation given by Eq. 2a is

$$\zeta_Z = (1 + \Delta\rho)\,(1 + \Delta_{3Q}), \tag{14a}$$

where, for leading order with M_t = 175 GeV and M_h = 160 GeV, the isospin-breaking loop corrections to the $W^{\pm}$ and Z propagators and Z self-energy terms, respectively, are

$$\Delta\rho \approx \frac{\alpha}{\pi}\frac{M_t^2}{M_z^2} - \frac{\alpha}{4\pi}\ln\frac{M_h^2}{M_z^2} \sim 0.0084 \quad , \tag{14b}$$

and

$$\Delta_{3Q} \approx \frac{\alpha}{9\pi}\ln\frac{M_h^2}{M_z^2} \sim 0.00031. \tag{14c}$$

The overall electroweak correction to the Z-mass relation is dominated by the quadratic M_t term and has the magnitude, $\Delta\zeta_z/\zeta_z \sim 0.0087$. The corresponding correction for M_W (Eq. 2b) is

$$\zeta_W = 1 - \Delta r_{ew}. \tag{15a}$$

where

$$\Delta r_{ew} \approx -\frac{c_0^2}{s_0^2}\Delta\rho \sim -0.028 \tag{15b}$$

with $s_0^2 c_0^2 = \pi\,\alpha_{em}(M_z^2)/\sqrt{2}\,G_F M_z^2$, yielding $\Delta\zeta_W/\zeta_W \sim 0.028$. The correction to the $Z^0 \rightarrow b\bar{b}$ partial width is given by

$$\zeta_{b\bar{b}} = (1+\delta_{QCD})\,(1+\delta_{vb}). \tag{16a}$$

The QCD component to next-to-leading order is

$$\delta_{QCD} \approx \frac{\alpha_s(M_z^2)}{\pi} + 1.41\left(\frac{\alpha_s(M_z^2)}{\pi}\right)^2 \sim 0.038 \tag{16b}$$

for $\alpha_s(M_Z^2) = 0.12$. The vertex correction, δ_{vb}, is

$$\delta_{vb} \approx -\frac{20}{13}\frac{\alpha}{\pi}\left(\frac{M_t^2}{M_z^2} + \frac{13}{6}\ln\frac{M_t^2}{M_z^2}\right) \sim -0.025. \tag{16c}$$

In addition to the QCD and vertex corrections for Γ_{bb}, there are QED and finite mass terms. From the list above, the largest correction to the mass relations of Eqs. 2a and 2b is the running of α_{em} from the atomic scale to the M_z scale. Other corrections are quadratic in M_t, and logarithmic in M_h. Two-loop electroweak top corrections are being reconsidered[21] and may be important.

8 Global Test of the Standard Model

Within the context of the SM different electroweak observables are sensitive to different radiative correction parameters (M_t, M_h, α_s). For example, R_b has a strong sensitivity to M_t and little to M_h, M_W is sensitive to both M_t and M_h, and σ_{had}^0 depends on α_s. These different sensitivities are exploited to determine the value of M_t and set limits on M_h or with a constrained value of M_t determine (still within large errors) the value of M_h. Table 10 shows the results of a fit (ZFITTER 5.10)[22] to all the world's electroweak data [M_W, R_V, Γ_Z, σ_{had}^0, R_l, A_{FB}^l, A_e, A_t, A_{LR}, A_{FB}^b, A_{FB}^c, R_b, R_c, Q_{FB}, Q_W(Cs), and Q_W (Tl)].

Table 10: Results of a fit to electroweak data at Z^0 pole.

Parameter	Constraint	Best Fit
M_Z (GeV)	91.187±0.002	91.187
$1/\alpha_{em}(M_Z^2)$	128.928±0.023	128.928±0.023
M_t (GeV)	173.8 ± 5.0	172.2±4.8
$\alpha_s(M_Z^2)$	----------	0.119±0.003
$\log_{10}(M_h)$	----------	$1.981^{+0.216}_{-0.242}$
M_h (GeV)	----------	$95.6^{+61.6}_{-40.9}$
χ^2/DoF	----------	15.2/14 (37%)

The fitted value of M_t is in agreement with the direct observation at FNAL[23] of M_t = 173.8 ± 5.0 GeV. The limits on M_h are > 90 GeV from direct searches and < 280 GeV at 95% CL.[24]

It is interesting to note that the SLD value of $\sin^2\theta_w^{eff}$ implies a Minimal Standard Model (MSM) Higgs scalar of ~ 40 GeV and is ~ 1 σ in contradiction with the present direct search limit of > 90 GeV at the 95% CL, whereas the LEP value is consistent with M_h ~ 220 GeV and is not excluded by direct searches.

A more sophisticated way of looking at the data is shown in Figure 5 as a comparison of the world's data with the variables S and T of Peskin and Takeuchi.[25] The variables S and T are normalized to (0,0) at a nominal SM point—determined by the measured value of M_t and a nominal value of M_h. The contributions to isospin-violating mass differences beyond the set point are described by T, which is roughly quadratic in M_t and logarithmic in M_h (see Eqs. 14 and 15 above). The parameter S is sensitive to isospin-independent terms which would grow systematically with the size of a new sector. The Peskin-Takeuchi variable U is assumed to be 0 in this application. The 68% confidence region of the LEP and SLD data is indicated by the oval region in the figure.

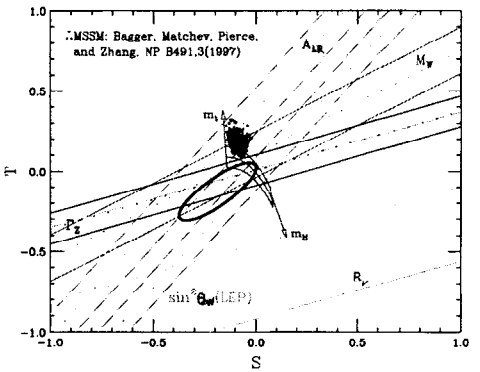


Figure 5: The S and T constraints provided by LEP, SLD, and FNAL are indicated. The SLD measurement of A_{LR} seems to prefer the MSSM of Pierce *et al.*[26]

The S-T region predicted by the MSM is indicated by the banana-region centered about the (0,0) point of the figure. In that region, the right-hand edge corresponds to M_h = 88 GeV and M_t = 173.9 ± 5.2 GeV. Increasing the Higgs mass up to 1 TeV is displayed by the width of the region. An indication of supersymmetry would be a shift of the experimental overlap region from the expectations of the MSM with S becoming slightly negative and T positive. The MSSM of Pierce *et al.*[26] is indicated by the ensemble of dots in the figure—each representing a choice of the five parameters of the model. We note the SLD data[27] favors a supersymmetric world, whereas LEP data disfavors the model by about 2 σ.

9 Summary and Conclusions

The Standard Model works remarkably well with only one parameter, A_b, out of some 18 measurements differing significantly from theory prediction.[28] The $SU(2)_L$ X U(1) structure of the theory is quite well established through one-loop electroweak radiative corrections, although the b-quark sector continues to exhibit troublesome inconsistencies. The largest discrepancies are $\Delta \sin^2\theta_w^{eff}$ (SLD versus LEP) ~ 2.2 σ and ΔA_b (SLD+LEP versus SM) ~ 3.2 σ. Most of the $\Delta \sin^2\theta_w^{eff}$ (SLD versus LEP) is resolved by considering only purely leptonic determinations of the electroweak mixing angle, isolating the puzzle of why A_b does not agree with the SM. Analyses of A_b are involved and, to a greater or lesser extent, rely on Monte Carlo simulation to estimate the experimental resolution, efficiency, and background. These simulations may be flawed in a subtle way and therefore contribute to the discrepancy—or there may be "new" physics at long last becoming visible in the b-sector. If the purely leptonic value (SLD and LEP) of $\sin^2\theta_w^{eff}$ is ultimately preferred, then the precision electroweak data suggest that the Higgs scalar is light and direct searches for it become quite tantalizing.

The SLD has proposed to take more data at the Z^0 pole, which will continue to probe the SM, especially in refining the measurement of $\sin^2\theta_w^{eff}$ and b-quark parameters. Data from LEP-II will decrease the uncertainty in M_W, explore the triple vertex coupling, and search for exotic particles beyond the SM. Tevatron at FNAL will decrease the M_t and M_W errors. And "modern" low energy data (Novosibirsk, BES, DAPHNE, B-factories) should improve the evaluation of $\Delta\alpha_{em}(M_Z^2)$ and sharpen the radiative corrections.

Acknowledgments

I would like to thank the organizers and participants of this conference for a very stimulating time. Special thanks go to my SLD colleagues and to the SLC Accelerator Department for their outstanding performance.

References

[1] M. Breidenbach, SLAC-PUB-3798 (1985); K. Abe *et al.*, Phys. Rev. D **53** (1996) 1023; M. Woods, SLAC PUB-7320 (1996) [hep-ex/9611006]; E. Torrence, SLAC-REPORT-509 (1997); A. Lath, SLAC-REPORT-454 (1994).

[2] K. Abe *et al.*, Nucl. Instrum. and Meth., A **400**, (1997), 287-343.

[3] S. Weinberg, Phys. Rev. Lett. **19** (1967) 1264; A. Salam, *Elementary Particle Theory*, edited by N. Svartholm, p. 367.

[4] See for example: D. C. Kennedy and B. W. Lynn, Nucl. Phys. B **322** (1989), 1.

[5] K. Abe *et al.*, Phys. Rev. Lett. **78** (1997) 2075-2079.

[6] K. Abe *et al.*, Phys. Rev. Lett. **79**, 804-808 (1997); R. S. Panvini, SLAC-PUB-7375; A. K. McKemey, in *Proceedings of ICHEP'96 Conference.*

[7] K. Abe et al., SLAC-PUB-7878 (1998).

[8] For recent LEP data see M. Martinez, *Precision Electroweak Results at LEP I and LEP II,* in these proceedings; see also D. Abbaneo *et al.*, ElectroWeak Working Group, Internal Memo LEPEWWG/97-01; CERN-PPE/97-154; J. Timmermans, in *Proceedings of XVIII Int. Sym. on Lepton-Photon Int.*, Hamburg (1997).

[9] SLD A_S: K. Abe *et al.*, SLAC-PUB-7825 (1998).

[10] SLD A_C: K. Abe *et al.*, SLAC-PUB-7879 (1998).

[11] A_b analyses from the SLD are discussed in K. Abe *et al.*, SLAC-PUB-7886 (1998); K. Abe *et al.*, SLAC-PUB-7959 (1998); see also Dong Su, in Heavy

Quark Couplings to the Z^0, presented at the *XVIII International Symposium on Lepton Photon Interactions* (1997), Hamburg, Germany.

[12] D. Jackson, Nucl. Instrum. and Meth., A **338**, (1997) 247.

[13] See for example: A. Blondel and C. Verzegnassi, Phys. Lett. B **311** (1993) 346–356.

[14] ALEPH: I. R. Tomalin, in *Proceeding of ICHEP'96*, Warsaw, Poland, p. 1340 (1996).

[15] SLD R_c: K. Abe *et al.*, SLAC-PUB-7880 (1998).

[16] SLD R_b: K. Abe *et al.*, Phys. Rev. Lett. 80 (1997) 660.

[17] T. Takeuchi, A. Grant and J. Rosner, p. 1231 in *Proceedings of 8th Meeting of the DPF* (Albuquerque, N.M. 1994), edited by S. Seidel.

[18] G. Altarelli, R. Barbieri, and S. Jadach, Nucl. Phys. B **369**, 3 (1992); W. Marciano, these proceedings.

[19] M. Swartz, Phys. Rev. D **53** (1996) 5268; H. Burkhardt and B. Pietrzyk , Phys. Lett. B **356** (1995) 398.

[20] R. D. Peccei, DESY 89-060.

[21] G. Degrassi et al., hep-ph/9412380.

[22] The fit uses ZFITTER, D. Bardin *et al.*, CERN-TH 6443/92 (May 1992); hep-ph/9412201.

[23] See R. Brock, in *Proceedings of Fund. Particles and Interactions,* p. 189, Nashville, TN, edited by R. S. Panvini and T. J. Weiler (1997); J. Kotcher, in these proceedings.

[24] D. Karlen, in *Proceedings of the 1998 Int. Conf. on High Energy Physics,* ICHEP'98, Vancouver, BC (1998).

[25] M. Peskin and T. Takeuchi , Phys. Rev. D **46**, 381-409 (1992).

[26] D. Pierce *et al.*, Nucl. Phys. B **491** (1997) 3-67.

[27] "Request for SLD Run Extension," SLD Collaboration (Aug. 1998).

[28] An earlier comparison of LEP-SLC data can be found in F. E. Taylor, in *Proceedings of Fund. Particles and Interactions*, p. 166, Nashville, TN, edited by R. S. Panvini and T. J. Weiler (1997).

PRECISION ELECTROWEAK MEASUREMENTS AND "NEW PHYSICS"

William J. Marciano*
Brookhaven National Laboratory
Upton, New York 11973

ABSTRACT

The status of several precisely measured electroweak parameters is reviewed. Natural relations among them are shown to constrain the Higgs mass, m_H, as well as various "New Physics" effects. Indications of an anomalous $Zb\bar{b}$ coupling are discussed. Constraints on excited W^* bosons are given.

*This manuscript has been authored under contract number DE-AC02-98CH10886 with the U.S. Department of Energy. Accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

©1998 by William J. Marciano.

1 Fundamental Parameters and Natural Relations

The $SU(2)_L \times U(1)_Y$ electroweak sector of the standard model contains 17 or more fundamental parameters. They include gauge and Higgs field couplings as well as fermion masses and mixing angles. In terms of those parameters, predictions can be made with high accuracy for essentially any electroweak observable. Very precise measurements of those quantities can then be used to test the standard model, even at the quantum loop level, or search for small deviations from expectations which would indicate "New Physics."

Some fundamental electroweak parameters have been determined with extraordinary precision. Foremost in that category is the fine structure constant α. It can best be obtained by comparing the measured[1] anomalous magnetic moment of the electron, $a_e \equiv (g_e - 2)/2$,

$$a_e^{\rm exp} = 1159652188(3) \times 10^{-12}, \tag{1}$$

with the calculated four loop QED prediction,[2]

$$\begin{aligned} a_e^{\rm th} &= \frac{\alpha}{2\pi} - 0.328478444\left(\frac{\alpha}{\pi}\right)^2 + 1.181234\left(\frac{\alpha}{\pi}\right)^3 - 1.5098\left(\frac{\alpha}{\pi}\right)^4 \\ &\quad +1.66 \times 10^{-12}, \end{aligned} \tag{2}$$

where the 1.66×10^{-12} comes from small hadronic and weak loop effects. Assuming no significant "new physics" contributions to $a_e^{\rm th}$, it can be equated with Eq. (1) to give

$$\alpha^{-1} = 137.03599959(40). \tag{3}$$

That precision is already quite extraordinary. Further improvement by a factor of ten appears to be technically feasible and should certainly be undertaken. However, at this time such improvement would not further our ability to test QED. QED tests require comparable measurements of α in other processes. Agreement between two distinct α determinations tests QED and probes for "new physics" effects. After a_e, the next best (direct) measurement of α comes from the quantum Hall effect,

$$\alpha^{-1}(qH) = 137.03600370(270), \tag{4}$$

which is not nearly as precise. Nevertheless, the agreement of Eqs. (3) and (4) (at the 1.5 σ level) is a major triumph for QED up to the four loop quantum level.

In terms of probing "new physics," one can search for a shift in a_e by m_e^2/Λ_e^2 where Λ_e is the approximate scale of some generic new short-distance effect. Current comparison of $a_e \to \alpha$ and $\alpha(qH)$ explores $\Lambda_e \lesssim 100$ GeV. To probe the much more interesting $\Lambda_e \sim \mathcal{O}$ (TeV) region would require an order of magnitude improvement in a_e and about two orders of magnitude error reduction in some direct precision determination of α such as the quantum Hall effect. Perhaps the most likely possibility is to use the already very precisely measured Rydberg constant in conjunction with a much improved m_e determination to obtain an independent α.

The usual fine structure constant, α, is defined at zero momentum transfer as is appropriate for low energy atomic physics phenomena. However, that definition is not well suited for short-distance electroweak effects. Vacuum polarization loops screen charges, such that the effective (running) electric charge increases at short distances. One can incorporate those quantum loop contributions into a short-distance[4] $\alpha(m_Z)$ defined at $q^2 = m_Z^2$. The main effect comes from lepton loops, which can be very precisely calculated, and somewhat smaller hadronic loops. The latter are not as theoretically clean and must be obtained by combining perturbative calculations with results of a dispersion relation which employs $\sigma(e^+e^- \to \text{hadrons})$ data. A recent study by Davier and Höcker finds[5]

$$\alpha^{-1}(m_Z) = 128.933(21), \tag{5}$$

where the uncertainty stems from low energy hadronic loops. Although not nearly as precise as α^{-1}, the uncertainty quoted in Eq. (5) is impressively small and a tribute to the effort that has gone into reducing it. (When I first studied this issue in 1979, I estimated[4] $\alpha^{-1}(m_Z) \simeq 128.5 \pm 1.0$.) However, the error in Eq. (5) is still somewhat controversial, primarily because of its reliance on perturbative QCD down to very low energies. For comparison, an earlier study by Eidelman and Jegerlehner,[6] which relied less on perturbative QCD and more on e^+e^- data found

$$\alpha^{-1}(m_Z) = 128.896(90) \qquad \text{(E \& J 1995)}. \tag{6}$$

That estimated uncertainty is often cited as more conservative and therefore employed in m_H and "new physics" constraints. As we shall see, the smaller uncertainty in Eq. (5) has very important consequences for predicting the Higgs mass. I note that a more recent study[7] by Eidelman and Jegerlehner finds

$$\alpha^{-1}(m_Z) = 128.913(35) \qquad \text{(E \& J 1998)}, \tag{7}$$

which is in good accord with Eq. (5) and also exhibits relatively small uncertainty. In my subsequent discussion, I employ the result in Eq. (5), but caution the reader that a more conservative approach would expand the uncertainty, perhaps even by as much as a factor of four or five.

A related short-distance coupling, $\alpha(m_Z)_{\overline{MS}}$, can be defined by modified minimal subtraction at scale $\mu = m_Z$. It is particularly useful for studies of coupling unification in grand unified theories (GUTS) where a uniform comparitive definition ($\overline{MS}$) of all couplings is called for.[8] The quantities $\alpha(m_Z)$ and $\alpha(m_Z)_{\overline{MS}}$ differ by a constant, such that[9]

$$\alpha^{-1}(m_Z)_{\overline{MS}} = \alpha^{-1}(m_Z) - 0.982 = 127.951(21). \tag{8}$$

In weak interaction physics, the most precisely determined parameter is the Fermi constant, G_μ, as obtained from the muon lifetime. One extracts that quantity by comparing the experimental value,

$$\tau_\mu = 2.197035(40) \times 10^{-6} s, \tag{9}$$

with the theoretical prediction,

$$\begin{aligned} \tau_\mu^{-1} = \Gamma(\mu \to \text{all}) &= \frac{G_\mu^2 m_\mu^5}{192\pi^3} f\left(\frac{m_e^2}{m_\mu^2}\right)(1 + \text{R.C.})\left(1 + \frac{3}{5}\frac{m_\mu^2}{m_W^2}\right) \\ f(x) &= 1 - 8x + 8x^3 - x^4 - 12x^2 \ell n x. \end{aligned} \tag{10}$$

In that expression R.C. stands for Radiative Corrections. Those terms are somewhat arbitrary in the standard model. The point is that G_μ is a renormalized parameter which is used to absorb most loop corrections to muon decay. Those corrections not absorbed into G_μ are explicitly factored out in R.C. For historical reasons and in the spirit of effective field theory approaches, R.C. has been chosen to be the QED corrections to the old V-A four fermion description of muon decay.[10] That definition is practical, since the QED corrections to muon decay in the old V-A theory are finite to all orders in perturbation theory. In that way, one finds

$$\text{R.C.} = \frac{\alpha}{2\pi}\left(\frac{25}{4} - \pi^2\right)\left(1 + \frac{\alpha}{\pi}\left(\frac{2}{3}\ell n\frac{m_\mu}{m_e} - 3.7\right) + \left(\frac{\alpha}{\pi}\right)^2\left(\frac{4}{9}\ell n^2\frac{m_\mu}{m_e} - 2.0\ell n\frac{m_\mu}{m_e} + C\right)\cdots\right). \tag{11}$$

The leading $\mathcal{O}(\alpha)$ terms in that expression have been known for a long time from the pioneering work of Kinoshita and Sirlin[11] and Berman.[12] Coefficients of the higher order logs can be obtained from the renormalization group constraint,[13]

$$\begin{aligned} \left(m_e\frac{\partial}{\partial m_e} + \beta(\alpha)\frac{\partial}{\partial\alpha}\right)\text{R.C.} &= 0 \\ \beta(\alpha) = \frac{2}{3}\frac{\alpha^2}{\pi} + \frac{1}{2}\frac{\alpha^3}{\pi^2}\cdots. \end{aligned} \tag{12}$$

The -3.7 two loop constant in parentheses was very recently computed by van Ritbergen and Stuart.[14] It almost exactly cancels the leading log two loop correction obtained from the renormalization group approach (or mass singularities argument) of Roos and Sirlin.[13] Hence, the original $\mathcal{O}(\alpha)$ correction in Eq. (9) is a much better approximation than one might have guessed. Comparing Eqs. (9) and (10), one finds

$$G_\mu = 1.16637(1) \times 10^{-5}\ \text{GeV}^{-2}. \tag{13}$$

There have been several experimental proposals to reduce the uncertainty in τ_μ and G_μ by a factor of ten. Such improvement appears technically feasible and, given the fundamental nature of G_μ, should certainly be undertaken. However, from the point of view of testing the standard model, the situation is similar to α. G_μ is already much better known than the other parameters it can be compared with; so, significant improvement must be made in other quantities before a more precise G_μ is required. This point should become clearer subsequently when I describe other indirect Fermi constant determinations and their uncertainty (about 100 times worse than Eq. (13)).

Let me emphasize the fact that lots of interesting loop effects have been absorbed into the renormalization of $g_{2_0}^2/4\sqrt{2}m_W^{0^2}$ which we call G_μ. Included are top quark[15] and Higgs loop corrections[16] to the W boson propagator as well as potential "new physics" from SUSY loops, Technicolor, etc. Even tree level effects of possibly more massive gauge bosons such as $W^{*\pm}$ bosons are effectively incorporated into G_μ. To uncover those contributions requires comparison of G_μ with other precisely measured electroweak parameters which have different quantum loop (or tree level) dependences. Of course, those quantities must be related to G_μ in such a way that short-distance divergences cancel in the comparison.

Fortunately, due to an underlying global $SU(2)_V$ symmetry in the Standard Model, there exist natural relations among various bare parameters[17]

$$\sin^2\theta_W^0 = \frac{e_0^2}{g_{2_0}^2} = 1 - (m_W^0/m_Z^0)^2. \tag{14}$$

Each of those bare unrenormalized expressions contains short-distance infinities, but the divergences are the same. Therefore, those relations continue to hold for renormalized quantities, up to finite, calculable radiative corrections.[17] The residual radiative corrections contain very interesting effects such as m_t and m_H dependence, as well as possible "new physics." So, for example, one can relate

$$G_\mu = \frac{\pi\alpha}{\sqrt{2}m_W^2(1 - m_W^2/m_Z^2)}(1 + rad.\ corr.) \tag{15}$$

and test the predicted radiative corrections, if m_Z and m_W are also precisely known.

Gauge boson masses are not as well determined as G_μ, but they have reached high levels of precision. In particular, the Z mass has been measured with high statistics Breit-Wigner fits to the Z resonance at LEP with the result

$$m_Z = 91.1867(21)\ \text{GeV}. \tag{16}$$

That determination is so good that one must be very precise regarding the definition of m_Z. (Remember the Z has a relatively large width $\sim$ 2.5 GeV.) The quantity in Eq. (16) is related to the real part of the Z propagator pole, m_Z (pole), and full width, Γ_Z, by[18]

$$m_Z^2 = m_Z^2(\text{pole}) + \Gamma_Z^2. \tag{17}$$

The two mass definitions m_Z and m_Z (pole) differ by about 34 MeV, which is much larger than the uncertainty in Eq. (16). Hence, one must specify which definition is being employed in precision studies. I note that the m_Z in Eq. (16) is also more appropriate for use in low energy neutral current amplitudes.

In the case of the $W^\pm$ bosons, the renormalized mass, m_W, is similarly defined by

$$m_W^2 = m_W^2(\text{pole}) + \Gamma_W^2. \tag{18}$$

That quantity is obtained from studies at $p\bar{p}$ colliders, $m_W = 80.41(9)$ GeV, as well as $e^+e^- \to W^+W^-$ at LEPII, $m_W = 80.37(9)$ GeV. Together they give

$$m_W = 80.39(6)\ \text{GeV}\ . \tag{19}$$

The current level of uncertainty, $\pm$60 MeV, is large compared to Δm_Z. It is expected that continuing efforts at LEPII and Run II at Fermilab's Tevatron should reduce that error to about $\pm$30 MeV. A challenging but worthwhile goal for future high energy facilities would be to push Δm_W to $\pm$10 MeV or better. At that level, all sorts of interesting "new physics" effects are probed (as I later illustrate). I note that the m_W defined in Eq. (19) is also the appropriate quantity for low energy amplitudes such as muon decay.

Another important quantity for precision standard model tests is m_t, the top quark mass. Measurements from CDF and D$\emptyset$ at Fermilab give

$$m_t(\text{pole}) = 174.3 \pm 5.1\ \text{GeV}. \tag{20}$$

Reducing that uncertainty further is important as we shall subsequently see. Future Tevatron efforts are expected to reduce the uncertainty in m_t to about $\pm$2 GeV. LHC and NLC studies should bring it well below $\pm$1 GeV.

In addition to masses, the renormalized weak mixing angle plays a central role in tests of the standard model. That parameter can be defined in a variety of ways, each of which has its own advocates. I list three popular examples:

$$\begin{aligned} &\sin^2\theta_W(m_Z)_{\overline{MS}} && (\overline{MS}\ \text{definition at}\ \mu = m_Z), && (a) \\ &\sin^2\theta_W^{\text{eff}} && (Z\mu\bar{\mu}\ \text{vertex}), && (b) \\ &\sin^2\theta_W \equiv && 1 - m_W^2/m_Z^2. && (c) \end{aligned} \tag{21}$$

They differ by finite $\mathcal{O}(\alpha)$ loop corrections. The $\overline{MS}$ definition is particularly simple, being defined as the ratio of two $\overline{MS}$ couplings $\sin^2\theta_W(m_Z)_{\overline{MS}} \equiv e^2(m_Z)_{\overline{MS}}/g_2^2(m_Z)_{\overline{MS}}$. It was introduced for GUT studies,[8] but is useful for most electroweak analyses. The effective, $\sin^2\theta_W^{\text{eff}}$, weak angle was invented for Z pole analyses. Roughly speaking, it is defined by the ratio of vector and axial-vector components (including loops) for the on-mass-shell $Z\mu\bar{\mu}$ vertex $\to$ 1–4 $\sin^2\theta_W^{\text{eff}}$. Although conceptually rather simple, analytic electroweak radiative corrections expressed in terms of $\sin^2\theta_W^{\text{eff}}$ are complicated and ugly. Numerically, it is close to the $\overline{MS}$ definition,[19]

$$\sin^2\theta_W^{\text{eff}} = \sin^2\theta_W(m_Z)_{\overline{MS}} + 0.00028, \tag{22}$$

but the analytic structure of the difference is quite complicated. For those intent on employing $\sin^2\theta_W^{\text{eff}}$, a strategy might be to calculate radiative corrections in terms of

$\sin^2\theta_W\ (m_Z)_{\overline{MS}}$ and then translate to $\sin^2\theta_W^{\rm eff}$ via Eq. (22). But why not simply use $\sin^2\theta_W\ (m_Z)_{\overline{MS}}$?

Currently, Z pole studies at LEP and SLAC give

$$\begin{aligned}\sin^2\theta_W(m_Z)_{\overline{MS}} &= 0.23100 \pm 0.00022,\\ \sin^2\theta_W^{\rm eff} &= 0.23128 \pm 0.00022.\end{aligned} \tag{23}$$

That result includes measurements of the left-right asymmetry, A_{LR}, at SLAC as well as the various lepton asymmetries at LEP and SLAC. The A_{LR} contribution had for some time given a relatively low value for the weak mixing angle, but as statistics have increased it has moved pretty much in line with Eq. (23). Currently, the $Z \to b\bar{b}$ forward-backward asymmetries at LEP give a higher $\sin^2\theta_W^{\rm eff}$ and would bring up the average, if included. However, the $Zb\bar{b}$ coupling appears to be somewhat anomalous; so, one should be cautious when including such results. I will return to this problem later.

Future higher statistics running at SLAC could reduce the uncertainty in $\sin^2\theta_W^{\rm eff}$ below ± 0.0002, mainly from improvements in A_{LR}. There are very good reasons to do even better. One could imagine redoing A_{LR} at a future polarized lepton-lepton (e^+e^- or $\mu^+\mu^-$) collider, but with very high statistics. In principle, one might reduce the uncertainty in $\sin^2\theta_W^{\rm eff}$ to ± 0.00004 or lower, an incredible achievement if possible.

The so-called on-shell or mass definition[20] in Eq. (21c) also has its advocates. It can be directly obtained from m_W and m_Z determinations. Indeed, at hadron colliders, the ratio m_W/m_Z can have reduced systematic uncertainties. One could imagine that the current uncertainty in

$$\sin^2\theta_W = 1 - m_W^2/m_Z^2 = 0.2228 \pm 0.0012 \tag{24}$$

might be reduced by a factor of about four at the LHC. Such a reduction is extremely interesting since the comparison of $\sin^2\theta_W$ and $\sin^2\theta_W(m_Z)_{\overline{MS}}$ provides a clean probe of "new physics." It is also possible (because of a subtle cancellation of certain loop effects[21]) to measure $\sin^2\theta_W$ in deep-inelastic $\nu_\mu N$ scattering. Indeed, a recent Fermilab experiment found[22]

$$\sin^2\theta_W = 0.2253 \pm 0.0019 \pm 0.0010 \tag{25}$$

where the first error is statistical and the second systematic. That single measurement is quite competitive with Eq. (24) and complements it nicely. One might imagine a future high statistics effort significantly reducing the error in Eq. (25), but that would require a new high energy neutrino beam.

All of the above precision measurements can be collectively used to test the standard model, predict the Higgs mass, and search for "new physics" effects. That ability stems from the natural relations in Eq. (14) and calculations[20,23] of the radiative corrections to them. Parametrizing those radiative corrections by Δr, $\Delta r(m_Z)_{\overline{MS}}$, and $\Delta\hat{r}$, one finds[24]:

$$\frac{\pi\alpha}{\sqrt{2}G_\mu m_W^2} = \left(1 - \frac{m_W^2}{m_Z^2}\right)(1 - \Delta r), \tag{a}$$

$$\frac{\pi\alpha}{\sqrt{2}G_\mu m_W^2} = \sin^2\theta_W(m_Z)_{\overline{MS}}(1 - \Delta r(m_Z)_{\overline{MS}}), \tag{b}$$

$$\frac{4\pi\alpha}{\sqrt{2}G_\mu m_Z^2} = \sin^2 2\theta_W(m_Z)_{\overline{MS}}(1 - \Delta\hat{r}). \tag{c}$$

(26)

Those expressions all contain one loop corrections to α, muon decay, m_W, m_Z, and $\sin^2\theta_W(m_Z)_{\overline{MS}}$, and incorporate some leading two loop contributions. The quantities Δr and $\Delta\hat{r}$ are particularly interesting because of their dependence on m_t and m_H. In addition, all three quantities provide probes of "new physics."

Numerically, all three radiative corrections in Eq. (26) contain a significant contribution from vacuum polarization effects[4] in α, about $+7\%$. They are basically the same as the corrections that enter into the evolution of α to $\alpha(m_Z)$. Leptonic loops contribute a significant part of that effect and can be very accurately computed. Hadronic loops are less clean theoretically and lead to a common uncertainty in Δr, $\Delta r(m_Z)_{\overline{MS}}$, and $\Delta\hat{r}$ of

$$-\alpha\Delta\alpha^{-1}(m_Z). \tag{27}$$

For $\Delta\alpha^{-1}(m_Z) = 0.021$ as in Eq. (5), that amounts to a rather negligible ± 0.00015 error. However, for $\Delta\alpha^{-1}(m_Z) = \pm 0.090$ as in Eq. (6), it increases to ± 0.00066. That large an uncertainty would impact precision tests. If one wishes to avoid that low energy hadronic loop uncertainty, dependence on α can be circumvented by considering

$$\sin^2\theta_W(m_Z)_{\overline{MS}} = \left(1 - \frac{m_W^2}{m_Z^2}\right)(1 - \Delta r + \Delta r(m_Z)_{\overline{MS}}). \tag{28}$$

Currently, that comparison is not competitive in constraining m_H. However, future significant improvements in m_W could make it very interesting.

Using $m_t = 174.3 \pm 5.1$ GeV as input, one can compute the radiative corrections in Eq. (26) as functions of m_H. Those results are illustrated in Table 1. Note that Δr is most sensitive to changes in m_H but also carries the largest uncertainty from $\Delta m_t = \pm 5.1$ GeV (± 0.0020). Hence, efforts to determine m_H from m_W will require a better measurement of m_t. On the other hand, determining m_H from $\sin^2\theta_W(m_Z)_{\overline{MS}}$ via $\Delta\hat{r}$ is less sensitive to Δm_t but more sensitive to $\Delta\alpha^{-1}(m_Z)$. Those dependences are illustrated by the following approximate relations[25] obtained from Eqs. (26a) and (26c)

$$m_W = (80.385 \pm 0.032 \pm 0.003 \text{ GeV}) \left(1 - 0.00072 \ell n \left(\frac{m_H}{100 \text{ GeV}}\right) - 1 \times 10^{-4} \ell n^2 \left(\frac{m_H}{100 \text{ GeV}}\right)\right), \quad (29)$$

$$\sin^2\theta_W(m_Z)_{\overline{MS}} = (0.23112 \pm 0.00016 \pm 0.00006) \left(1 + 0.00226 \ell n \left(\frac{m_H}{100 \text{ GeV}}\right)\right) \quad (30)$$

where the errors correspond to $\Delta m_t = \pm 5.1$ GeV and $\Delta\alpha^{-1}(m_Z) = \pm 0.021$ respectively. Note that increasing $\Delta\alpha^{-1}(m_Z)$ to ± 0.090 would significantly compromise the utility of $\sin^2\theta_W(m_Z)_{\overline{MS}}$ for determining m_H but would have less of an impact on m_W. Predictions for m_W and $\sin^2\theta_W(m_Z)_{\overline{MS}}$ are illustrated in table 2 for various m_H values.[26]

Table 1. Values of Δr, $\Delta r(m_Z)_{\overline{MS}}$, and $\Delta\hat{r}$ for various m_H. A top quark mass of 174.3 ± 5.1 GeV and $\alpha^{-1}(m_Z) = 128.933(21)$ are assumed.

m_H (GeV)	Δr $\pm 0.0020 \pm 0.0002$	$\Delta r(m_Z)_{\overline{MS}}$ $\pm 0.0001 \pm 0.0002$	$\Delta\hat{r}$ $\pm 0.0005 \pm 0.0002$
75	0.03402	0.06914	0.05897
100	0.03497	0.06937	0.05940
125	0.03575	0.06955	0.05974
150	0.03646	0.06964	0.06000
200	0.03759	0.06980	0.06042
400	0.04065	0.07005	0.06144

Table 2. Predictions for m_W and $\sin^2\theta_W(m_Z)_{\overline{MS}}$ for various m_H values.

m_H (GeV)	m_W (GeV)	$\sin^2\theta_W(m_Z)_{\overline{MS}}$
75	80.401	0.23097
100	80.385	0.23112
125	80.372	0.23124
150	80.360	0.23133
200	80.341	0.23148
400	80.289	0.23184

Employing $m_W = 80.39 \pm 0.06$ GeV and $\sin^2\theta_W(m_Z)_{\overline{MS}} = 0.23100 \pm 0.00022$, one finds:

$$m_H = 92^{+141+75+5}_{-67-35-5} \text{ GeV} \quad (\text{from } m_W), \quad (31)$$

$$m_H = 79^{+41+28+9}_{-27-21-8} \text{ GeV} \quad (\text{from } \sin^2\theta_W(m_Z)_{\overline{MS}}), \quad (32)$$

where the second and third errors correspond to $\Delta m_t = \pm 5.1$ GeV and $\Delta\alpha^{-1}(m_Z) = \pm 0.021$. Several features of those predictions are revealing. The first is that $\sin^2\theta_W(m_Z)_{\overline{MS}}$ currently gives a very good (best) determination of m_H. Note, however, that the uncertainties scale as the central value; so, the relatively small value, 79 GeV, helps reduce the uncertainties. Also, a larger $\Delta\alpha^{-1}(m_Z) = \pm 0.090$ would significantly increase the overall uncertainty.[27] In the case of m_W, one needs a better measurement of that parameter along with improvement in m_t, if it is to pinpoint m_H.

Taken together, Eqs. (31) and (32) are very suggestive of a relatively light Higgs scalar not far from the current LEPII bound from non-observation of $e^+e^- \to ZH$:

$$m_H > 89.8 \text{ GeV} \quad (\sqrt{s} = 183 \text{ GeV data}). \quad (33)$$

Preliminary studies of $\sqrt{s} = 189$ GeV e^+e^- data indicate that the bound will soon rise to ~ 95 GeV. Future upgrades to $\sqrt{s} \simeq 200$ GeV will push the Higgs discovery potential to ~ 105 GeV. In addition, searching for the Higgs via associated $W^\pm H$ and ZH at the Fermilab $p\bar{p}$ collider during Run II promises discovery up to $m_H \sim 115$ GeV, perhaps even higher. The Higgs discovery may soon be at hand.

2 The $Zb\bar{b}$ Problem

Currently, the only real anomaly in Z pole measurements seems to involve $b\bar{b}$ final states. The LEP $b\bar{b}$ forward-backward asymmetry and SLAC $b\bar{b}$ left-right, forward-backward asymmetry are consistent with a 3σ deviation for A_b from the Standard Model expectation:

$$\begin{aligned} A_b^{\rm exp}/A_b^{\rm theory} &= 0.94 \pm 0.02 \\ A_b &= \frac{g_L^2(b) - g_R^2(b)}{g_L^2(b) + g_R^2(b)} \end{aligned} \tag{34}$$

where $g_L(b)$ and $g_R(b)$ are the Z couplings to left and right-handed b quarks, normalized to have standard model values (at tree level):

$$\begin{aligned} g_L(b) &= -\frac{1}{2} + \frac{1}{3}\sin^2\theta_W \simeq -0.423, \\ g_R(b) &= \frac{1}{3}\sin^2\theta_W \simeq +0.077. \end{aligned} \tag{35}$$

At the same time, the quantity $R_b \equiv \Gamma(Z \to b\bar{b})/\Gamma(Z \to \text{hadrons})$ exhibits very good accord with standard model expectations:

$$R_b^{\rm exp}/R_b^{\rm theory} = 1.004 \pm 0.004, \tag{36}$$

and thereby leads to the constraint

$$g_L^2(b) + g_R^2(b) = 0.1858 \pm 0.0010. \tag{37}$$

Solving Eqs. (34) and (37) together gives

$$\begin{aligned} g_L^2(b) &= 0.1746 \pm 0.0020, \\ g_R^2(b) &= 0.0112 \pm 0.0018. \end{aligned} \tag{38}$$

In terms of deviations from the standard model, $\delta g_L(b)$ and $\delta g_R(b)$, one finds (ignoring the negative $g_R(b)$ and positive $g_L(b)$ solutions):

$$\begin{aligned} \delta g_L(b) &= 0.005 \pm 0.002, \\ \delta g_R(b) &= 0.0287 \pm 0.0088. \end{aligned} \tag{39}$$

The $\delta g_L(b)$ deviation on its own amounts to only a -1% shift and could probably be interpreted as a "new physics" quantum loop correction; however, such a large $\delta g_R(b)$ shift of 40% is very difficult to explain. For that reason, most theorists have dismissed the above 3σ deviation as experimental in origin, i.e., stemming from a statistical or systematic effect, rather than an indication of "new physics." Nevertheless, it is amusing to contemplate other potential consequences of non-zero $\delta g_L(b)$ and $\delta g_R(b)$ of the magnitude in Eq. (39). First, I note that deviations of similar magnitude cannot occur in $\delta g_L(d)$ and $\delta g_R(d)$; otherwise, they would have been observed in atomic parity violation and $\nu_\mu N$ experiments. Furthermore, it is unlikely that they are present in $\delta g_R(s)$ and $\delta g_L(s)$. If that were the case, one would expect (but could avoid) induced $s \to d$ flavor-changing weak neutral currents which could significantly enhance $K_L \to \mu^+\mu^-$, $K^+ \to \pi^+\nu\bar{\nu}$, etc.; and that seems not to be the case. If one concludes that the anomaly occurs only in $Zb\bar{b}$, it is still likely that related new flavor changing $b \to s$, $b \to d$, and $s \to d$ weak neutral currents would occur. The predicted magnitude of those effects depends on the degree and nature of quark mixing; however, generically interesting observable consequences almost certainly result. It will be interesting to see if anomalies in $Z \to b\bar{b}$ asymmetries persist as the data is further scrutinized and whether FCNC b (and K) decays will be in accord with Standard Model expectations or also exhibit anomalies.

3 Muon Decay and the S, T, U Parameters

As previously discussed, muon decay provides a very precise determination of $G_\mu = 1.16637(1) \times 10^{-5}\,\text{GeV}^{-2}$ which contains within it potential "new physics" effects. For example, heavy chiral fermions present in fourth generation models or technicolor theories would contribute to gauge boson self energies. Those loop effects would show up in the Δr, $\Delta r(m_Z)_{\overline{MS}}$ and $\Delta\hat{r}$ of Eq. (26) as additional contributions. One way to unveil or constrain such effects is to define Fermi constants in terms of α, $\sin^2\theta_W(m_Z)_{\overline{MS}}$, m_W, and m_Z via Eq. (26):

$$G_F^{(1)} = \frac{\pi\alpha}{\sqrt{2}m_W^2(1 - m_W^2/m_Z^2)(1 - \Delta r)}, \tag{a}$$

$$G_F^{(2)} = \frac{\pi\alpha}{\sqrt{2}m_W^2 \sin^2\theta_W(m_Z)_{\overline{MS}}(1 - \Delta r(m_Z)_{\overline{MS}})}, \tag{b (40)}$$

$$G_F^{(3)} = \frac{4\pi\alpha}{\sqrt{2}m_Z^2 \sin^2 2\theta_W(m_Z)_{\overline{MS}}(1-\Delta\hat{r})}. \qquad (c)$$

Comparison of those quantities with G_μ tests the standard model and probes for possible "new physics" in the Δr, $\Delta r(m_Z)_{\overline{MS}}$, and $\Delta\hat{r}$. If there is no "new physics" one should find $G_\mu = G_F^{(1)} = G_F^{(2)} = G_F^{(3)}$.

To examine the situation, we take $m_H = 125$ GeV as our central value and allow for the range $75 < m_H < 200$ GeV. Then from the values in table 1 and the measured m_W, m_Z, $\sin^2\theta_W(m_Z)_{\overline{MS}}$ and α, one finds:

$$\begin{aligned} G_F^{(1)} &= 1.1676(55) \times 10^{-5}\,\text{GeV}^{-2}, && (a) \\ G_F^{(2)} &= 1.1671(21) \times 10^{-5}\,\text{GeV}^{-2}, && (b) \quad (41) \\ G_F^{(3)} &= 1.1672(14) \times 10^{-5}\,\text{GeV}^{-2}, && (c) \end{aligned}$$

where the uncertainties reflect errors in m_t, m_H, $\Delta^{-1}\alpha(m_Z)$, m_W, and $\sin^2\theta_W(m_Z)_{\overline{MS}}$. The excellent agreement between those quantities and $G_\mu = 1.16637(1) \times 10^{-5}\,\text{GeV}^{-2}$ obtained from muon decay is quite remarkable. It shows no indication of "new physics." Note also that the uncertainty in even the most precise $G_F^{(3)}$ is more than 100 times the current error in G_μ. Hence, improving G_μ further would not sharpen such tests; improving $\sin^2\theta_W(m_Z)$, m_W, m_t and measuring m_H would.

As an example of the utility of Eq. (41), consider the deviations expected from heavy chiral fermion doublets. The appendage of such particles to the standard model modifies gauge boson self-energies. Those effects shift the radiative corrections in Eq. (26). Such shifts are conveniently parametrized by the S, T, and U parameters of Peskin and Takeuchi:[28,29]

$$\begin{aligned} \delta\Delta r &= 0.0166S - 0.0258T - 0.0195U, && (a) \\ \delta\Delta r(m_Z)_{\overline{MS}} &= 0.0084(S+U), && (b) \quad (42) \\ \delta\Delta\hat{r} &= 0.011S - 0.00782T. && (c) \end{aligned}$$

Hence, if S, T and $U \neq 0$, one expects the relationships:

$$\begin{aligned} G_\mu &= G_F^{(1)}(1 + 0.0166S - 0.0258T - 0.0195U), && (a) \\ G_\mu &= G_F^{(2)}(1 + 0.0084(S+U)), && (b) \quad (43) \\ G_\mu &= G_F^{(3)}(1 + 0.011S - 0.00782T). && (c) \end{aligned}$$

In technicolor models, one has the generic prediction[28] $S \sim \mathcal{O}(+1)$ which would lead to about a one percent difference between the G_μ and $G_F^{(i)}$. However, Eq. (41) exhibits no such effect. In fact it constrains that quantity at $\mathcal{O}(0.1\%)$. A global fit to all electroweak data (for $m_H \simeq 100$ GeV) gives[30]:

$$\begin{aligned} S &= -0.17^{+0.17}_{-0.12}, && (a) \\ T &= -0.16^{+0.15}_{-0.18}, && (b) \quad (44) \\ U &= 0.19 \pm 0.21. && (c) \end{aligned}$$

Those constraints are consistent with Eqs. (43) and (41). If one assumes $m_H \sim \mathcal{O}(1$ TeV) as would be more appropriate for technicolor, one finds $S = -0.29 \pm 0.14$, which is even more incompatible with $S \sim \mathcal{O}(+1)$. Therefore, if dynamical electroweak symmetry breaking is to be consistent with precision measurements, the dynamics must be very novel to render $S \sim 0$.

4 Extra Dimensions and $W^{*\pm}$ Bosons

Another interesting "new physics" scenario involves excited $W^{*\pm}$ bosons which may arise in theories with extra compact dimensions[31] (Kaluza-Klein excitations) or models with composite gauge bosons. Assuming fermionic couplings to $W^{*\pm}$ identical to those of $W^\pm$ and $g_2^* = g_2$, direct searches at the Tevatron lead to the bound[32]

$$m_{W^*} > 720\,\text{GeV} \qquad (95\%\,\text{CL}). \qquad (45)$$

If such bosons exist, they would also contribute to low energy charged current processes such as muon decay and be incorporated into G_μ. They would replace g_2^2/m_W^2 in the decay amplitude by $g_2^2/\langle m_W^2\rangle$ where

$$\frac{1}{\langle m_W^2\rangle} = \frac{1}{m_W^2} + \frac{(g_2^*/g_2)^2}{m_{W^*}^2} + \frac{(g_2^{**}/g_2)^2}{m_{W^{**}}^2} + \cdots \qquad (46)$$

As long as the signs are positive, the effective low energy mass $\langle m_W\rangle$ is always smaller than m_W, since the reciprocal sum acts like resistors in parallel. Therefore, if W^* bosons exist, G_μ should be larger than the $G_F^{(i)}$ in Eq. (41). However, there is no such indication. Quantitatively, one expects

$$G_\mu = G_F^{(i)} \left(1 + C \left(\frac{g_2^*}{g_2}\right)^2 \frac{m_W^2}{m_{W^*}^2}\right), \tag{47}$$

where

$$C = 1 + \left(\frac{g_2^{**}}{g_2^*}\right)^2 \frac{m_{W^*}^2}{m_{W^{**}}^2} + \cdots > 1. \tag{48}$$

In the simplest extra dimension theory, one might typically expect $C = \sum_{n=1}^{\infty} 1/n^2 = \pi^2/6$. More realistic scenarios can lead to even larger C. Here, I am interested only in lower bounds on m_{W^*}; so, C will not enter as long as $C \geq 1$.

Comparing Eqs. (47) and (41b) gives (for $m_H \lesssim 200$ GeV):

$$m_{W^*} > 1.67 \left(\frac{g_2^*}{g_2}\right) \text{ TeV} \qquad (95\% \text{ CL}) \tag{49}$$

which is very constraining. It suggests that the radius of the extra dimensions $R \simeq 1/m_{W^*} < 1 \times 10^{-17} (g_2/g_2^*)$ cm. and that continuing searches for W^* bosons at the Tevatron are likely to yield null results. Of course, the bound can be significantly relaxed if $g_2^* << g$.

One can improve the bound in Eq. (49) by employing $G_F^{(3)}$ rather than $G_F^{(2)}$ in comparison with G_μ. One finds

$$m_{W^*} > 2.27 \left(\frac{g_2^*}{g_2}\right) \text{ TeV} \qquad (95\% \text{ CL}). \tag{50}$$

However, that bound is subject to a larger dependence on m_t, m_H, and "new physics" effects. Further improvements in $\sin^2\theta_W(m_Z)_{\overline{MS}}$ and m_W could push the m_{W^*} sensitivity to $\mathcal{O}(5$ TeV), which is competitive with LHC capabilities.

5 Conclusion

Precision electroweak measurements have tested the standard model at the $\pm 0.1\%$ level. As a byproduct, they have been used to predict the large top quark mass and now suggest a relatively light Higgs,[30]

$$m_H < 255 \text{ GeV} \qquad (95\% \text{ CL}), \tag{51}$$

with values around 100 GeV favored. Discovery of the Higgs scalar may be close.

The good agreement between theory and experiment severely constrains the possible "new physics" one can append to the standard model. For example, the S parameter must be near zero. That finding leaves little room for additional chiral fermion doublets, such as a fourth generation of fermions, and requires dynamical symmetry breaking scenarios to exhibit novel dynamics which respects that constraint (a difficult task). Other types of "new physics," such as relatively large extra dimensions, SUSY, Z' bosons, etc., are also being constrained by such measurements. So far, there are no signs of "new physics." Nevertheless, we must continue to probe shorter distances and search for new phenomena. Surprises are certainly waiting to be unveiled.

References

[1] R. S. van Dyck Jr., P. B. Schwinberg, and H. G. Dehmelt, Phys. Rev. Lett. **59**, 26 (1987).

[2] A. Czarnecki and W. Marciano, preprint hep-ph/9810512.

[3] G. Gabrielse and J. Tan, in *Cavity Quantum Electrodynamics*, edited by P. R. Berman (Academic Press, San Diego, 1994), p. 267.

[4] W. Marciano, Phys. Rev. D**20**, 274 (1979).

[5] M. Davier and A. Höcker, Phys. Lett. B**439**, 427 (1998).

[6] S. Eidelman and F. Jegerlehner, Z. Phys. C**67**, 585 (1995).

[7] F. Jegerlehner in *Proceedings of the IVth International Symposium on Radiative Corrections*, Barcelona, Spain (1998).

[8] W. Marciano and A. Sirlin, Phys. Rev. Lett. **46**, 163 (1981).

[9] J. Erler, Phys. Rev. D **59**, 054008 (1999).

[10] S. Berman and A. Sirlin, Ann. Phys. **20**, 20 (1962).

[11] T. Kinoshita and A. Sirlin, Phys. Rev. **113**, 1652 (1959).

[12] S. Berman, Phys. Rev. **112**, 267 (1958).

[13] M. Roos and A. Sirlin, Nucl. Phys. B**29**, 296 (1971).

[14] T. van Ritbergen and R. Stuart, Phys. Rev. Lett. **82**, 488 (1999).

[15] M. Veltman, Nucl. Phys. B**123**, 89 (1977).

[16] W. Marciano, Nucl. Phys. B**84**, 132 (1975).

[17] C. Bollini, J. Giambiagi, and A. Sirlin, Nuovo Cimento A**16**, 423 (1973).

[18] A. Sirlin, Phys. Rev. Lett. **67**, 2127 (1991); S. Willenbrock and G. Valencia, Phys. Lett. B**259**, 373 (1991); R. Stuart, Phys. Lett. B**262**, 113 (1991).

[19] P. Gambino and A. Sirlin, Phys. Rev. **49**, R1160 (1994).

[20] A. Sirlin, Phys. Rev. D**22**, 471 (1980).

[21] W. Marciano and A. Sirlin, Nucl. Phys. B**189**, 442 (1981).

[22] CCFR Collaboration, K. McFarland *et al.*, Eur. Phys. J. C**1**, 509 (1998).

[23] W. Marciano and A. Sirlin, Phys. Rev. D**22**, 2695 (1980).

[24] G. Degrassi, S. Fanchiotti, and A. Sirlin, Nucl. Phys. B**351**, 49 (1991).

[25] G. Degrassi, P. Gambino, M. Passera, and A. Sirlin, Phys. Lett. B**418**, 209 (1998).

[26] G. Degrassi, P. Gambino, and A. Sirlin, Phys. Lett. B**394**, 188 (1997).

[27] P. Gambino in *Proceedings of RADCOR98*, preprint hep-ph/9812332.

[28] M. Peskin and T. Takeuchi, Phys. Rev. Lett. **65**, 964 (1990); Phys. Rev. D**46**, 381 (1992).

[29] W. Marciano and J. Rosner, Phys. Rev. Lett. **65**, 2963 (1990).

[30] J. Erler and P. Langacker, preprint hep-ph/9809352.

[31] V. A. Kostelecky and S. Samuel, Phys. Lett. B**270**, 21 (1991); I. Antoniadis, Phys. Lett. B**246**, 377 (1990); I. Antoniandis, K. Benakli, and M. Quiros, Phys. Lett. B**331**, 313 (1994); G. Chapline and R. Slansky, Nucl. Phys. B**209**, 461 (1982).

[32] PDG Tables 1998; Euro. Phys. J. C**3**, 1 (1998).

Evidence for a Cosmological Constant and for the Expansion of the Universe*

GERSON GOLDHABER for the Supernova Cosmology Project†
Institute for Nuclear and Particle Astrophysics
Lawrence Berkeley National Laboratory
& Center for Particle Astrophysics
University of California at Berkeley
Berkeley, California 94720
gerson@lbl.gov

*This work was supported in part by the United States Department of Energy, contract numbers DE-AC03-76SF00098, CfPA, and NSF contract number AST-9120005.

†G. Aldering,[a] B. J. Boyle,[b] P. G Castro,[a] W. J. Couch,[c] S. Deustua,[a] S. Fabbro,[d] R. S. Ellis,[e] A. V. Filippenko,[f] A. Fruchter,[g] G. Goldhaber,[a] A. Goobar,[h] D. E. Groom,[a] I. M. Hook,[i] M. Irwin,[e] A. G. Kim,[j] M. Y. Kim,[a] R. A. Knop,[a] J. C. Lee,[e] T. Matheson,[f] R. G. McMahon,[e] H. J. M. Newberg,[n] C. Lidman,[o] P. Nugent,[a] N. J. Nunes,[a] R. Pain,[d] N. Panagia,[g] C. R. Pennypacker,[a] S. Perlmutter,[a] R. Quimby,[a] P. Ruiz-Lapuente,[k] B. Schaefer,[l] and N. Walton.[m] (*a*) LBNL, Berkeley, California, USA, (*b*) Anglo-Australian Observatory, Sydney, Australia, (*c*) University of New South Wales, Sidney, Australia, (*d*) LPNHE, CNRS-IN2P3, University of Paris VI & VII, Paris, France, (*e*) Institute of Astronomy, Cambridge, United Kingdom, (*f*) Dept. of Astronomy, Univ. of California, Berkeley, USA, (*g*) Space Telescope Science Inst., Baltimore, Maryland, USA, (*h*) Dept of Physics, Stockholm Univ., Sweden, (*i*) Inst. of Astronomy, Univ. of Edinburgh United Kingdom, (*j*) L.P.C.C. College de France, Paris, France, (*k*) Dept. of Astronomy, Univ. of Barcelona, Spain, (*l*) Dept. of Astronomy, Yale Univ., New Haven, Connecticut, USA, (*m*) Isaac Newton Group, La Palma, Spain, (*n*) FNAL, Batavia, Illinois, USA, (*o*) ESO, La Scilla, Chile.

© 1998 by Gerson Goldhaber.

ABSTRACT

A search for cosmological supernovae has discovered over 75, most of which are type Ia supernovae (SNe Ia). There is strong evidence from measurements of nearby type Ia supernovae that they can be considered as distance indicators or "standard candles" after correction for the width (time-scale stretch parameter) of the individual light curves. Measurements and analysis are completed on 42 of these distant, $z = 0.18$ to 0.83, supernovae. These supernovae, together with 18 "nearby" $z < 0.11$ supernovae from the Calán/Tololo Supernova Survey, allow us to measure the ratio of the matter density of the universe to the critical density Ω_M, together with the normalized cosmological constant Ω_Λ, the energy density of the vacuum. For a flat universe, i.e., $\Omega_M + \Omega_\Lambda = 1$, we obtain $\Omega_M = 0.28\ ^{+0.09}_{-0.08}(stat.)\ ^{+0.05}_{-0.04}(syst.)$. The data are strongly inconsistent with a $\Lambda = 0$ flat cosmology, the simplest inflationary universe model. An open, $\Lambda = 0$ cosmology also does not fit the data well: the data strongly suggest that the cosmological constant is nonzero and positive, with a confidence of $P(\Lambda > 0) > 99\%$. By treating 36 of our high-z SNe Ia with $0.30 < z < 0.70$ as a group, we find that the expected light curve widths, based on the nearby width distribution, are in agreement with the observed distribution. In particular, we demonstrate that when normalized, our SNe fit a single curve with a single parameter stretch, s. We also obtain a direct confirmation of the hypothesis that extragalactic redshifts are due to cosmological expansion by observing the expected "time dilation" Doppler shift $1 + z$ of SNe Ia light curves at high redshift.

1. The Matter Density and Vacuum Energy Density of the Universe

When Einstein introduced his General Theory of Relativity (1916) the universe was assumed to be static. To accomplish such a static universe, Einstein had to introduce a repulsive force corresponding to a "cosmological constant," Λ, to compensate for the gravitational attraction of the matter in the universe. Later, when Hubble discovered that the universe was actually expanding (1924), Einstein called the cosmological constant "his greatest mistake." All the same, since Λ is consistent with Einstein's theory, one need not assume that $\Lambda = 0$. Thus, the value of Λ becomes an experimental question. Here Ω_Λ, the vacuum energy density, is given by $\Omega_\Lambda = \Lambda/(3H_0^2)$ and Ω_M is the matter density of the universe. In our current study, we primarily obtain a linear combination of these two densities.

When we started out with this study ten years ago, we were planning to measure the deceleration of the universe q_0. On the then prevalent assumption (prejudice) that $\Omega_\Lambda = 0$, we expected a positive deceleration of the universe due to matter density. With our present observations we find, to a high probability $P(\Lambda > 0) > 99\%$, that q_0 is actually negative, corresponding to an acceleration of the expansion of the universe! Our best fit corresponds to the relation $0.8\Omega_M - 0.6\Omega_\Lambda \approx -0.2 \pm 0.1$. For a flat cosmology, $\Omega_M + \Omega_\Lambda = 1$, this corresponds to $\Omega_M = 0.28\ ^{+0.09}_{-0.08}(stat.)\ ^{+0.05}_{-0.04}(syst.)$.

Before our present work started, estimates of Ω_M, based on a variety of experiments, ranged from 0.05 to 1.5. Since the light emitting matter contributes roughly 0.01 to Ω_M, a measurement of Ω_M also determines the contribution of dark matter to the universe.

2. Type Ia SNe as Calibrated Standard Candles?

There is good evidence that type Ia supernovae (SNe Ia), the brightest of all the different types of SNe, can be calibrated to have a standard brightness. A plausible explanation for this behavior is that SNe Ia are the consequence of the explosion of a white dwarf star as it approaches a critical mass of 1.4 solar masses, the Chandrasekhar limit.

A white dwarf is a star that has burnt all of its hydrogen and helium by fusion primarily to carbon and oxygen, and as a result has collapsed under the gravitational force to a degenerate electron gas in which the C and O nuclei are embedded. The white dwarf thus has the mass of the order of the mass of the sun, but a radius comparable to that of the earth. If this white dwarf is in a binary system with another star, a common occurrence, the companion star can transfer matter to the white dwarf until it approaches the Chandrasekhar limit. Near this mass, the electron degeneracy pressure can no longer support the star and a runaway thermonuclear explosion occurs.

In this supernova explosion, the temperature and density reach the point at which the C and O nuclei fuse and produce higher mass nuclei. The fusion process stops at ^{56}Ni, a radioactive isotope with an equal number of protons and neutrons. In fact, an enormous amount of ^{56}Ni is produced, typically 0.6 of a solar mass. In the explosion, the newly produced material is ejected with velocities of about 10,000 km/sec to 30,000 km/sec. At first, the exploding star is too small to be seen. Over a period of a few days, however, it expands rapidly and its brightness reaches a maximum value, in about 18 days, in the rest system. It is during this period that we first discover the SN as we will describe below. The light observed is produced by ionization from the decay products of ^{56}Ni and its daughter isotope ^{56}Co. The ^{56}Ni isotope has a half-life of about six days and ^{56}Co has a half-life of 77 days, so it is primarily the ^{56}Co decay that is powering the supernova at maximum light. Finally, ^{56}Co decays to stable ^{56}Fe.

All the same, not all SNe Ia have exactly the same brightness at the peak of the light curve. There is a dispersion of about 0.24 to 0.5 magnitudes (depending on the sample selection criteria) in the maximum brightness distribution.[1] Phillips[2] noted a correlation between the brightness and the decline rate of the SNe Ia. We express this correlation (Perlmutter *et al.*, 1997, Ref. 3) in terms of a stretch parameter s applied to the time axis in the light curve of the SN. Thus, s affects both the rise rate and the decline rate of the SN. The greater s, meaning the wider the SN light curve, with reference to the standard Leibundgut[4] curve, the brighter it is. We measure typical values for s from 0.8 to 1.2 with a few outliers above and below these values. We find a correction to the B-band absolute magnitude M of $-\alpha(s-1)$. In our analysis, we apply this correction to the apparent magnitude m_B giving an effective magnitude $m_B^{\text{effective}} = m_B + \alpha(s-1)$, which then corresponds to a unique value of the absolute magnitude M with an uncertainty of approximately 0.17 magnitudes or ~17% in flux. (This dispersion can be improved still further with multiple filter-band constraints on extinction; see discussions in Perlmutter *et al.*[5] and references therein.) In theoretical models, it has been suggested that this light curve width-luminosity relation can be understood in terms of varying amounts of ^{56}Ni produced, and hence, variation in the temperature of the ejecta.

3. The Observation of 42 SNe

In an ongoing systematic search we have, over the last six years, discovered and studied 75 SNe. The majority of these were the most distant SNe ever observed. We have now completed the analysis of 42 of these SNe. The analysis of the first seven of our SNe has already been published (Perlmutter, 1997).[3] An additional SN at $z = 0.83$ has been measured with the Hubble Space Telescope and added to this sample (Perlmutter, 1998a, Ref. 6). The first analysis of a 40-supernova sample was presented at the January 1998 meeting of the American Astronomical Society (Perlmutter, 1998b, Ref. 7); The current, more extensive analysis[5] agrees with those results—as well as with a subsequent analysis of 16 SNe by the Hi-Z Supernova Search Team (Riess *et al.* 1998).[8]

When we began our program, only one high-redshift supernova had been discovered (18 days past maximum).[9] We developed a technique that made it possible to discover *batches* of high-redshift supernovae, while still on the rise to maximum light in a systematic, predictable way. About twice a year, during two sets of two nights at the telescope, our procedure is to take about 50–70 different CCD images, depending on the depth in redshift we are trying to achieve. Each image covers about 15 × 15 square minutes of arc on the sky. The total area covered is thus about 3-4 square degrees. The first set of images are considered as "reference" images. The second set observed about three weeks later, are the "search" images in which we hunt for the new light due to a supernova. The timing relative to the moon is critical, since we deal with such dim objects that they are difficult to observe at or near full moon. We take our reference images just after the new moon and our search images just before the next new moon. After the discovery, we take follow-up images of the SNe as well as the spectra.

The new supernova in this sample of 42 were all discovered while still brightening, using the Cerro Tololo four-meter telescope with the 2048^2 prime-focus CCD camera or the 4×2048^2 big throughput camera. The supernovae were followed with photometry over the peak of their light curves, and approximately two to three months further (~ 40–60 days in their restframe) using the CTIO 4-m, WIYN 3.6-m, ESO 3.6-m, INT 2.5-m, and WHT 4.2-m telescopes. (SN 1994ap and other more recent supernovae have also been followed with HST photometry.) The supernova redshifts and spectral identifications were obtained using the

Keck I and II 10-m telescopes with LRIS and the ESO 3.6-m telescope. We have observed spectra of the host galaxies for all of these. SNe spectra were observed for almost all SNe after the first six SNe. The photometry coverage was most complete in Kron-Cousins R-band, with Kron-Cousins I-band photometry coverage ranging from two or three points near peak to relatively complete coverage paralleling the R-band observations.

4. The Light Curve and K-Corrections

We compare our measurements to nearby SN light curves measured with a blue filter. Because of the large redshift, the spectral features captured in the blue filter appear in the "red." We thus carry out our measurements with a red filter and then translate our measurements into the blue. As it turned out, this correction—called a K-correction—was particularly straightforward for our first distant SN because $1 + z = 1.458$ gives the ratio of the redshifted wavelength to the wavelength at emission. This value turned out to be the ratio of the central wavelength transmitted by our red filter to that of the blue filter, used for the nearby standard light curve. The acceptance width of the two filters were also roughly in this same ratio. For other z values the K-correction is a function of the epoch on the light curve and the stretch. The procedure modifies the light curve to allow for changes in the portion of the spectrum captured by each filter, Kim *et al.*[11] and Nugent *et al.*[12]

5. Our Method for Obtaining Ω_M and Ω_Λ

Our method for obtaining both Ω_M and Ω_Λ is discussed in Perlmutter *et al.*[3,6] and Goobar and Perlmutter.[13]

To obtain Ω_M and Ω_Λ, we make use of the following measurements:

- The redshift of the supernova, or parent galaxy, z is related to the wavelength change $\lambda - \lambda_o$ of the receding object. Here, z is defined as $z = (\lambda - \lambda_o)/\lambda_o$ and λ_o is the wavelength in the rest system, while λ is the measured wavelength for identified lines. This requires taking a spectrum of the SN to identify type Ia SN whenever possible, as well as of the galaxy, for an accurate redshift determination.
- The apparent magnitude m_B of the high-redshift SN at the peak of its light curve, calibrated to a standard light curve width ($s = 1$). This is based on the measurement of the SN light intensity as a function of time (generally in two filters), the photometry calibration, K-corrections, and a fit to the light curve to obtain the peak brightness and the light-curve width s. (The two-filter measurements make it possible to characterize the extinction due to dust that reddens the light.)
- The same measurements obtained for a group of 18 well-measured nearby Type Ia SNe from the Calán/Tololo Supernova Survey.[14–16]

The relation between the measured apparent magnitude m_R and the absolute magnitude M_B for type Ia SN, is given by

$$m_R = 5 \log D_L + M_B + 25 + K_{RB}. \quad (1)$$

Here, K_{RB} is the K-correction factor between red and blue SNe magnitudes and $D_L = D_L(z; \Omega_m, \Omega_\Lambda, H_o)$ is the luminosity distance given in Mpc (≈ 3 million light years). Here, H_o is the Hubble constant, c the velocity of light, z is the measured redshift, usually obtained from the spectrum of the host galaxy. At first sight it appears that D_L depends on the Hubble constant H_o as well. However, it turns out that the Hubble constant cancels because H_o also appears in the determination of M_B obtained from type Ia SN measurements in nearby galaxies Hamuy *et al.*[14–16]

Figure 1 shows $m_B^{\text{effective}}$ for our 42 SNe, as well as the 18 Calán/Tololo Supernova Survey SNe, plotted against redshift. The curves represent the Hubble diagram for a variety of cosmological models as indicated on the figure. The fit to the expression for D_L gives an approximately linear relation between Ω_M and Ω_Λ in the region of interest ($\Omega_M < 1.5$). Our best fit is illustrated in Fig. 2. We have carried out a series of fits under somewhat different data selections, e.g., (a) all the data, (b) with two outlying SNe removed, (c) with two reddened and two extreme stretch SNe removed, etc. (see, Perlmutter *et al.*, 1998c, for details).[5] Our results are quite stable with respect to these various sample selections. In particular, one fit was made with *no* stretch corrections, and this yielded a fit completely consistent with our results *with* stretch corrections. We have used the

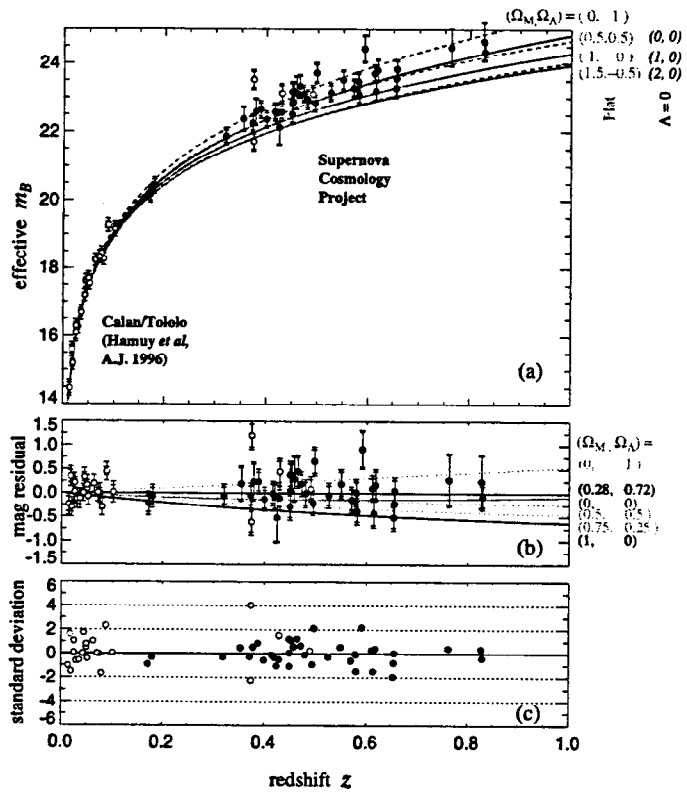


Figure 1: (a) Hubble diagram for 42 high-redshift SNe Ia from the Supernova Cosmology Project, and 18 low-redshift SNe Ia from the Calán/Tololo Supernova Survey, after correcting both sets for the SN Ia light curve width-luminosity relation. The inner error bars show the uncertainty due to measurement errors, while the outer error bars show the total uncertainty when the intrinsic luminosity dispersion, 0.17 mag, of light curve-width-corrected SNe Ia is added in quadrature. The unfilled circles indicate supernovae not included in Fit C. The solid curves are the theoretical $m_B^{\rm effective}(z)$ for a range of cosmological models with zero cosmological constant as indicated on the figure. (b) The magnitude residuals from the best-fit flat cosmology for the Fit C supernova subset, $(\Omega_{\rm M}, \Omega_\Lambda) = (0.28, 0.72)$. The curves are for a range of flat cosmological models as indicated on the figure. (c) The uncertainty-normalized residuals from the best-fit flat cosmology for the Fit C supernova subset, $(\Omega_{\rm M}, \Omega_\Lambda) = (0.28, 0.72)$.

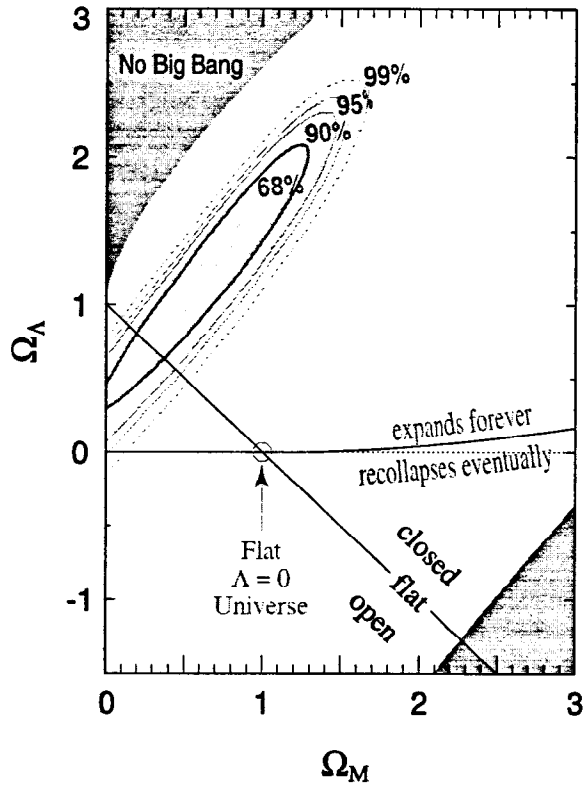


Figure 2: The 68%, 90%, 95%, and 99% confidence regions in the $\Omega_{\rm M}$–Ω_Λ plane from Perlmutter *et al.*[6] (see this reference for details of the fit procedure). (The table of this two-dimensional probability distribution is available at *http://www-supernova.lbl.gov/*.) In cosmologies above the near-horizontal line, the universe will expand forever while below this line the expansion of the universe will eventually come to a halt and recollapse. This line is not quite horizontal because at very high mass density there is a region where the mass density can bring the expansion to a halt before the scale of the universe is big enough that the mass density is dilute with respect to the cosmological constant energy density. The upper-left shaded region, labeled "no Big Bang," represents "bouncing universe" cosmologies with no Big Bang in the past. The lower right-shaded region corresponds to a universe that is younger than the oldest heavy elements, for any value of $H_0 \geq 50$ km s^{-1} Mpc^{-1}.

two-filter color measurements to study the possible effect of extinction due to dust that reddens the supernova light, and find that our results are robust with respect to this effect [again, see Perlmutter (Ref. 5) for details].

Our measurements have to be carried out at cosmological distances i.e., large values of z, since for small z values $z < 0.11$, the dependence on Ω is negligible and the luminosity distance becomes $D_L = cz/H_o$.

6. Type Ia Supernovae as Clocks at Cosmological Distances

We will now make use of the "standard" nature of the SNe Ia light curve shapes. This feature allows us to consider the SNe Ia as "clocks" at cosmological distances. As discussed above, SNe Ia are as a class highly homogeneous, since they are thermonuclear explosions that are apparently triggered under very similar physical conditions.

The work of Phillips,[2] Hamuy[14–16] and Riess, Press and Kirshner,[17] have emphasized the *inhomogeneity* of Type Ia light curve shapes, particularly for the "non-normal" redder Type Ia's. Perlmutter[3,5] provided a single-parameter characterization of these light curve differences, which is simply a time-axis stretch factor, s, which stretches (or compresses) the Leibundgut template light curve. From the study based on *nearby* SNe, it was shown that this stretch factor s extends over a range of 0.80 to 1.20 with an occasional outlier outside these values. Light curve shapes have been found to be correlated with peak magnitudes, UV colors, spectral features, etc.[3] We will now show data obtained by treating 36 of these SNe with $0.30 < z < 0.70$ as a group.

Using supernova light curves to test the cosmological expansion was first suggested by Wilson, 1939 (Ref. 18). Over the last decade, it has become clear that SNe Ia are superb and precise clocks, located both nearby and at cosmological distances.

We presented the first clear observation of this time-dilation effect based on our first seven high z type Ia SNe at the 1995 conference at Aigua Blava, Spain, Goldhaber.[19] Later, Leibundgut *et al.*[20] gave confirmatory evidence for time dilation, based on a single, high-z supernova. More recently, Riess *et al.*, 1997 (Ref. 21), showed evidence that the spectral features of SNe Ia can be timed sufficiently well to measure the time interval between two spectra taken ten days apart in the observer system. Applying this method to one SN gave results consistent with the time-dilation factor $1 + z$ at the 96.4% confidence level.

7. Fit to the Light Curve

In our data analysis, we fit each of our observed SNe light curve points to R-band template light curves using the fitting program MINUIT (James and Roos, 1995)[10] . Each R-band template is constructed from the B-band rest system, the Leibundgut template, and the appropriate cross filter K-correction.[11,12]

Each SN is fit for three variables: the day and magnitude of peak light and the width factor $w = s(1+z)$ by which the Leibundgut curve had to be expanded. In the present analysis, we make use of the day at peak to align all the 36 SNe light curves, and we normalize the flux of each SN to unity. As these SNe are measured in the R-band, we now translate the points to the B-band. The individual photometry points are K-corrected[11,12] from the measured R-band photometry points to the B-band, where lower z data is available for comparison.[14–16] Two SNe at $z = 0.18$ and three at $z > 0.7$ have to be K-corrected to the V-band and U-band, respectively, and are thus not included here. One distinct outlier SN97O is also left out here.

8. Analysis of 36 Supernovae Treated as a Group

The stretch factor affects our ability to test time dilation, since the expected dilation factor $1 + z$ is modified by the stretch factor s. The observable effect is then $w = s(1 + z)$. Figure 3(a) shows the distribution of w for SCP data (solid histogram) and Calán/Tololo data (dashed histogram). Figure 3(b) shows the corresponding histograms for s.

If no independent s corrections are made to our width measurements, an additional dispersion of w is observed, assuming a similar population of SNe at high redshift.

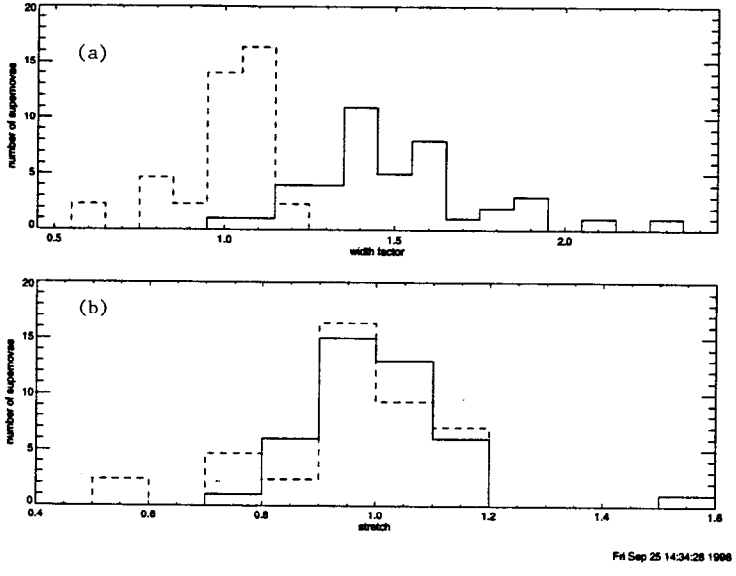


Figure 3: (a) The distribution of the observed light curve width w for the 18 of the Calán/Tololo SNe (broken line histogram) with an area normalized to 42 and the SCP data for 42 fully analyzed SNe (solid line histogram). (b) The corresponding stretch distributions s defined as the observed light curve width divided by $1+z$ for each SN.

Figure 4(a) shows the distribution of w versus $1+z$ for both the Calán/Tololo data and the SCP data. As can be noted, except for two outliers in the SCP data and three in the Calán/Tololo data, $0.8 < s < 1.2$. In Fig. 4(b), we show the stretch distribution s versus $1+z$.

In Fig. 5(a), we show the R-band photometry points, individually K-corrected to the B-band for the SNe with $0.30 < z < 0.70$, aligned at maximum light in the observer system. The measured points for SNe with $0.45 \leq z < 0.70$ are shown in red, the points from SNe with $0.30 < z < 0.45$ are shown in green. Also shown are points (violet) from the 18 SNe from the Calán/Tololo study in the B-band (Hamuy, 1996) with $z < 0.11$, these are also K-corrected. These 18 SNe were chosen out of 29 SNe by having been discovered prior to five days after maximum light. Figure 5(b) shows the same data averaged over all SNe in each z band. Figure 5(b) also shows the same data with the SCP data averaged over the entire $0.30 < z < 0.70$ interval and averaged over two days. Finally, Fig. 5(b) shows the rest system Leibundgut curve extended down to the explosion time (Nugent *et al.*, 1999, Ref. 22, Goldhaber *et al.*, 1999, Ref. 23). These figures illustrate the time dilation effect clearly. The width of the light curves are z dependent in the observer system. The nearby SNe (violet points) approximate the rest system curve, while the high-redshift data (green and red points) give wider distributions. This time dilation is particularly clear in Fig. 5(c). The (violet) low-z points follow the curve while the (green) outer points are clearly much wider, indicating time-dilation.

We now transform the time axis from the observer frame to the rest frame by dividing the appropriate $1+z$ for each data point of each of the SNe, and then divide the time axis by s as well. In Fig. 6(a) and 6(b), we note excellent overlap of all the individual photometry points. Thus with one parameter s and the time dilation $1+z$, all the normalized and aligned SNe fit a single curve. The relations between stretch, color and brightness as noted by Riess *et al.*[19] will be discussed in detail in terms of stretch in an upcoming publication (Kim *et al.*, 1999, Ref. 23).

9. Time Dilation and Doppler Shift–An Example

The time dilation effect can be understood in terms of the Doppler shift. We present a simple example of this.

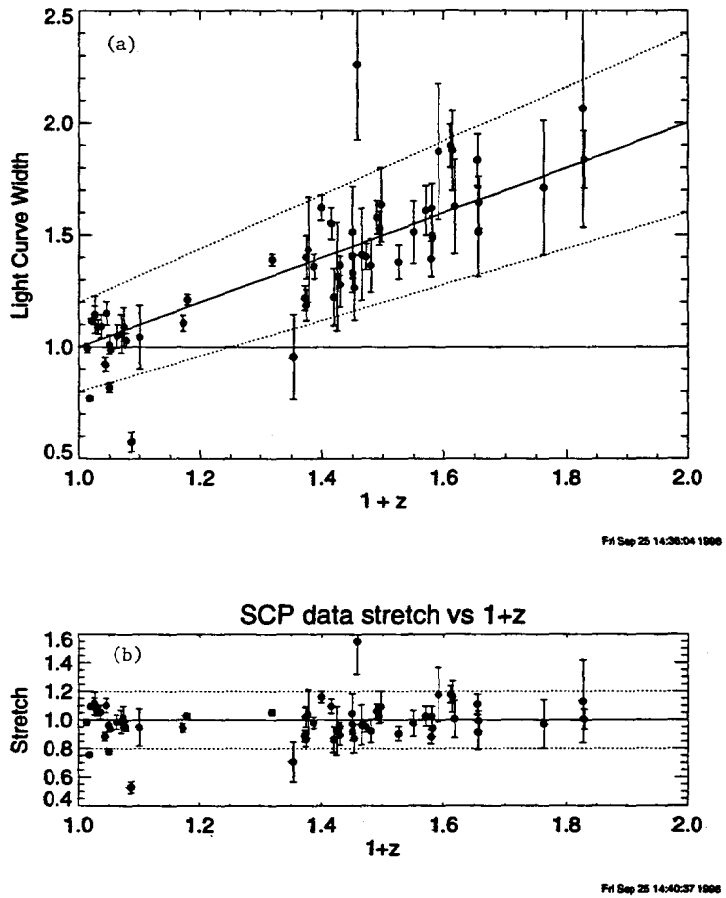


Figure 4: (a) The observed light curve width vs $1+z$. The (violet) points for $z < 0.11$ are the Calán/Tololo SNe, the (red) points for $z > 0.15$ are the SCP data for 42 fully analyzed SNe. The band delineated by the dotted lines corresponds to stretch values 0.8 to 1.2 which encompass the bulk of the data, except for a few outliers. (b) The stretch s vs $1+z$. Stretch is defined as the observed light curve width w divided by $1+z$ for each SN. The points and lines are defined as in (a).

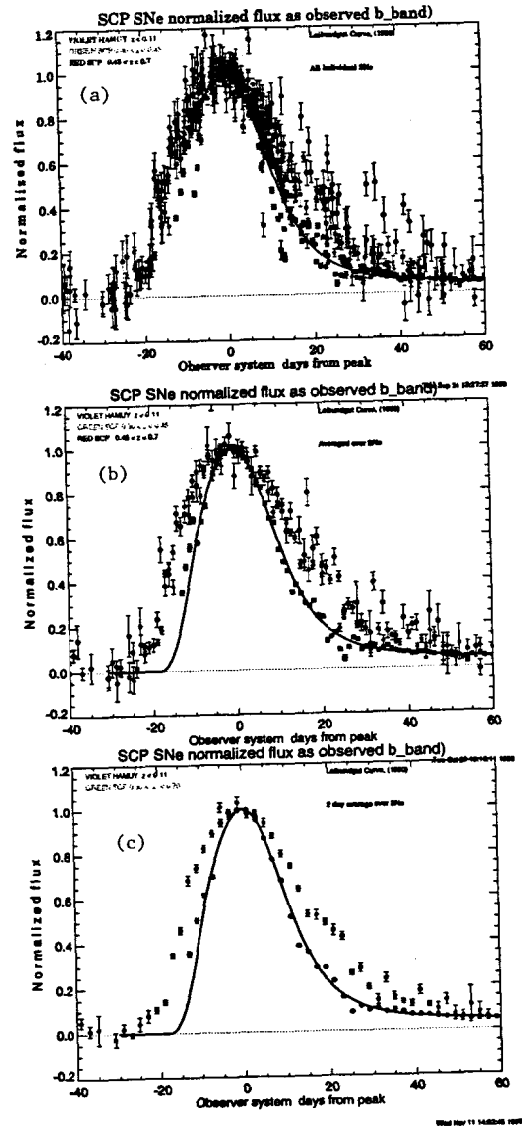


Figure 5: (a) The photometry points for the 34 SCP and 18 Calán/Tololo SNe, flux normalized to unity and aligned at the peak in the observer system. (b) The same data as in (a) averaged over each day. The Leibundgut light curve extended to explosion time[23,24] is also shown, to indicate the rest system. (c) The SCP data for $0.30 < z < 0.70$ and the Calán/Tololo SNe averaged over two days, compared with the extended Leibundgut light curve.

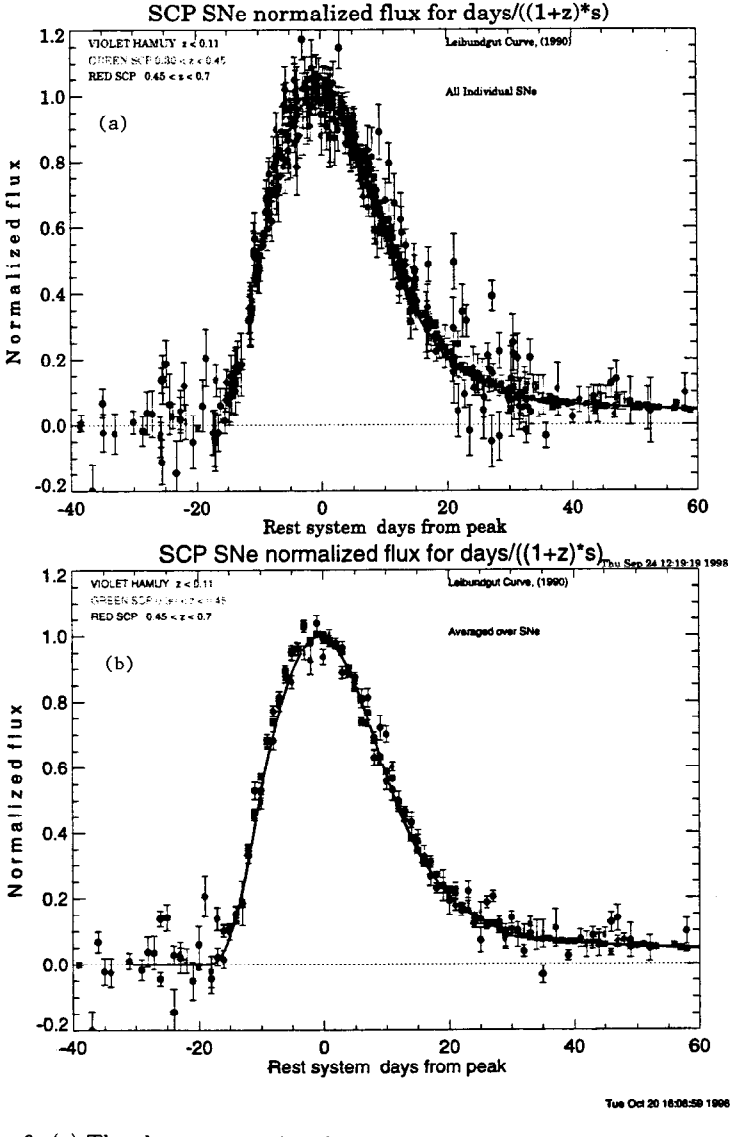


Figure 6: (a) The photometry points for the same three sets of SNe, transformed to the rest system and corrected for the stretch factor s. (b) The same data as in (a) averaged over each day. All the points are consistent with the extended Leibundgut curve[24] indicating that the single parameter s, combined with the expansion factor $1+z$, falls on a unique curve.

We note that the upper limit of the observed redshift $z = 0.7$ corresponds to $\beta = 0.49$ and that $\beta = 0.5$ corresponds to $z = 0.73$. Consider two points in time on the light curve symmetric around the peak and four weeks apart, in the SN rest system for $z = 0.73$. When the light emitted from the first point reaches us, we start our clock in the observer frame. When the light is emitted from the second point the SN will, by the expansion of the universe, be at a greater distance from our telescope. With $\beta = 0.5$, this extra distance will be two light weeks. From this classical effect alone, the two points in the observer system will be six weeks apart, a time dilation of 1.5. Thus, time-dilation measures the expansion of the universe directly. The relativistic version is given by $1 + z = \sqrt{(1+\beta)/(1-\beta)}$ which equals $\gamma(1+\beta)$, where $\gamma = 1/\sqrt{1-\beta^2}$. For the redshift values we are dealing with, the γ factor is only a small correction 1.15 to give 6.93 weeks in the observer frame. The main effect we are observing is thus the direct expansion of the universe. The same argument also holds for the redshift of the light emitted by the SN, where we now think of the two points as two adjacent crests of the light wave.

REFERENCES

1. D. Branch and G.A. Tammann, 1992. Ann. Rev. Astron. Astrophys. **30**, 359–89 (1992).
2. M. M. Phillips, Astrophys. J. **413**, L105 (1993).
3. S. Perlmutter *et al.*, Astrophys. J. **483**, 565 (1997).
4. B. Leibundgut, Supernova, edited by S. E. Woosley (Springer, New York, 1991) 751.
5. S. Perlmutter *et al.*, 1998c. Also available at *www-supernova.LBL.gov* and *astro-ph* (submittted to Astrophys. J.)
6. S. Perlmutter *et al.*, Nature **391**, 51 and erratum (on author list), **392**, 311 (1998a).
7. S. Perlmutter *et al.*, presentation at the *January 1998 Meeting of the American Astronomical Society,* available at *www-supernova.LBL.gov* and *astro-ph*; referenced in Bull. Am. Astron. Soc., vol. 29, p. 1351, 1997, (submittted to Astrophys. J.).
8. A. G. Riess *et al.*, Astrophys. J **116**, 1009 (1998).
9. Norgaard-Nielsen *et al.*, Nature **339**, 523 (1989).
10. F. James and M. Roos, Comput. Phys. Commun. **10**, 343 (1975).
11. A. Kim, A. Goobar, and S. Perlmutter, Publ. Astron. Soc. Pac. **108** 190 (1996).
12. P. Nugent *et al.*, Publ. Astron. Soc. Pac., 1999 (in preparation).
13. A. Goobar, and S. Perlmutter, Astrophys. J. **450**, 14 (1995).

14. M. Hamuy *et al.*, Publ. Astron. Soc. Pac. **105**, 787 (1993).
15. M. Hamuy *et al.*, Astron. J. **106**, 2392 (1993b).
16. M. Hamuy *et al.*, Astrophs. J. **109**, 1 (1995).
17. A. G. Riess, W. H. Press, and R. P. Kirshner, Astrophys. J. **438** L17 (1995). *ibid* **473**, 88 (1996).
18. O. C. Wilson, Astrophys. J. **90**, 634 (1939).
19. Goldhaber, G., *et al.*, in presentations at the *NATO ASI in Aiguablava*, Spain, (1995), LBL-38400, page III.1. Also published in Thermonuclear Supernova, edited by P. Ruiz-Lapuente, R. Canal, and J. Isern; Dordrecht, Kluwer, page 777 (1997).
20. B. Leibundgut *et al.*, Astrophs. J. Lett. **466**, L21 (1996).
21. A. G. Riess *et al.*, Astron. J. **114(2)**, 722 (1997).
22. P. Nugent *et al.*, Astrophys. J., 1999 (in preparation).
23. G. Goldhaber *et al.*, Astrophys. J., 1999 (in preparation).
24. A. Kim, G. Aldering *et al.* Astrophys. J., 1999 (in preparation).

DEEP INELASTIC SCATTERING AT HERA

Gregorio Bernardi*
LPNHE Paris
4 Place Jussieu, Tour 33 R.d.C.
Universités Paris VI et VII
75232 Paris CEDEX 05, France

Representing the H1 and ZEUS Collaborations

ABSTRACT

The inclusive measurements performed between 1994 and 1997 in Deep Inelastic Scattering at HERA are reviewed. The subjects covered include the measurement of the proton structure function $F_2(x, Q^2)$ at low and high Q^2, the comparison of F_2 with Next to Leading Order DGLAP evolution, and the determination of the longitudinal structure function F_L at high y. The study of the charm density in the proton and the extraction of F_2^c are presented, as well as different extractions of the gluon density $xg(x, Q^2)$ in the proton. The measurements of the double differential cross-sections at high Q^2 are shown both in neutral and charged current with an emphasis on the high x region. Measurements of the W propagator mass are presented.

*e-mail: gregorio@mail.desy.de

© 1998 by Gregorio Bernardi.

1 Introduction

The first two years after the commissioning of the first *ep* collider HERA in May 1992 were devoted in deep inelastic scattering (DIS) to the observation of the new processes which would become the main subject of study nowadays in DESY. The observation of the rise of the F_2^p proton structure function at low x[1-4] was the first of this type of measurement which is still progressing in terms of precision. The structure function results provide the opportunity to test QCD in new kinematic regimes, to extract the gluon and the charm density in the proton, and in the future, to measure α_S. The measurement of these quantities can also be made at HERA in an exclusive way by measuring the charm induced processes, or the rates of events with two or three jets. The observation of high Q^2 charged and neutral current events,[5,6] with very low statistics at that time, but in an excellent experimental environment, allowed a glimpse of the future of the HERA collider. The subsequent years (1995,1996,1997) are covered in this report. The increased integrated luminosity and the improvements of the detector allowed for the precise measurements of the phenomena mentioned above, besides many others. Indeed in 1995, the H1 collaboration upgraded the backward detectors of the experiment, introducing in particular a more precise drift chamber (BDC) to measure the polar angle of the scattered electron, and a new SPACAL calorimeter with better hadronic containment, better granularity, and better angular acceptance. Similarly, ZEUS upgraded the backward part of its detector by introducing a beam pipe calorimeter (BPC) which allows for the measurement of the scattered positron at Q^2 as low as 0.1 GeV^2. The results obtained have fulfilled the expectations and a new step in precision could thus be reached for the low Q^2 physics.

In this report, after an introduction in section 2 to deep inelastic scattering and its basic measurement issues, the following points will be reviewed: in section 3 the results on cross-sections and structure functions at low Q^2 are presented in the context of the transition between the DIS and photoproduction regime. In section 4, the structure function F_2 results in the intermediate Q^2 region (a few GeV^2) are given followed in section 5 by its QCD analysis. Section 6 is devoted to the longitudinal structure function F_L determination, section 7 to the measurement of the charm component of the structure function, and section 8 to the measurements of the gluon density $xg(x, Q^2)$ in the proton. In section 9 are presented the cross-section measurements at high Q^2 and the extraction of M_W from these data.

2 Deep Inelastic Scattering at HERA

2.1 Basics of DIS

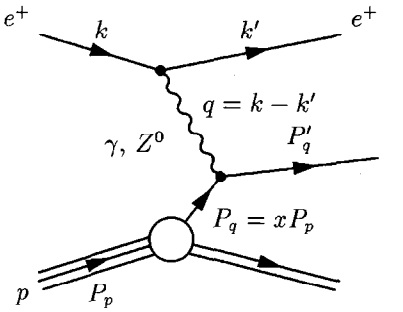
To define the kinematic variables used in DIS measurements, let us use the following graph which is an illustration of the basic DIS process. The exchanged boson is a photon in most of the cases, but the exchange of Z^0 or $W^\pm$ becomes sizable at high Q^2. The graph represents a neutral current event (NC). For a charged current event (CC) which occurs via $W^\pm$ exchange, the outgoing lepton becomes a neutrino, and the kinematics can thus be only reconstructed from the hadronic final state. The variables x, y, Q^2, and W^2 are defined in the following way:

4-momentum transfer:
$Q^2 = -q^2 = -(k - k')^2$

parton momentum fraction:
$x = Q^2/(2p \cdot q)$

fractional energy transfer:
$y = p \cdot q/(p \cdot k)$

mass^2 of hadronic system:
$W^2 = (p + q)^2 \approx Q^2/x.$



At low Q^2, *i.e.*, in the region in which the interaction proceeds essentially via photon exchange, the structure function F_2 of the proton can be derived from the double differential cross-section. In the naive Quark Parton Model (QPM) we have

$$\frac{d^2\sigma^{e^\pm p}}{dxdQ^2} = \frac{2\pi\alpha^2}{Q^4 x}\left[1 + (1-y)^2\right] F_2(x, Q^2) = \frac{2\pi\alpha^2}{Q^4 x} Y_+ F_2(x, Q^2) \qquad (1)$$

and F_2 is related to the quark densities:

$$F_2(x) = x\sum_{i=1}^{n_f} e_i^2 (q_i(x) + \overline{q}_i(x)). \qquad (2)$$

The longitudinal structure function F_L is equal to 0 in the QPM, following the Callan-Gross relation. In Quantum Chromodynamics (QCD) the presence of spin 1 partons (the gluons) changes the picture: F_2 becomes a function of x and Q^2, and a non-zero

F_L must be taken into account:

$$\left.\frac{d^2\sigma^{e^\pm p}}{dx\,dQ^2}\right|_{Born} = \frac{2\pi\alpha^2}{xQ^4}\left[Y_+F_2(x,Q^2) - y^2F_L(x,Q^2)\right]. \quad (3)$$

Not shown in the previous equations is the influence of the QED radiative corrections which are precisely known, with a precision of the order of 1%, and for which the cross-section is corrected. The case of DIS at higher Q^2 which involves W and Z exchange and electroweak radiative corrections will be treated in section 9.

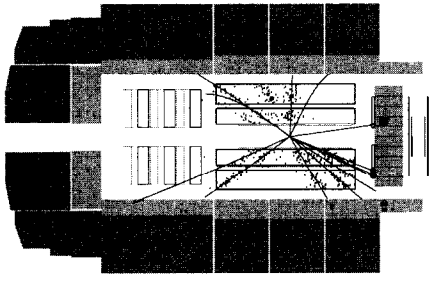
2.2 Kinematics and Measurement

Measured quantities in **inclusive** processes:

e^+: energy E'_e
polar angle θ_e

hadrons: $\Sigma = \sum_i (E_i - p_{z,i})$
$p_{T,h} = \sum_i p_{T,i}$
$\gamma_h = 2 \cdot \arctan \Sigma / p_{T,h}$

(The sums runs over all particles of the hadronic final state).



The selection of NC DIS events at HERA requires the detection of an electromagnetic cluster representing the scattered electron* above a certain energy threshold which is typically around 5 GeV, and the presence of a reconstructed interaction vertex in order to reject the beam associated background and to improve the reconstruction quality. Further requirements are applied depending on the analyses but we refer to the original papers for more details on this matter.[3,4] The selection of CC DIS events requires the detection of missing transverse energy which represents the energy carried away by the scattered neutrino. This selection is more difficult than in the NC case, since the energy of the hadronic final state is shared among many softer particles, hence the difficulty to have a low energy trigger, and to reject unambiguously the background which can come either from external sources (cosmic muons, parasitic muons produced by the off-beam protons) or from ep interactions, such as photoproduction events with jet production.

*HERA can run both with e^+ and e^-, but here "electron" is used both for e^+ and e^-.

The kinematics can be reconstructed at HERA with different methods for the neutral current (NC) events [electron only (e), Sigma (Σ), Double Angle (DA) or some combinations of these methods], but only one (hadrons only, h) for the charged current (CC). In general H1 uses for the NC the e, the Σ and the $e\Sigma$ methods, since they complement each other, while ZEUS uses the e method at low Q^2 and the DA or the PT methods at higher Q^2 since these last two methods are less sensitive to the reconstruction of the positron energy. The methods used for kinematic reconstruction are reviewed[7] and defined in the following way:

- Electron Method: $y_e = 1 - E'_e/E_e \sin^2\theta_e/2 \quad Q^2_e = 4E'_eE_e\cos^2\theta_e/2.$
  It is most precise at low x but quite sensitive to QED radiation, with corrections to the cross-section of the order of 30% in the absence of specific kinematic cuts. It has also an excellent Q^2 resolution in the full kinematic range, but a poor resolution in x at low y which necessitates the use of an alternative method in this region.
- Hadron Method[8]: $y_h = \Sigma/2E_e \qquad Q^2_h = p^2_{T,h}/(1-y_h).$
  This method does not allow for a precise determination of the kinematics but it is the only inclusive method for charged current events.
- Σ Method[9]: $y_\Sigma = \Sigma/(\Sigma + E'_e(1-\cos\theta_e)) \quad Q^2_\Sigma = E^2_e\sin^2\theta_e/(1-y_\Sigma).$
  The Σ method can be used in the full kinematic range since it has good x and Q^2 resolution both at high and low y. Furthermore, it is essentially independent of QED initial state radiation, thereby allowing an experimental cross-check of the radiative corrections by a comparison with the cross-section measurement done with the electron method.
- Double Angle method:[10] x_{DA}, Q^2_{DA} are determined from θ_e, γ_h.
  This method, which uses only the angle of the electron and the inclusive hadronic angle, leads to a high precision at high Q^2, but it is sensitive to QED radiation, and becomes imprecise at low γ_h, *i.e.*, at low y. Its strong point is to be independent of the electron energy and, in first order, of the hadronic energy scale. It is thus also used for calibration purposes.
- $e\Sigma$ Method[9]: $x_{e\Sigma} = x_\Sigma \qquad Q^2_{e\Sigma} = Q^2_e.$
  The simple combination of the advantages of the electron and Σ methods is very precise over the whole kinematic range, allowing in particular for an extension of the measurement to high x.
- PT Method[4]: x_{PT}, Q^2_{PT} are obtained from the DA method in which γ_h is

replaced by $\gamma_{PT} \equiv 2 \cdot \arctan \frac{2E_e \cdot \Sigma}{[\Sigma + E_e(1-\cos\theta_e)]p_{T,e}}$, not using therefore the unprecise measurement of $p_{T,h}$ from the hadronic final state.

Once the kinematics of the events are determined, the measurements rely on the following main points: The NC data are binned in (x, Q^2) with about 5 bins/order of magnitude in x and 10 bins/order of magnitude in Q^2. The CC data due to their lower resolution are binned with about 3 bins/order of magnitude in x and Q^2. The purity and stability[†] of an (x, Q^2) bin are requested to be typically greater than 30%, for the measurement to be done in that bin. From the selected events the *ep* background (mainly due to misinterpreted photoproduction events) is then subtracted. The acceptance corrections are performed using the simulation which describes well the data. The systematic errors associated with the measurement are split into a correlated and an uncorrelated contribution for each x, Q^2 bin. The total errors vary for NC cross-sections between 5% and 25% from low to high Q^2, for which the statistical error dominates. For the CC cross-sections, which are based on much lower statistics, the error range is between about 15% to 30%. The luminosity is known within 1.5–2.5% at HERA, depending on the year and the various beam conditions. The radiative corrections are known either from Monte Carlo or from analytical programs. Thus, after applying bin center corrections, the Born cross-section $\frac{d^2\sigma}{dxdQ^2}\Big|_{Born}$ can be measured at a given x and Q^2.

2.3 Overview of the F_2 Measurements at HERA

With the first HERA data taking in 1992, H1 and ZEUS have observed the strong rise of $F_2(x, Q^2)$ at low x at fixed Q^2. This rise has been confirmed with higher precision in 1993 and 1994,[3,4] and a good agreement between HERA and the older fixed target experiments has been observed. Next to Leading Order (NLO) QCD has been shown to give a good description of the data at Q^2 as low as 1.5 GeV^2 and has allowed for the extraction of the gluon density, which was also observed to rise towards low x.

H1 and ZEUS have brought with these measurements, and with the more recent measurements which will be described in more detail below, a major extension of the kinematic range of the F_2 measurement, as can be observed in Fig. 1, permitting a significant contribution to the understanding of the proton structure and of the theory of strong interactions.

[†]The stability (purity) is defined as the fraction of events which originate from a bin and which are reconstructed in it, divided by the number of generated (reconstructed) events in that bin.

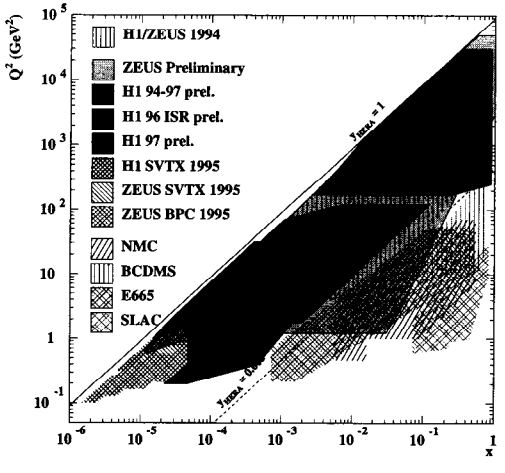


Figure 1: *Kinematic plane indicating the x and Q^2 regions of proton structure function measurements from HERA and fixed target experiments.*

- The extension to low x, below 10^{-5}, has permitted us to test successfully the low x limit of perturbative QCD.
- The extension to high x, up to 0.65 at $Q^2 > 500$ GeV^2, made possible the probing of the valence quark region at high Q^2.
- Reaching y as low as 0.005 provided F_2 measurements in overlap with fixed target experiment ones, which were thus shown to be consistent without extrapolation.
- The high y region measured at medium Q^2 ($\simeq$ 3–100 GeV^2), up to 0.82, displayed the sensitivity of the cross-section to F_L.
- By measuring very low Q^2 events, down to $Q^2 = 0.1$ GeV^2, the transition region from photoproduction to DIS has been explored.
- The increasing integrated luminosity permits us to approach the HERA kinematic limit, and at the highest Q^2 presently reached (30000 GeV^2) in NC events, the sensitivity to electroweak effects ($\gamma - Z^0$ interference) has been demonstrated.

We will now review the recent results obtained in '95, '96, and '97 in three main regions of Q^2, the low Q^2 region which extends from 0.1 to a few GeV^2, the intermediate Q^2 region, from a few GeV^2 to about 100 GeV^2, and the high Q^2 region, up to 30000

GeV^2. This subdivision corresponds to three main physics regimes of deep inelastic scattering.

3 The Low Q^2 Region

Experimentally, since Q^2 is well determined by the scattered electron measurement with $Q^2 = 4E'_e E_e \cos(\frac{\theta_e}{2})$, we can see that this kinematic regime can be optimally explored by[11,12]:

- Shifting the interaction vertex towards the forward direction (defined as the proton direction), by slightly changing the parameters of the collider, hence increasing the angular acceptance for the electron.
- Adding dedicated low angle detectors (BPC in ZEUS) or upgrading the backward calorimeter to improve, among others, its hermeticity (SPACAL in H1).
- Selecting "radiative" events, *i.e.*, events in which the incident lepton has emitted a photon before the deep inelastic collision, which have thus a lower center of mass energy, hence a lower Q^2, for a given measured angle and energy. The results using this technique are not reviewed in this paper.

These three techniques have been used by the two HERA experiments to measure F_2 at low Q^2 ($Q^2 \lesssim 5$ GeV^2).

3.1 $F_2(x, Q^2)$ at Low Q^2

In Fig. 2 are shown the measurements taken in shifted vertex position by H1[11] and ZEUS[12] in 1995. The main observations are that F_2 rises at low x, even at the lowest Q^2 = 0.11 GeV^2 measured at HERA. The lowest x point is reached at 6×10^{-6}. The rise becomes less pronounced when Q^2 decreases, and no discontinuous behavior can be observed from the data.

From a theoretical point of view, two approaches are attempting to describe this low Q^2 regime.

- The Regge approach as in the Donnachie-Landshoff model.[15] Here the total cross-section $\sigma^{tot} \propto W^{2\alpha_P}$, α_P is Q^2 independent and determined from hadron data ("soft Pomeron hypothesis"). This approach describes successfully the photoproduction data.

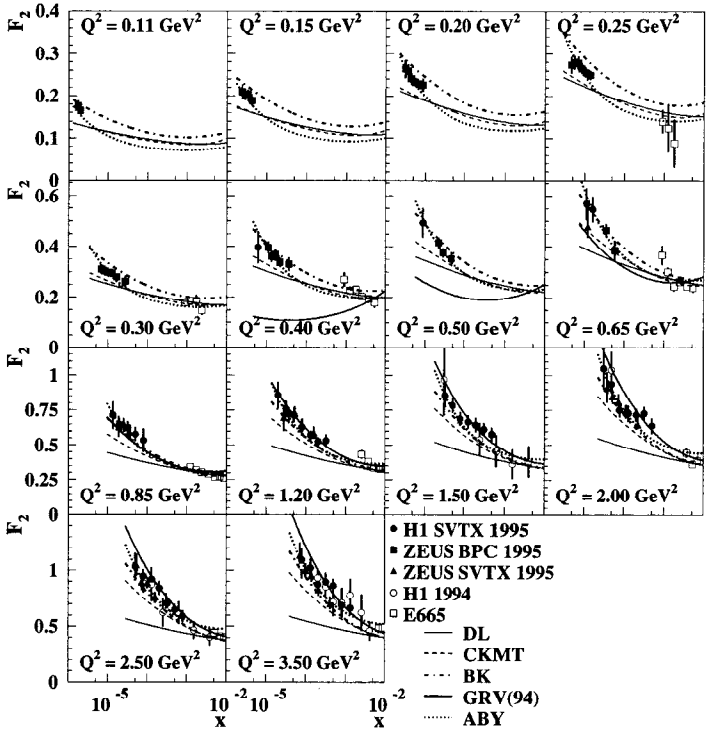


Figure 2: *F_2 measurements at very low Q^2 compared to various parametrizations. SVTX refers to data taken in shifted vertex position. BPC refers to data taken in shifted vertex position with the BPC detector (see text).*

- The perturbative QCD (pQCD) approach which is well known to describe successfully the DIS regime at high Q^2. It has been conjectured in the GRV[16] model that using valence-like partons at a very low starting scale ($Q_0^2 \sim 0.4$ GeV^2) and evolving them according to the DGLAP[17] evolution equation may describe successfully the DIS data at low Q^2.

The following conclusions can be drawn from the comparisons in Fig. 2 of the results with the GRV and DL models (the comparison to the other models will be commented on in the next section).

- pQCD (*i.e.*, GRV) describes the data down to $Q^2 \lesssim 1$ GeV^2. The breakdown of perturbative QCD happens at lower Q^2 than expected, but is visible at $Q^2 \simeq 0.5$–0.8 GeV^2.
- The DL Regge Model prediction is too low at low x, even at very low Q^2, but it describes the data at $Q^2 \lesssim 2$ GeV^2 and $x \gtrsim 10^{-3}$.

Thus there is no obvious way to reconcile these two approaches, and the transition region is not clearly visible in these measurements.

A better picture can be achieved by interpreting these F_2 measurements as a cross-section measurement of the virtual photon – proton ($\gamma^* p$) interaction. Indeed in the Hand convention for the flux of virtual photons,[13] the total $\gamma^* p$ cross-section can be related to the F_2 structure since at low x the lifetime of the virtual photon in the proton rest frame is large compared to the $\gamma^* p$ interaction time.[14] The ep scattering is thus interpreted as a $\gamma^* p$ scattering with total cross-section:

$$\begin{aligned} \sigma_{tot}^{\gamma^* p}(W^2, Q^2) &= \sigma_L^{\gamma^* p} + \sigma_T^{\gamma^* p} \simeq \frac{4\pi\alpha^2}{Q^2} F_2 \equiv \sigma_{eff}^{\gamma^* p}(W^2, Q^2) \\ W^2 &\approx Q^2/x \quad \text{at low } x. \end{aligned}$$

These measurements on $\sigma_{eff}^{\gamma^* p}$ at low Q^2 ($\approx 0.1 - 5$ GeV^2) can be compared to the photoproduction total cross-section $\sigma_{tot}^{\gamma p}$ determined at HERA at $Q^2 \approx 0$ in order to understand the transition between the photoproduction and the DIS regimes.

The transition between these two descriptions of the data is visible in Fig. 3a. The $\gamma^* p$ cross-section decreases steeply in the DIS regime, while it is expected to stay constant in the photoproduction regime. The smooth turning point happens at Q^2 between 0.1 and 1 GeV^2 as can be seen by the measurements at W between 115 and 245 GeV (corresponding, at $Q^2 = 1$ GeV^2, to x between $1.7 \cdot 10^{-5}$ and $0.8 \cdot 10^{-4}$). Further results at $Q^2 < 0.1$ GeV^2 will become available with the analysis of more recent data and will complement these measurements and provide a complete coverage of the transition region.

A further requirement on the models is to be able to describe the W^2 behavior of the total cross-section, as shown in Fig. 3b.

3.2 Models for the Transition DIS $\rightarrow \gamma$p

The new data obtained at HERA revived the theoretical attempt to describe the DIS $\rightarrow \gamma$p transition and additional models were developed. Two different non-perturbative

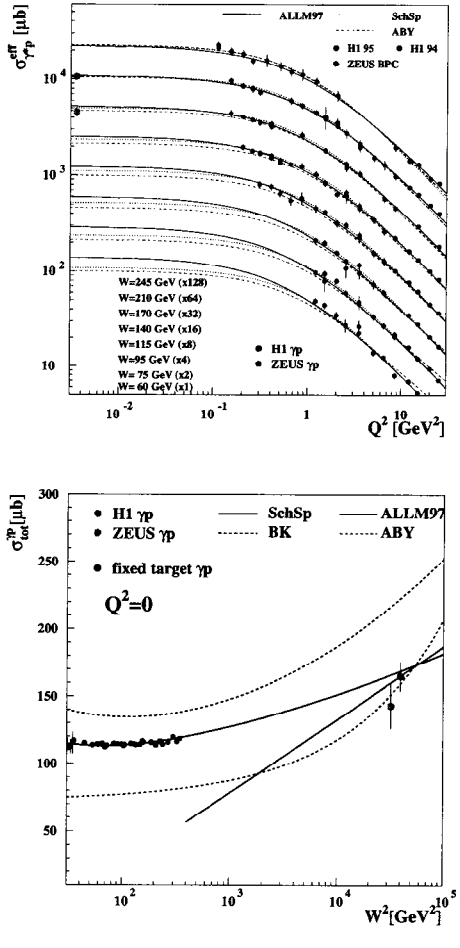


Figure 3: *$\gamma^* p$ cross-section as a function of Q^2 for different values of W. The transition between the photoproduction and the DIS behavior is visible between 0.1 and 1 GeV^2 (left). Total γp cross-section as a function of W^2 (right). The data (points) are compared with different parameterizations (see text).*

approaches have been considered to understand this transition, namely Regge theory (as briefly sketched above) and the vector meson dominance model (VMD). VMD relates the hadronic interactions of the photon to a sum over interactions of the ρ^0, ω and ϕ vector meson states.[18,19] To accommodate deep inelastic scattering data, the sum has to be extended to an infinite number of vector mesons giving the generalized vector dominance model (GVMD).[20]

Since neither the non-perturbative VMD and DL approaches nor pQCD are describing the Q^2 behavior of F_2 over the complete range from photoproduction to very large Q^2 deep inelastic scattering (DIS), the following models mix these approaches.

- The Badelek and Kwiecinski (BK) model uses a VMD approach at low Q^2 and pQCD at Q^2 above 1 GeV^2. It gives a good description in the transition region (Fig. 3a), but predicts a too large photoproduction cross-section (Fig. 3b).
- Schildknecht and Spiesberger (SchSp)[21] have revived the GVMD approach. The domain of application of their model is expected to be $0<x<0.05$, $0<Q^2<350$ GeV^2 and $W>30$ GeV. However, the photoproduction behavior is not described in this model, as shown in Fig. 3b.
- Adel, Barreiro, and Yndurain (ABY)[23] assume that pQCD is applicable to the lowest values of Q^2. To modify the behavior of F_2 which becomes flat as a function of x for Q^2 values below 1 GeV^2, a hard contribution ($\propto x^{-\lambda_s}, \lambda_s = 0.48$) is introduced, preventing the flattening at low x even at low Q^2 values. This model gives a fair description of the transition region, but fails in the photoproduction region.
- Abramowicz *et al.* (ALLM97)[22] assume that the total $\gamma^* p$ cross-section consists of two contributions which distinguish Reggeon and Pomeron exchange, and the power λ is assumed to vary with Q^2, in a similar way as pQCD in the high Q^2 region. This model currently describes all available data, but, contrarily to the others, makes use also of the photoproduction data to constrain its parameters.
- A Regge fit (including Pomeron and Reggeon contributions)[24] gives a good description of the very low Q^2 ZEUS data (which have been extrapolated to $Q^2 = 0$ using a GVDM fit) and of the low W γp data (see Fig. 4a). The resulting α_P is 1.010 ± 0.002. The photoproduction total cross-sections measured at HERA are not well described by this fit (Fig. 4b).

In conclusion, the Regge or pQCD approach cannot describe the data on its full range, but models such as ABY,[23] SchSp,[21] or ALLM97,[22] which are decribed in the next section and which all include the HERA data when fitting their parameters can give a good description of the data but for the total photoproduction cross-section. However, the ALLM97 parametrization, which makes use also of the low W photoproduction data, can give a good description of the complete W dependence of the γp cross-sections, as can be seen in Fig. 3.

3.3 Evolution of the Slope of $F_2(x, Q^2)$ vs x

The effective x slope at fixed Q^2 has been fitted at low x assuming $F_2 \propto x^{-\lambda}$ (for $x < 0.1$ in H1 or $x < 0.01$ in ZEUS). We thus have

$$\lambda \approx \frac{d \ln F_2}{d \ln 1/x}. \tag{4}$$

The precision obtained on λ with the recent data (H1 1997) is typically $\simeq 1\%$(stat) $\oplus 5\%$(syst). The rise of F_2 is thus quantified, and this effective slope decreases from values of 0.3 at $Q^2 \simeq 30$ GeV^2 down to 0.1 at $Q^2 \simeq 1$ GeV^2 displaying a smooth transition between the pQCD regime where λ is rapidly evolving as a function of Q^2 and the γp regime where λ is essentially constant.

4 Measurements of $F_2(x, Q^2)$ in the DIS Regime

New high precision measurements of $F_2(x, Q^2)$ with the NC data collected in 1996-97 by H1 and ZEUS have been released in a preliminary form in 1998.[25,26] Depending on the kinematic region and on the experiment, the integrated luminosity of the data varies between 7 and 27 pb^{-1}. The analyses follow closely the strategy adopted for the 1994 published papers; however, the higher statistics have made possible a substantial reduction of the total errors on the measurements. Below 100 GeV^2, they are typically below $\approx 1\%$ for the statistical error and about 4% for the systematic error, except at high y for which the systematic error grows up to about 12%. The main sources and the magnitude of systematic errors are:

- the electron energy scale: $\frac{\delta E_e}{E_e} = 0.5$–$1\%$,
- the hadronic energy scale: $\frac{\delta E_h}{E_h} = 2$–$3\%$,
- the electron identification: 1–2%,
- the photoproduction background: 1–2% at high y,

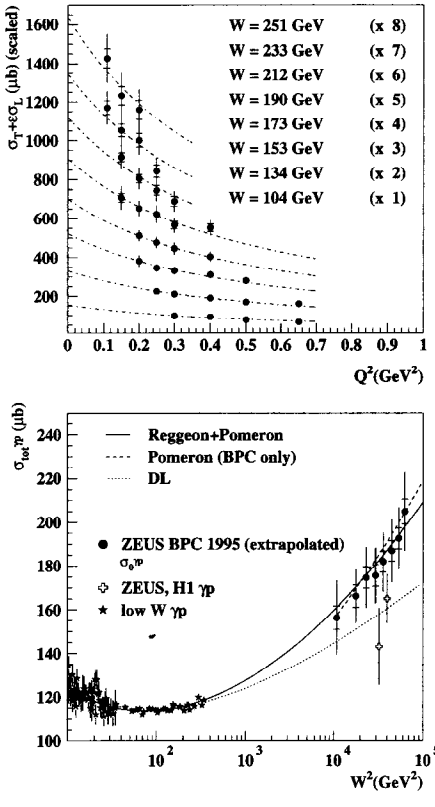


Figure 4: *Upper plot.* ZEUS BPC measurements of the total cross-section $\sigma_T + \epsilon\sigma_L$ in bins of W as a function of Q^2 and the GVMD fit to the data. *Lower plot.* $\sigma_{tot}^{\gamma p}$ as a function of W^2. The ZEUS BPC95 points are extrapolated to $Q^2 = 0$ using the GVMD fit. Also shown are the direct measurements of the total photoproduction cross-section from H1, ZEUS, and earlier experiments at low energies. The curves show Regge fits: the original DL fit[15] to the low W data (dotted); the Pomeron only fit to the BPC $\sigma_0^{\gamma p}$ data (dashed) and the Pomeron+Reggeon fit to the low W and BPC $\sigma_0^{\gamma p}$ data (full). See the ZEUS Collaboration[24] for more details.

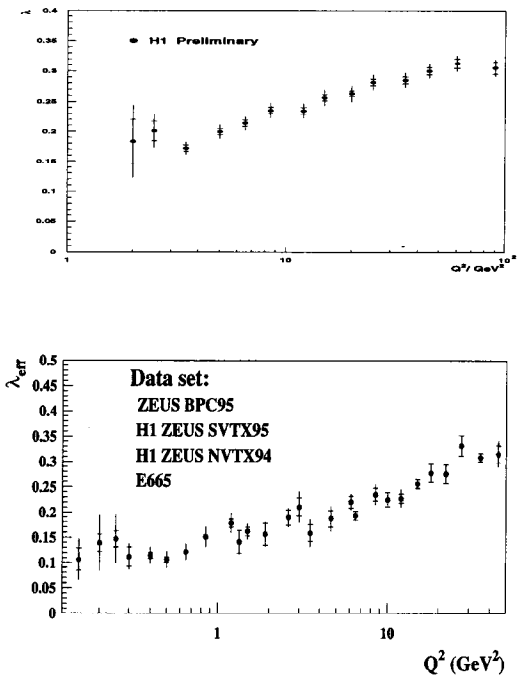


Figure 5: *Effective slopes of $F_2 \propto x^{-\lambda}$ for $x < 0.1$ (H1) and $x < 0.01$ (ZEUS).*

- the radiative corrections: 1–2%,
- the tracking efficiencies and vertex reconstruction 1–2%.

The ZEUS preliminary results are shown in Fig. 6 at fixed x as a function of Q^2. The coverage in x has been extended to higher x, yielding an increased overlap with the fixed target experiments, in which the agreement is good. The measurements are now performed over four orders of magnitude in Q^2 and x with relative precision better than 10% in most of the bins. Similar results have been obtained by H1, and are discussed below in conjunction with their QCD interpretation. In Fig. 7 are shown the ZEUS preliminary results at fixed Q^2 as a function of x for Q^2 between 1.5 and 27 GeV^2 (see also section 9 for results at higher Q^2). The rise in F_2 for $x \to 0$ is measured with improved precision.

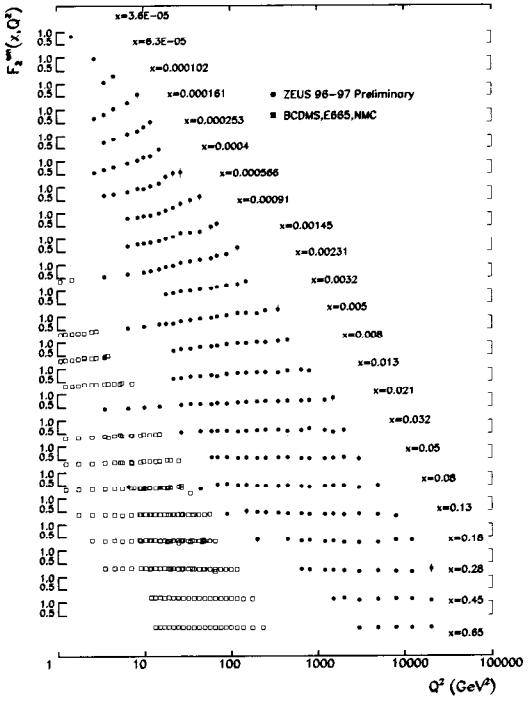


Figure 6: *Preliminary ZEUS measurements of F_2 as a function of Q^2 for fixed x from the 1996–1997 data (solid dots). The open rectangles show the results from the fixed target experiments.*

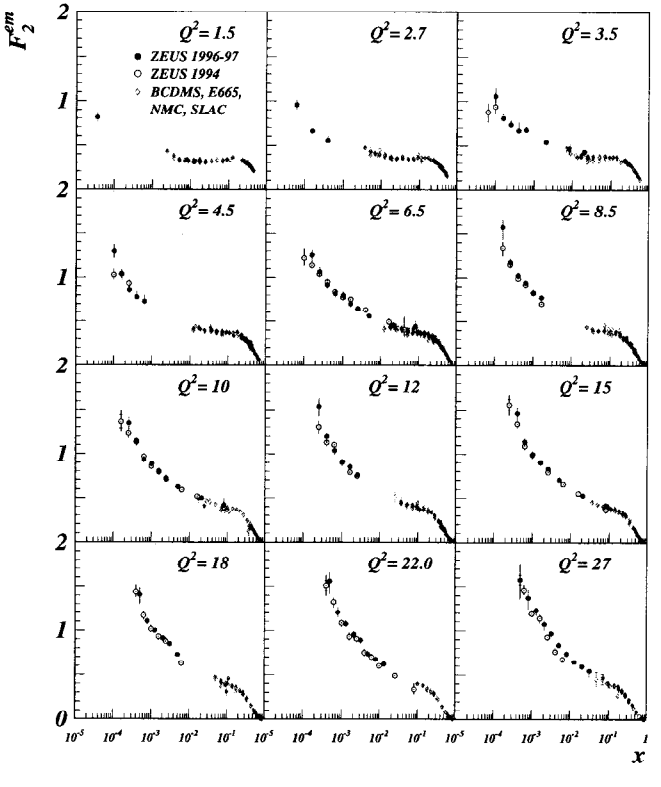


Figure 7: *Preliminary ZEUS measurements of F_2 as a function of x for fixed Q^2 values below 30 GeV^2 from the 1996-1997 data (solid dots). The diamonds show the results from the fixed target experiments. The inner error bars show the statistical errors, the outer error bars the quadratic sum of statistical and systematic errors.*

The extended precision obtained at HERA with the larger integrated luminosity allows for precise tests of perturbative QCD once they are used in combination with the fixed target experiment results. For a better understanding of the subsequent data interpretation, we will introduce the basics of QCD evolution equations and the QCD analyses performed by the two collaborations.

4.1 DGLAP Equations in QCD

Meaningful perturbative QCD tests necessitate going from the leading order expression for F_2 to its NLO version which relates it to the gluon density in the proton $xg(x)$:

$$\begin{array}{lll} \text{Leading Order:} & F_2(x) & = x\sum_{i=1}^{nf} e_i^2(q_i+\bar{q}_i) \\ \text{NLO } (\overline{\text{MS}}): & F_2(x,Q^2) & = x\sum_{i=1}^{n_f} e_i^2 C_q \otimes (q_i+\overline{q_i}) + C_g \otimes g \end{array}$$

in which the coefficient functions $C_{q,g}$ have been calculated to NLO. The singlet and non-singlet parton densities are defined as

$$\begin{array}{lll} x\Sigma & = \sum_{i=1}^{n_f}(xq_i + x\overline{q_i}) & \text{Singlet} \\ xq_{NS} & = \sum_{i=1}^{n_f}(xq_i - x\overline{q_i}) = xu_{val} + xd_{val} & \text{Non Singlet} \end{array}$$

and the DGLAP evolution equations of parton densities are given by:

$$\begin{array}{lll} \frac{\partial}{\partial t} q_{NS}(x,t) & = & \frac{\alpha_s(t)}{2\pi} P_{qq}^{NS} \otimes q_{NS}(y,t); \quad t = \log\frac{Q^2}{\Lambda^2} \\ \frac{\partial}{\partial t}\left(\Sigma(x,t), g(x,t)\right) & = & \frac{\alpha_s(t)}{2\pi}\left(P_{qq}, P_{qg}, P_{qg}, P_{gg}\right) \otimes \left(\Sigma(y,t), g(y,t)\right). \end{array}$$

The splitting functions P_{ij} are also known to NLO. A major property of these equations is the prediction of the Q^2 slope of $F_2(x,Q^2)$ for given parton densities at $Q^2 = Q_0^2$. However, no prediction on the x dependence at Q_0^2 is available from the theory, and this dependence must be obtained from a QCD fit to the data.

4.2 The H1 and ZEUS NLO-QCD Fit Programs

These programs use several assumptions which are often common to other widespread QCD analysis programs, like those from the MRS[59] or the CTEQ[48] groups.

- u, d, s are considered massless partons.
- the c density is obtained from Boson Gluon Fusion (BGF) calculated at NLO (Riemersma *et al.*) in the $\overline{MS}$ scheme, with the charm mass $m_c = 1.5$ GeV and the scale $\mu^2 = Q^2 + m_c^2$.

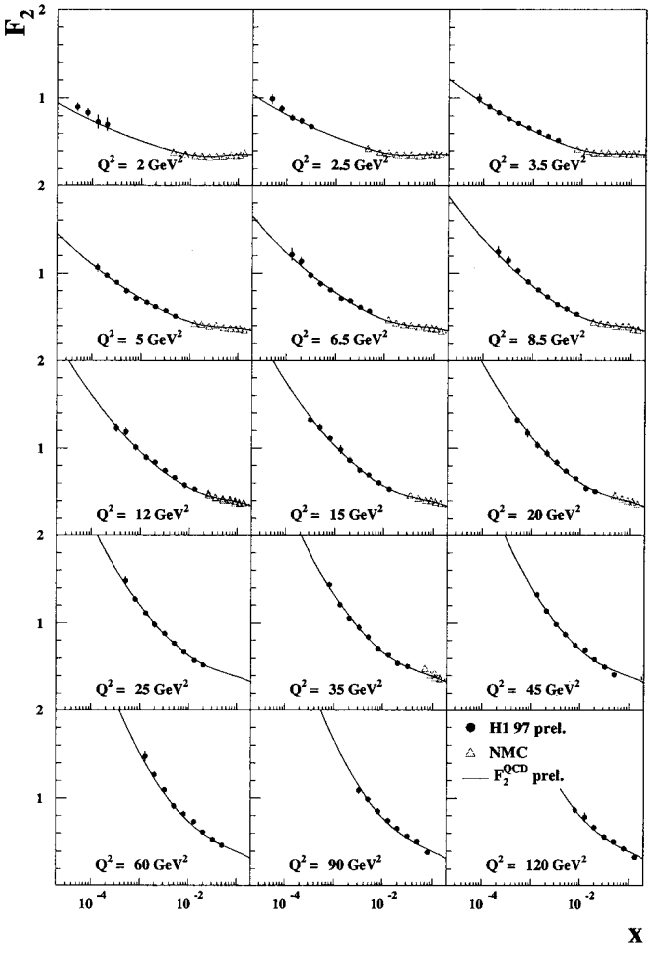


Figure 8: *H1 preliminary measurements on the 1997 data compared to the H1 NLO DGLAP fit which describes the data in the complete kinematic range.*

- A complete treatment of the correlated systematic errors is done, a particularity of the H1 and ZEUS QCD analysis.
- The theoretical uncertainties from α_s and m_c are taken into account.
- The Momentum Sum rule is assumed to be true: $\int_0^1 dx\, x(g+\Sigma) = 1$.
- The Quark Counting Rules are applied: $\int_0^1 dx\, u_v = 2$, $\int_0^1 dx\, d_v = 1$.
- The sea parton densities $\bar{u}$ and $\bar{d}$ are assumed equal. The strange parton density is assumed to be equal to 20% of the total sea parton density at Q_0^2.
- The data with x > 0.5 at low Q^2 are not used.

The input parton densities are parameterized as:

- $xg(x,Q_0^2) = A_g x^{B_g}(1-x)^{C_g}(1+D_g x+E_g\sqrt{x})$,
- $xu_v(x,Q_0^2) = A_u x^{B_u}(1-x)^{C_u}(1+D_u x+E_u\sqrt{x})$,
- $xd_v(x,Q_0^2) = A_d x^{B_d}(1-x)^{C_d}(1+D_d x+E_d\sqrt{x})$,
- $xS\,(x,Q_0^2) = A_s x^{B_s}(1-x)^{C_s}(1+D_s x+E_s\sqrt{x})$.

The datasets used in the QCD analysis are

- for ZEUS: ZEUS data (94+95) at $Q^2 > 1$ GeV^2, NMC[45] and BCDMS[60] p, d data,
- for H1: H1 data (94+95+96+97) at $Q^2 > 2$ GeV^2, NMC[45] and BCDMS[60] p, d data.

The H1 (ZEUS) experiment uses a Q_0^2 starting scale 2 (7) GeV^2. In both cases $\alpha_s(M_Z^2)$ is set to 0.118. For the H1 fit the gluon parametrization uses only three parametes, *i.e.*, $D_g = 0$ and $E_g = 0$ is assumed.

5 $F_2(x,Q^2)$ Compared to NLO QCD

To extract F_2 from the NC double differential cross-section, an assumption on F_L is made, based on QCD, which has however only a small influence on the results. In Fig. 8, the H1 preliminary measurements on the 1997 data are shown together with the NLO DGLAP fit which is describing them. The description is good in the whole phase space, but for the two lowest x points at 2 and 2.5 GeV^2 in Q^2 which display a tendency to lie above the QCD fit. Final results from H1 and ZEUS and final QCD analyses are needed however to draw strong conclusions from this observation. Also visible in the figure is the now well-known behavior of the rise of F_2 at low x, which becomes more

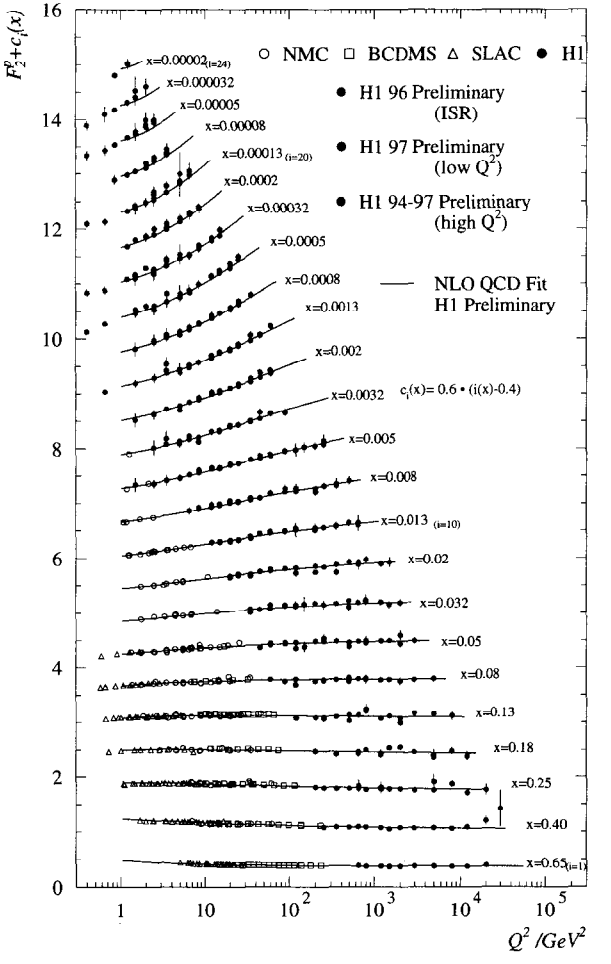


Figure 9: *Measurement of F_2 as a function of Q^2 in bins of fixed x. The H1 preliminary data are compared to the fixed target data, and to the H1 NLO QCD fit.*

steep when Q^2 increases. From this consequence of the DGLAP equations, Glück *et al.*[16] predicted the rise at low x of F_2 before any data at HERA was collected and without having to invoke BFKL effects.[27]

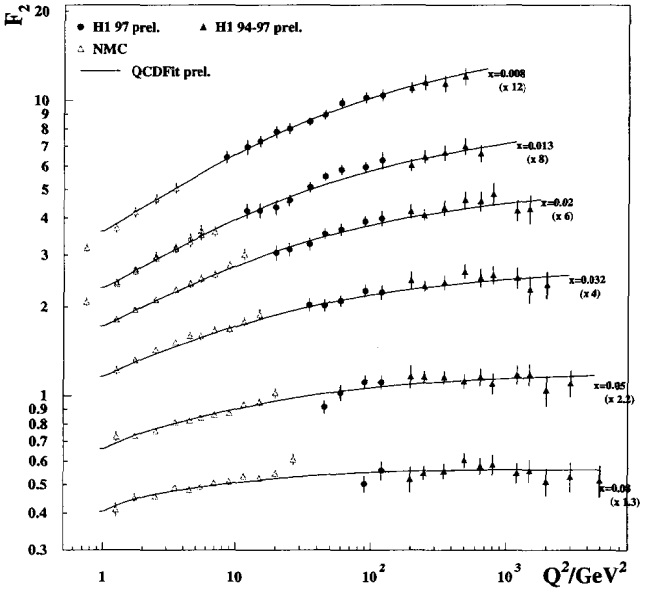


Figure 10: *Measurement of F_2 as a function of Q^2 in bins of fixed x between $x = 0.008$ and $x = 0.08$. The H1 preliminary data are compared to the NMC data, and to the H1 NLO QCD fit.*

In Fig. 9 the H1 data are shown at fixed x as a function of Q^2 and compared to the QCD fit. Also included in the plot are the high Q^2 preliminary measurements at high Q^2 on the 1994-97 data which will be described below. A good description of the data by the NLO DGLAP fit is realized over more than four orders of magnitude in x and Q^2. The figure shows clearly the famous scaling pattern at $x \simeq 0.1$ and the scaling violations which becomes stronger when going to low x due to the larger gluon radiation.

In Fig. 10 an enlargement of the same plot is made in the region where scaling is starting to be realized, around $x \simeq 0.1$. On this plot we can observe that the precision on F_2 at HERA is now comparable to the one obtained at fixed target experiments (a few percent), and that a good agreement between all these measurements using very different techniques can be observed. With the extension to high Q^2 brought by HERA, the scaling properties at $x \simeq 0.1$ are now observed up to $Q^2 = 5000$ GeV^2.

6 Extraction of F_L from the Cross-Sections at High y

The cross-section measurement at low Q^2 and at y close to 1 is particularly difficult at HERA since it requires the identification of the scattered positron having a few GeV in energy, while the photoproduction events which are occurring at a much higher rate often produce fake electrons in this energy range. At these low energies the usual trigger requirement of energy deposited in the calorimeter has to be complemented by a track requirement in order to get rid of beam induced background. Moreover, at low energies, even for $Q^2 \simeq 10$ GeV^2, the positron is scattered at relatively low angles ($\theta_e \simeq 160°$). Therefore at low Q^2 and at the largest y, the track of the positron is in the acceptance of the tracking system. This permits the subtraction of the photoproduction background using techniques based on the measured charge of these tracks.[25] Since at low energies the positron has less energy than the hadronic final state which is also scattered backwards, the positron identification procedure has also to be changed. Thus the unique features of the high y analysis are a modified positron identification and the requirement of a track with correct charge pointing to the calorimeter cluster.

The result for the measured DIS reduced cross-section for Q^2 between 12 and 35 GeV^2 is shown in Fig. 11 which compares the published 1994 data[28] with the preliminary 1996[29] and 1997[25] data. The reduced cross-section is defined as

$$\tilde{\sigma} = \frac{1}{\kappa}\frac{d^2\sigma}{dxdQ^2} = F_2 - \frac{y^2}{Y_+}F_L \text{ with } \kappa = \frac{2\pi\alpha^2 Y_+}{Q^4 x} \text{ and } Y_+ = 1 + (1-y)^2 \qquad (5)$$

when the Z^0 exchange is neglected. The error bars comprise both statistical and systematic errors added in quadrature. Both cross-section measurements agree well. For four Q^2 points the y range was extended with measurements up to $y = 0.82$ which considerably enhances the sensitivity to F_L.

Figure 11 shows also calculations of the cross section using the QCD fit to F_2, described in Section 4, and three different assumptions on the longitudinal structure function F_L. The measured cross-section is in agreement with the NLO QCD calcula-

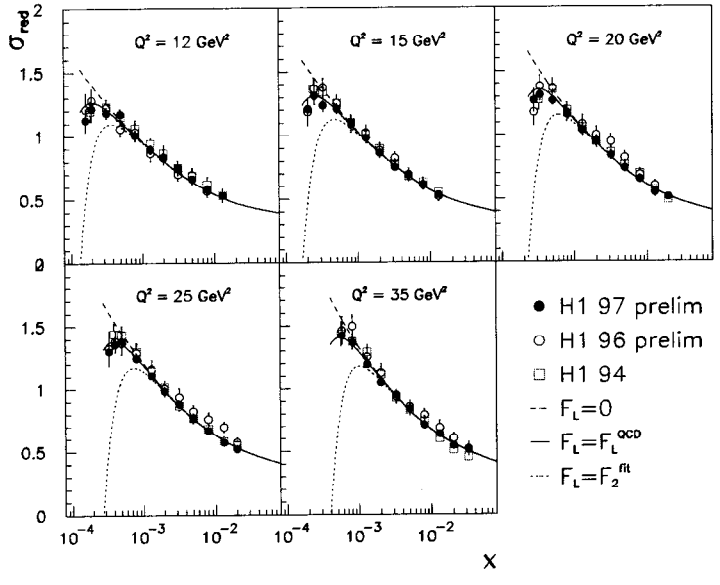


Figure 11: *Measurement of the reduced DIS cross-section at low Q^2. Closed circles: H1 1997 preliminary measurement. Open circles: H1 1996 preliminary measurement. Open squares: H1 data from the 1994 run[28] taken with a former backward setup. The lines use the QCD calculation of F_2 and of F_L (solid line) or the extreme assumptions $F_L = F_2$ (dotted line) and $F_L = 0$ (dashed line). The largest y value is 0.82.*

tion apart from large y, for $12 \leq Q^2 \leq 25$ GeV^2, where the measured points tend to be lower than the QCD lines (solid lines in Fig. 11).

F_L is then determined at $y = 0.68, 0.82$ by subtracting to the reduced cross-section, the extrapolated value of F_2 as predicted by a QCD fit to the data at lower y (< 0.35), on which the influence of F_L is negligible.

$$F_L = \frac{Y_+}{y^2}\left(F_2^{fit} - \frac{1}{\kappa}\frac{d^2\sigma^{exp}}{dxdQ^2}\right). \tag{6}$$

These results are shown and compared to another determination of F_L based on the measured $\partial\tilde{\sigma}/\partial\log y$ derivatives, in the next two sections.

6.1 Determining F_L using $\partial\tilde{\sigma}/\partial \ln y$

The cross-sections were used to measure the derivatives $\partial\tilde{\sigma}/\partial\log y$ at fixed Q^2 in the whole y range. This derivative is defined at fixed Q^2 as:

$$\frac{\partial\tilde{\sigma}}{\partial \ln y} = -\frac{\partial F_2}{\partial \ln x} - F_L \cdot 2y^2 \cdot \frac{2-y}{Y_+^2} + \frac{\partial F_L}{\partial \ln x}\cdot\frac{y^2}{Y_+}. \tag{7}$$

In Fig. 12 the cross-section derivative measurements are compared with the NLO QCD calculation using different assumptions on F_L. The derivative is seen to be sensitive to the longitudinal structure function. For $y \to 1$ the cross-section derivative tends to the limit $-\frac{\partial F_2}{\partial \ln x} - 2\cdot F_L$ with a negligible contribution from the derivative of F_L. This is in contrast to the F_L influence on the non-differentiated cross-section $\tilde{\sigma}$, see Fig. 11, where the contribution of F_2 dominates for all y and thus has to be controlled with high precision.

The measurement of $\frac{\partial\tilde{\sigma}}{\partial \ln y}$ can be used to determine F_L in a new way. Assuming that $\partial F_2/\partial \ln y = A \ln y + B$, straight line fits were made to $\partial\tilde{\sigma}/\partial \ln y$ in Q^2 bins on the measured points at $y < 0.2$. The extrapolation of those fits are taken to represent the contribution of F_2 at high y, and the uncertainties of the straight line fit are included in the systematic error. The small contribution of $\frac{\partial F_L}{\partial \ln y}$ to the derivative is corrected for using NLO QCD. This new method gives access to a lower Q^2 than the subtraction method seen above, since no QCD extrapolation at low x and low Q^2 are needed here.

6.2 Results on the Longitudinal Structure Function F_L

The results of these two methods of F_L determination are shown in Fig. 13 and they are compared to the NLO QCD expectation. The subtraction method ($\bullet$) and the derivative method ($\star$) give consistent results in good agreement with QCD. These two methods use the same data but different characteristics of the F_2 and F_L behavior. Note however that the systematic errors of the points at the same y are strongly correlated, implying that the tendency of the data points at $y = 0.82$ to lie above the QCD expectation is not significant.

In the future F_L will be measured directly by changing beam energies.

7 The Charm Contribution to F_2

The F_2 structure function is related to the quark and gluon densities of the proton. Thus, beyond the inclusive study of the proton structure it is also possible to pin down

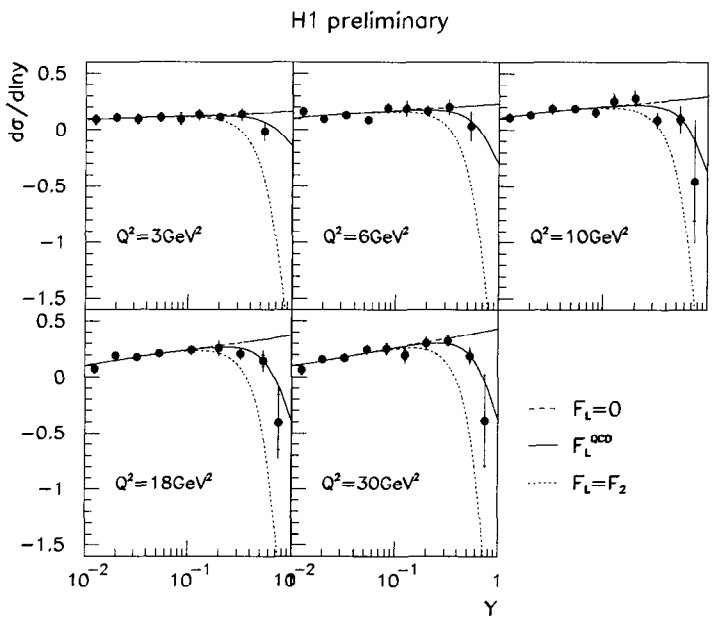


Figure 12: *Measurement of $\partial\tilde{\sigma}/\partial \ln y$ at fixed Q^2 using the1997 H1 data. The influence of F_L on this measurement is displayed by the curves which give the QCD expectation ($F_L = F_L^{QCD}$) and two extreme cases ($F_L = 0$), ($F_L = F_2$).*

exclusively the different contributions arising from the gluon or a given quark flavor, by separating the contribution from the light quarks, u, d, and s from the contribution of heavy quarks, in particular the charm contribution which is accessible with the present statistics in the kinematic domain of HERA.

The charm production in DIS occurs predominantly via photon-gluon fusion (BGF), which has been calculated analytically in NLO.[30] Other contributions to the charm production arising from diffractive heavy flavor production,[31] from the scattering of the virtual photon off a charm sea quark (flavor excitation process),[32] from charmed hadron production after the decay of b flavored hadrons[33] or from the production of $c\bar{c}$ in fragmentation,[34] are expected to have much smaller cross sections than BGF. The

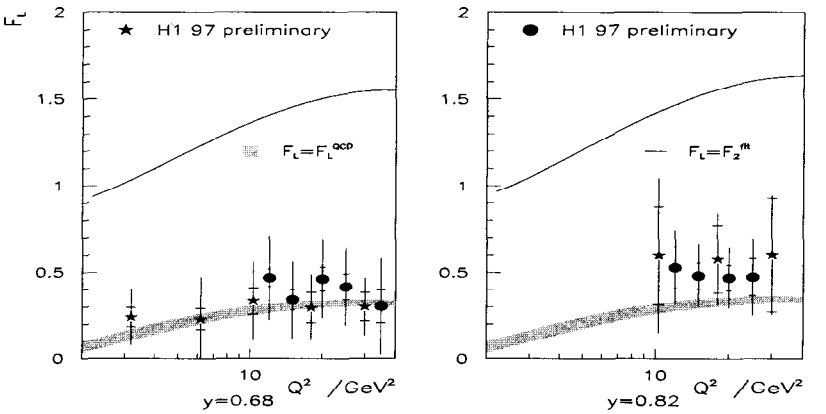


Figure 13: *Determination of the longitudinal structure function F_L using the two methods described in the text. The two determinations use the same data set. The narrow band represents the uncertainty range of F_L when calculated using the gluon and quark distributions determined with the H1 NLO QCD.*

possible contribution of intrinsic charm would give rise to charm production at large x,[35,32] and is currently beyond reach at HERA. Figure 14 illustrates the contribution from light quarks to the total F_2, as obtained by a NLO QCD fit[4] to the NMC and ZEUS F_2 measurements in bins of x as a function of Q^2. At high x, approaching the valence quark region, no contribution from the heavy quark flavors is expected, as well as at low Q^2, where the production of heavy flavors is suppressed due to their heavy mass. However at low x and relatively high Q^2, the charm contribution to F_2 amounts, according to the QCD fit results, to up to 25%.

In the NLO QCD fit programs used by H1 and ZEUS, the charm contribution is generated dynamically from BGF,[36,37] where both the renormalization and the factorization scales are taken as $\mu = \sqrt{Q^2 + 4m_c^2}$. The largest uncertainties in this calculation are related to that of the charm quark mass to which a typical uncertainty of about 30% is assigned. The direct measurement of the charm contribution to F_2 therefore allows for a test of this approach, and the close relation between BGF and the gluon density of the proton provides another independent way to measure the gluon density via the charm production.

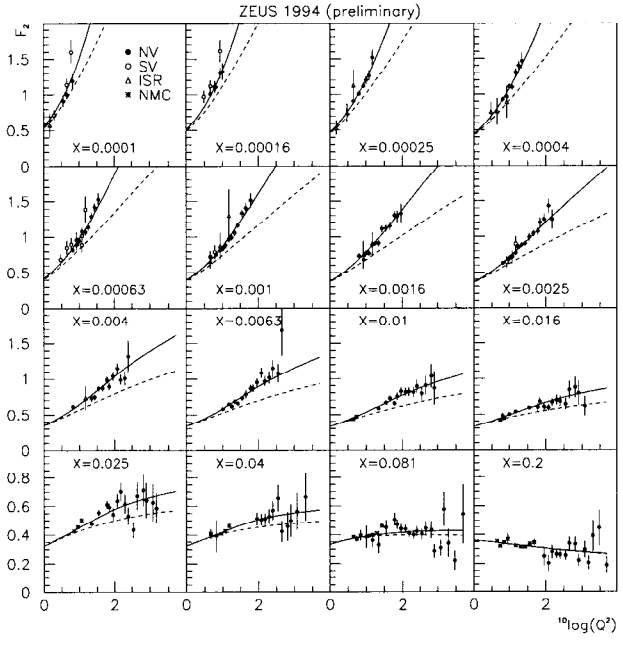


Figure 14: *The contribution from the charm quark to the inclusive F_2 from NLO-QCD fit. The full line corresponds to $F_2 = F_2^{u,d,s} + F_2^c$, whereas the dashed line accounts only for the contribution from the light quarks $F_2 = F_2^{u,d,s}$. The datasets used in the fit correpond to the ZEUS measurements on '94 data at nominal vertex (NV)[4] and shifted vertex (SV) events with initial state photon radiation (ISR)[12] and to the NMC data.[45]*

7.1 Charm in DIS

So far two ways to tag charm in DIS were explored at HERA: the open charm production of D^* or D^0 mesons,[38,39] and the identification of electrons from semileptonic charm decays.[40] The initial DIS sample reaches from $1 < Q^2 < 600$ GeV^2 and $0.02 < y < 0.7$. The D^* mesons are reconstructed using the decay

$$D^{*+} \to D^0\pi^+ \to (K^-\pi^+)\pi^+ + c.c., \tag{8}$$

which has a branching ratio of 2.62%.[41] The decay products are detected in the central trackers leading to an acceptance of $|\eta| < 1.5$ and of $1.5 < p_T(D^*) < 15$ GeV. No particle identification is performed and therefore the tracks assigned to the charged kaon and pion are required to produce an invariant mass $M(K\pi)$ between 1.80–1.92 GeV and a difference $M(K\pi\pi) - M(K\pi)$ between 143–148 MeV.

From the number of events obtained the derived total cross-section is

$$\sigma(ep \to eD^*X) = 8.55 \pm 0.40^{+0.30}_{-0.24}\ \text{nb}. \tag{9}$$

The differential cross-sections as functions of Q^2, x, W, $p_T(D^*)$, $\eta(D^*)$, and $x(D^*)$ are shown in Fig. 15. The measurement is compared to the prediction using the Monte Carlo-like program HVQDIS,[42] where the charm production is calclulated from BGF and the obtained production cross-section is convoluted with the Peterson fragmentation function[43] in order to obtain the visible D^* cross-section. The light parton distributions are taken from the GRV94 HO parton distributions. The error band represents a variation in the charm mass between 1.2 and 1.6 GeV and a good agreement of this "massive" NLO pQCD calculation with the measurement is observed, except for some small deviations at high-η and low-x of the D^* meson.

A component of direct electrons in jets can be ascribed to the semileptonic decay of charmed hadrons. In order to identify the electrons, the dE/dx measurement from the central tracking system is used. Supplementary cuts on the shower shape variables of calorimetric measurement associated to these tracks allow for a rejection of the charged pion background, and algorithms on mutually tangential tracks differentiate the electrons from photon conversion from the prompt electrons from the signal. The electron tracks identified by this procedure are only considered for $p > 1.2$ GeV and in the central tracking system with $0.65 < \theta < 2.5$. Acceptance corrections for these cuts as well as trigger and detection efficiencies are obtained from detailed Monte Carlo studies. Furthermore the inclusive $c \to e$ branching ratio $9.8 \pm 0.9 \pm^{0.6}_{0.5}$ % as measured by ARGUS[44] is applied in order to unfold the cross-section.

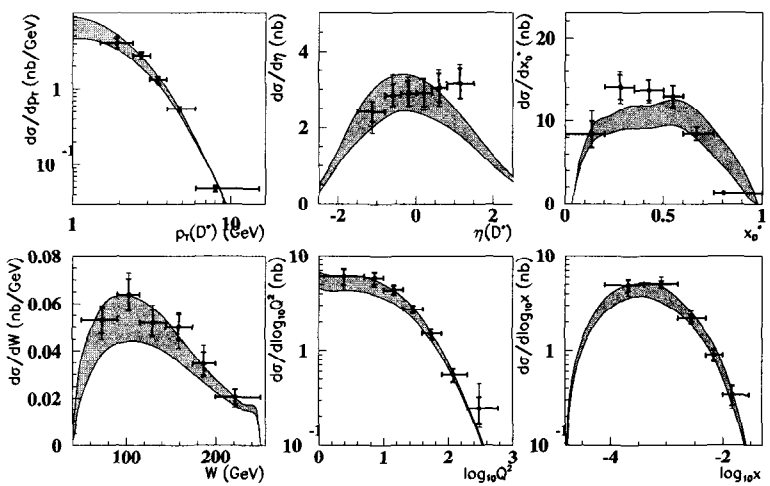


Figure 15: *Differential D^* cross-sections as functions of $\log Q^2$ (a), $\log x$ (b), W (c), $p_T(D^*)$ (d), $\eta(D^*$ (e), and $x(D^*)$ (f). The data points are compared to a NLO QCD fit, where the error band corresponds to a variation in the charm mass between 1.2 GeV (upper boundary) and 1.6 GeV (lower boundary).*

7.2 Measurement of F_2^c

The charm structure function F_2^c is related to the differential cross-section for $c\bar{c}$ production by

$$\frac{d^2\sigma_{c\bar{c}X}}{dx\,dQ^2} = \frac{2\pi\alpha^2}{xQ^4}[1+(1-y)^2]F_2^c(x,Q^2) \qquad (10)$$

considering in the given kinematic range only one photon exchange and neglecting an F_L^c contribution, which is estimated from NLO predictions to be lower than 1% in the measured Q^2, y region. For the extraction of F_2^c the cross-sections measured in a restricted p_T and η region are extrapolated to the full phase space using the HVQDIS program. Typical extrapolation factors are between 4 at low Q^2 and 1.5 at high Q^2. These extrapolated cross-sections are then translated into $c\bar{c}$ cross-sections using the constant hadronization fraction of charm to D^{*+}, $f(c \to D^{*+}) = 0.222$.

Figure 16 is showing the preliminary results of the F_2^c measurement in various bins of Q^2 as a function of x for the D^* and the semileptonic analyses on datasets taken

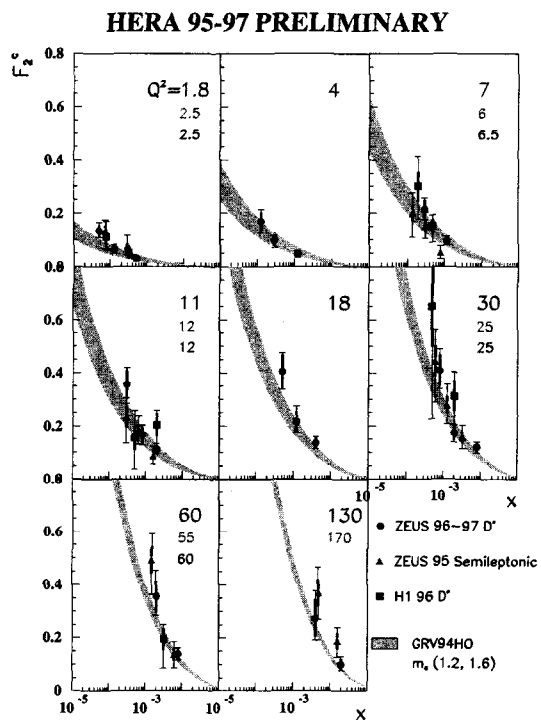


Figure 16: *F_2^c from D^* and semileptonic decays compared to the prediction from GRV94HO with a variation of the charm mass between 1.2 and 1.6 GeV.*

during the '95, '96, and/or '97 running periods. The results from the different analysis methods are in good agreement with a precision of about 15–20 %, which is still largely dominated by the statistics, but significantly more precise than the earlier measurements. The behavior of F_2^c shows a steep rise towards low x, which can be understood as being directly correlated to the steep rise of the gluon density established by the former inclusive measurements of F_2 at low x. The indirect prediction from perturbative QCD of the charm contribution is in good agreement with the direct measurements.

8 The Gluon Density of the Proton

One of the major results of the first years of the running of HERA was to establish the steep rise of the proton structure function at low x and its relation to the increase of the gluon density, which is extracted from NLO QCD fits to the structure function measurement. But the gluon density can also be extracted from processes invoking photon-gluon fusion (BGF), like the charm production, giving a cross check of the global measurement and access to different kinematic ranges.

8.1 The Gluon Density from NLO QCD Fit

The NLO QCD fits used to extract the gluon densities from the inclusive F_2 measurements are described in section 4.1. As the HERA data do not constrain the high x region, proton and deuteron structure function data from NMC[45] and BCDMS[60] were included in the fit.

Figure 17 shows the gluon densities obtained by the H1[46] and the ZEUS[24] fits at $Q^2 = 20\text{GeV}^2$. Both results are in good agreement and show a steep rise of $xg(x)$ for decreasing x, which is mainly due to the asymptotic behavior in $\log 1/x$ of the solutions to the DGLAP equations. The error bands shown take into account the statistical errors, a full treatment of the correlated systematic errors and possible variations of α_S and of the charm mass. The precision obtained for the gluon density at $x \sim 5 \cdot 10^{-4}$ is about 15%. The results from the HERA experiments allow for a smooth transition towards the results from NMC in the high x domain and are consistent with the gluon density of the MRSR1[47] and CTEQ4M[48] parametrizations obtained from fits to a wider number of datasets. The gluon density from the GRV[16] parametrization gives a somewhat higher result, which is understandable since it assumes a lower α_S value than the other fits.

The non-singlet quark momentum distribution $xq_{NS} = \sum_{i=1}^{n_f}(xq_i - x\bar{q}_i) = xu_{val} + xd_{val}$, which expresses the sum of the valence quark densities, evolves independently of the gluon distribution, while the singlet quark distribution $x\Sigma = \sum_{i=1}^{n_f}(xq_i + x\bar{q}_i)$ is coupled to the gluon distribution, since at small values of x, $x\Sigma$ is dominated by the contribution of the sea $q\bar{q}$ pairs.

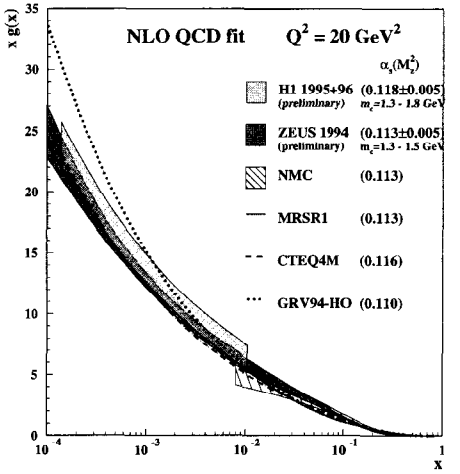


Figure 17: *The gluon density of the proton as obtained from NLO order fit to the inclusive F_2 measurements at $Q^2 = 20$ GeV^2 by the H1 and ZEUS collaborations compared to the results of the NMC Collaboration at high x and results from the global analyses by MRSR and CTEQ as well as the prediction from GRV94. The α_s values used in the fits are quoted between brackets.*

Figure 18 is comparing the singlet quark distribution with the gluon distribution at $Q^2 = 20$ GeV^2, $Q^2 = 7$ GeV^2, and $Q^2 = 1$ GeV^2. In the two higher Q^2 bins, the xg shows a steeper rise at low x than the sea quark densities. At $Q^2 = 1$ GeV^2, the gluon density shows an almost flat behavior at $Q^2 = 1$ GeV^2, indeed compatible with 0, while $x\Sigma$ is still slightly rising. This difference in behavior may indicate a difference in the dynamics when approaching the transition region, at $Q^2 \approx 1$ GeV^2, from perturbative to non-perturbative QCD.

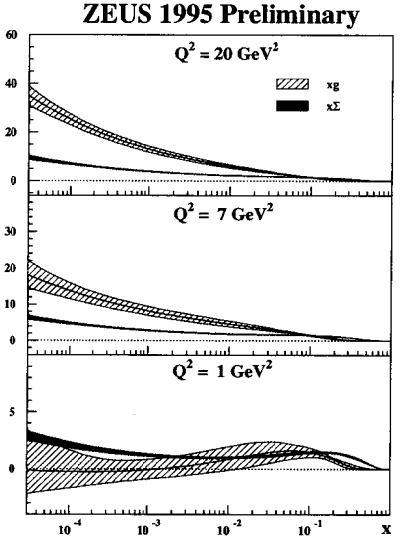


Figure 18: *The quark singlet momentum distribution $x\Sigma$ (shaded) and the gluon momentum distribution $xg(x)$ (hatched) as a function of x at different values of Q^2.*

8.2 The Gluon Density from the Charm

In a similar way to how the F_2^c contribution is extracted from the open charm production, the unfolding can be also done directly for the gluon density.[49] This is possible, not only in DIS, but also from D^* events in photoproduction, where $Q^2 \simeq 0$, using the FMNR program[50] to calculate the cross-sections. For the unfolding procedure, the gluon momentum fraction x_g is determined from the momentum carried by the D^* meson using an iterative procedure. Each value of x_g corresponds to a different scale, and the results is shown in Fig. 19 at the average scale $\mu^2 = 4m_c^2 + \langle Q \rangle^2 = 25$ GeV^2.

The results of both D^* analyses are in good agreement, and are consistent at low x with the determination of the gluon density from the QCD fit to F_2. The measured values are also well described by the CTEQ4F3[48] parametrization, both at low and high x.

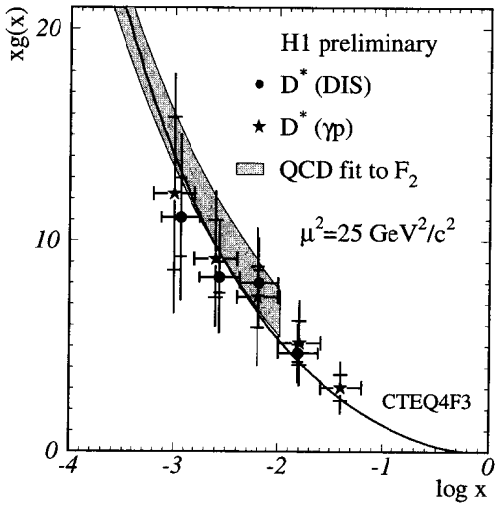


Figure 19: *The gluon density obtained from the D^* production in DIS and photoproduction compared to the results of the gluon extraction from the inclusive F_2 measurement at $\mu^2 = 25$ GeV^2 and the CTEQ4F3 parametrization.*

9 DIS Events at High Q^2

With the increased luminosity produced by HERA during the past four years, the high Q^2 region, *i.e.*, Q^2 typically above 100 GeV^2, became accessible for precision measurements, for the observation of electroweak effects in ep scattering and for the exploration of the very high Q^2 region and the search for exotic particles.[51] In Fig. 20b is shown a high Q^2 Neutral Current event detected in the ZEUS detector, in which the electron is backscattered in the forward direction. In Fig. 20a is shown a charged current event in the H1 detector, in which the outgoing neutrino escapes undetected, leading to a large missing transverse momentum.

From 1994 until 1997, HERA was colliding positrons off protons. At low Q^2 the e^-p and e^+p cross-sections are identical, but at high Q^2 the increasing influence of the weak contributions gives rise to a destructive interference of the electromagnetic and

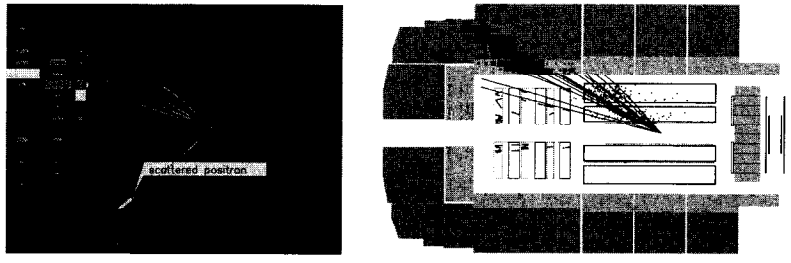


Figure 20: *Neutral (left) and Charged (right) Current events seen in the ZEUS and in the H1 detectors. The incident positron comes from the left, the proton from the right.*

the weak contribution and decreases the e^+p cross-section compared to the e^-p one.[5,6] The comparison of Neutral and Charged Current cross-sections as functions of Q^2 as shown in Fig. 21 displays their similar behavior at $Q^2 > 10000$ GeV^2, a region where the weak contribution becomes important, whereas at lower Q^2, the NC cross-section is rising steeply, contrarily to the CC cross-section, which increases mildly at lower Q^2. The remaining differences at high Q^2 are due to the weak coupling to the different quark flavors. At very high Q^2, above 15000 GeV^2, is the region where the excess of events was reported[52,53] on the 1994 to 1996 data. This excess is still visible, but has lowered in significance after including the 1997 data.[54,55]

9.1 Neutral Current Cross-Sections

The NC cross-section for unpolarized e^+p scattering can be written as a linear combination of the F_2, F_3, and F_L structure functions:

$$\frac{d^2\sigma^{e^+p}}{dx\,dQ^2} = \frac{2\pi\alpha^2}{xQ^4}\left[Y_+F_2(x,Q^2) - y^2F_L(x,Q^2) - Y_-xF_3(x,Q^2)\right]. \qquad (11)$$

In this equation α is the fine structure constant, and the helicity dependence of the electroweak interactions is contained in the function Y defined as $Y_\pm(y) = 1 \pm (1-y)^2$. The generalized structure function F_2 can be decomposed as

$$F_2 = F_2^{em} + \frac{Q^2}{(Q^2+M_Z^2)}F_2^{int} + \frac{Q^4}{(Q^2+M_Z^2)^2}F_2^{wk} = F_2^{em}(1+\delta_Z) \qquad (12)$$

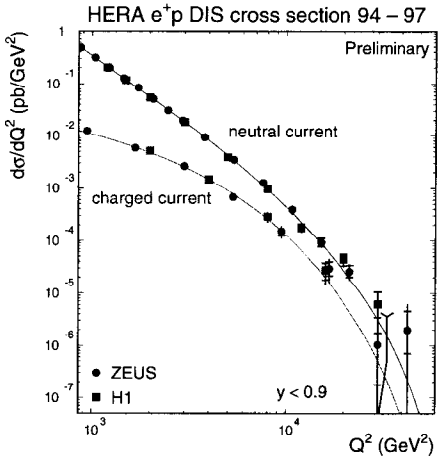


Figure 21: *Single differential cross-section as a function of Q^2 for Neutral and Charged Currents.*

where F_2^{em} contains the contribution from the pure photon exchange, F_2^{wk} the one of the Z^0 exchange, and F_2^{int} the γZ^0 interference. The reduced cross-section is defined in eq. 5 as

$$\tilde{\sigma}(e^+p) \equiv \frac{xQ^4}{2\pi\alpha^2}\frac{1}{Y_+}\frac{\mathrm{d}^2\sigma}{\mathrm{d}x\mathrm{d}Q^2} \qquad (13)$$

and can be written as $F_2^{em}(1+\delta_Z-\delta_3-\delta_L)$ with δ_Z, δ_3, and δ_L being small at low Q^2. The contributions from F_2^{em} and from F_2 to the reduced NC cross-section for $Q^2 \geq 3000$ GeV^2 are shown in Fig. 22, which depicts the effect of the γZ destructive interference in e^+p collisions, which is most important at high y. The contribution from $\delta_Z - \delta_3$ is below 1% at $Q^2 < 1500$ GeV^2 and about 10% at $Q^2 = 5000$GeV^2 and $x = 0.08$. The influence of the longitudinal structure function δ_L is negligible at $y < 0.5$ and up to 5% at $y = 0.9$. At $Q^2 \geq 15000$ GeV^2 no structure function has a dominant contribution. All the NC cross-sections presented have been corrected for QED radiation from the lepton line, whereas the radiation from the quark line can be safely neglected.

The preliminary results from H1 on the reduced NC cross-section are shown in Fig. 23 for 200 GeV$^2 \leq$ Q$^2 \leq$ 30000 GeV^2 and $0.005 \leq x \leq 0.65$. The NLO QCD fit gives a good description of the data in the whole Q^2 and x range. Above $Q^2 = 1000$

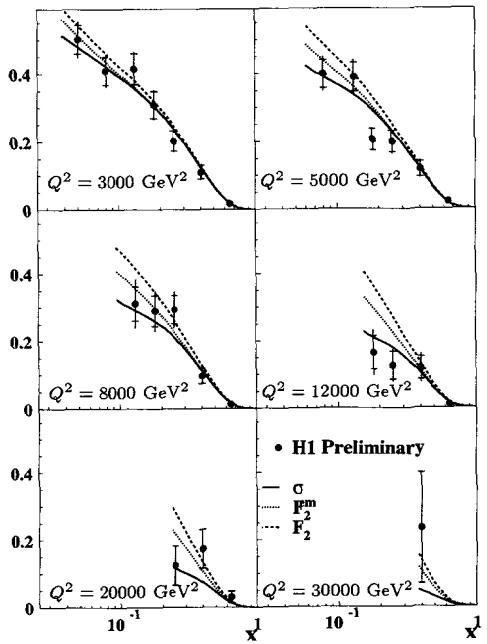


Figure 22: *The reduced NC cross-section for $Q^2 \geq 3000$ GeV^2. Indicated are the contributions from the electromagnetic F_2^{em} and the generalized F_2 as obtained in the NLO QCD fit to the data.*

GeV^2, the statistical error grows typically to 5-15% and dominates the total error. The total error (quadratic sum of the systematic and of the statistical error) is below 10% at $Q^2 \lesssim 1000$ GeV^2.

In Fig. 24 is shown the ZEUS preliminary F_2 measurement on the 1996-1997 data at $Q^2 \geq 800$ GeV^2 which is compared to the previous measurement on the 1994 data done on statistics about ten times lower. The new measurement extends in range the previous one and improves its precision.

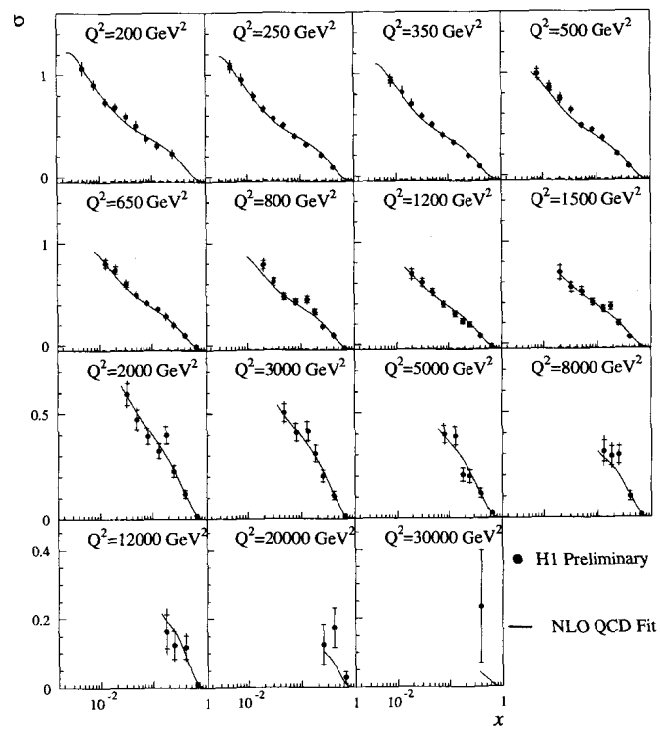


Figure 23: *Preliminary 1994-1997 H1 results on the reduced NC cross-section compared to the NLO QCD fit.*

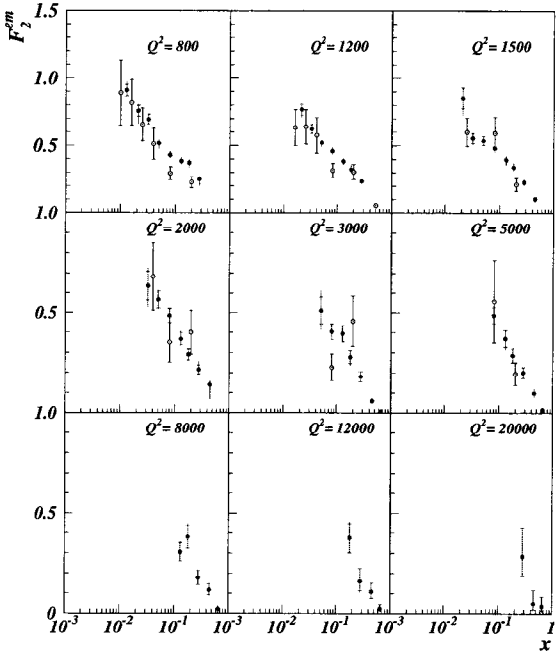


Figure 24: *ZEUS preliminary results (1996-1997 data) on F_2 measured at high Q^2 compared to the ZEUS 1994 measurement.*

9.2 The Weak Contribution to the NC Cross-Section

The effect of the γZ interference is visible at HERA for the first time, as can be shown on the NC cross-section $d\sigma/dx$. In Fig. 25 is shown the $d\sigma/dx$ cross-section at $Q^2 > 1000$ GeV^2. The measurement extends from $1.3 \cdot 10^{-2}$ to 0.65. The maximum of the distribution is obtained at $\approx 2 \cdot 10^{-2}$, *i.e.*, at this low Q^2, the cross-section is still dominated by low x, "sea" partons (the sharp drop of the cross-section at $x < 2 \cdot 10^{-2}$ comes from the kinematic inaccessibility of this region at high Q^2). No difference in the prediction is observed with respect to the Standard Model if only γ exchange is taken into account. In contrast, at $Q^2 > 10000$ GeV^2 the two predictions are significantly different, and the data which have now their maximum in the valence region ($x \approx 0.2$),

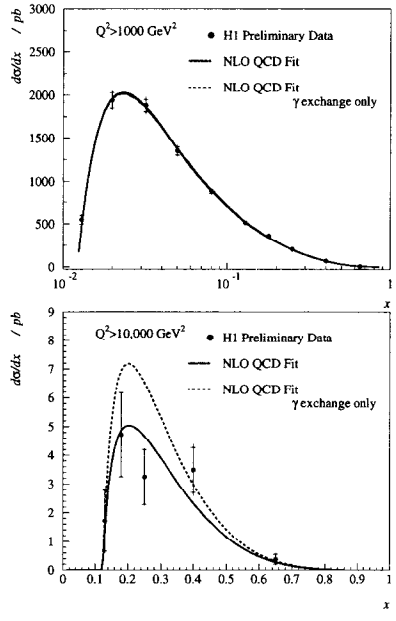


Figure 25: *The neutral current cross-section $d\sigma/dx$ at high Q^2 compared with the Standard Model prediction and with the prediction when the coupling to the Z boson is not taken into account.*

agree with the Standard Model prediction and disagree at the three standard deviation level if only γ exchange is taken into account.

9.3 Reduced Cross-Section at High x

In Fig. 26 the emphasis is put on the high x region, and the reduced cross-section is shown as a function of Q^2 at fixed values of x between $x = 0.07$ and $x = 0.65$. At $x = 0.45$ some excess of the cross-section over the Standard Model prediction is visible for the highest Q^2 values, which corresponds to the accumulation of events around an inclusive invariant mass of the lepton quark system of about 200 GeV, and which was already pointed out in the 1994 to 1996 data.[52] The significance of this excess has decreased with the enhanced statistics which include the 1997 data.

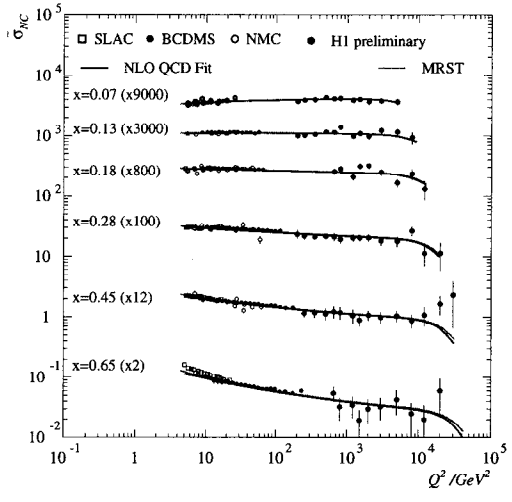


Figure 26: *The reduced neutral current cross-section at high x compared with Standard Model predictions from MRST and the H1 NLO QCD fit. The error bars represent the total error of the measurements.*

The cross-section is compared to the MRST[59] parton density and the NLO QCD fit. At the lowest x value of 0.07, both parameterizations are in good agreement, but at higher x the new data, which are included in the QCD fit, lower the prediction from the fit by up to 5%. The direct comparison with the BCDMS measurement[60] shows the same trend: from $x = 0.28$ up to $x = 0.65$, the H1 data are slightly below the extrapolation obtained from the BCDMS data.

9.4 Charged Current Cross-Sections

The Born double differential CC cross-section can be expressed as in the NC case

$$\left(\frac{d^2\sigma_{CC}}{dx\, dQ^2}\right)_{Born} = \frac{G_F^2}{2\pi x} \cdot \left(\frac{M_W^2}{M_W^2 + Q^2}\right)^2 \cdot \phi_{CC}(x, Q^2) \qquad (14)$$

with G_F being the Fermi coupling constant, M_W the mass of the W boson, and $\phi_{CC}(x, Q^2)$ the CC structure function term. In the special case of leading order QCD, and neglecting the effect of quark mixing, this cross-section is related to the quark parton densities (xu, xd, xs, xc) which are functions of x and Q^2:

$$\left(\frac{d^2\sigma_{CC}}{dxdQ^2}\right)_{LO} = \frac{G_F^2}{2\pi}\left(\frac{M_W^2}{M_W^2 + Q^2}\right)^2 \left((\bar{u} + \bar{c}) + (1-y)^2(d+s)\right). \qquad (15)$$

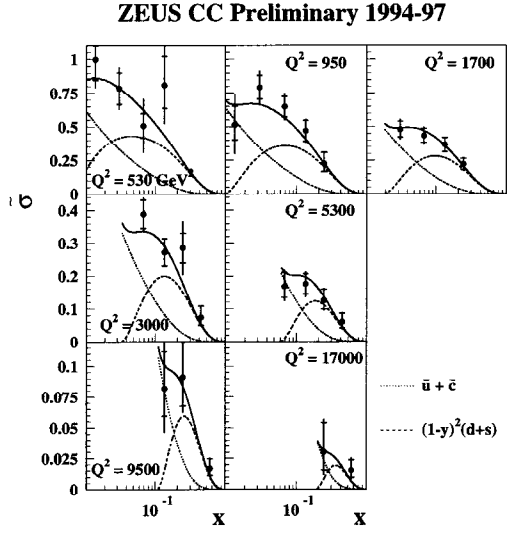


Figure 27: *The reduced charged current cross-section as a function of x for different Q^2 compared with the Standard Model prediction (using CTEQ4D parton distributions) and the xd contribution to it.*

The reduced charged current cross-section is defined in analogy with the NC one:

$$\tilde{\sigma}_{CC}(x, Q^2) \equiv \frac{x \cdot 2\pi}{G_F^2}\left(\frac{M_W^2 + Q^2}{M_W^2}\right)^2 \frac{d^2\sigma_{CC}}{dxdQ^2}. \qquad (16)$$

Note that with this definition, apart for the small weak radiative corrections which have not been applied, the CC reduced cross-section is equal to the structure function term, while the NC reduced cross-section has an extra $1/Y_+$ term.

The double differential CC cross-section has been measured for the first time at high Q^2 by H1 and ZEUS. The ZEUS preliminary results are displayed in Fig. 27. They are

still largely dominated by the statistical error. The most important systematic error is the uncertainty on the energy scale of the hadronic final state. Within these errors, the measurement is in good agreement with the predicted cross-section from the Standard Model using the MRST[59] parton distributions. Also shown is the contribution to the cross-section of the term $(1-y)^2 x(d+s)$ which is dominated in the Standard Model by the d valence quark at high x. The strong suppression of the d and s contributions

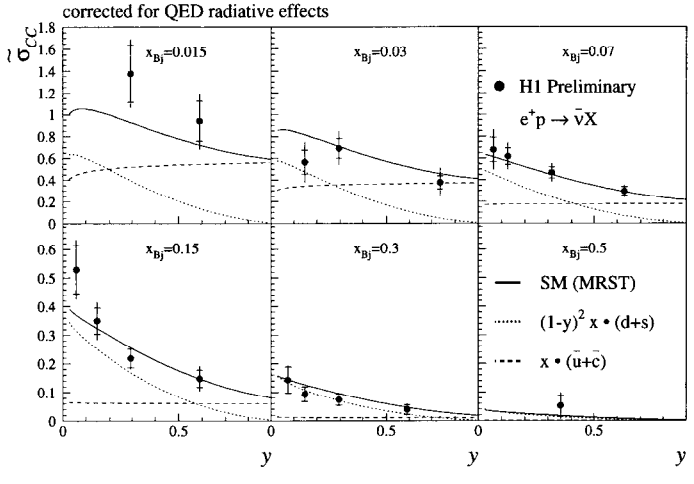


Figure 28: *The reduced charged current cross-section as a function of y for different values of x (lower plot) compared with the Standard Model prediction (using MRST parton distributions) and the $(d+s)$ and sea antiquarks' contribution to it.*

in the double differential CC cross-section at high y due to the $(1-y)^2$ term, and the small contribution of the sea quarks at high x, are better studied when replotting the reduced cross-section at fixed x as a function of y, as can be seen in Fig. 28. The relative contribution given by the MRST parametrization are also shown. As expected, the reduced cross-section grows at low y, because the $(1-y)^2$ term increases towards one, and it grows at low x since the sea quark contribution increases in this domain. At $x = 0.3$, the shape of the measured distribution, which decreases quickly towards zero, indicates that the sea quark contribution is indeed small at large x.

9.5 The Propagator Mass M_W

In the Standard Model the CC cross-section is directly sensitive to the Fermi coupling constant (G_F) and the W mass (M_W). It is visible from the leading order CC cross-section formula that the constraint on M_W comes from the propagator factor, $[(M_W^p)^2/(Q^2+(M_W^p)^2)]^2$, where in the SM, the propagator mass $M_W^p \equiv M_W$. Since CC interactions arise from the t channel exchange of a virtual W boson, the value of M_W^p extracted from the propagator factor can be interpreted as a direct measurement of the W mass in the spacelike regime. It is a crucial test of the universality of the Standard Model to compare such a measurement with timelike and indirect determinations from other experiments.

In order to construct a fit of the CC cross-section that is sensitive only to the value of M_W^p from the propagator term, the Standard Model expectation for the CC cross-section is obtained in which M_W^p is allowed to vary, and is decoupled from the value of M_W entering the theoretical expectation for $G_F(\alpha, M_W, M_Z, M_{top}, M_H)$.

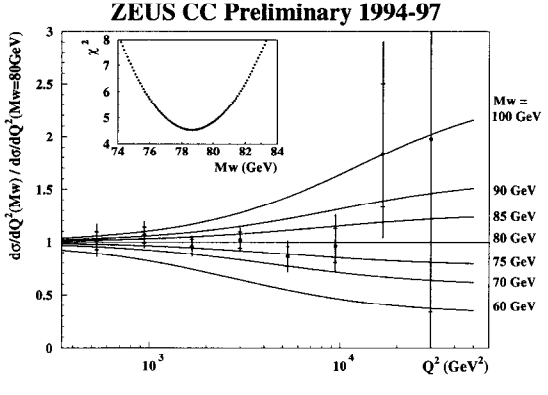


Figure 29: *The points show the ratio of the measured cross-section to the Standard Model prediction. The solid curves show the expected deviations of the cross-section from the nominal $M_W = 80$ GeV scenario for various indicated values of M_W. Inset is the χ^2 fit of the data to the Standard Model plotted againt M_W.*

In Fig. 29 is shown the result of the ZEUS fit, where the ratio of $d\sigma(M_W)/dQ^2$ to $d\sigma(M_W = 80\ \text{GeV})/\text{dQ}^2$ is plotted as a function of Q^2 together with the χ^2 of the

fit to the ZEUS cross-section measurement, as a function of the propagator mass. The propagator mass results from the HERA[56,58] experiments are

$$81.2\ ^{\pm 3.3 \pm 4.3}\ \text{GeV for the H1 experiment} \tag{17}$$

$$78.6\ ^{+2.5+3.3}_{-2.4-3.0}\ \text{GeV for the ZEUS experiment.} \tag{18}$$

Both results are in agreement with the world average $M_W = 80.41 \pm 0.10$ GeV.[41]

9.6 Comparison of the NC and CC Cross-Sections at High x

At $x \geq 0.3$ the CC cross-section is largely dominated by the d valence quark. The comparison in the high x domain of the NC structure function term (*i.e.*, $Y_+\tilde{\sigma}_{NC}$) with the CC one (*i.e.*, $\tilde{\sigma}_{CC}$) as shown in Fig. 30, in a region in which the γZ interference term is still small ($Q^2 \lesssim 5000$ GeV^2), gives direct information on the relative size of u_v and d_v at high Q^2. From the 2000 and 4000 GeV^2 points of Fig. 30, we can observe that the u valence quark distribution is between 1.5 to 3 times larger than the d at $x = 0.3$.[56]

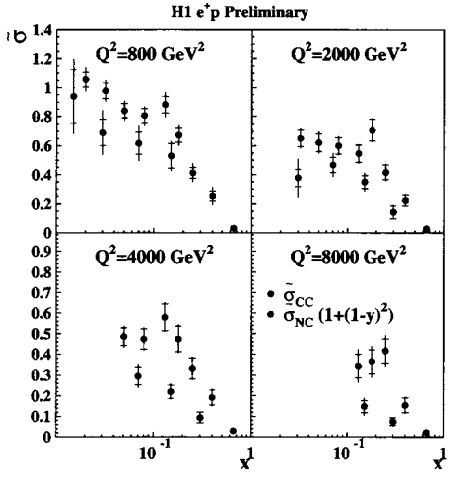


Figure 30: *Comparison of the structure function terms of the NC and CC double differential cross-sections (i.e., of $Y_+\tilde{\sigma}_{NC}$ and $\tilde{\sigma}_{CC}$) as a function of x in bins of Q^2.*

10 Summary

In its first years of operation, HERA has been testing the Standard Model, in particular QCD and the structure of the proton, in new kinematic domains. A largely improved knowledge of the parton distributions has been achieved, in particular at low x. QCD has been tested extensively and successfully, from the breakdown of perturbative QCD at low Q^2, up to the highest Q^2 where the small excess of events observed by H1 and ZEUS remains compatible with the Standard Model.

The future will bring new exciting results since about 50 pb^{-1} of e^-p data are expected in 1999-2000 which will complement the e^+p results presented here. New measurements of α_S, xg, and of the parton distribution functions will become available. The slight excess observed at high Q^2 will be also tested in e^-p.

Afterwards with the luminosity upgrade of HERA and its detectors in between 2000 and 2005, 500-1000 pb^{-1} of integrated luminosity are expected. Thus this program will continue with higher precision and with an extended potential for searching for deviations to the Standard Model.

Acknowledgments

I would like to thank the organizers of the SLAC topical workshop for their invitation to talk at this inspiring meeting, and all my colleagues of HERA and of the H1 and ZEUS collaborations for the results presented in this paper. Thanks also to Martin Erdmann and Stefan Schlenstedt for the careful reading of the manuscript. A very special thank to Ursula Bassler for her great help in the preparation of the talk and in the writing of the manuscript.

References

[1] H1 Collab., I. Abt *et al.*, Nucl. Phys. **B407** (1993) 515; H1 Collab., T. Ahmed *et al.*, Nucl. Phys. **B439** (1995) 471.

[2] ZEUS Collab., M. Derrick *et al.*, Phys. Lett. **B316** (1993) 412; ZEUS Collab., M. Derrick *et al.*, Z. Phys. **C65** (1995), 379.

[3] H1 Collab., S. Aid *et al.*, Nucl. Phys. **B470** (1996) 3.

[4] ZEUS Collab., M. Derrick *et al.*, Z. Phys. **C72** (1996) 399.

[5] H1 Collab., T. Ahmed *et al.*, Phys. Lett. **B324** (1994) 241; H1 Collab., S. Aid *et al.*, Z. Phys. **C67** (1995) 565; H1 Collab., S. Aid *et al.*, Phys. Lett. **B379** (1996) 319.

[6] ZEUS Collab., M. Derrick *et al.*, Phys. Rev. Lett. **75** (1995) 1006; ZEUS Collab., M. Derrick *et al.*, Z. Phys. **C72** (1996) 47.

[7] U. Bassler and G. Bernardi, Nucl. Instrum. Methods **A426** (1999) 583.

[8] A. Blondel and F. Jacquet, *Proceedings of the Study of an ep Facility for Europe*, ed. U. Amaldi, DESY 79/48 (1979) 391.

[9] U. Bassler and G. Bernardi, Nucl. Instr. Meth. **A361** (1995) 197.

[10] S. Bentvelsen *et al.*, *Proceedings of the Workshop Physics at HERA*, vol. 1, eds. W. Buchmüller and G. Ingelman, DESY (1992) 23.
C. Hoeger, ibid., 43.

[11] H1 Collab., C. Adloff *et al.*, Nucl. Phys. **B497** (1997) 3.

[12] ZEUS Collab., M. Derrick *et al.*, Z. Phys. **C69** (1996) 607; ZEUS Collab., J. Breitweg *et al.*, Phys. Lett. **B407** (1997) 432.

[13] L. N. Hand, Phys. Rev. **129** (1963) 1834; S. D. Drell and J. D. Walecka, Ann. Phys. **28** (1964) 18; F. J. Gilman, Phys. Rep. **4C** (1972) 95.

[14] B. L. Ioffe, Phys. Lett. **40** (1969) 123; B. L. Ioffe, V. A. Khoze, and L. N. Lipatov, *Hard Processes*, North Holland (1984) 185.

[15] A. Donnachie and P. V. Landshoff, Z. Phys. **C61** (1994) 139.

[16] M. Glück, E. Reya, and A. Vogt, Z. Phys. **C67** (1995) 433.

[17] Yu. L. Dokshitzer, Sov. Phys. JETP **46** (1977) 641; V. N. Gribov and L. N. Lipatov, Sov. J. Nucl. Phys. **15** (1972) 438 and 675;
G. Altarelli and G. Parisi, Nucl. Phys. **B126** (1977) 297.

[18] J. J. Sakurai, Ann. Phys. **11** (1960) 1.

[19] T. Bauer *et al.*, Rev. Mod. Phys. **50** (1978) 261.

[20] J. J. Sakurai and D. Schildknecht, Phys. Lett. **B40** (1972) 121.

[21] D. Schildknecht and H. Spiesberger, BI-TP-97-25 (1997) [hep-ph/9707447].

[22] H. Abramowicz *et al.*, Phys. Lett. **B269** (1991) 465; H. Abramowicz and A. Levy, DESY 97-251 (1997) [hep-ph/9712415].

[23] K. Adel, F. Barreiro, and F. J. Yndurain, Nucl. Phys. **B495**, 221 (1997).

[24] ZEUS Collab., Eur. Phys. J. **C7** (1999) 609.

[25] H1 Collab., "Precision Measurement of the Inclusive Deep Inelastic *ep* Scattering Cross-Section at Low Q^2 at HERA," ICHEP 98-Vancouver, 98-534.

[26] ZEUS Collab., "Measurement of the Proton Structure Function F2 in e+p Collisions at HERA," ICHEP 98-Vancouver, 98-769.

[27] E. A. Kuraev, L. N. Lipatov, and V. S. Fadin, Sov. Phys. JETP **45** (1977) 199; Y. Y. Balitsky and L. N. Lipatov, Sov. J. Nucl. Phys. **28** (1978) 822.

[28] H1. Collab., C. Adloff *et al.*, Phys. Lett. **393B** (1997) 452.

[29] H1 Collab., "A Measurement of the Inclusive Deep Inelastic *ep* Scattering Cross-Section at low Q^2 at HERA," IECHEP 97-Jerusalem, 97-260 (1997).

[30] B. W. Harris and J. Smith, Nucl. Phys. **B452** (1995) 109; B. W. Harris and J. Smith, Phys. Lett. **B353** (1995) 535.

[31] P. Bruni, G. Ingelman, and A. Solano, *Proceedings of the Workshop Physics at HERA*, vol. 1, eds. W. Buchmüller and G. Ingelman, DESY (1992) 363.

[32] G. Ingelman, L. Jönsson, and M. Nyberg, Phys. Rev. **D47** (1993) 4872.

[33] A. Ali *et al.*, *Proceedings of the Workshop Physics at HERA*, vol. 1, eds. W. Buchmüller and G. Ingelman, DESY (1992) 393; A. Ali *et al.*, *Proceedings of the Workshop Physics at HERA*, vol. 2, eds. W. Buchmüller and G. Ingelman, DESY (1992) 667.

[34] M. L. Mangano and P. Nanson, Phys. Lett. **B285** (1988) 160; M. H. Seymour, Z. Phys. **C63** (1994) 99; M. H. Seymour, Nucl. Phys. **B436** (1995) 16.

[35] S. J. Brodsky *et al.*, Phys. Lett. **B93** (1980) 451; S. J. Brodsky *et al.*, Nucl. Phys. **B369** (1992) 519.

[36] E. Laenen *et al.*, Nucl. Phys. **B392** (1993) 162, 229; E. Laenen *et al.*, Phys. Lett. **B291** (1992) 325; S. Riemersma, J. Smith, and W. L. van Neerven, Phys. Lett. **B347** (1995) 143.

[37] M. Glück, E. Reya, and M. Stratmann, Nucl. Phys. **B422** (1994) 37.

[38] H1 Collab., C. Adloff *et al.*, Z. Phys. **C72** (1996) 563; H1 Collab., "$D^{*\pm}$ Meson Production in DIS at HERA," ICHEP 98-Vancouver, 98-540.

[39] ZEUS Collab., J. Breitweg *et al.*, Phys. Lett. **B407** (1997) 402; ZEUS Collab., "Measurement of the D^* Cross-Section and the Charm Structure Function fo the Proton in Deep Inelastic Scattering at HERA," ICHEP 98-Vancouver, 98-768.

[40] ZEUS Collab., "Measurement of the Charm Structure Function of the Proton in Deep Inelastic Scattering using Semileptonic Charm Decays," ICHEP 98-Vancouver, 98-772.

[41] Particle Data Group, R. M. Barnett *et al.*, Phys. Rev. **D54** (1996) 1.

[42] B. W. Harris and J. Smith, hep-ph/9706334.

[43] C. Peterson *et al.*, Phys. Rev. **D27** (1983) 105.

[44] ARGUS Collab., H. Albrecht *et al.*, Phys. Lett. **B278** (1992) 202.

[45] NMC Collab., M. Arneodo *et al.*, Nucl. Phys. **B483** (1997) 3.

[46] H1 Collab., "A Measurement of the Inclusive Deep-Inelastic *ep* Scattering Cross-Section at Low Q^2 at HERA," IECHEP-97, 97-260 (1997).

[47] A. D. Martin, R. G. Roberts, and W. J. Stirling, Phys. Lett. **B387** (1996) 419.

[48] H. Lai *et al.*, Phys. Rev. **D55** (1997) 1280.

[49] H1 Collab., "Determination of the Gluon Density in the Proton from Charm Electro- and Photo-Production using NLO QCD," ICHEP 98-Vancouver, 98-538.

[50] S. Frixione *et al.*, Phys. Lett. **B348** (1995) 3.

[51] U. Katz, "New HERA Results on Deep-Inelastic e^+p Scattering at Very High Q^2," *these proceedings*.

[52] H1 Collab., C. Adloff *et al.*, Z. Phys. **C74** (1997) 191.

[53] ZEUS Collab., J. Breitweg *et al.*, Z. Phys. **C74** (1997) 207.

[54] H1 Collab., "A Search for Leptoquark Bosons in DIS at High Q^2 at HERA", ICHEP 98-Vancouver, 98-580.

[55] ZEUS Collab., "Resonance Search in *ep* Scattering at High x and Q^2 Using the ZEUS Detector at HERA," ICHEP 98-Vancouver, 98-754.

[56] H1 Collab., "Measurement of Inclusive Cross-Sections for Neutral and Charged Current Interactions at High-Q^2," ICHEP 98-Vancouver, 98-533.

[57] ZEUS Collab., "Measurement of High-Q^2 Neutral-Current DIS Cross-Sections at HERA," ICHEP 98-Vancouver, 98-752.

[58] ZEUS Collab., "Measurement of High-Q^2 Charged-Current DIS Cross-Sections at HERA," ICHEP 98-Vancouver, 98-751.

[59] A. Martin *et al.*, Eur. Phys. J. **C4** (1998) 463.

[60] BCDMS Collab., A. C. Benvenuti *et al.*, Phys. Lett. **B223** (1989) 485; Phys. Lett. **B237** (1990) 592.

B PHYSICS AND QCD AT SLD

David J. Jackson*
Rutherford Appleton Laboratory,
Chilton, Didcot, Oxfordshire OX11 0QX, England

Representing the SLD Collaboration

ABSTRACT

Analysis of B decays and QCD phenomena at the SLD experiment at the SLC e^+e^- collider is presented. The results are based on a sample of 150,000 hadronic Z^0 decays collected between 1993 and 1995 together with 250,000 hadronic Z^0 decays collected from 1996 to 1998 with an upgraded vertex detector. The large longitudinal polarization of the SLC e^- beam (average $\langle P_e \rangle \sim 73\%$) was utilized by some of the measurements. The precise CCD pixel vertex detector allows efficient and pure identification of the B decay tracks.

The topics in QCD covered here include hadron production and correlations, the flavor independence of α_s, the structure of $b\bar{b}g$ events and the measurement of the B hadron energy. The B physics presented consists of measurements of the B^+ and B^0 lifetimes, a search for charmless and measurement of double-charm B decays, and B_s^0-$\overline{B_s^0}$ mixing analyses. The excluded values of the oscillation frequency Δm_s will be discussed. The results presented are preliminary.

*This work was supported in part by the US Department of Energy and the UK Particle Physics and Astronomy Research Council.

© 1998 by David J. Jackson.

1 Introduction

Electrons and positrons from the two-mile long Stanford Linear Accelerator are brought into head-on collision by the Stanford Linear Collider (SLC) at a centre-of-mass energy of 91.2 GeV. The SLC Large Detector (SLD) collects data from the Z^0 decays produced in such collisions. About 70% of the Z^0s decay into quark (u, d, s, c or b flavor) anti-quark pairs which in turn hadronise into final state jets of particles, hence the opportunity to perform tests of QCD. During the period 1993-1998 around 400,000 hadronic Z^0 decays were recorded.

While the Z^0 boson has been studied in detail by each of the four LEP experiments at CERN (each has recorded around 4,000,000 Z^0 events), the unique features of the SLC and SLD allow for complementary analysis. The high longitudinal polarization of the SLC e^- beam allows the selection of quark or anti-quark jets. In addition, the powerful vertex reconstruction capability of the SLD 3D CCD pixel vertex detector,[1] upgraded before the 1996 run, provides a unique opportunity to tag and study B decays.

1.1 B–Tagging

The excellent vertexing capabilities at SLD are exploited in the reconstruction of B decays in $Z^0 \to b\bar{b}$ events. The analysis begins with the selection of hadronic Z^0 events. A detailed description of the experimental procedure can be found elsewhere.[2] Topological vertex reconstruction is applied separately to the tracks in each hemisphere (defined with respect to the event thrust axis) as described in detail in Ref. 3. The vertices are reconstructed in 3D co-ordinate space by defining a vertex function $V(\mathbf{r})$ at each position $\mathbf{r}$ in terms of the track parameters such that it is large in regions of high track multiplicity. Maxima are found in $V(\mathbf{r})$ and clustered into resolved spatial regions. Tracks are associated with these regions to form a set of topological vertices. The efficiency for reconstructing at least one secondary vertex in a b hemisphere is $\sim$ 50% ($\sim$ 67%) for VXD2 (VXD3 - the upgraded vertex detector).

The mass M of the reconstructed vertex is calculated by assuming each track has the mass of a pion. The minimum transverse momentum P_T of the vertex tracks relative to the vertex axis (which is varied within the 1σ limits constraining the axis at the measured interaction point (IP) and reconstructed seed vertex) is calculated in order to determine the P_T corrected mass:

$$M_{P_T} = \sqrt{M^2 + P_T^2} + |P_T|. \quad (1)$$

This quantity is the minimum mass the decaying hadron could have in order to produce a vertex with the quantities M and P_T. The tag is applied to each event hemisphere, with the resulting M_{P_T} distribution for Monte Carlo and data events as shown in Fig. 1.

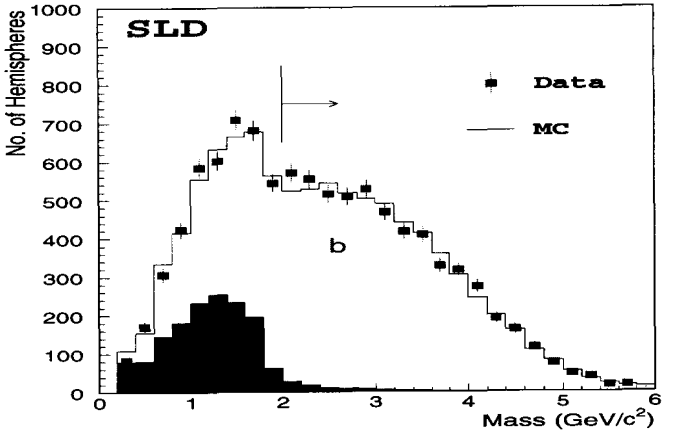


Fig. 1. The P_T corrected mass, M_{P_T}, for b (clear histogram) c (light shaded) and uds (dark shaded) Monte Carlo and data (points) events.

Requiring $M_{P_T} > 2$ GeV/c^2 (above the charm hadron mass) selects a sample of B hadrons with a purity of $\sim$ 98% and and efficiency of $\sim$ 50% using VXD3.[4] This topological vertex reconstruction is used in most of the B physics measurements as well as in some of the QCD analyses.

2 QCD

2.1 Hadron Production and Correlations[5]

The analysis in this section is based on the 1993–95 SLD data sample. The selection of tracks and hadronic events is described in Ref. 5. Samples of events enriched in light and b primary flavors were selected with a more primitive flavor

tagging technique based on signed impact parameters δ of charged tracks with respect to the IP in the plane transverse to the beam. For each event we define n_{sig} to be the number of tracks passing a set of impact-parameter quality cuts that have impact parameter greater than three times its estimated error, $\delta > 3\sigma_\delta$. Events with $n_{sig} = 0$ were assigned to the light-tagged sample and those with $n_{sig} \geq 3$ were assigned to the b-tagged sample. The remaining events were classified as a c-tagged sample. The light-, c- and b-tagged samples comprised 60.4%, 24.5% and 15.2% of the selected hadronic events, respectively.

Separate samples of hemispheres enriched in light-quark and light-antiquark jets were selected from the light-tagged event sample by exploiting the large electroweak forward-backward production asymmetry with respect to the beam direction. The event thrust axis was used to approximate the initial $q\bar{q}$ axis and was signed such that its z-component was along the electron beam direction, $\hat{t}_z > 0$. Events in the central region of the detector, where the production asymmetry is small, were removed by the requirement $|\hat{t}_z| > 0.2$, leaving 74% of the light-tagged events. The quark-tagged hemisphere in events with left- (right-)handed electron beam polarization was defined to comprise the set of tracks with positive (negative) momentum projection along the signed thrust axis. The remaining tracks in each event were defined to be in the antiquark-tagged hemisphere. For the selected event sample, the average magnitude of the polarization was 0.73. Using this value and assuming Standard Model couplings, a tree-level calculation gives a quark (antiquark) purity of 0.73 in the quark-(antiquark-)tagged sample.

2.1.1 Total Production Cross Sections

In this section we discuss details of the analysis for three categories of identified hadrons: charged tracks identified as $\pi^\pm$, $K^\pm$ or p/$\bar{\text{p}}$ in the SLD Cerenkov Ring Imaging Detector (CRID), see Fig. 2; K^0_s and $\Lambda^0/\bar{\Lambda}^0$ reconstructed in their charged decay modes and tagged by their long flight distance; and $K^{*0}/\bar{K}^{*0}$ and ϕ reconstructed in charged decay modes including one and two identified $K^\pm$, respectively.

We have integrated our differential cross sections over their respective measurement ranges, Fig. 3, taking into account the bin-to-bin correlations in the systematic errors. In order to quote total cross sections, we must extrapolate into the unmeasured regions of x_p, and we have done this using three MC models. From

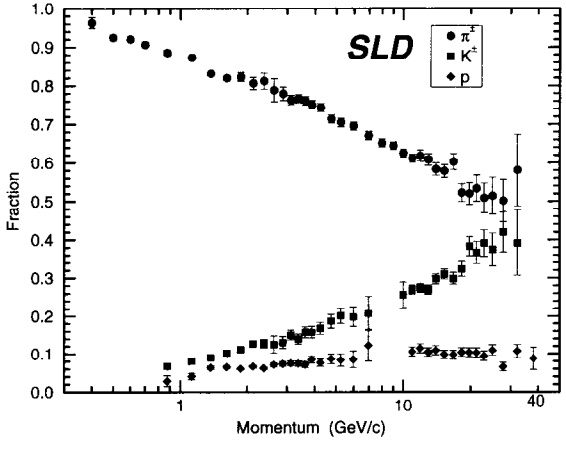


Fig. 2. Charged particle identification with the CRID.

the hadrons of each species generated using each of these models, we calculated the fraction that were generated with x_p in the range of our measurement. For each hadron species the three fractions were found to be similar, with the UCLA (HERWIG) fraction being typically 1% larger (1–2% smaller) than the JETSET fraction.

We applied the same procedure to our measurements for the three flavor categories. The three simulations were found to give similar flavor dependences, with the accepted fraction in b (c) events typically 0.02 (0.01) larger than that in light-flavor events.

2.1.2 Leading Particle Effects

We extended these studies to look for differences between particle and antiparticle production in light quark (as opposed to antiquark) jets, in order to address the question of whether e.g. a primary u-initiated jet contains more hadrons that contain a valence u-quark (e.g. π^+, K^+, p, Λ^0) than hadrons that do not (e.g. π^-, K^-, $\bar{\text{p}}$, $\bar{\Lambda}^0$). To this end we used the light quark- and antiquark-tagged hemispheres described in section 2.1.

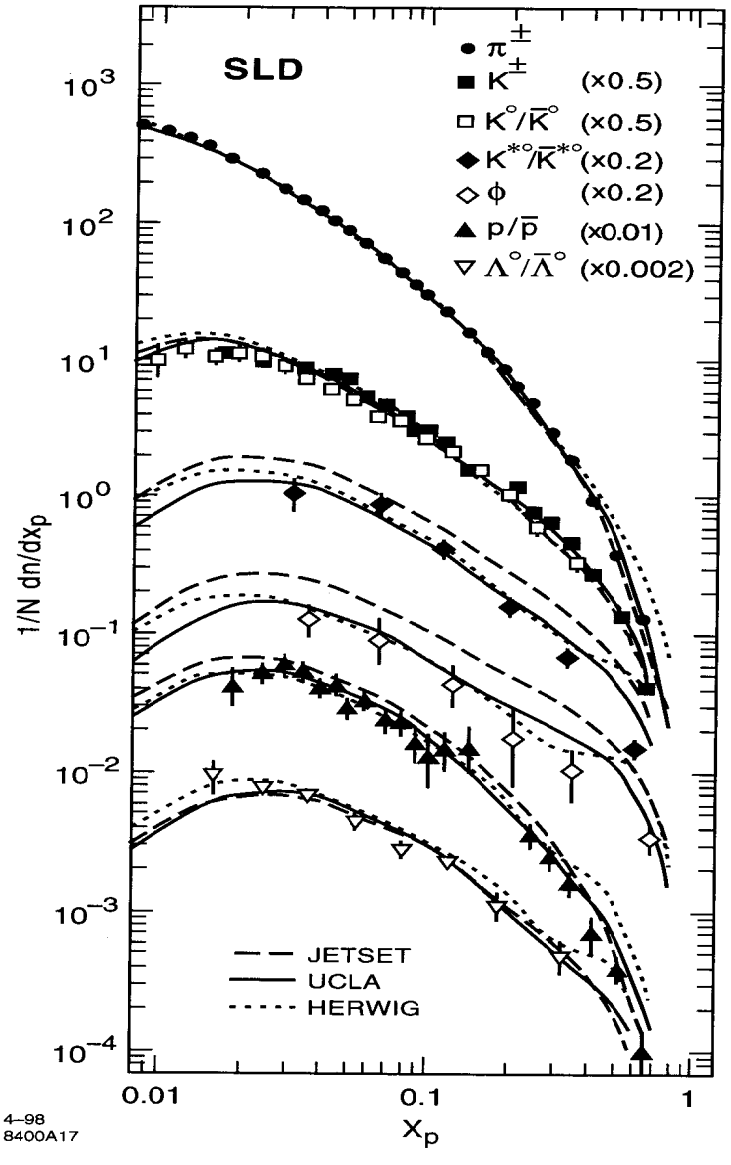


Fig. 3. Particle production in *uds* tagged events.

We measured the differential cross sections per light quark jet

$$R_h^q = \frac{1}{2N_{evts}} \frac{d}{dx_p} \left[N(q \to h) + N(\bar{q} \to \bar{h}) \right], \quad (2)$$

$$R_{\bar{h}}^q = \frac{1}{2N_{evts}} \frac{d}{dx_p} \left[N(q \to \bar{h}) + N(\bar{q} \to h) \right], \quad (3)$$

where: q and $\bar{q}$ represent light-flavor quark and antiquark jets respectively; N_{evts} is the total number of events in the sample; h represents any of the identified hadron species π^-, K^-, $\overline{K}^{*0}$, p, or Λ^0, and $\bar{h}$ indicates the corresponding antihadron. Then, for example, $N(q \to h)$ is the number of hadrons of species h in light quark jets. This formulation assumes CP symmetry, i.e. $N(q \to h) = N(\bar{q} \to \bar{h})$, which was found to be satisfied in the data in all cases.

The charged hadron fractions analysis was repeated on the sample of positively charged tracks in the quark-tagged jets and negatively charged tracks in the antiquark-tagged jets, yielding measured values of $R_{\pi^+}^q$, $R_{K^+}^q$, and R_{p}^q in the tagged samples. The same procedure applied to the remaining tracks yielded $R_{\pi^-}^q$, $R_{K^-}^q$, and $R_{\bar{\mathrm{p}}}^q$. The $K^{*0}/\bar{K}^{*0}$ and $\Lambda^0/\bar{\Lambda}^0$ analyses were applied similarly to the quark- and antiquark-tagged jets to yield $R_{\overline{K}^{*0}}^q$, $R_{K^{*0}}^q$, R_{Λ}^q and $R_{\bar{\Lambda}}^q$.

It is convenient to show the results in the form of the difference between hadron h and antihadron $\bar{h}$ production normalized by the sum:

$$D_h = \frac{R_h^q - R_{\bar{h}}^q}{R_h^q + R_{\bar{h}}^q}. \quad (4)$$

A value of zero corresponds to equal production of hadron and antihadron, whereas a value of +(−)1 corresponds to complete dominance of (anti)particle production. As shown in Fig. 4 in each case the difference is consistent with zero at low x_p. For charged pions it is also consistent with zero at high x_p, but for the other hadrons there are significant positive differences that appear to increase with increasing x_p.

The results for the baryons [Figs. 4(a) and 4(b)] afford the most straightforward interpretation. Since baryons contain valence quarks and not antiquarks, the observed excess of both protons and Λ^0s over their respective antibaryons for $x_p > 0.2$ is clear evidence for the production of leading baryons. The data suggest that the effect increases with x_p, however more data are needed to study the x_p dependence in detail. For $x_p < 0.2$ the data are consistent with equal production of baryons and antibaryons, however the contribution from fragmentation is very

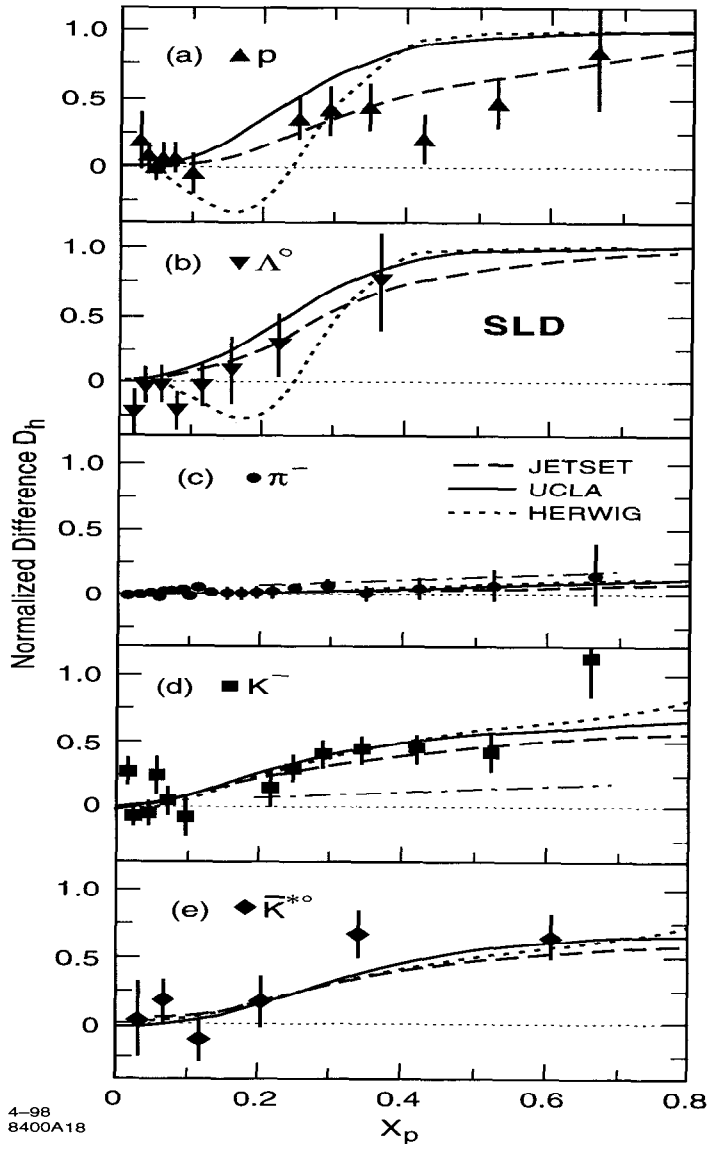


Fig. 4. Observation of leading particle effect.

high in this region and we cannot exclude that leading baryons are also produced at low x_p. Differences are also observed for both pseudoscalar and vector K-mesons, which indicate not only leading production of these two hadron species but also that leading strange mesons are produced more often from initial s quarks than from initial u or d quarks.

2.1.3 Hadron Correlations

We present a study of correlations in rapidity between pairs of identified pions, kaons and protons in hadronic Z^0 decays into light flavors. We observe excesses of opposite-charge over same-charge pairs for all pair types at low values of the absolute rapidity difference, $\mid y_1 - y_2 \mid$ in Fig. 5, indicating that there is a high degree of local conservation of baryon number, strangeness and electric charge in the fragmentation process.

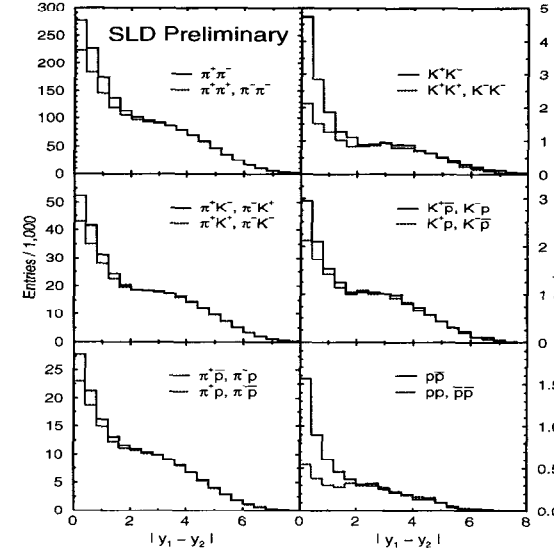


Fig. 5. Short range correlations due to local conservation of charge, strangeness and baryon number.

The range of these short-range correlations is found to be independent of momentum but to depend on the types of hadrons forming the pair. The predictions

of the JETSET fragmentation model are found to provide a qualitative description of the data, although they fail to describe the shapes of some of the correlations in detail. We observe a large excess of K^+K^- pairs over same-charge kaon pairs at large values of the absolute rapidity difference, and the effect is larger for higher momenta, as expected from leading kaon production in $s\bar{s}$ events. Considering only tracks with momentum greater than 9 GeV/c, we observe significant long-range correlations between opposite-charge $\pi\pi$, πK and Kp pairs, as well as KK pairs. The simulation provides a good description of the data in general, but predicts smaller long-range πK and Kp correlations than are observed, and predicts a small long-range πp correlation that is not observed.

The SLC electron beam polarization is used to tag the quark hemisphere in each event, allowing the first study of rapidities signed such that positive rapidity is along the quark rather than antiquark direction. Distributions of signed rapidities and of ordered differences between signed rapidities provide new insights into leading particle production and several new tests of fragmentation models.

We have studied distributions of rapidity signed so that positive rapidity corresponds to the quark (rather than antiquark) direction. Differences between signed rapidity distributions for positive and negative hadrons of all three types are observed, giving further evidence for leading production of charged pions, kaons, and protons. The distribution of the difference between the signed rapidities of p and $\bar{\text{p}}$ shows a large asymmetry at small values of the absolute rapidity difference indicating the ordering of baryons along the event axis. Ordered rapidity differences were also studied for other pair types, providing additional tests of fragmentation models.

2.2 Flavor Independence of α_s [6]

In this section we present a comparison of the strong coupling of the gluon to light ($q_l = u{+}d{+}s$), c, and b quarks, determined from multijet rates in flavor-tagged samples of hadronic Z^0 decays recorded at SLD between 1993 and 1995. Flavor separation among primary $q_l\bar{q}_l$, $c\bar{c}$, and $b\bar{b}$ final states was made on the basis of the topological vertex reconstruction described in section 1.1.

For the study of flavor-independence the jet structure of events was reconstructed in turn using six iterative clustering algorithms. We used the 'E', 'E0', 'P', and 'P0' variations of the JADE algorithm, as well as the 'Durham' ('D') and 'Geneva' ('G') algorithms.[7] In each case events were divided into two categories: those containing (i) two jets, and (ii) three or more jets. The fraction of the event sample in category (ii) was defined as the 3-jet rate R_3. This quantity is infrared- and collinear-safe and has been calculated to $\mathcal{O}(\alpha_s^2)$ in perturbative QCD.[7,8] For each algorithm we repeated the subsequent analysis successively across a range of values of the normalised jet-jet invariant-mass parameter y_c, $0.005 \leq y_c \leq 0.12$. In the final stage an 'optimal' y_c value was chosen for each algorithm so as to minimise the overall error on the analysis.

Each of the six jet-finding algorithms was applied to each tagged-event subsample j, ($j = 1,2,3$ for the uds, c, b subsamples) For each algorithm the 3-jet rate in each subsample was calculated. For the test of the flavor-independence of strong interactions it is convenient to consider the ratios (in which many systematic errors should cancel) of the 3-jet rates in heavy- and light-quark events, namely R_3^c/R_3^{uds} and R_3^b/R_3^{uds}. These were derived from the unfolded R_3^{uds}, R_3^c and R_3^b values.

The test of the flavor-independence of strong interactions can be expressed in terms of the ratios $\alpha_s^i/\alpha_s^{uds}$ ($i = c$ or b). Recalling that with our definition R_3 is the rate of production of 3 or more jets, $\alpha_s^i/\alpha_s^{uds}$ can be derived from the respective measured ratio R_3^i/R_3^{uds} using the next-to-leading-order perturbative QCD calculation:

$$\frac{R_3^i}{R_3^{uds}} = \frac{A^i\,\alpha_s^i + B^i(\alpha_s^i)^2 + O((\alpha_s^i)^3)}{A^{uds}\,\alpha_s^{uds} + B^{uds}(\alpha_s^{uds})^2 + O((\alpha_s^{uds})^3)} \tag{5}$$

where the coefficients A and B represent, respectively, the leading-order (LO) perturbative QCD coefficient for the 3-jet rate and the next-to-leading-order (NLO) coefficient for this rate.

For illustration, the measured ratios R_3^c/R_3^{uds} and R_3^b/R_3^{uds}, are shown in Fig. 6(a). R_3^b/R_3^{uds} lies above unity for the E, E0, P and P0 algorithms, and below unity for the D and G algorithms; note that all six data points are highly correlated with each other, so that the differences between algorithms are more significant than naively implied by the statistical errors displayed. For comparison, the corresponding QCD calculations of R_3^b/R_3^{uds} are also shown by the arrows in Fig. 6(a), under the *assumption* of a flavor-independent strong coupling with an input value of $\alpha_s(M_Z^2) = 0.118$, for $m_b(M_{Z^0}) = 3.0 \pm 0.5$ GeV/c^2. Under this assumption the calculations are in good agreement with the data, and the data clearly demonstrate the effects of the non-zero b-quark mass, which are larger than

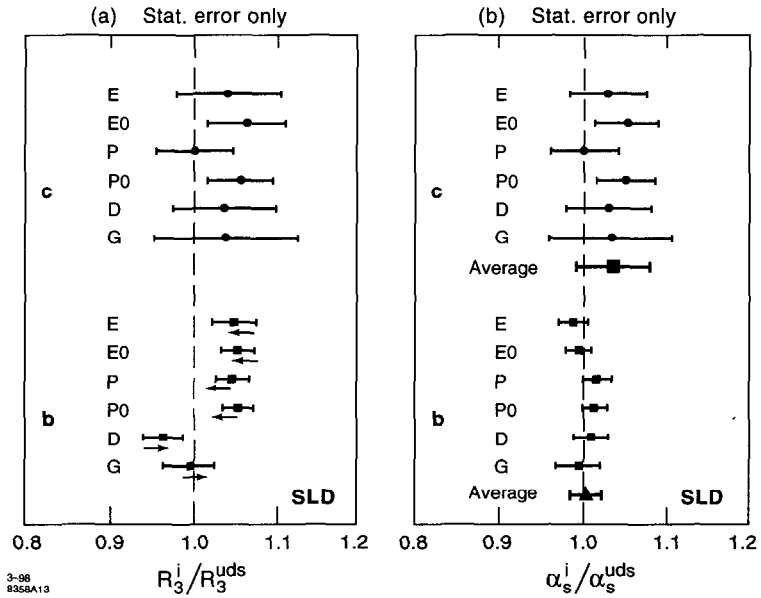


Fig. 6. Measured 3-jet rate ratios and corresponding α_s ratios for the 6 jet algorithms.

the statistical error. For the translation from R_3^b/R_3^{uds} to $\alpha_s^b/\alpha_s^{uds}$ we used a value of the running b-quark mass $m_b(M_{Z^0}) = 3.0$ GeV/c^2.

The $\alpha_s^c/\alpha_s^{uds}$ and $\alpha_s^b/\alpha_s^{uds}$ ratios are summarised in Fig. 6b. It can be seen that the ratios determined using the different jet algorithms are in good agreement with one another.

We obtained:

$$\alpha_s^c/\alpha_s^{uds} = 1.036 \pm 0.043(stat.)^{+0.041}_{-0.045}(syst.)^{+0.020}_{-0.018}(theory)$$
$$\alpha_s^b/\alpha_s^{uds} = 1.004 \pm 0.018(stat.)^{+0.026}_{-0.031}(syst.)^{+0.018}_{-0.029}(theory).$$

The theoretical uncertainties are only slightly smaller than the respective experimental systematic errors, and comprise roughly equal contributions from the hadronisation and translation uncertainties, as well as from the r.m.s. deviation over the six jet algorithms. We find that the strong coupling is independent of quark flavor within our sensitivity. Alternatively, under the assumption of the flavor independence of α_s, we derive $m_b(M_{Z^0}) = 3.23^{+1.02}_{-1.81}$ GeV/c^2.

2.3 Structure of $b\bar{b}g$ Events[9]

The observation of e^+e^- annihilation into final states containing three hadronic jets, and their interpretation in terms of the process $e^+e^- \to q\bar{q}g$, provided the first direct evidence for the existence of the gluon, the gauge boson of the theory of strong interactions, quantum chromodynamics (QCD). Following these initial observations, studies of the partition of energy among the three jets were performed at the DESY e^+e^- collider PETRA and SLAC e^+e^- storage ring PEP. Comparison of the data with leading-order QCD predictions, and with a model incorporating the radiation of spin-0 (scalar) gluons, provided qualitative evidence for the spin-1 (vector) nature of the gluon, which is a fundamental element of QCD.

In these studies the gluon jet was not explicitly tagged. Instead the jets were energy ordered and the lowest-energy jet was assumed typically to be the gluon jet. If the gluon jet could be tagged event-by-event more detailed studies of the structure of QCD could be performed. This is now possible at SLD using three-jet $b\bar{b}g$ events.

The chromomagnetic moment of the bottom quark is induced at the one-loop level in QCD and is of the order α_s/π. One can also write down an *ad hoc* Lagrangian[10] with a $b\bar{b}g$ coupling modified via anomalous chromoelectric and chromomagnetic moments:

$$\mathcal{L}^{b\bar{b}g} = g_s\bar{b}T_a\{\gamma_\mu + \frac{i\sigma_{\mu\nu}k^\nu}{2m_b}(\kappa - i\tilde{\kappa}\gamma_5)\}bG_a^\nu \qquad (6)$$

where g_s is the strong charge, T_a are the $SU(3)_c$ generators, m_b is the bottom quark mass, k is the outgoing gluon momentum, and κ and $\tilde{\kappa}$ parameterize the anomalous chromomagnetic and chromoelectric moments, respectively, which might arise from physics beyond the Standard Model. The effect of the former on three-jet observables has been calculated recently.[10] The latter is CP-violating, and in this analysis we have not attempted to discriminate between the b and $\bar{b}$ jets and are hence insensitive to non-zero values of $\tilde{\kappa}$ at leading order in $\tilde{\kappa}$. Non-zero values of κ would modify the gluon energy distribution in $b\bar{b}g$ events.

Charged tracks with large transverse impact parameters were used to tag $b\bar{b}$ events. Events with $N_{sig}^{evt} \geq 4$ were selected as $b\bar{b}$ events. This gave a $b\bar{b}$ sample with an estimated purity of 94.7%. Events were retained in which two jets were tagged as b or $\bar{b}$ by requiring them to have a number of significant tracks per jet $N_{sig}^{jet} \geq 2$. The remaining jet in each event was tagged as the gluon.

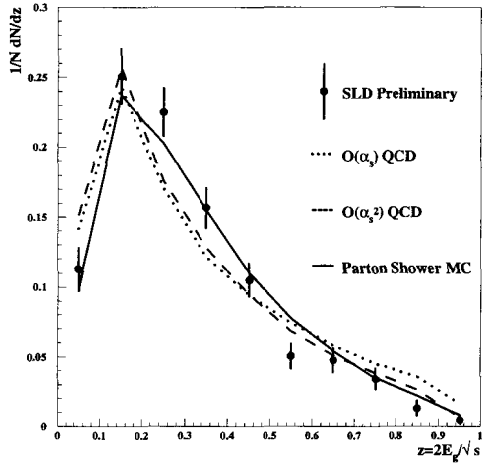


Fig. 7. Gluon energy spectrum. The Parton Shower (PS) MC agrees best with the data (e.g. around $z = 0.3$).

We studied the distribution of scaled gluon energy $z = 2E_{\rm gluon}/\sqrt{s}$. As the absolute value of κ is increased the gluon energy spectrum gets harder, producing an excess of gluon jets with high scaled energy. A difficulty arises in that the $O(\alpha_s)$ QCD calculation does not describe the data as well as the PS calculation, see Fig. 7, so that the higher-order QCD effects included in the PS prediction could be mimicked in the extended $O(\alpha_s)$ calculation by an artificially large anomalous moment κ.

A χ^2 minimisation fit was performed to the PS z distribution corrected with κ as the only free parameter. We find:

$$\kappa = -0.030^{+0.061}_{-0.062}(\text{stat.}) \tag{7}$$

We set 95% confidence-level upper limits of $-0.15 < \kappa < 0.09$.

2.4 B Hadron Energy Measurement[11]

The production of heavy hadrons (H) in e^+e^- annihilation provides a laboratory for the study of heavy-quark (Q) jet fragmentation. This is commonly characterised in terms of the observable $x_{E_H} \equiv 2E_H/\sqrt{s}$, where E_H is the energy of a B or D hadron containing a b or c quark, respectively, and $\sqrt{s}$ is the c.m. energy. In contrast to light-quark jet fragmentation, one expects[12] the distribution of x_{E_H}, $D(x_{E_H})$, to peak at an x_{E_H}-value significantly above 0. Since the hadronisation process is intrinsically non-perturbative $D(x_{E_H})$ cannot be calculated directly using perturbative Quantum Chromodynamics (QCD). However, the distribution of the closely-related variable $x_{E_Q} \equiv 2E_Q/\sqrt{s}$ can be calculated perturbatively and related, via model-dependent assumptions, to the observable quantity $D(x_{E_H})$; a number of such models of heavy quark fragmentation have been proposed.[13] Measurements of $D(x_{E_H})$ thus serve to constrain both perturbative QCD and the model predictions.

Here we describe the results of a new method for reconstructing the B-energy, inclusively, using only charged tracks from B decay vertices. An upper limit on the mass of the missing particles is found for each reconstructed B-decay vertex:

$$M_{0max}^2 = M_B^2 - 2M_B\sqrt{M_{ch}^2 + P_t^2} + M_{ch}^2. \tag{8}$$

where M_{ch} and P_t is the measured invariant mass and transverse momentum of the vertex tracks and M_B is the known B hadron mass. The true missing mass M_0 is hence bounded, $0 \leq M_0^2 \leq M_{0max}^2$, and the total B energy is well constrained when M_{0max} is small. In order to obtain an approximately x_B-independent selection efficiency we chose the *ad hoc* upper cut:

$$M_{0max}^2 < \left[1.1 + 6\frac{E_{beam} - E_B^{rec}}{1000} + 3.5e^{\frac{-(E_B^{rec}-5.5)}{3.5}}\right]^2$$

where the two terms that depend on the reconstructed energy E_B^{rec} increase the efficiency at lower B hadron energy. The overall efficiency for selecting B hadrons is 3.95% and the estimated B hadron purity is 99.5%.

We examined the B-energy resolution of this technique. The normalized difference between the true and reconstructed B hadron energies, $(x_B^{rec} - x_B^{true})/x_B^{true}$, for Monte Carlo events, was fitted by a double Gaussian, resulting in a core width of 10.4% and a tail width of 23.6%, with a core fraction of 83%.

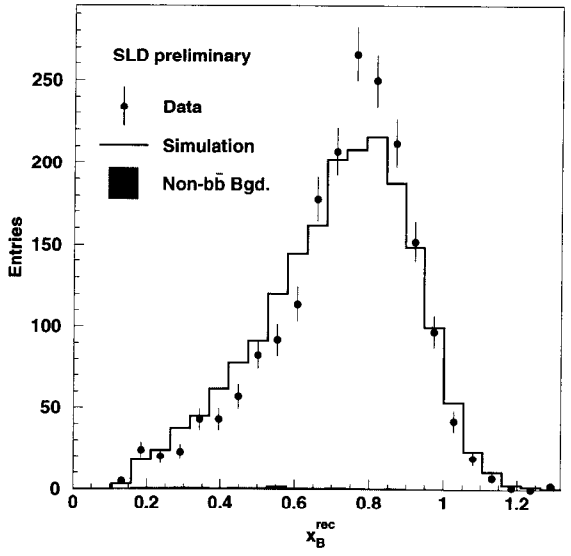


Fig. 8. Distribution of the reconstructed scaled B hadron energy for data (points) and Monte Carlo (histogram). The solid histogram shows the non-$b\bar{b}$ background.

A total of 1938 vertices in the data for 1996–97 satisfy all selection cuts. Figure 8 shows the distribution of the reconstructed scaled B hadron energy for data, $D^{data}(x_B^{rec})$, and for the Monte Carlo simulation, $D^{MC}(x_B^{rec})$. The distribution $D^{data}(x_B^{rec})$ was compared with various perturbative QCD, phenomenological and *ad hoc* model predictions. Each model was treated as an analytic functional form of $D(x_B^{rec})$. It is conventional to evaluate the mean of this distribution, $< x_B >$. For the three functions providing the best description of the data we took the average of the evaluated values of $< x_B >$ as the central result, and defined the unfolding uncertainty to be the r.m.s. deviation. Other systematic errors have not yet been included. We obtained:

$$< x_B > \; = 0.719 \pm 0.005(stat.) \pm 0.001(unfolding)$$

which is consistent with previous measurements. It was found that $< x_B >$ is relatively insensitive to the variety of allowed forms of the shape of the fragmentation function $D(x_B)$.

3 B Physics

3.1 B^+ and B^0 Lifetimes[14]

The spectator model predicts that the lifetime of a heavy hadron depends upon the properties of the constituent weakly decaying heavy quark Q and is independent of the remaining, or spectator, quarks in the hadron. This model fails for the charm hadron system for which the lifetime hierarchy $\tau_{D^+} \sim 2\tau_{D_s^+} \sim 2.5\tau_{D^0} \sim 5\tau_{\Lambda_c^+}$ is observed. Since corrections to the spectator model are predicted to scale with $1/m_Q^2$ the B meson lifetimes are expected to differ by less than 10% (Ref. 15). Hence a measurement of the B^+ and B^0 lifetimes provides a test of this prediction. In addition, the specific B meson lifetimes are needed to determine the element V_{cb} of the CKM matrix.

The analysis is performed using topologically reconstructed B decays in the 1993-98 data. The distance from the IP to the B decay location is the reconstructed decay length. Since the purity of the B charge reconstruction is lower for decays close to the IP, where tracks are more likely to be wrongly assigned, decay lengths are required to be > 1 mm. To avoid using the vertices of tracks originating from interactions with the detector material, the radial decay length is required to be < 24 mm for VXD2 (< 22 mm for VXD3), i.e. more than 1 mm inside the SLC beampipe.

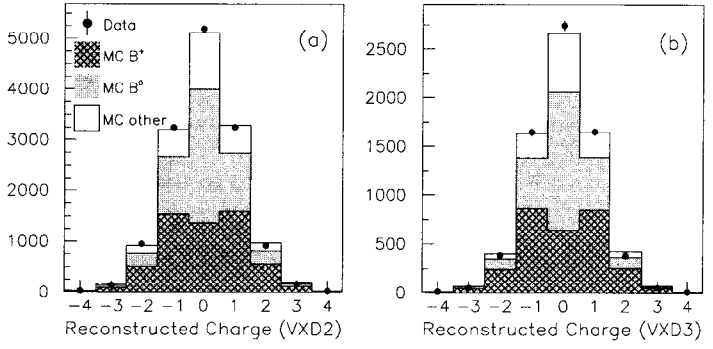


Fig. 9. Reconstructed vertex charge for data (points) and Monte Carlo (histogram) for (a) 1993-5 and (b) 1996.

Figure 9 shows a comparison of the reconstructed charge between data and

Monte Carlo for (a) VXD2 and (b) VXD3 (1996 only). At this stage the charged sample consists of 30028 vertices with vertex charge equal to $\pm$ 1,2 or 3, while the neutral sample consists of 19636 vertices with charge equal to 0 for the 1993-98 data. Monte Carlo studies indicate that for VXD3 the charged sample is 97.6% pure in B hadrons consisting of 55.1% B^+, 32.0% B^0, 8.6% B_s^0, and 4.3% B baryons. Similarly, the neutral sample is 98.2% pure in B hadrons consisting of 24.1% B^+, 54.1% B^0, 15.0% B_s^0 and 6.8% B baryons. The statistical precision of the measurement depends on the separation between the B^+ and B^0 in these samples.

The charge purity is a function of M_{P_T} and events are weighted accordingly. The initial state $b/\bar{b}$ tag (see section 3.5) is also used to enhance the charged sample purity by giving a higher (lower) weight to the B^+ hypothesis if the vertex charge agrees (disagrees) with the $b/\bar{b}$ tag.

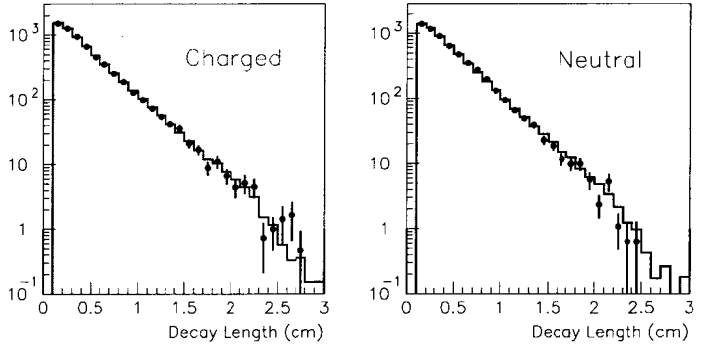


Fig. 10. Decay length distributions for 1997-8 data (points) and best fit Monte Carlo (histogram).

The B^+ and B^0 lifetimes are extracted from the decay length distributions of the B vertices in the charged and neutral samples using a binned χ^2 fit. These distributions are fitted simultaneously to determine the B^+ and B^0 lifetimes. Figure 10 shows the reconstructed decay length for data and best fit MC for the 1997-8 data and the charged and neutral samples.

The systematics are determined separately for the 1993-5, 1996 and 1997-8 measurements and then combined taking into account correlated uncertainties. The combined 1993-8 measurements yield the following result:

$$\begin{aligned} \tau_{B^+} &= 1.686 \pm 0.025(\text{stat}) \pm 0.042(\text{syst}) \text{ ps}, \\ \tau_{B^0} &= 1.589 \pm 0.026(\text{stat}) \pm 0.055(\text{syst}) \text{ ps}, \\ \frac{\tau_{B^+}}{\tau_{B^0}} &= 1.061 \pm^{0.031}_{0.029} (\text{stat}) \pm 0.027(\text{syst}). \end{aligned}$$

These results are consistent with the expectation that the B^+ lifetime is up to 10% greater than that of the B^0 and have the best statistical accuracy of the current measurements.

3.2 Search for Charmless B Decays

Four exclusive charmless hadronic channels listed in the first column of table 1 have been studied. The decays involve either $\rho^0 \to \pi^+\pi^-$ or $K^{*0} \to K^+\pi^-$ producing a 3-prong final state. Hence the initial selection requires a 3-prong vertex (reconstructed from all track combinations) with a total charge of ± 1. A Fischer discriminant function is constructed from eight kinematic variables, each parametrized for signal and background. The variables used include vertex mass and energy, ρ^0 or K^{*0} mass, and the centre-of-mass helicity angle. The resulting Fischer discriminant distribution for signal and background is shown in Fig. 11.

	Theory	CLEO	DELPHI	ALEPH	SLD	SLD ϵ
$B^+ \to \rho^0\pi^+$	0.4 – 1.4	5.8	16	3.2	8.2	0.50
$B^+ \to \rho^0 K^+$	0.01 – 0.06	1.4	12		9.2	0.45
$B^+ \to K^{*0}\pi^+$	0.6 – 0.9	3.9	39		7.9	0.52
$B^+ \to K^{*0}K^+$					9.0	0.46

Table 1. Theoretical prediction and experimental 90% C.L. upper limits (BR$\times 10^{-5}$). SLD signal efficiency, ϵ, is shown in last column.

The cut in the Fischer discriminant is chosen such as to minimize the expected upper limit using Monte Carlo. The ability to compete with other experiments is due to the high vertexing efficiency at SLD. The signal efficiencies are around 50% as listed to the right in table 1. No candidate signal events are seen in the 1993-98 SLD data from which the SLD limits shown in the table are derived.

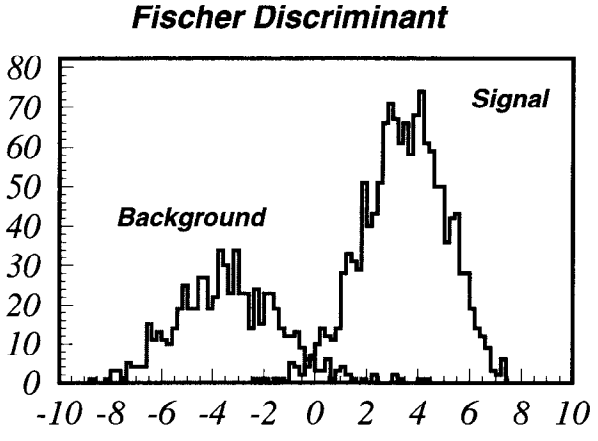


Fig. 11. Fischer discriminant based on eight kinematic variables.

3.3 Search for $b \to sg$ [16]

The low measured B semileptonic decay branching ratio (BR) and charm production deficit in B decays compared to theoretical expectations has been a persistent puzzle in B physics. This has left room for B decays from possible mechanisms beyond the Standard Model with BR as large as 10%. In the Standard Model, flavor changing neutral current decays of the type $b \to s\, g^{(*)}$ (including $b \to sg$ and $b \to sq\bar{q}$) have a total branching ratio of $\approx$ 1% (Ref. 17). One hypothesized scenario is that the $b \to sg$ decay is enhanced to a BR of 10% by new physics.[18] Such a scenario is consistent with all existing experimental constraints, including the measured BR($b \to s\gamma$), and would nicely explain the above puzzles.

The signal for $b \to sg$ is a high momentum kaon from the B decay. In experiments where the B's have significant boost, such as SLD at the Z^0 pole, it is rather difficult to derive the momentum of the B decay kaon in the B rest frame due to the uncertainty in the B boost. We circumvent this difficulty by using the kaon momentum transverse to the B flight direction (p_t). A key factor making this analysis viable at SLD is the excellent particle identification capability of the SLD Cerenkov Ring Imaging Detector to allow efficient use of $K^{\pm}$ production. The analysis is performed using the 300,000 Z^0 decays collected during 1993-1997.

For the evaluation of analysis sensitivity to the $b \to sg$ process, we used the JETSET inspired model[19] with a specific set of parameter choices as a bench mark test. Figure 12 shows the projected p_t spectrum for $K^{\pm}$ produced in standard $b \to c$ decays and in $b \to sg$ decays for BR($b \to sg$) = 10% and 15% from this model. For $p_t > 1.8$, the ratio of $b \to sg$ to $b \to c$ approaches unity for $BR(b \to sg) \approx 10\%$ so that a $b \to sg$ signal should appear as a clear excess in this region.

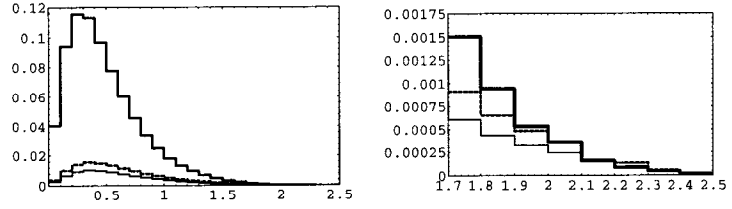


Fig. 12. $K^{\pm}$ p_t spectra (GeV/c) for standard $b \to c$ (heavy line) and for $b \to sg$ with BR=10% (light line) and BR=15% (medium line). The first plot shows the whole spectrum and the second shows the high-p_t region.

Charm lifetimes are long enough that tracks arising from their cascade decays will often be inconsistent with the B-decay vertex. In contrast, all tracks associated with a $b \to sg$ decay should fit to a single vertex. The data is therefore split into two samples: "1-Vertex" where all tracks identified as coming from the B decay cascade fit to a single vertex, and "2-Vertex" where they do not; $b \to sg$ events should show up primarily in the 1-Vertex sample.

$K^{\pm}$,$p_t > 1.8$	1 vertex	All
	# events	# events
Data	37.0	93.0
$b \to c$ MC	29.0	78.2
Difference	8.0 ± 6.1	14.8 ± 9.6
BR($b \to sg$)=10%	24.0	33.3

Table 2. Comparison of number of high p_t $K^{\pm}$ expected and observed in the 1- and 2-vertex samples.

Figure 13 shows the p_t spectra for identified $K^{\pm}$ in the 1- and 2-vertex samples. The Monte Carlo was normalized to the data by the number of tagged

b-hemispheres. Table 2 gives the total number of $K^{\pm}$ with $p_t > 1.8$ Gev in both samples. Small excesses are seen in both the 1- and 2-vertex samples indicating that no significant signal for $b \to sg$ is observed.

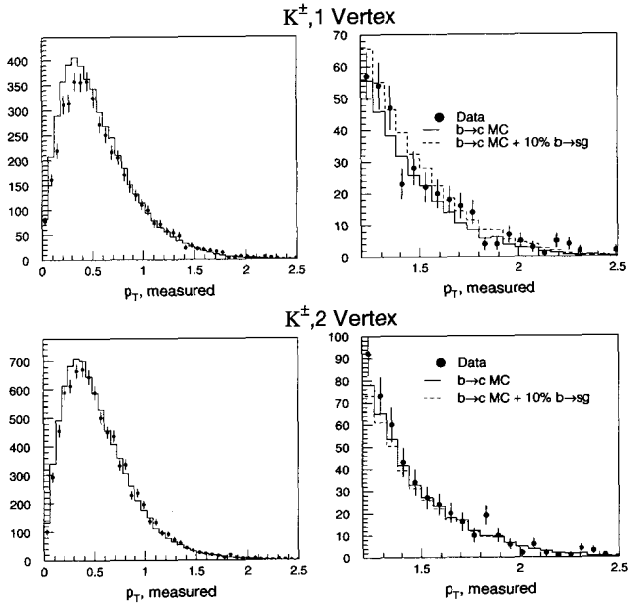


Fig. 13. Comparison of data to Monte Carlo p_t spectrum of identified $K^{\pm}$ in the 1- and 2- vertex samples. The solid line shows the Monte Carlo predicted spectrum of standard $b \to c$ decays. The dashed line shows the predicted spectrum for BR($b \to sg$)=10%.

The observation of kaons with $p_t > 1.8$ GeV/c in the 1-vertex sample yields a result of $8.0 \pm 6.1 \pm 5.9$ events over an expected background of 29 events. After correcting for $K^{\pm}$ identification efficiency, this yields a high-p_t production excess of $(5.0 \pm 3.8 \pm 3.6) \times 10^{-4}$ per B decay. This corresponds to a statistical precision of $\pm 2.5\%$ on the $b \to sg$ branching ratio for a particular set of choices of $b \to sg$ model parameters.

We have demonstrated that this measurement technique already has interesting statistical precision to test the hypothesis of a largely enhanced $b \to sg$ branching ratio at the 10% level, even with a rather small input data sample. Data taken in 1998 will be added in the near future, and this will approximately double the data statistics.

3.4 Double-Charm B Decays

The excellent vertex resolution at SLD also allows a measurement of the process $B \to D\overline{D}X$ resulting from the quark transition $b \to c\bar{c}s$. The tracks from the topologically reconstructed B decays are required to have a 1-vertex fit probability of $< 5\%$ and are separated into a secondary and tertiary vertex. The task is then to estimate whether these two vertices are the B and D vertices of a $B \to$ 1-charm decay or the $\overline{D}$ and D vertices of a $B \to$ 2-charm decay. Two techniques are employed. The first counts the number of identified kaons in each vertex. For $B \to D\overline{D}X$ decays a greater proportion of events will have a kaon in the secondary (D decay) vertex compared with $B \to$ 1-charm decays (where the secondary is the B decay). By comparison with Monte Carlo the fraction of double charm decays in the data $f_{D\overline{D}}$ is extracted, see table 3.

	CLEO	DELPHI	ALEPH	SLD Kaon	SLD D charge
$f_{D\overline{D}}$	18 ± 4	17 ± 4	$25 \pm 4 \pm 4$	$18.8 \pm 2.5 \pm 5.9$	19.5 ± 4.5

Table 3. Measurements of the double charm branching fraction $B \to D\overline{D}X$ (%).

The second technique utilizes the beam polarization and the jet charge to tag the $b/\bar{b}$ flavor of the B. The correlation between this tag and the charge of the tertiary vertex is again sensitive to $f_{D\overline{D}}$. This second measurement gives a result consistent with the kaon analysis, as can be seen in table 3. Current work is aimed at combining these techniques and reducing the systematic error.

3.5 B_s Mixing[20]

Transitions between B^0 and $\overline{B^0}$ mesons take place via second order weak interactions. In the Standard Model, a measurement of the oscillation frequency Δm_d for B^0_d–$\overline{B^0_d}$ mixing determines, in principle, the value of the Cabibbo-Kobayashi-Maskawa matrix element $|V_{td}|$, which is parameterized in terms of the Wolfenstein parameters ρ and (the CP-violating phase) η, both of which are currently poorly

constrained. However, theoretical uncertainties in calculating hadronic matrix elements are large ($\sim$ 25% [Ref. 21]), and thus, limit the current usefulness of precise Δm_d measurements. These uncertainties are significantly reduced ($\sim$ 6–10%) for the ratio between Δm_d and Δm_s. Thus, combining measurements of the oscillation frequency of both B_d^0–$\overline{B_d^0}$ and B_s^0–$\overline{B_s^0}$ mixing translates into a measurement of the ratio $|V_{td}|/|V_{ts}|$ and provides a stronger constraint on the parameters ρ and η.

Experimentally, a measurement of the time dependence of B^0–$\overline{B^0}$ mixing requires three ingredients: (i) the B decay proper time has to be reconstructed, (ii) the B flavor at production (initial state $t = 0$) needs to be determined, as well as (iii) the B flavor at decay (final state $t = t_{\rm decay}$). At SLD two methods using the same initial state flavor tag but using different techniques to reconstruct the B decay and tag its final state flavor have been studied. The data consists of 250,000 hadronic Z^0 decays collected with the upgrade vertex detector (VXD3) during 1996-98.

The probability that a meson created as a B_s^0 ($\overline{B_s^0}$) will decay as a $\overline{B_s^0}$ (B_s^0) after proper time t can be written as

$$P_m(t) = \frac{\Gamma}{2}\, e^{-\Gamma t}\, [1 - \cos(\Delta m_s\, t)] \ . \tag{9}$$

where Δm_s is the mass difference between the mass eigenstates, Γ is the average decay width of the two states and P_m denotes the probability to 'mix'. The effects of CP violation are assumed to be small and are neglected.

The probability for a decay to be tagged as mixed is expressed taking into account the fraction of each b-hadron type (B^+, B_d^0, B_s^0, and b-baryon), as well as their lifetimes and mistag probabilities, and $udsc$ background (this gives $\mathcal{P}_m(t, \Delta m_s)$). Detector and vertex selection effects are then introduced by convoluting the above probability functions with a proper time resolution function $\mathcal{R}(T, t)$ and a time-dependent efficiency function $\varepsilon(t)$:

$$P_m(T, \Delta m_s) = \int_0^\infty \mathcal{P}_m(t, \Delta m_s)\ \mathcal{R}(T, t)\ \varepsilon(t)\ dt \ , \tag{10}$$

where t is the "true" time and T is the reconstructed time. Again, a similar expression applies to the unmixed probability P_u. The resolution function is parameterized by the sum of two Gaussians with the core fraction set to 60%. The proper time resolution is a function of proper time and also depends on the measured boost $\gamma\beta = p_B/m_B$, its resolution $\sigma_{\gamma\beta}$ and on the estimate of the decay length resolution σ_L.

A tag of the initial state flavor is provided by the large polarized forward-backward asymmetry for $Z^0 \to b\,\bar{b}$ decays:

$$\tilde{A}_{FB} = 2A_b\ \frac{A_e - P_e}{1 - A_e P_e}\ \frac{\cos\theta_T}{1 + \cos^2\theta_T} \ , \tag{11}$$

where $A_b = 0.94$ and $A_e = 0.150$ (Standard Model values), P_e is the electron beam longitudinal polarization, and θ_T is the angle between the thrust axis and the electron beam direction (the thrust axis is signed such that it points in the same hemisphere as the reconstructed vertex). Thus, left- (right-) polarized electrons tag b ($\bar{b}$) quarks in the forward hemisphere, and $\bar{b}$ (b) quarks in the backward hemisphere.

The tag is enhanced by the addition of other flavor-sensitive quantities derived from tracks in the opposite hemisphere, utilizing: a jet charge technique, the total track charge and charge dipole (see below) of the opposite hemisphere vertex (if found) and the charge of identified kaons or leptons in the opposite hemisphere. The various tags are combined to form an overall initial state tag b-quark probability. The average correct tag probability is 0.80 with full efficiency.

3.5.1 Lepton+D Analysis

The lepton+"D" analysis aims at reconstructing the B and D vertex topologies of semileptonic B decays. It proceeds by first selecting event hemispheres containing an identified lepton (e or μ) with $|\cos\theta| < 0.7$. Then, a D vertex candidate is reconstructed using a similar topological technique as that described earlier. Several cuts are added to clean up the D vertex candidate and reduce the contamination from cascade ($b \to c \to l$) charm semileptonic decays. The final state B^0 or $\overline{B^0}$ flavor is tagged by the sign of the lepton charge.

The B decay vertex is reconstructed by intersecting the lepton and D trajectories. To enhance the fraction of B_s^0 decays, the sum of lepton + D vertex track charges is required to be $Q = 0$. This enhances the B_s^0 fraction to 15.3% of selected b hadrons in the $Z^0 \to b\bar{b}$ MC (the B_s^0 production fraction in the $Z^0 \to b\bar{b}$ MC is 11.5%). A sample of 1492 decays is thus obtained in the 1996-98 data.

Reconstruction of the b-hadron boost uses both tracking and calorimeter information. The overall performance of the decay length and boost measurements

for B_s^0 decays proceeding via the direct $(b \to l)$ transition is shown in Fig. 14.

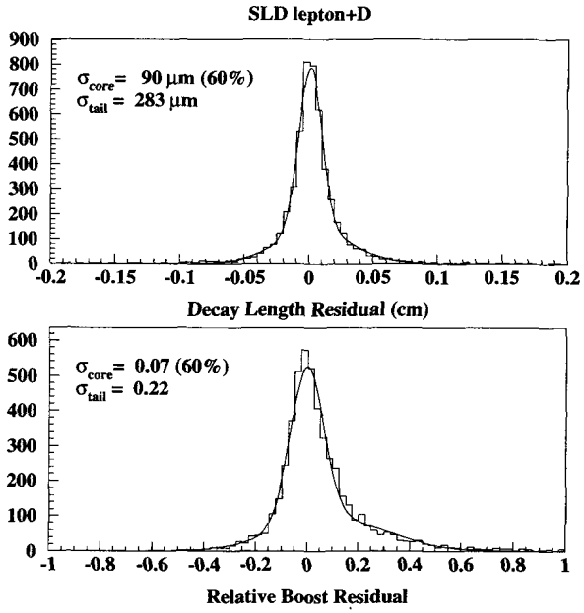


Fig. 14. Distributions of the decay length and relative boost residuals for B_s^0 $(b \to l)$ decays in the simulation.

For each decay, the resolution σ_L is computed from the vertex fit and IP position measurement errors. The relative boost residual $\sigma_{\gamma\beta}/\gamma\beta$ is parameterized as a function of the lepton + D vertex total track energy, with parameters extracted from the MC simulation. Similarly, the efficiency $\varepsilon(t)$ is parameterized using the MC simulation. All parameterizations are performed separately for each b-hadron type.

3.5.2 Vertex Charge Dipole Analysis

The Charge Dipole analysis aims at reconstructing the B and D vertex topologies in inclusive decays, and tags the B_s^0 or $\overline{B_s^0}$ decay flavor based on the charge difference between the B and D vertices. This analysis technique is unique to SLD. Hemispheres containing an inclusive topological vertex with $M_{P_T} > 2$ GeV/c^2 are selected and the total vertex track charge Q is required to be 0. To select decays with non-negligible separation between the B and D decay points, the probability for fitting all tracks to a single vertex is required to be less than 1%. The tracks are then rearranged into various two-vertex combinations and the combination with the lowest overall χ^2 is selected. The vertex that is closer to the IP is labelled "B" and that further away is labelled "D."

A "Charge Dipole" is defined as $\delta Q \equiv D_{BD} \times SIGN(Q_D - Q_B)$, where D_{BD} is the distance between the two vertices and Q_B (Q_D) is the charge of the B (D) vertex. Positive (negative) values of δQ tag $\overline{B_s^0}$ (B_s^0) decays and the correct tag probability increases with increasing $|\delta Q|$.

For all data and MC events, hemispheres already containing a vertex selected by the lepton+D analysis are removed such that the two analyses are statistically uncorrelated. A total of 5719 decays remain in the dipole sample. Figure 15 displays the distribution of charge dipole δQ for the data sample and also indicates the separation between b hadrons containing b or $\bar{b}$ quarks in the MC.

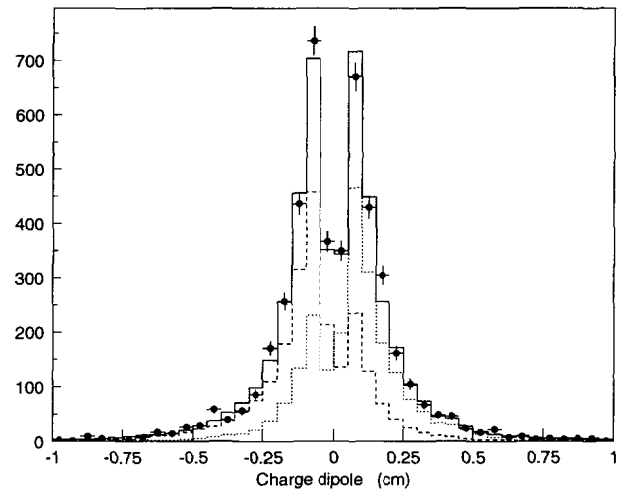


Fig. 15. Distribution of the vertex charge dipole for data (points) and Monte Carlo (solid histogram). Also shown are the contributions from b hadrons containing a b quark (dotted histogram) or a $\bar{b}$ quark (dashed histogram).

3.5.3 Amplitude Fit

The study of the time dependence of B_s^0–$\overline{B_s^0}$ mixing is carried out using the amplitude method described in Ref. 22. Instead of fitting for Δm_s directly, the analysis is performed at fixed values of Δm_s and a likelihood fit to the amplitude A of the oscillation is performed, i.e. in the expression for the mixed and unmixed probabilities, one replaces $[1 \pm \cos(\Delta m_s t)]$ with $[1 \pm A\cos(\Delta m_s t)]$. This method is similar to Fourier transform analysis and has the advantage of facilitating the combination of results from different analysis techniques and different experiments.

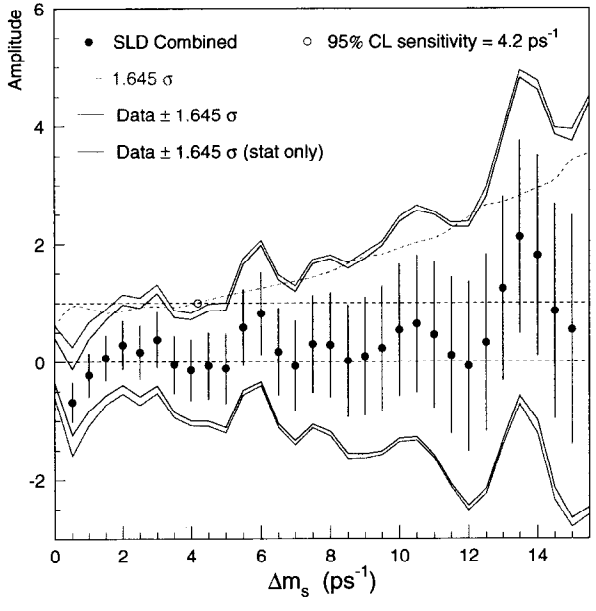


Fig. 16. Measured amplitude as a function of Δm_s the combined analyses.

The lepton+D and Charge Dipole analyses are combined taking into account correlated systematic errors. Figure 16 shows the measured amplitude as a function of Δm_s for the combination. The measured values are consistent with $A = 0$ for the whole range of Δm_s up to 15 ps^{-1} and no evidence is found for a preferred value of the mixing frequency. The following ranges of B_s^0–$\overline{B_s^0}$ oscillation frequencies are excluded at 95% C.L.: $\Delta m_s < 1.7$ ps^{-1} and $3.3 < \Delta m_s < 5.0$ ps^{-1}, i.e., the condition $A + 1.645\,\sigma_A < 1$ is satisfied for those values. The combined sensitivity to set a 95% C.L. lower limit is found to be at a Δm_s value of 4.2 ps^{-1}.

It is worth noting that the overall sensitivity is expected to improve rapidly as the rest of the 1998 data and more analysis techniques are added.

4 Summary and Conclusions

The unique features of SLD and SLC have helped produce world best or competitive results in B physics and QCD. Most of the QCD results are yet to be updated using the 1996-98 data with the upgraded vertex detector. It is the flavor tagging abilities at SLD that have resulted in competitive studies of the flavor independence of α_s, the structure of $b\bar{b}g$ events and a measurement of the B hadron energy. By utilizing particle identification and the e^- beam polarisation new tests of fragmentation models have been performed.

The precise vertexing has been most beneficial in the studies of B physics. The worlds best measurement of the lifetime ratio of the B^+ and B^0 has been made and unique analysis of charmless and double-charm B decays has been possible. A competitive limit for B_s oscillations already exists and will improve with updated tagging techniques. In particular, with more data from the proposed SLD run extension a strong constraint on the CKM unitarity triangle will be made, complementary to the results from the B-factories.

References

[1] K. Abe *et al.*, *Design and Performance of the SLD Vertex Detector, a 307 Mpixel Tracking System*, Nucl. Inst. Methods A **400**, 287 (1997).

[2] K. Abe *et al.*, Phys. Rev. D **53**, 1023 (1996).

[3] D. J. Jackson, *A Topological Vertex Reconstruction Algorithm for Hadronic Jets*, Nucl. Inst. Methods A **388**, 247 (1997).

[4] K. Abe *et al.*, *A Measurement of R_b using a Vertex Mass Tag.*, Phys. Rev. Lett. **80**, 660 (1998).

[5] K. Abe *et al.*, *Production of π^+, K^+, K^0, K^{*0}, ϕ, p and Λ^0 in Hadronic Z^0 decays*, Phys. Rev. D **59**, 052001 (1999).

K. Abe *et al.*, *A Study of Correlations between Identified Charged Hadrons in Hadronic Z^0 decays*, SLAC-PUB-7824; June 1998.

[6] K. Abe *et al.*, *An Improved Test of the Flavor Independence of Strong Interactions*, Phys. Rev. D **59**, 012002 (1999).

[7] S. Bethke *et al.*, Nucl. Phys. B **370** (1992) 310;
erratum: Nucl. Phys. B **523**, 681 (1998).

[8] Z. Kunszt *et al.*, CERN 89–08 Vol. I, 373 (1989).

[9] K. Abe *et al.*, *A Preliminary Study of the Structure of $b\bar{b}g$ Events using Z^0 decays*, SLAC-PUB-7822; June 1998.

[10] T. Rizzo, Phys. Rev. D **50** (1994) 4478.

[11] K. Abe *et al.*, *A Preliminary Improved Measurement of the B Hadron Energy Distribution in Z^0 decays*, SLAC-PUB-7826; June 1998.

[12] See, for example, J.D. Bjorken, Phys. Rev. D **17**, 171 (1978).

[13] M.G. Bowler, Z. Phys. C **11**, 169 (1981); C. Peterson, D. Schlatter, I. Schmitt and P.M. Zerwas, Phys. Rev.D **27**, 105 (1983); B. Andersson, G. Gustafson, G. Ingelman, T. Sjöstrand, Phys. Rep. **97**, 32 (1983).

[14] K. Abe *et al.*, *Measurement of the B^+ and B^0 Lifetimes using Topological Vertexing at SLD*, SLAC-PUB-7868; July 1998.

[15] See, for example, I. I. Bigi *et al.*, in *B Decays*, edited by S. Stone (World Scientific, New York, 1994), p. 132.

[16] K. Abe *et al.*, *Inclusive Search for $b \to sg$*, SLAC-PUB-7896; July 1998.

[17] A. Lenz, U. Nierste, and G. Ostermaier, Phys. Rev. D **59**, 034008 (1999); A. Lenz, U. Nierste, and G. Ostermaier, Phys. Rev. D **56**, 7228 (1997).

[18] A.L. Kagan, Phys. Rev. D **51**, 6196 (1995); M. Ciuchini, E. Gabrielli and G. F. Giudice, Phys. Lett. B **388**, 353 (1996); B.G. Grzadkowski and W.-S. Hou, Phys. Lett. B **272**, 383 (1991).

[19] A. Kagan and J. Rathsman, `hep-ph/9701300`.

[20] K. Abe *et al.*, "Time Dependent B_s^0–$\overline{B_s^0}$ Mixing Using Inclusive and Semileptonic B Decays at SLD," SLAC-PUB-7885; July 1998.

[21] P. Paganini, F. Parodi, P. Roudeau, and A. Stocchi, Phys. Scripta **58**, 556 (1998) and `hep-ph/9802289`.

[22] H.-G. Moser and A. Roussarie, Nucl. Inst. Methods A **384**, 491 (1997).

A SELECTION OF RECENT CLEO RESULTS: B PHYSICS, SILICON DETECTOR, AND MORE...

Roy A. Briere*
Harvard University
HEP Lab, 42 Oxford St.
Cambridge, MA 02138

Representing the CLEO Collaboration

ABSTRACT

Some recent results from the CLEO experiment are reviewed. The first portion highlights our efforts studying the CKM matrix with measurements of semi-leptonic B decays and our comprehensive survey of rare B decay modes. Next, some of the first results from our new silicon vertex detector are presented, along with a glimpse of related ongoing analyses. Finally, a few examples of the excellent physics CLEO does outside the B system are mentioned.

*Supported by DOE Grant DE-FG02-91ER40654.

© 1998 by Roy A. Briere.

1 Introduction

The CLEO detector operates at the Cornell Electron Storage Ring. This 768 meter ring is filled with electrons and positrons, each comprising nine trains of three bunches with a current of more than 200 mA/beam. The trains are 300 ns apart, while bunches are 28 ns apart within each train. Out peak luminosity now exceeds $6 \times 10^{32}\mathrm{cm}^{-2}\mathrm{s}^{-1}$; this surpasses our most recent goal and is the highest at any collider. The most luminosity actually logged to tape by CLEO in one day (month) is 30 (480) pb^{-1}.

After detector and accelerator upgrades are completed in 1999, CESR will run with nine trains of five bunches at a 14 ns bunch spacing. The goal is peak luminosities of $1.7 \times 10^{33}\mathrm{cm}^{-2}\mathrm{s}^{-1}$, yielding 15 fb^{-1} per year with a realistic operating schedule.

CESR has run on all of the Υ resonances in the past. The $\Upsilon(4S)$ is special, being just above $B\bar{B}$ threshold. Standard running now includes only two energies: on the $\Upsilon(4S)$ resonance and in the continuum region just below it.

The off-resonance, or "continuum" energies are typically 60 MeV below the $\Upsilon(4S)$ peak, at $E_{cm} = 10.52$ GeV. The hadronic cross-section is about 3 nb. Fermion-anti-fermion pairs are produced in the decay of one 10.6 GeV off-shell photon at rest: $e^+e^- \to \gamma^\star \to \tau^+\tau^-, c\bar{c}$, etc. Processes such as two-photon collisions, $\gamma\gamma \to X$, also occur.

The on-resonance, or "On4S," running takes place at $E_{cm} = 10.58$ GeV. The effective cross-section for $\Upsilon(4S)$ production is almost 1.1 nb. The reaction $e^+e^- \to \Upsilon(4S) \to B\bar{B}$ results in the decay of two 5.28 GeV B mesons almost at rest. This process is also accompanied by all the continuum processes listed above. The $B\bar{B}$ portion is partly separable via event shape. The jetty hadronization of quark jets is usually quite distinct from the round symmetric decay of two slow B mesons. One can also use the off-resonance data to study backgrounds to the $B\bar{B}$ sample.

Two-thirds of our luminosity is taken On4S. With this ratio, 1.4 fb^{-1} total (On4S + continuum) data yields 10^6 $B\bar{B}$ pairs or equivalently 10^6 B^+ and 10^6 $\bar{B}^0$. The *total* luminosity of the data used for each of our results will be quoted.

Please note that all results not yet published are preliminary and hence subject to minor changes.

2 The CLEOII Detector

The CLEOII.V detector is shown in Fig. 1. The basic detector characteristics are summarized below; more information may be found in the references.[1,2]

CLEO has tracking over more than 90% of 4π in a solenoidal field, B = 1.5 Tesla. Tracks curl inside the chamber for $p_T < 250$ MeV/c; the resolution is $\sigma_p/p \sim 0.6\%$ at 2 GeV/c. Specific ionization information (dE/dx) from 49 of the wire chamber layers gives K/π separation of about 2σ for 2.6 GeV/c (the momentum of stiff tracks from interesting rare two-body B decays).

Particle identification (PID) is supplemented by a Time-of-Flight (TOF) system. The resolution of the barrel system is about 150 ps; TOF separation of πs and Ks deteriorates for momenta greater than about 1 GeV/c.

Our calorimeter consists of 7800 CsI(Tl) crystals located in a barrel inside the super-conducting magnet coil and in two endcaps and covers 98% of 4π. The best barrel section, with the least material in front of it, comprises 70% of the coverage. Here, the typical π^0 mass resolution is 6 MeV.

The Muon ID system consists of iron absorbers interspersed with streamer counters covering 85% of the solid angle. The efficiency turn-on for a five interaction length penetration cut is at 1.4 GeV/c, with 1% fakes.

The pre-SVX CLEOII detector, featuring the new CsI calorimeter, operated from 1989-1995, collecting 5 fb^{-1} of data. Many discoveries were made with this data. During this era, CESR saw many innovations in its running, including the transition to one interaction region, pretzel orbits, a crossing angle, and multi-bunch operations. All of this data is now re-analyzed; tracking in particular is improved.

The start of the CLEOII.V era in 1995 is marked by the installation of our first silicon vertex detector (SVX), also the first to be operated at the $\Upsilon(4S)$. The upgrade included a new beam pipe and interaction region. The main drift chamber gas was changed from 50-50% Ar-C_2H_6 to 60–40% He-C_3H_8; this reduces multiple scattering and the Lorentz angle which improved resolution and efficiency.[3] Running over the period 1995-1999 will produce over 8 fb^{-1} of physics data; more than 7 fb^{-1} is already on tape as of October 1998.

The CLEOIII upgrade, scheduled for installation in 1999, includes a Ring-Imaging Cherenkov (RICH) detector for greatly improved particle identification. In addition, a new drift chamber, silicon vertex detector, beam pipe and interaction region will be installed. The concurrent CESR machine upgrade includes super-conducting RF cavities,

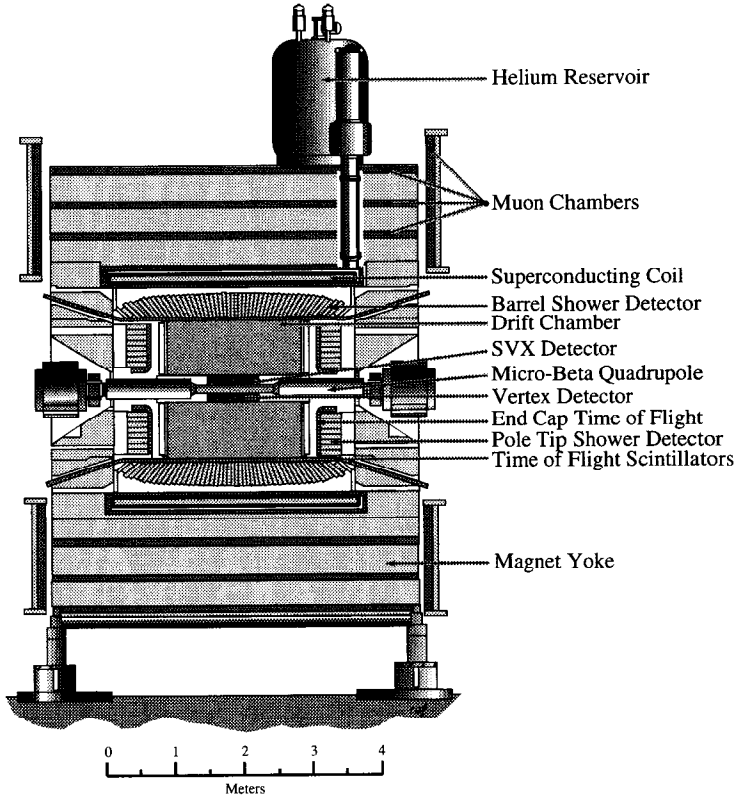


Fig. 1. The CLEOII.V Detector.

and new super-conducting focusing quadrupoles. CLEO expects to collect 15 fb^{-1} of data per year.

3 Kinematics at the $\Upsilon(4S)$

CESR has symmetric beam energies such that the energy of B mesons is given by $E_B = E_{beam}$, where E_{beam} is well-known from accelerator parameters. The average B meson momentum is quite small, $|P_B| \simeq 325$ MeV/c.

When all daughters of the decay are measured, there are two very useful variables. The first is (sums are over daughter particles):

$$\Delta E \equiv \sum E_i - E_{beam}. \tag{1}$$

This simply expresses energy conservation, peaking at zero for real events. It is sensitive to missing particles and also to K-π mis-identification which both affect the energy. A typical resolution is about 25 MeV for $D\pi$; however, it is much improved, to 8 MeV, for final states like $D^{*+}D^{*-}$ and $D\bar{D}\bar{K}$ with large rest-mass. The second key variable is the "beam-constrained" mass:

$$M_B \equiv \sqrt{E_{beam}^2 - |\sum \vec{P}_i|^2}. \tag{2}$$

This expresses momentum conservation. Using E_{beam} instead of $\sum E_i$ improves the mass resolution by a factor of 10, to 2.5 MeV, for two-body modes like $D\pi$.

Often these are the final two variables examined for a peaked excess indicative of a signal. This technique is used extensively for hadronic decays where the decay products are generally all accessible as charged tracks and photons. It is also used for some semi-leptonic modes; this is possible when the neutrino is inferred from global four-momentum balance. More details will be given below when discussing $|V_{ub}|$.

Another variant on the full reconstruction may be termed semi-inclusive reconstruction. Part of the final state is common to all channels, but a varying number of other particles are added in order to find the best B candidate. This is used, for example, in the $b \to s\gamma$ analysis discussed later, where one requires a stiff photon, a kaon, and a varying number of added pions.

There are many other techniques commonly used at CLEO, for example, partial reconstruction of a D^{*+} from the slow transition π^+ only, but there is not time to discuss these further.

4 Magnitudes of CKM Matrix Elements at CLEO

Recall the basic pattern of CKM mixing:

$$\begin{pmatrix} |V_{ud}| & |V_{us}| & |V_{ub}| \\ |V_{cd}| & |V_{cs}| & |V_{cb}| \\ |V_{td}| & |V_{ts}| & |V_{tb}| \end{pmatrix} \sim \begin{pmatrix} 1 & \lambda & \lambda^3 \\ \lambda & 1 & \lambda^2 \\ \lambda^3 & \lambda^2 & 1 \end{pmatrix} \quad (3)$$

where only the approximate magnitudes by order in the parameter $\lambda = \sin\theta_C \simeq 0.22$ are shown.

CLEO does best at measuring $|V_{ub}|$ from $B \to \pi\ell\nu$, with theory to normalize the form factor, and $|V_{cb}|$ from $B \to D^*\ell\nu$ with HQET to constrain the form factor. CLEO can also, in principle, very cleanly get $|V_{ub}|$ from $B \to \ell\nu$, but the very small rate will make the mere observation of this process difficult.

CLEO has also measured $B\bar{B}$ mixing, which scales with $|V_{td}|^2$, while the $b \to s\gamma$ process can yield some information on $|V_{ts}|$. The ratio $|V_{td}/V_{ts}|$ will eventually be accessible from the comparison of $B \to \rho\gamma$ and $B \to K^*\gamma$, but the former process remains undetected and only a limit on $|V_{td}/V_{ts}|$ is possible. Competing limits are available elsewhere from comparisons of B_s and B_d mixing.

We will return to the issue of phases of the CKM matrix elements, such as V_{ub} and V_{td}, later. First, some recent results involving the magnitudes are surveyed.

4.1 Introduction to Semi-Leptonic B Decays

Given the $V-A$ structure of the weak interaction and the fact that leptons do not interact strongly with the hadronic decay products, semi-leptonic decays, such as $B \to D^{(*)}\ell\nu$, are almost as simple as two-body decays. The only complication is that q^2, the $\ell - \nu$ invariant mass, is not fixed. The non-trivial dependences of the matrix element on q^2 are contained in functions known as form factors.

One often replaces q^2 with the HQET-inspired variable

$$w = (M_B^2 + m_M^2 - q^2)/2M_B m_M \quad (4)$$

for $B \to M\ell\nu$ decay, which is just the Lorentz γ of the meson M in the B rest-frame. Clearly, $w = 1$ corresponds to q^2_{max} where M is at rest ('zero-recoil'), and the ℓ and ν are back to back while w_{max} corresponds to $q^2 = 0$ where M recoils against the $\ell\nu$ system.

The form-factors are often parameterized as:

$$\mathcal{F}(w) = \mathcal{F}(1) \times (1 - \hat{\rho}^2(w-1) + c(w-1)^2 + ...). \quad (5)$$

There is one relevant form factor for $P\ell\nu$ and three for $V\ell\nu$, where P and V represent pseudo-scalar and vector mesons, respectively, such as the D and D^*. One can use HQET to inter-relate the form factors if desired.

One important feature of semi-leptonic decays is Luke's Theorem: there are no $1/m_b$ corrections in HQET at q^2_{max} for the $D^*\ell\nu$ process. Note that this is *not* true for other q^2 or for $D\ell\nu$. Of course, nature is not overly kind, and experiments suffer from lower rates and efficiencies and higher backgrounds at q^2_{max}.

Lattice QCD calculations promise to offer increasing precision on the form factor for $B \to D\ell\nu$, these studies faring better for the ground-state D rather than D^*. Once again Nature confounds us somewhat since experiments suffer higher background, notably from D^* and D^{**} feed-down, a lower branching ratio, and proportionately less rate near zero recoil for $D\ell\nu$ relative to $D^*\ell\nu$.

Measurements of V_{ub} are even more difficult for experimenters due to low rates and high backgrounds, and also for theorists. On the lattice, the low energy scale is troublesome, while off the lattice, the niceties of HQET are now absent. Nonetheless, reliable results have been extracted; these more recent achievements are presented first.

4.2 Measurements of $|V_{ub}|$

The $b \to u$ transition was first observed at CLEO by studying the lepton momentum spectrum near the kinematic endpoint. Leptons from these transitions can have momenta beyond the $b \to c$ endpoint. Since only a tiny fraction of the spectrum is examined, model-dependence is the dominant error in the most recent such result:[4]

$$|V_{ub}/V_{cb}| \simeq 0.06 - 0.10. \quad (6)$$

A newer neutrino-reconstruction technique enabled CLEO to observe the exclusive semi-leptonic $b \to u$ final states $\pi\ell\nu$, $\rho\ell\nu$, and $\omega\ell\nu$. The neutrino four-momentum is inferred from the known beam energy and careful measurement of all tracks and unmatched showers in the rest of the event (both Bs). Exactly one lepton is required, so that only one neutrino is missing, and good charge-balance among the tracks is enforced. Care is taken to avoid mistaking hadronic shower fragments for photons to prevent double-counting. After much hard work, the resolution on the inferred neutrino

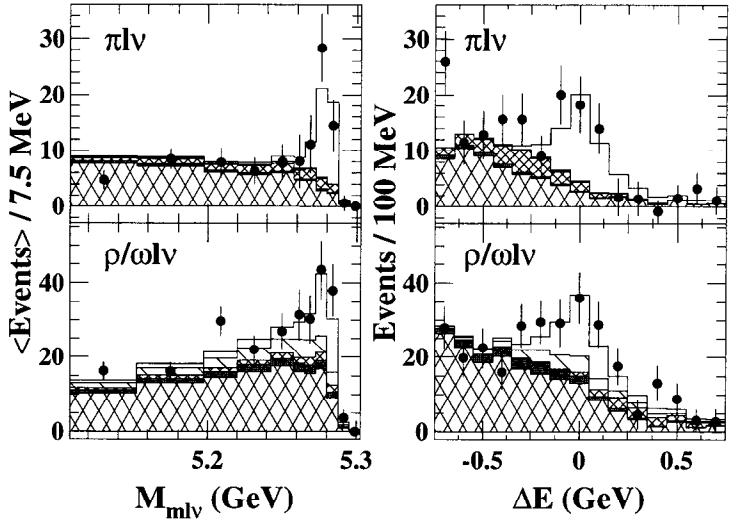


Fig. 2. The beam-constrained mass and energy-balance distributions for the $\pi\ell\nu$ and $\rho/\omega\ell\nu$ modes. The points with errors represent the data after continuum and fake subtraction, and the histogram is the total fit to the data. The open area under the histogram represents the signal component, and the various shaded areas are backgrounds, including cross-feed among modes and feed-down from other $b \to u\ell\nu$ processes. The largest background, from $b \to cX$, is cross-hatched.

is excellent, $\sigma_{p_\nu} \sim 110$ MeV/c. With the neutrino in hand, full reconstruction of the B meson is possible. The resulting ΔE and M_B distributions are displayed in Fig. 2.

The results from 4 fb^{-1} of CLEOII data are:[5]

$$\mathcal{B}(B \to \pi^-\ell^+\nu) = (1.8 \pm 0.4(stat) \pm 0.3(syst) \pm 0.2(model)) \times 10^{-4}, \quad (7)$$

$$\mathcal{B}(B \to \rho^-\ell^+\nu) = (2.5 \pm 0.4(stat)^{+0.5}_{-0.7}(syst) \pm 0.5(model)) \times 10^{-4}, \quad (8)$$

$$|V_{ub}| = (3.3 \pm 0.2^{+0.3}_{-0.4} \pm 0.7) \times 10^{\ 3}, \quad (9)$$

where isospin constraints are employed between the two $\pi\ell\nu$ modes and the three $\rho\ell\nu, \omega\ell\nu$ modes, leading to the extraction of two average branching ratios.

The limiting factors when extracting $|V_{ub}|$ are a 12% error from the branching ratios, mostly from neutrino reconstruction efficiency, and a 20% model dependence. More statistics will help further constrain the models.

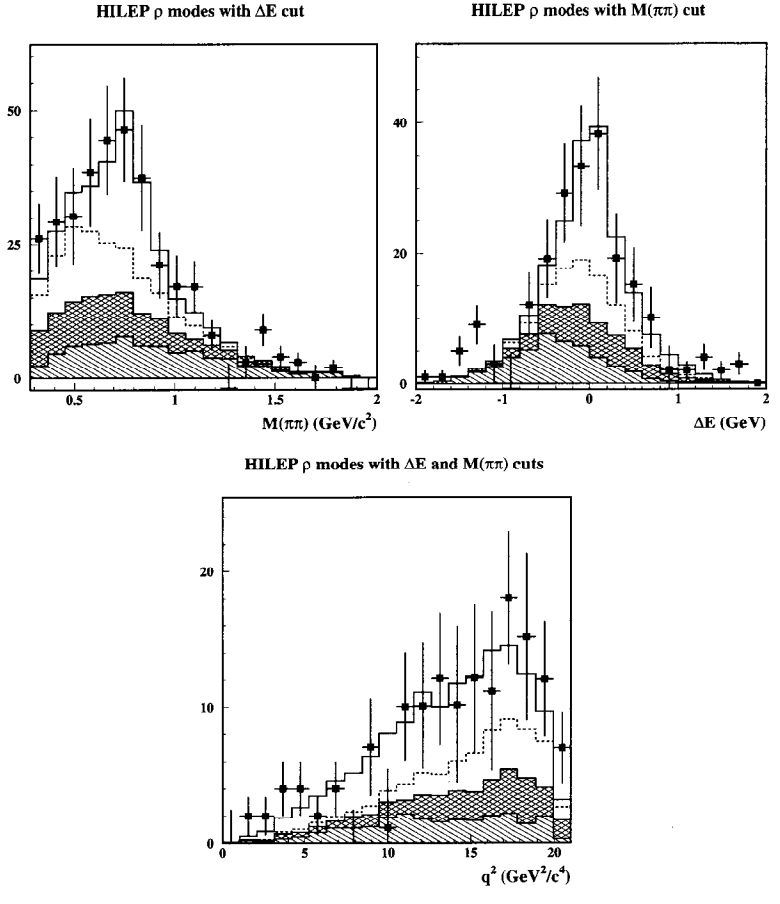


Fig. 3. Preliminary results from the fit to continuum-subtracted data $B \to \rho\ell\nu, \omega\ell\nu$ data in the 2.3–2.7 GeV/c lepton momentum bin. Points are data; the fit is a sum of $b \to c$ (single-hatch), other $b \to u$ (double-hatch), cross-feed (dashed histogram), and signal (solid histogram).

A new $b \to u\ell\nu$ analysis on the full 5 fb^{-1} of CLEOII data is now available. The decays $B \to \rho\ell\nu, \omega\ell\nu$ were measured in an analysis using a simpler, "loose-cut" neutrino reconstruction. Leptons are required to have $p > 1.7$ GeV/c to control background, and the decay kinematics are required to be consistent with the neutrino hypothesis. The most sensitive portion of the momentum spectrum is the so-called "HILEP" bin: leptons from 2.3–2.7 GeV/c. Background comes from other $b \to u$, $b \to c$, and continuum.

The analysis uses a likelihood fit to the 2π or 3π mass (for the ρ or ω channel), ΔE, background levels of other $b \to u$, $b \to c$, continuum, fake ℓ, and finally branching ratios for $\pi^+/\pi^0/\rho^+/\rho^0/\omega\ell\nu$. The $\pi\pi$ mass, ΔE and q^2 distributions are shown in Fig. 3. The preliminary results are:[6]

$$\begin{aligned} \mathcal{B}(B^0 \to \rho^-\ell^+\nu) &= (2.8 \pm 0.4 \pm 0.4 \pm 0.6) \times 10^{-4}, \\ |V_{ub}| &= (3.2 \pm 0.3^{+0.2}_{-0.3} \pm 0.6) \times 10^{-3}, \\ \rho^2 &= 0.52 \pm 0.11 \pm 0.09 \pm 0.05. \end{aligned}$$

These are consistent with the previous neutrino reconstruction results; they also benefit from improved models of $b \to u\ell\nu$ decays spurred by that earlier result.

4.3 Measurements of $|V_{cb}|$

One major goal of studying $b \to c\ell\nu$ processes is the measurement of $|V_{cb}|$. It is also desirous to gain a complete understanding of the structure of these decays; they are the simplest of B decays and also those with the largest exclusive branching ratios.

The first method for determining $|V_{cb}|$ begins with the inclusive branching ratio $\mathcal{B}_{SL} \equiv \mathcal{B}(B \to X_c\ell\nu)$. Combined with τ_B, this yields a partial width which can be compared to a theoretical expression. Naively, the semi-leptonic width, Γ_{SL}, is simply determined by the decay of a free b quark, akin to muon decay: $\Gamma_{SL} \sim |V_{cb}|^2 m_b^5$.

There are some subtleties to contend with, however. First, what does one insert into the formula for m_b? This involves issues of quarks vs hadrons and "duality." Furthermore, the rate superficially behaves as m_b^5 and is thus quite sensitive to the value used. Much theoretical work has been done attempting to clarify these issues. Another potential concern is that the LEP and CLEO results on the inclusive semi-leptonic rate historically have hinted at disagreement. The largest background to the inclusive lepton spectrum comes from secondary $D \to X\ell\nu$ decays. By looking at both like- and unlike-sign lepton pairs in $B\bar{B}$ decays, one can separate the primary leptons from B decay and those leptons from D decay (after including a straightforward correction for B mixing). Such an analysis was performed on 3 fb^{-1} of CLEOII data. This dilepton analysis uses one stiff lepton, ℓ, as a tag to isolate $B\bar{B}$ events. Then, by looking at additional inclusive electrons and counting like- and unlike-sign ℓe pairs, one extracts the lepton spectra for both $b \to Xe\nu$ and $c \to Xe\nu$; the final result is displayed in Fig. 4. The numerical results are:[7]

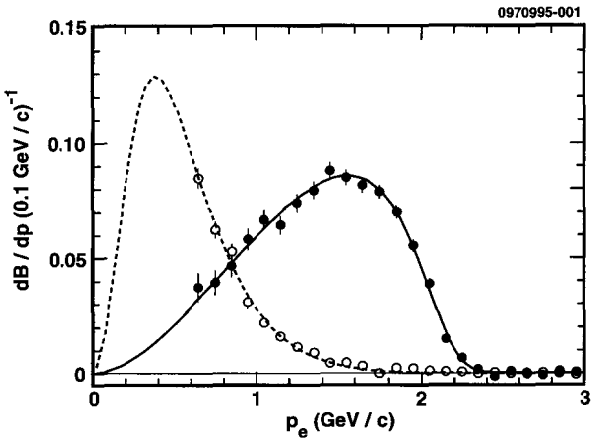


Fig. 4. Lepton momentum spectra derived from the dilepton analysis. Solid circles denote $B \to Xe\nu$; open circles, $D \to Xe\nu$.

$$\mathcal{B}(B \to Xe\nu) = 10.49 \pm 0.17(stat) \pm 0.43(syst)\%, \quad (10)$$

$$|V_{cb}| = 0.040 \pm 0.001(exp) \pm 0.004(theory). \quad (11)$$

The experimental error on this branching ratio is already quite small, so the limitation is currently theoretical. Some may feel that theoretical error on $|V_{cb}|$ is already better than 10%. CLEO is working to update this analysis with more data, and improved reconstruction and Monte Carlo. Some possibilities for aiding theorists with new measurements are also discussed below. Another technique for extracting $|V_{cb}|$ is found in the marriage of exclusive modes and HQET and involves measuring $d\Gamma/dq^2$ for $D^*\ell\nu$. Then, extrapolating in q^2 to zero recoil yields $|V_{cb}|\mathcal{F}(1)$, where $\mathcal{F}(1)$ is the form-factor at zero recoil ($w = 1$). Invoking HQET allows fairly accurate evaluations of $\mathcal{F}(1)$. Sensitivity to the shape of the form factor complicates this extrapolation. There are also

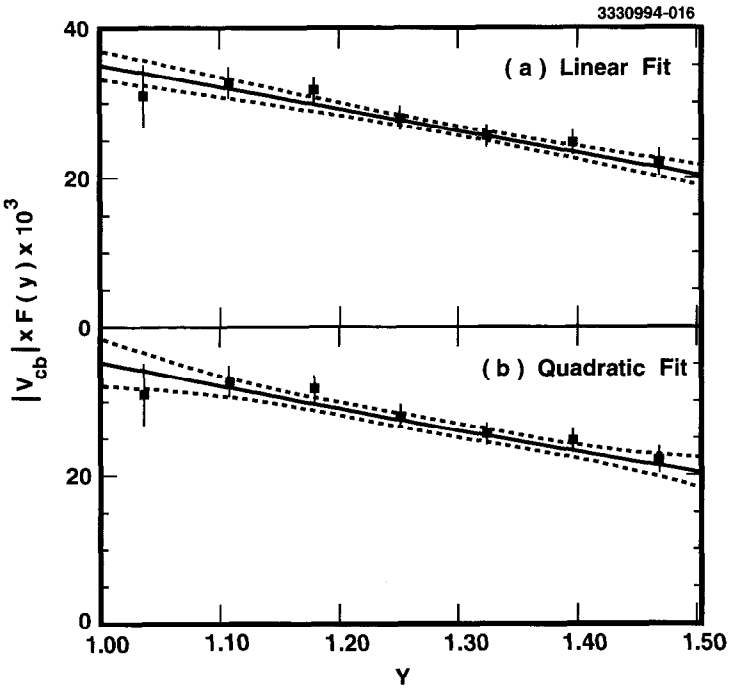


Fig. 5. Fits to the form-factor data for $D^*\ell\nu$ decays versus Y, which is essentially the HQET parameter w (see Ref. 8 for details). Both linear and quadratic fits are shown; in each, the solid line shows the central fit enclosed in the dashed error bands corresponding to one-sigma variations of the fit parameters. Error bars are statistical only.

corrections to the $m_Q = \infty$ limit of HQET; the mentioned earlier.

An analysis was performed on both the D^* CLEOII data. The use of both charge states a of charged and neutral B production fractions, and the extrapolation to zero-recoil, shown in

$$|V_{cb}|\mathcal{F}(1) = 0.0351 \pm 0.0019(stat)$$

A typical theoretical value for the form-factor

Finally, let us return to the method of det leptonic decay rate. The main problem invol level picture and the reality of hadronic bound s with help from the operator-product expansio new parameters. The OPE parameters are $\bar{\Lambda}$, (or 'brown muck'), λ_1, the kinetic energy of hyperfine chromo-magnetic interaction. The p perimental value of $m_{B^*} - m_B$. Further experi parameters $\bar{\Lambda}$ and λ_1, aiding theorists with inc quarks bound in B mesons.

In particular, CLEO is working to constrai ments of semi-leptonic decay variables. As a moment': it is just a sum over all events with ments, such as a mean or RMS, can constrain the lepton energy moments:

$$< E_\ell >, < (E_\ell -$$

are used, where E_ℓ is the lepton energy and $<$ These, along with corresponding hadronic ma which may be used to constrain theory.

Preliminary analyses are in progress to de moments.[9] They should provide accurate en them from being a serious source of error wh that even the $D^*\ell\nu$ HQET technique requires form factor at zero recoil and will also benefit.

The analysis obtaining the hadronic moments uses neutrino reconstruction on 5 fb^{-1} of CLEOII data. One can get the hadronic mass from just the measured ν and ℓ by taking the B to be at rest, and folding in the effects of the small motion as a correction. This avoids having to partition tracks and showers between the two overlapping B decays which would be necessary, and very difficult, to explicitly reconstruct the recoiling hadronic system. The leptonic moment analysis uses 3 fb^{-1}, employing the same dilepton technique as the inclusive branching ratio measurement described above. The lepton spectrum is cleanly determined down to 600 MeV/c; corrections for the small unseen tail are straightforward.

Given theoretical calculations for the moments in terms of $\bar{\Lambda}$ and λ_1, each experimental moment result constrains these parameters to lie in a band. The resulting constraints are displayed in Fig. 6. There is an apparent disagreement between the hadronic and leptonic constraints. It is not clear at this time how the blame is to be divided among statistics, measurement, and theoretical calculations. Since this method in principle promises very accurate results for $|V_{cb}|$, much effort is being expended to clarify the situation.

4.4 Outlook for $|V_{ub}|$ and $|V_{cb}|$

The $b \to u$ neutrino reconstruction analysis will focus on the q^2 dependence next; three times more data than used for the published results will soon be available. There remains one crucial need from theory: the absolute normalization of the form-factor, at *any* q^2 value, in order to tie the exclusive measurement to $|V_{ub}|$. Novel techniques may allow even lattice calculations to help out, despite the low-mass mesons.

An improved determination of $|V_{cb}|$ from the full reconstruction of $D^*\ell\nu$ events is also in progress. More data is available to study the form-factor curvature. The $D^{*+} - D^0$ mass difference resolution is now improved, reducing background in both old and new data. There is also a large effort to better understand the slow pion efficiency in order to reduce that source of systematic error. CLEO will also update the inclusive branching ratio while trying to resolve the issues involved in using the moment measurements to constrain the OPE parameters.

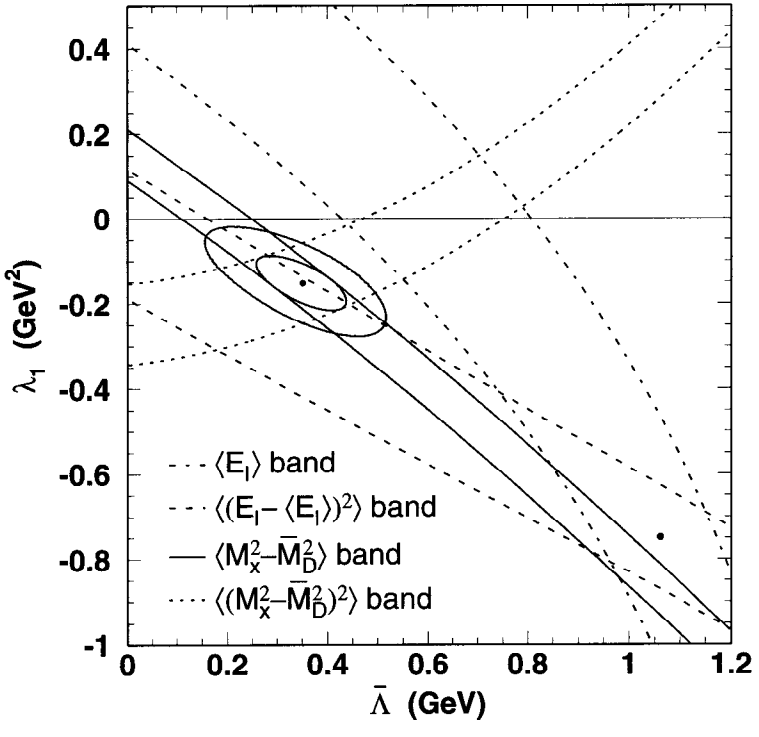


Fig. 6. Preliminary results on constraints from the four measured hadronic and leptonic moments are shown in the $\lambda_1 - \bar{\Lambda}$ plane. Two dots show the central solutions for the hadronic and leptonic moments separately. For the hadronic bands, $1, 2\sigma$ error ellipses are shown surrounding the central intersection point.

5 Rare Decays and CP-Violation

We now turn from the magnitudes of the CKM matrix elements to their phases. So far, CP violation has only been observed in neutral kaon decay. The one free phase parameter in the standard CKM matrix may be the sole source of this effect, but confirmation of model predictions in B physics is needed to provide the required independent test. The key modes include rare charmless decays as well as Cabibbo-suppressed decays to final states with charm. Some of the Feynman diagrams responsible for the rare charmless B decays are shown in Fig. 7. CP-violation can be observed in various ways. One

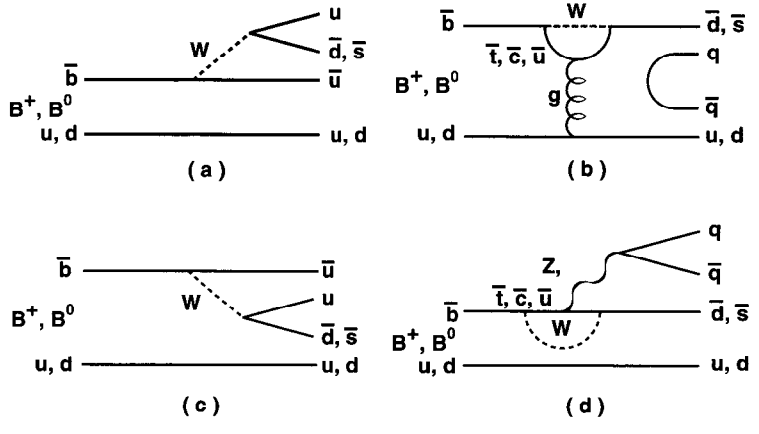


Fig. 7. Important Feynman diagrams responsible for rare charmless B decays: (a) external (color-allowed) spectator, (b) gluonic penguin, (c) internal (color-suppressed) spectator, and (d) electroweak penguin.

popular theme involves looking at time-dependent asymmetries in the decay rate to a CP-eigenstate, f_{CP}. Here, one utilizes mixing to interfere the decay paths $B \to f_{CP}$ and $B \to \bar{B} \to f_{CP}$ measuring, in the Standard Model, the phase of V_{td} present in the mixing process. CP-violation may also show itself via rate asymmetries between B and $\bar{B}$ decays to a final state f and its conjugate $\bar{f}$, respectively. These rate asymmetries require two interfering diagrams with a relative weak phase as well as a strong phase from final state interactions (FSI). The weak phase may come, for example, from the phase of V_{ub} in the spectator diagram relative to the gluonic penguin.

Many papers have analyzed the various decays and suggested ways of isolating the interesting information. CLEO is now cataloging many of the interesting final states and is usually observing them for the first time. Of key importance is the relative size of certain classes of diagrams, i.e., penguin and $b \to u$ spectator processes.

As an example of one general feature, note that penguin processes easily produce kaons, since CKM factors favor $b \to s$ over $b \to d$, while $b \to u$ spectator diagrams favor pions, due to Cabibbo-suppression of $W^- \to \bar{u}s$ relative to $W^- \to \bar{u}d$. As summarized below, CLEO observes the K modes rather than the π modes, hinting that penguins are large. In fact, $B^0 \to \pi^+\pi^-$ is not yet observed. This will make planned CP studies using this channel perhaps more difficult than originally suspected.

5.1 Rare Cabibbo-Suppressed $b \to c$ Decays

Let's begin with a Cabibbo-suppressed decay which is related to the well-established Cabibbo-allowed decay $B^0 \to D_s^+ D^{*-}$. The decay $B \to D^{*+}D^{*-}$ is of interest for time-dependent CP-violation studies, just like those planned for $\psi K^{(*)}$. One will have to determine the CP-content of the final state, as CLEO did for ψK^*,[10] in order to make real use of this mode.

The analysis is a standard full reconstruction of the entire B decay, using 8.5 fb^{-1} of CLEOII data. One noteworthy cut examines the significance of the vertex separation between the two Ds; this works particularly well with the subset of data (about 40%) taken with the SVX detector in place.

The final ΔE-M_B plot shows a clear excess in the signal region; see Fig. 8. The background is estimated in several ways, almost exclusively based on clever uses of the data, to be 0.31 ± 0.10. The probability that the four candidate signal events arise from a background fluctuation is estimated as 1.1×10^{-4}. This first observation yields the preliminary branching ratio:[11]

$$\mathcal{B}(B^0 \to D^{*+}D^{*-}) = (6.2^{+4.0}_{-2.9} \pm 1.0) \times 10^{-4} \qquad (14)$$

consistent with the Cabibbo-suppression, and other simple factors, relative to $D_s^* D^{*-}$.

CLEO also has another recent observation of a Cabibbo-suppressed decay of interest for CP-violation studies, $B^- \to D^0 K^-$, with 5 fb^{-1} of CLEOII data:[12]

$$\mathcal{B}(B^- \to D^0 K^-) = (2.57 \pm 0.65 \pm 0.32) \times 10^{-4}. \qquad (15)$$

Current analyses are working on other Cabibbo-suppressed modes such as $D^{*+}D^-$, D^+D^-, and D^0K^{*-}.

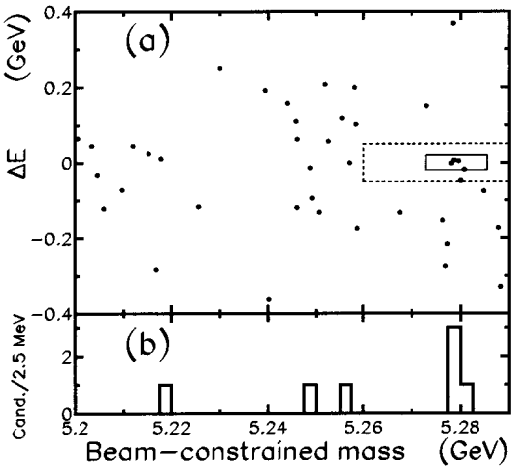


Fig. 8. The ΔE vs M_B distribution for the $B \to D^{*+}D^{*-}$ candidates. The solid box is the signal region; the sideband outside the dashed box is used to help evaluate background levels. Shown below is the mass projection with the ΔE cut used for signal.

5.2 Angular Analysis of $B \to D^*\rho$ and Search for FSI

Final state interactions are required in order for inclusive decay-rate asymmetries to be non-zero. It is sometimes argued that in decays such as $B \to K\pi$, the decay products separate so fast that such phases must be small, but the truth is far from clear. Given this, it is interesting to look for evidence of FSI phases anywhere in B physics. In B decays to charm, perhaps the slower relative motion of the decay products will admit an observable effect.

A new analysis examines the $D^*\rho$ final state; it has relatively good statistics, and a non-trivial angular distribution. This is the first full fit to this angular distribution; previously, CLEO has done a similar analysis of ψK^* decays.[10]

A maximum likelihood fit to three helicity amplitudes, H_+, H_-, H_0, is performed. Factorization can be checked by comparing $\Gamma_L/\Gamma = |H_0|^2/\sum|H_i|^2$ with Γ_L/Γ for $B \to D^*\ell\nu$ at $q^2 = m_\rho^2$. Finally, one can look for hints of final state interactions in the relative phases of the H_i. The analysis uses 5 fb^{-1} of CLEOII data. The preliminary

$B^0 \to D^{*-}\rho^+$	magnitude
H_0	0.936
H_-	$0.317 \pm 0.052 \pm 03$ (
H_+	$0.152 \pm 0.058 \pm 07$]
$B^+ \to \bar{D}^{*0}\rho^+$	magnitude
H_0	0.932
H_-	$0.283 \pm 0.068 \pm 09$]
H_+	$0.228 \pm 0.069 \pm 06$ (

Table 1. Preliminary results of the fit to thelicit The phase of H_0 is zero by convention and thmpli $\sum|H_i|^2 = 1$.

result for Γ_L/Γ, along with some relevant compsons

$D^{*-}\rho^+$: $\Gamma_L/\Gamma = 0.8 \pm 0.$

CLEO $D^*\ell\nu$: $\Gamma_L/\Gamma = 0.9 \pm 0.$

Theory : $\Gamma_L/\Gamma \simeq 0.8$- 0.8

Table 1 summarizes the preliminary results of tlletai

Factorization appears to be working, withine er the phases, but they are not yet significant bason tl phases are fixed at zero. It is possible to look fonases angular terms with $Im(H)$ explicit in them, bthis current statistics. Work is in progress to add mcdata

5.3 Analysis of Rare Charmless B Eays

Let us now turn to the rare charmless B decays.:fore is useful to briefly discuss some common aspeof the

First, in order to keep efficiencies as high asssibl fits to extract yields. These fits include quantitisuch and helicities, particle ID information, shape vaples

When plotting M_B, for example, to illustrate resu other variables beyond those used to select eveifor t

to see sidebands. Thus, the full statistical power of the data is not evident in these fit projections. The likelihood results are also cross-checked with simple cut-and-count analyses, looking at events above background in a ΔE-M_B signal box after all cuts.

Due to our limited ability to separate the stiff charged Ks and πs produced in two-body charmless B decays, simultaneous fits are used. For example, the rates for $B^+ \to \eta' K^+$ and $B^+ \to \eta' \pi^+$ are extracted from one fit. In such cases, the likelihood contours as a function of the two yields (or branching ratios) are displayed. The K-π separation at 2.6 GeV/c from dE/dx is about 2σ and an additional, independent 1.7σ is obtained from the mass-dependence inherent in calculating ΔE. The notation h is sometimes used to denote an undifferentiated combination of charged πs and Ks.

CLEO recently published three major papers, all based on 5 fb^{-1} of CLEOII data. These publications covered $K\pi$ and related final states,[14] final states with an η or η',[15] and final states containing an ω or ϕ.[16] These results are summarized graphically in Figs. 9 and 10. A few peculiar* modes from the three publications are omitted from these graphics. Some other modes are shown which have not been updated recently; these results are contained in an earlier paper.[17]

Now we summarize the recent, unpublished updates to these results.

5.4 Recent Updates for Rare Charmless B Decays

A preliminary update of modes with η''s was given at ICHEP98.[18] The analyses were performed with 8.5 fb^{-1} of CLEOII data. As discussed earlier, the $\eta' K^+$ and $\eta' \pi^+$ modes are fit simultaneously; the resultant likelihood contours for $\eta' h^+$ are shown in Fig. 11. The data indicates only K^+ in the final state, with no hint of π^+ yet. Figure 12 shows ΔE and M_B projections for these $\eta' h^+$ final states as well as for $\eta' K_S$. These new η' results are summarized in Table 2.

ICHEP98 further witnessed a preliminary update of $B^0 \to K^+\pi^-$ and related modes also using 8.5 fb^{-1} of data.[19] Figure 13 shows the likelihood contours for the $K^+\pi^-$ and $\pi^+\pi^-$ yields, and displays the clear evidence for a signal in the key projection plots. The yields for $K^+\pi^-, \pi^+\pi^-$ are extracted simultaneously, and only $K^+\pi^-$ is significant. This is a sign that penguins are large. Also, the contours in Fig. 13 indicate that it is unlikely for the $K^+\pi^-$ and $\pi^+\pi^-$ modes to have equal branching ratios.

Analogous plots for $h^\pm K_S$ and $h^\pm \pi^0$ modes are shown in Figs. 14 and 15. The new

*These are $\phi\pi$, $\phi\rho$, $\phi\omega$, and $\phi\eta^{(\prime)}$. All have an $s\bar{s}$ in the final state and are not easily accessible via W-emission or penguin diagrams and so are expected to have very small branching ratios.

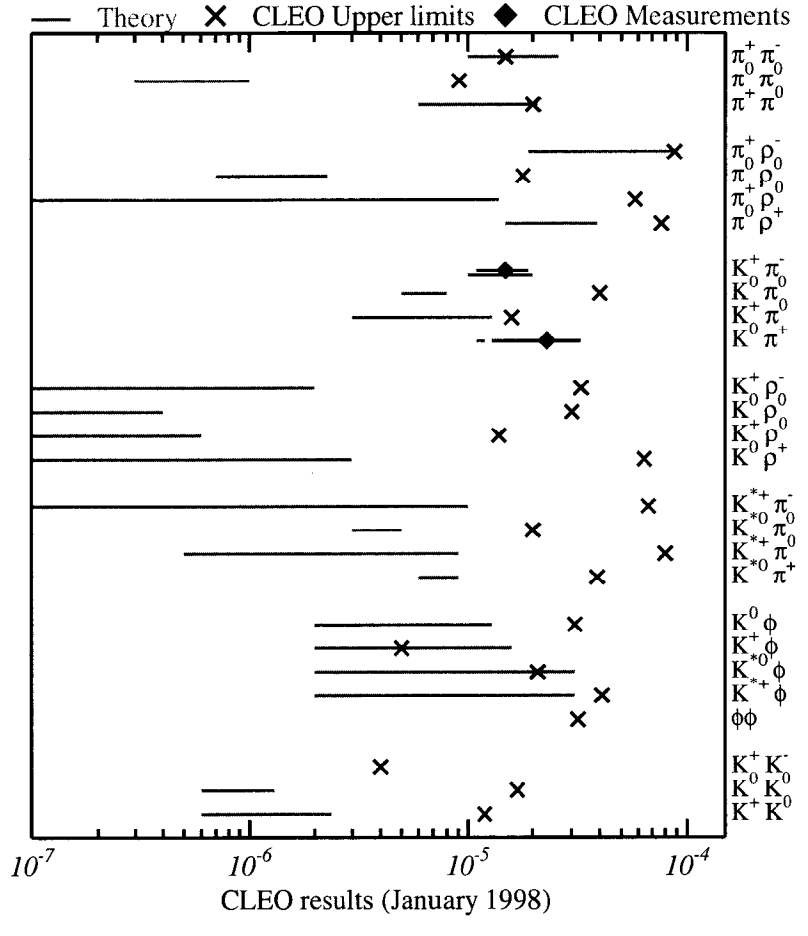


Fig. 9. A graphical summary of *published* CLEO rare charmless B decay results. All results are based on 5 fb^{-1} of CLEOII data, except modes of the form $h\rho$ and $K^*\pi$, which use 3.6 fb^{-1}. Note that some updated conference results are presented later, including the first observation of $B^+ \to K^+\pi^0$.

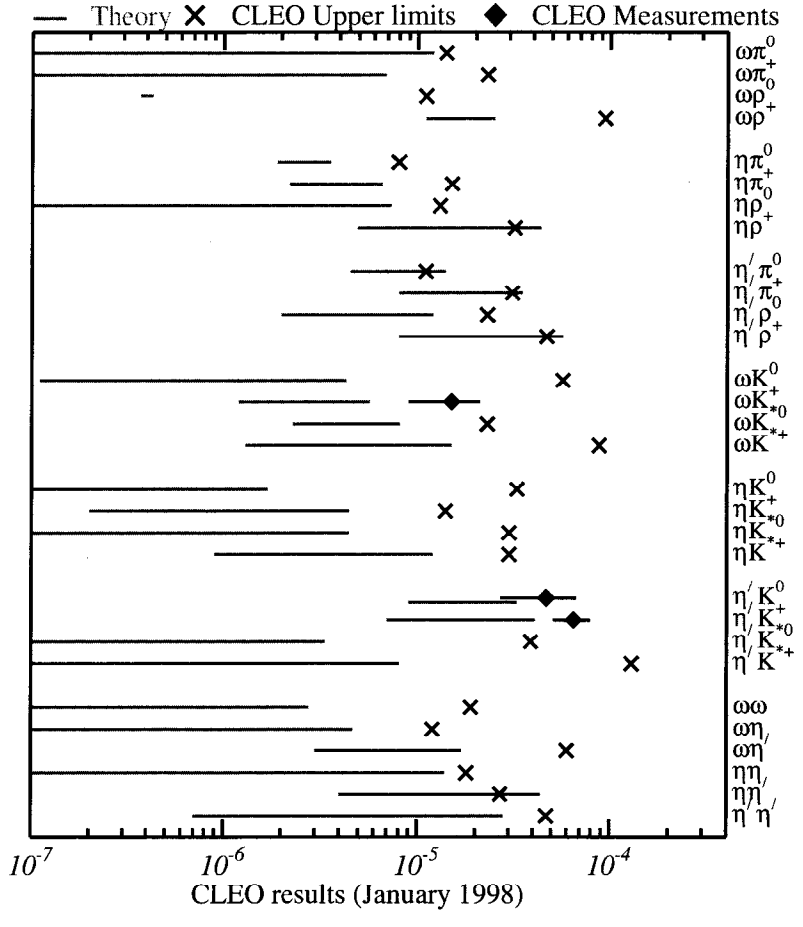


Fig. 10. A graphical summary of *published* CLEO rare charmless B decay results. All results are based on 5 fb^{-1} of CLEOII data. Note that some updated conference results are presented later.

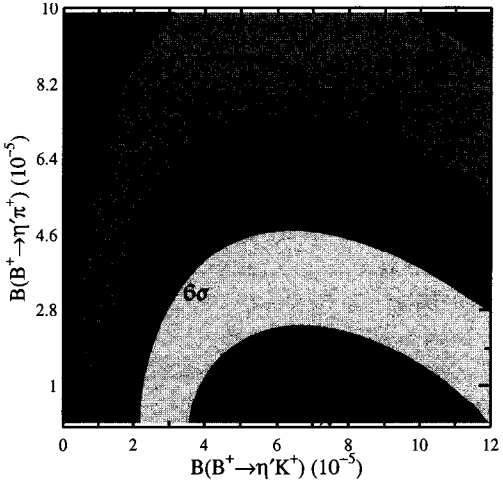


Fig. 11. Contours from the likelihood fit for the $B \to \eta' K^+$ and $B \to \eta' \pi^+$ branching ratios.

$KK, K\pi, \pi\pi$ results are summarized in Table 3; note that $B^+ \to K^+\pi^0$ is a new first observation.

Further updates on many modes, including more $B \to PV$ (pseudo-scalar vector) modes, like πK^* and $\pi\rho$, are expected to become available in the near future.

Mode	$\epsilon(\%)$	$N_{\rm signal}$	Signif.	$\mathcal{B}(\times 10^{-5})$
$\eta' K^+$	$29-36$	68.6	12.7	$7.4^{+0.8}_{-1.3} \pm 1.0$
$\eta' K^0$	$28-33$	16.1	7.3	$5.9^{+1.8}_{-1.6} \pm 0.9$
$\eta'\pi^+$	$29-36$	1.0	-	< 1.2

Table 2. A summary of new preliminary results for modes including an η'. Note that these results are NOT included in Figs. 9 and 10.

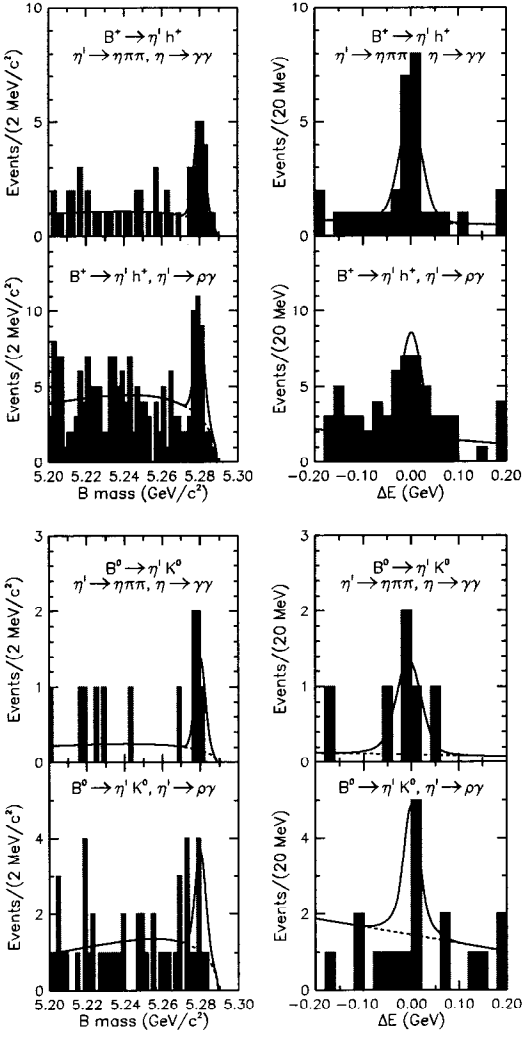


Fig. 12. Projections of ΔE and M_B for η' modes with an observed signal. The top four plots show projections for $\eta' h^{\pm}$ separated into two different η' decay channels; the bottom four plots are the analogous ones for $\eta' K_S$.

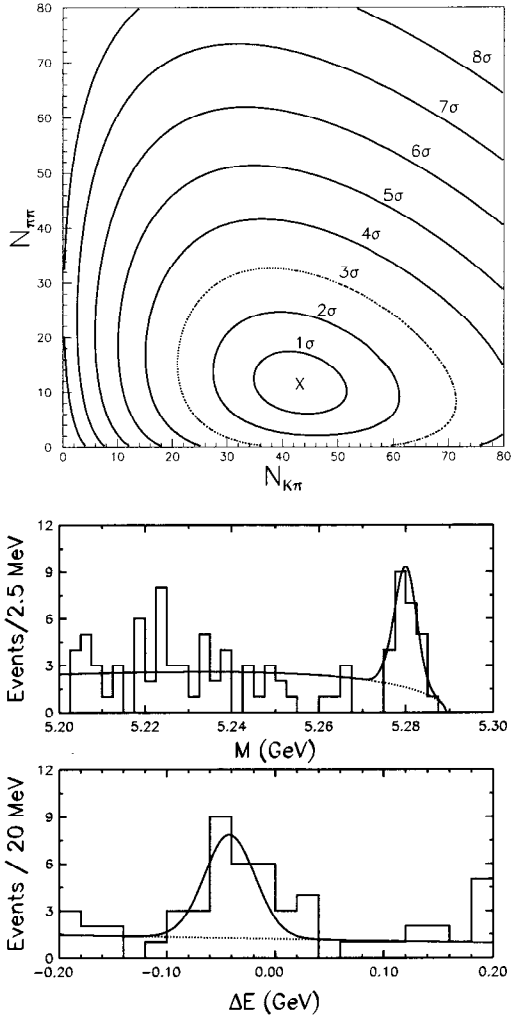


Fig. 13. The top plot shows the contours from the likelihood fit to $K^{\pm}\pi^{\mp}$ and $\pi^{\pm}\pi^{\mp}$ yields. The two lower plots are ΔE and M_B projections for the $h^{\pm}\pi^{\mp}$ signal. Note that our reconstruction assumes the π mass for all charged tracks, so ΔE shifts to -42 MeV for true $K^{\pm}\pi^{\mp}$.

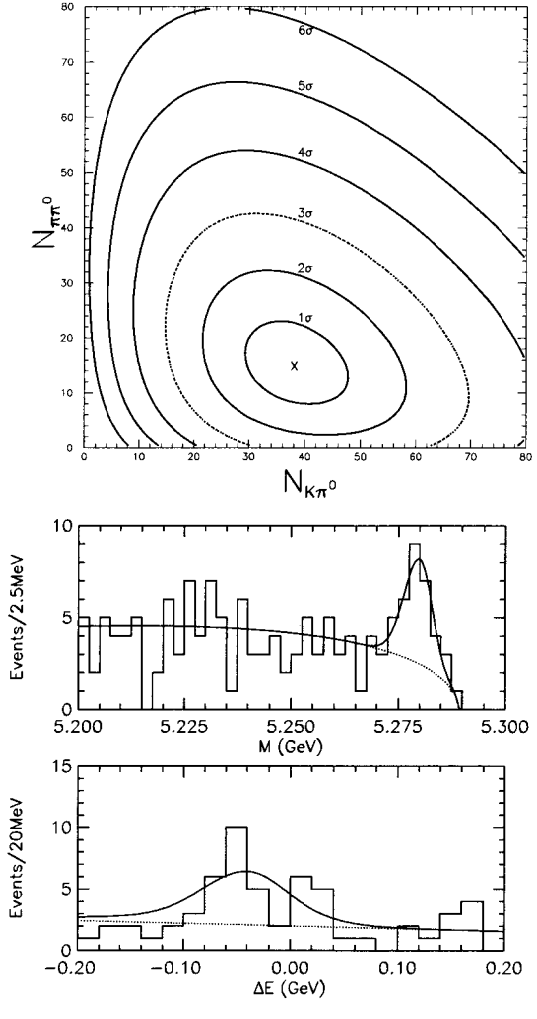


Fig. 14. The top plot shows the contours from the likelihood fit to $\pi^{\pm}\pi^0$ and $K^{\pm}\pi^0$ yields. The two lower plots are ΔE and M_B projections for the $h^{\pm}\pi^0$ signal. Note that our reconstruction assumes the π mass for all charged tracks, so ΔE shifts to -42 MeV for true $K^{\pm}\pi^0$.

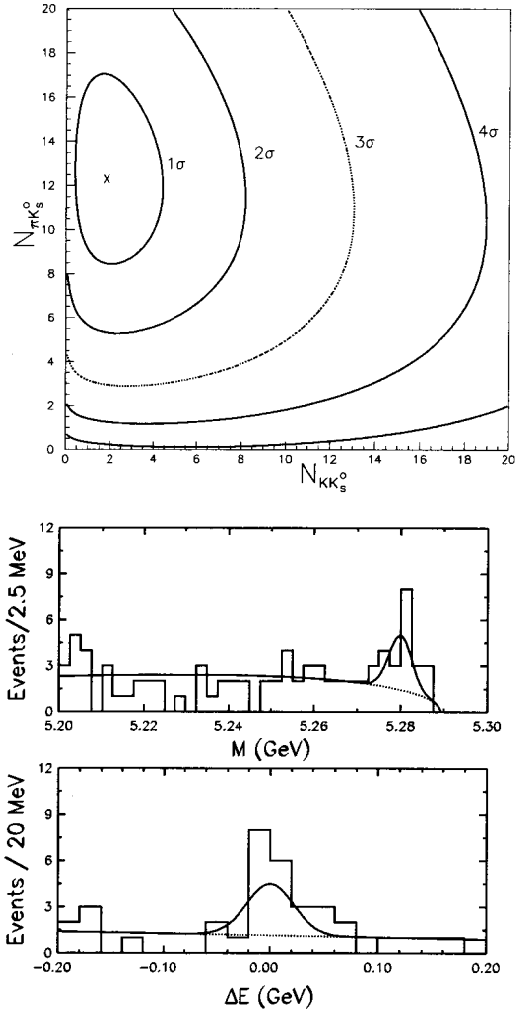


Fig. 15. The top plot shows the contours from the likelihood fit to the $\pi^{\pm}K_S$ and $K^{\pm}K_S$ yields. The lower plots are ΔE and M_B projections for the $h^{\pm}K_S$ signal.

Mode	$\mathcal{E}(\%)$	$\mathcal{B}$ (10^{-5})	Theory $\mathcal{B}$ (10^{-5})
$\pi^+\pi^-$	53 ± 5	< 0.84	0.8–2.6
$\pi^+\pi^0$	42 ± 4	< 1.6	0.4–2.0
$K^+\pi^-$	53 ± 5	$1.4 \pm 0.3 \pm 0.2$	0.7–2.4
$K^+\pi^0$	42 ± 4	$1.5 \pm 0.4 \pm 0.3$	0.3–1.3
$K^0\pi^+$	15 ± 2	$1.4 \pm 0.5 \pm 0.2$	0.8–1.5
K^+K^-	53 ± 5	< 0.24	–
$K^+\bar{K}^0$	15 ± 2	< 0.93	0.07–0.13

Table 3. A summary of new preliminary results for PP modes, where $P = \pi^\pm, \pi^0, K^\pm$, or K_S. Note that these results are NOT included in Figs. 9 and 10.

5.5 $B \to \eta_c K$

CLEO has observed large rates for the $B \to \eta' K$ decays. Some explanations for these rates have posited a $c\bar{c}$ component in the η'. This has motivated a search for the related $\eta_c K$ mode.

The analysis uses 5 fb^{-1} of CLEOII data in a maximum likelihood fit, just as is done for the charmless rare B decays. Only the $\eta_c \to \phi\phi$ decay is used; it has a small branching ratio ($\sim 0.7\%$), but is very clean. Figure 16 shows some of the kinematic distributions for the candidate events. There are two candidates with very low background; this gives a 3.9σ significance.

The preliminary result for the measured branching ratio is consistent with expectations based on the ψK rate:[20]

$$\mathcal{B}(B^- \to \eta_c K^-) = (1.54^{+1.39}_{-0.87}(stat) \pm 0.15(syst) \pm 0.60(\eta_c BR)) \times 10^{-3}. \quad (19)$$

Limits on some related modes are given in Table 4. Adding the $\eta_c \to K_S K\pi$ decay mode in the future will further increase our sensitivity.

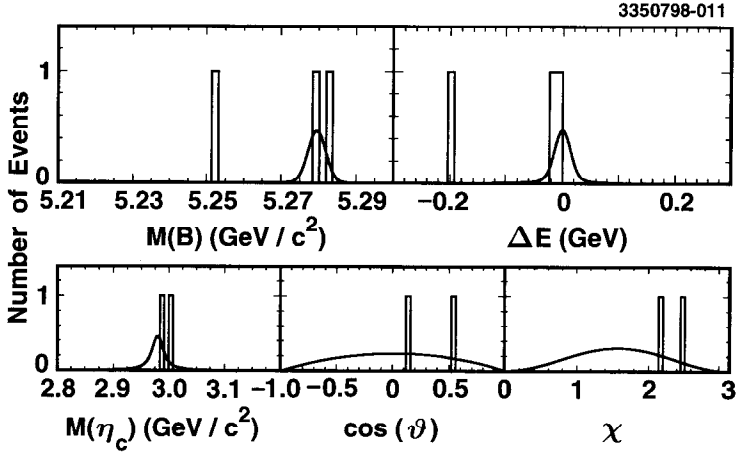


Fig. 16. The ΔE and M_B projections of the $\eta_c K^-$ fit. Also shown are the η_c mass, and the η_c decay angles.

Decay Channel	Upper Limit (90% C.L.)
$B^0 \to \eta_c K^0$	6.8×10^{-3}
$B^0 \to \eta_c K^{*0}$	5.95×10^{-3}
$B^\pm \to \eta_c K^{*\pm}$	18.5×10^{-3}

Table 4. Preliminary upper limits for modes related to the $B^- \to \eta_c K^-$ signal.

6 Rare Decays from Electroweak Penguins

Decays mediated by electroweak penguins are of considerable interest. The inclusive rate for $b \to s\gamma$ was first measured by CLEO.[21] It is a very important constraint on physics beyond the Standard Model; for example, it may be used to constrain the charged Higgs in SUSY models. This analysis has recently been updated.

The largest backgrounds come from $q\bar{q}$ continuum processes and initial-state radiation. It is difficult for generic B decays to make photons as stiff as those in the signal process. CLEO's analysis uses two techniques which are combined for the final result; only a brief summary of the techniques is given here.

Method one is a "shape analysis" which looks for only the stiff photon. A sophisticated combination of variables, including event shape and energy flow, is used to suppress background. The end result is an event weight which characterizes the chance that a given event is signal or background.

Method two utilizes a modified full reconstruction. A stiff photon, a kaon ($K^{\pm}$ or K_S), and from zero to four pions (where at most one π is a π^0) are combined in search of a viable B candidate. For each event, the best such candidate is picked based on a grand χ^2 which includes ΔE, M_B, particle ID information, and any K_S or π^0 masses. Finally, an event weight is produced as in method one; it is a function of ΔE, M_B, and $cos\theta_{tt}$ (where θ_{tt} is the angle between the candidate B thrust axis and the rest of the event).

A combined event weight is calculated when an event is present in both methods; the reconstruction analysis events are a subset of the shape analysis events. This technique yields greater statistical power than our previous, published analysis. More data and a lower photon energy cut (2.1 vs 2.2 GeV) are also used for this update. The measured E_γ spectrum is detailed in Fig. 17, showing a clear excess in the energy range expected from a quasi-two body γX_s final state. Our new preliminary result based on 5 fb^{-1} is:[22]

$$\mathcal{B}(b \to s\gamma) = (3.15 \pm 0.35(stat) \pm 0.32(syst) \pm 0.26(model)) \times 10^{-4}. \quad (20)$$

This is consistent with our old result and also very consistent with theory:[23]

$$\mathcal{B}(b \to s\gamma) = (3.28 \pm 0.33) \times 10^{-4}. \quad (21)$$

There are many related electroweak penguin modes also under study. The $B \to K^*\gamma$ decay was discovered by CLEO,[24] however $B \to \rho\gamma$ has yet to be seen. The ratio of branching fractions for these two modes can give $|V_{td}/V_{ts}|$. The latest results[25] use only 3.5 fb^{-1} of data; work on an update is underway.

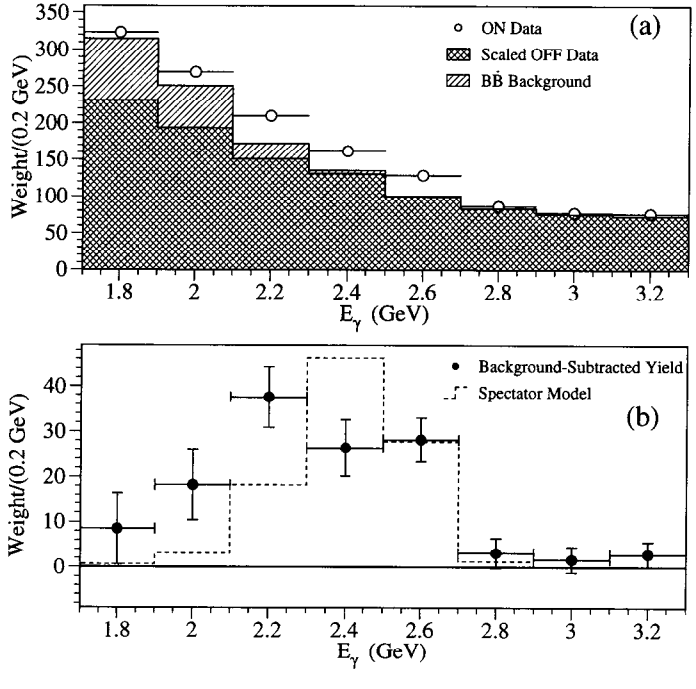


Fig. 17. The observed E_γ spectrum from the $b \to s\gamma$ analysis is shown in (a). The dominant background is from continuum events, but $B\bar{B}$ events come into play at lower photon energies. The background-subtracted plot in (b) shows a clear excess which is compared to a simple hadronization model shown as the dashed histogram.

Modes where the photon materializes as a lepton pair are sensitive to more terms in the effective Hamiltonian than $b \to s\gamma$ itself and thus are of interest also. Both an inclusive $b \to s\ell^+\ell^-$ and an exclusive $B \to K^{(*)}\ell^+\ell^-$ analysis have been performed with 5 fb^{-1} of CLEOII data. The inclusive results are:[26]

$$\mathcal{B}(b \to s\, e^+e^-) < 5.7 \times 10^{-5}, \quad (22)$$

$$\mathcal{B}(b \to s\, \mu^+\mu^-) < 5.8 \times 10^{-5}, \quad (23)$$

$$\mathcal{B}(b \to s\, e^\pm\mu^\mp) < 2.2 \times 10^{-5}, \quad (24)$$

$$(25)$$

while the preliminary exclusive results are:[27]

$$\mathcal{B}(B \to K^*\ell^+\ell^-) < 0.68 \times 10^{-5}, \quad (26)$$

$$\mathcal{B}(B \to K\ell^+\ell^-) < 0.70 \times 10^{-5}. \quad (27)$$

All are 90% C.L. limits.

Both the inclusive and exclusive dilepton modes have sensitivities which are approaching the theoretical expectations. Perhaps some of these modes, and the exclusive $\rho\gamma$ final state, will be seen in the next round of B experiments, providing us with further insights and constraints.

7 *D* Lifetimes with the Silicon Vertex Detector

CLEO now has the opportunity to make precise measurements of charm meson and baryon lifetimes due to the excellent resolution of the new silicon vertex detector. The first such analysis to explore these possibilities uses 3.7 fb^{-1} of CLEOII.V data and measures the D meson lifetimes.

The charm mesons were reconstructed via the following decay modes:

$$D^0 \to K^-\pi^+, \quad K^-\pi^+\pi^0, \quad K^-\pi^-\pi^+\pi^+, \quad (28)$$

$$D^+ \to K^-\pi^+\pi^+, \quad (29)$$

$$D_s \to \phi\pi^+ \quad (\text{with } \phi \to K^+K^-). \quad (30)$$

For the D^0 (D^+), a slow π^+ (π^0) was required to form a good $D^{*+} \to D^0\pi^+$ ($D^{*+} \to D^+\pi^0$) candidate in order to reduce background. No tagging was required for the D_s mesons.

The D flight path is extrapolated to the flat, well-known beam spot to determine the production point. The beam spot size is given by $\sigma_y \sim 7\,\mu$m, $\sigma_x \sim 350\,\mu$m, and is determined on a run-by-run basis. This construction is illustrated in Fig. 18. Fragmentation tracks, which may bias the interaction point, are avoided. The $D^0 \to K^-\pi^+$ vertex resolution varies from about 60–110 μm, along the flight direction, compared to a typical flight of $< \gamma\beta c\tau > = 200\mu$m. Only the two-dimensional projection of the decay in the $r - \phi$ plane is used. This is done by choice; the silicon z measurements are available if desired.

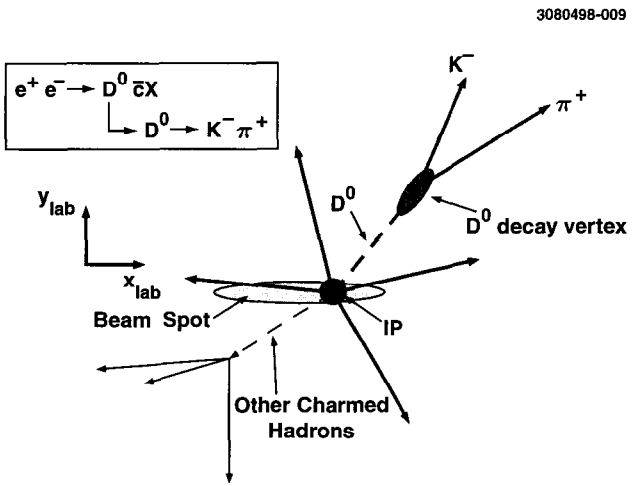


Fig. 18. An illustration of the technique used in determining charm lifetimes; see the text for further description.

An unbinned maximum-likelihood fit with seven parameters is used to extract results. The fit parameters are: the D lifetime, the fraction of background with lifetime, the lifetime of this background, the fraction of mis-reconstructed events and their average resolution, the fraction of "flat" background (with a fixed resolution of 6 ps), and a scale factor for the predicted measurement error (found to be 1.13 ± 0.02). Figure 19 shows the fit to the proper lifetime data.

The preliminary results for the lifetimes of the D^0, D^+, and D_s^+ mesons are:[28]

$$D^0 \;:\; 408.5(stat) \pm 4.1^{+3.0}_{-2.9}(syst) \text{ fs}, \quad (31)$$

$$D^+ \;:\; 1033.6(stat) \pm 22.1^{+7.4}_{-10.9}(syst) \text{ fs}, \tag{32}$$

$$D_s \;:\; 486.3(stat) \pm 15.0^{+3.8}_{-4.0}(syst) \text{ fs}. \tag{33}$$

The D^0 lifetime reported here is comparable in precision with the world average and the D_s lifetime is more precise than the world average. The excellent $D^* - D$ mass-difference resolution in the CLEOII.V data is illustrated in Fig. 20; the resolution is 210 KeV. This is achieved with a three-dimensional vertex fit to improve the slow-pion parameters. This resolution directly improves the signal to noise in certain key analyses. In particular, CLEO is re-measuring the doubly-Cabibbo-suppressed decay (DCSD) $D^0 \to K^+\pi^-$, first seen by CLEO.[29] With the silicon detector, CLEO will also be able to directly separate the DCSD process from possible $D - \bar{D}$ mixing, which is negligible in the Standard Model. Such excellent mass-difference resolution also begs for a study of the intrinsic width of the D^*, which may be directly accessible for the first time. Clean signals are also available for several D^0 decays to CP-eigenstates; preliminary peaks with loose cuts are shown in Fig. 20. More data and further background suppression are in hand already. These modes will be used to look for lifetime differences between the two CP-eigenstates of the D^0. Finally, efforts are underway to use the silicon for measuring charm baryon lifetimes.

8 Charm Baryons

Charm baryon spectroscopy is an area in which CLEO has recently displayed great prowess. Eight new states have been discovered since the 1996 PDG listing, all of them at CLEO. The newest among these are the two Ξ_c' states,[30] shown in Fig. 21, and the first two $L = 1$ Ξ_c states,[31] shown in Fig. 22. All these preliminary results use 5 fb^{-1} of data. In each case, the observation of the two expected charge states greatly aids our confidence in the existence and interpretation of these new states.

The world knowledge of charm baryon spectroscopy is summarized in Table 5. In addition to the ground state multiplets only the established $L = 1$ excitations are listed. CLEO has discovered 11 states, all other experiments combined have found six, and there are two $L = 0$ states (and likely many more narrow $L = 1$) states still to be found.†

†One might argue that the accounting is 10 and 7, rather than 11 and 6, if a single candidate Σ_c^+ is taken as a convincing discovery.

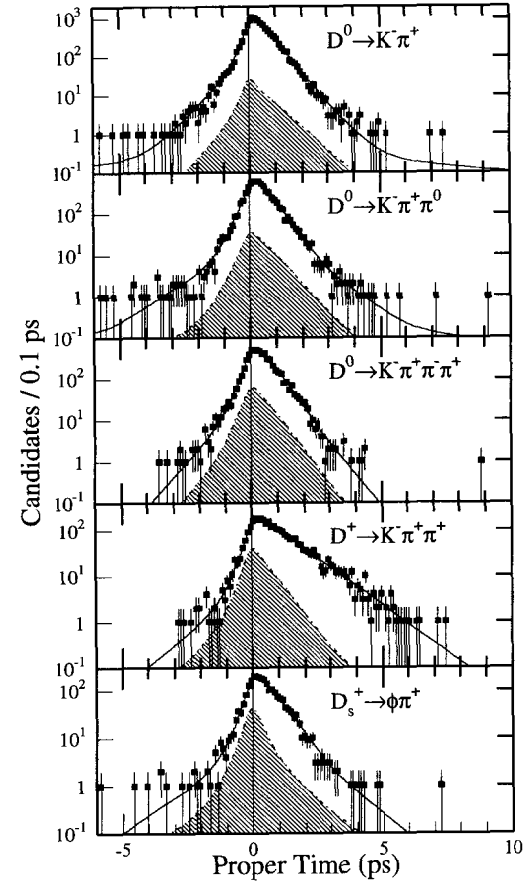


Fig. 19. The results of the D lifetime fits for each decay channel, shown as solid curves overlaid on the data points. The shaded histograms indicate the background level and shape.

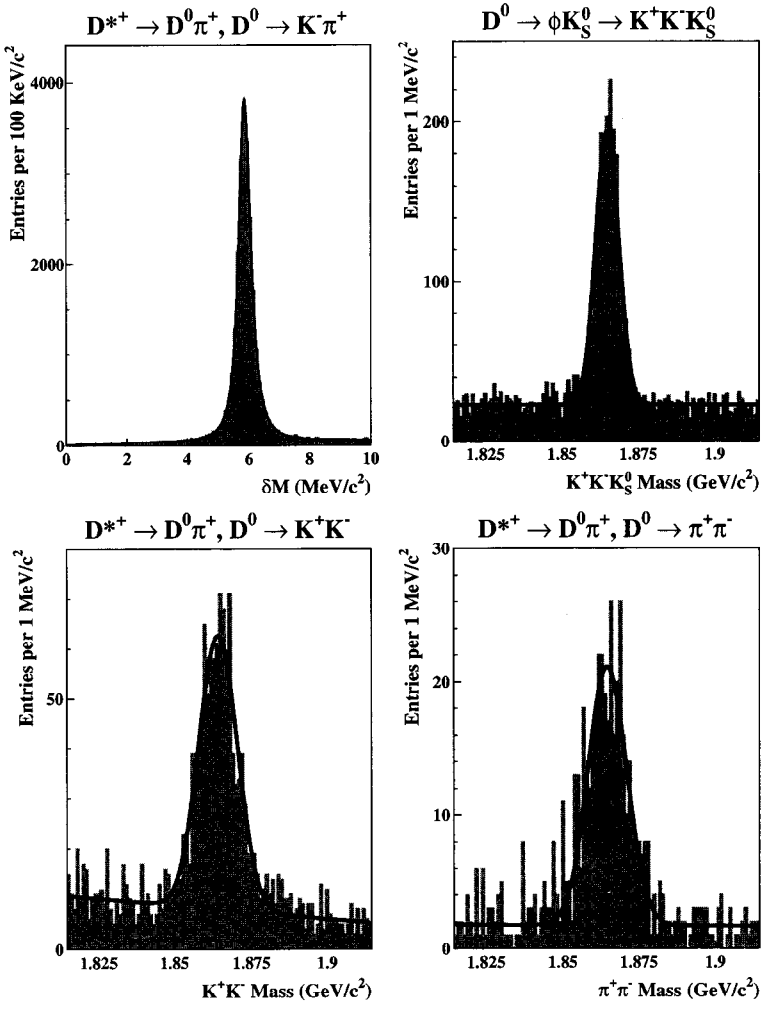


Fig. 20. The first plot demonstrates our excellent $D^* - D$ mass difference resolution; mass($K\pi\pi$) - mass($K\pi$) - m_π is plotted. The remaining plots show mass peaks for some decay modes of the D^0 into CP-eigenstates.

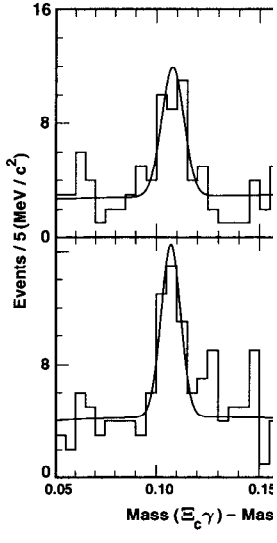


Fig. 21. Plots of the $\Xi_c\gamma - \Xi_c$ mass f

CLEO hopes to find even more narrow L : be increasingly careful when hunting for new ous correlated soft πs and γs from transitions underway working to catalog new decay mode as well.

In the near future, CLEO will measure cl vertex detector. An inclusive single-experim service. Finally, one crucial need in this field is $pK\pi$). Attempts are underway to contribute he

9 Tau Physics

CLEO has the world's highest statistics tau sa topics. The particular analysis discussed here a light hadron physics.

$L=0$ $J^P=1/2^+$			Ω_c^0		
	$\underline{\Xi_c^0}$	$\underline{\Xi_c^+}$	$\underline{\Xi_c^{0\prime}}$	$\underline{\Xi_c^{+\prime}}$	
	Λ_c^+	Σ_c^0	$\underline{\Sigma_c^+}$	$\underline{\Sigma_c^{++}}$	
$L=0$ $J^P=3/2^+$			$\underline{\underline{\Omega_c^{0*}}}$		
		$\underline{\underline{\Xi_c^{*0}}}$		$\underline{\underline{\Xi_c^{*+}}}$	
		$\underline{\Sigma_c^{*0}}$	$\underline{\underline{\Sigma_c^{*+}}}$	$\underline{\Sigma_c^{*++}}$	
$L=1$ $J^P=1/2^+$	$\underline{\Lambda_c(2593)}$				
$L=1$ $J^P=3/2^+$	$\Lambda_c(2625)$ $\underline{\Xi_c^+(2815)}$ $\underline{\Xi_c^0(2820)}$				

Table 5. A graphic illustration of CLEO's contributions to charm baryon spectroscopy. The top two sections are arranged according to $SU(3)$ multiplets. States underlined were first observed by CLEO. States not underlined were first seen by other experiments. States underlined twice are not yet established. Only the observed, narrow $L=1$ states are listed; the J^P assignments for many of the states are educated guesses.

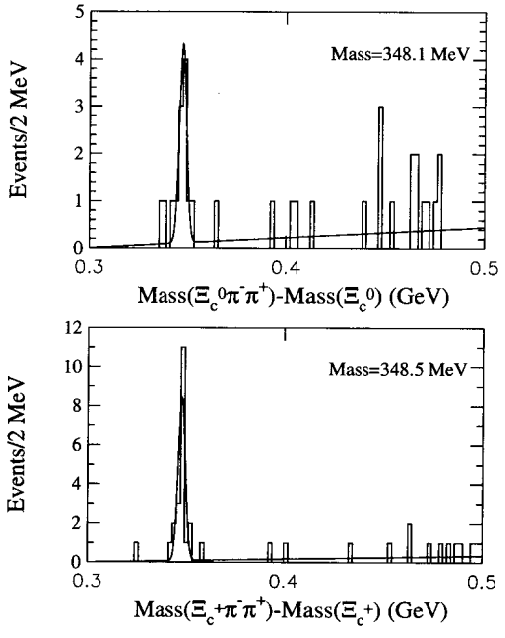


Fig. 22. Plots of the $\Xi_c\pi^-\pi^+ - \Xi_c$ mass for the $L=1$ Ξ_{c1}^0 and Ξ_{c1}^+ candidates.

A very thorough analysis of $\tau \to \pi^0\pi^0\pi^-$ was performed on 5 fb^{-1} of CLEOII data.[32] The decay is dominated by the s-wave $a_1(1260) \to \rho\pi$ component. Details of the $a_1(1260)$ state are poorly understood and tau decay is a perfect laboratory in which to study it. Recall also that the hadronic mass shape near the endpoint of the 3π mode is of interest for neutrino mass limit analyses.

The $\pi^-\pi^0\pi^0$ final state is superior to the $\pi^+\pi^+\pi^-$ one since it has lower hadronic and tau feed-across background. Also, there is only one $I=0$ combination of πs, making it ideal for looking at scalar resonances in the decay substructure.

The 3π Dalitz plots are fit for several components relative to the dominant s-wave $\rho\pi$ term. When this is done in bins of 3π mass, it is found that the data is well-described with resonance parameters which are independent of this mass; therefore we take these parameters as constant. A portion of the information extracted about the complex couplings for each channel is summarized in Table 6; more information is contained in the conference paper.[32]

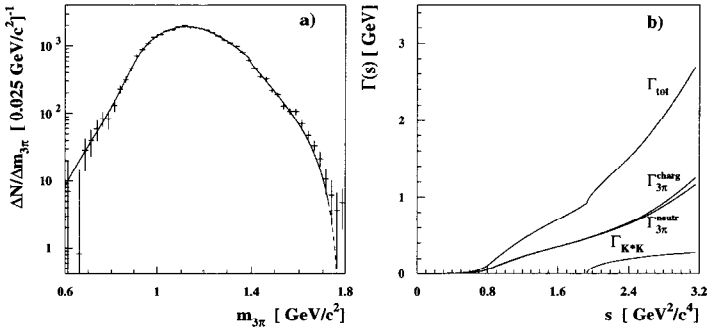


Fig. 23. The left plot shows the fit to the 3π invariant mass from $\tau \to \pi^0\pi^0\pi^-$ decays. On the right are the mass-dependent channel widths corresponding to this fit.

With the substructure determined, one can examine the 3π mass spectrum itself, which is dominated by the a_1 lineshape. The preliminary fit is shown in Fig. 23. Note that the high-mass structure visible there is partly from the K^*K threshold, which must of course be present. The coupling to this channel is allowed to float in the fit, with the turn-on shape taken from theory.

Resonance	L	Signif.	$\mathcal{B}$ fraction(%)
$\rho(1450)$	s-wave	1.4σ	$0.30 \pm 0.64 \pm 0.17$
ρ	d-wave	5.0σ	$0.36 \pm 0.17 \pm 0.06$
$\rho(1450)$	d-wave	3.1σ	$0.43 \pm 0.28 \pm 0.06$
$f_2(1275)$	p-wave	4.2σ	$0.14 \pm 0.06 \pm 0.02$
σ	p-wave	8.2σ	$16.18 \pm 3.85 \pm 1.28$
$f_0(1370)$	p-wave	5.4σ	$4.29 \pm 2.29 \pm 0.73$

Table 6. Results from preliminary fits to the substructure in the 3π system in $\tau \to \pi^0\pi^0\pi^-$ decays. The strength and significance of each component is noted. Note that the branching ratios do not sum to one due to interference.

This analysis also yields signed neutrino helicity via a decay asymmetry due to interference between the two $\pi^-\pi^0$ combinations present. The analysis yields $h_{\nu_\tau} = -1.02 \pm 0.13 \pm 0.03$, as compared to the expectation of -1.

A corresponding analysis looking at the $\pi^+\pi^-\pi^-$ mode has already produced some promising results.[32] Using the $\pi^-\pi^0\pi^0$ fit results in an appropriately modified current suitable for the $\pi^+\pi^-\pi^-$ final state, once can generate a Monte Carlo event sample. The agreement with the $\pi^+\pi^-\pi^-$ data is excellent, indicating that consistent results will be extracted from the two modes.

10 Two-Photon Physics

CLEO also makes good use of the two-photon physics reach in our dataset. One recent study involves determining the two-photon coupling of a glueball candidate.

From previous measurements of $\psi \to gg\gamma \to X\gamma$ done elsewhere, one can extract information on the coupling of a state X to gluons. Searching for X in two-photon production can measure or limit the coupling to $\gamma\gamma$.

One can then define for particle X a "stickiness," S:

$$S \sim \Gamma(\psi \to \gamma X)/\Gamma(X \to \gamma\gamma) \tag{34}$$

where the factors omitted compensate for effects of differing coupling constants, phase space and spins. Note that the branching ratio of X to the detected final state can be canceled. The value of S should be of $\mathcal{O}(1)$ for $q\bar{q}$ mesons, since valence quarks couple well to both glue and photons. CLEO has looked for two-photon production of the $f_J(2220)$ in both the K_SK_S and $\pi^+\pi^-$ final states. The invariant mass spectra of the candidates is shown for both searches in Fig. 24; in both cases, limits were extracted. The final combined result for the stickiness, S, is:[33,34]

$$S(f_J(2220)) > 102 \qquad 95\% \text{ C.L.}, \tag{35}$$

making the $f_J(2220)$ the stickiest state yet evaluated and strengthening the case for its interpretation as a glueball. One potential out would be to question the very existence of this state; although it has been seen by several experiments in several decay modes, some questions still remain.[35]

11 Other Recent Analysis Topics

There are many new summer conference papers which were not discussed, due to length constraints. These include:

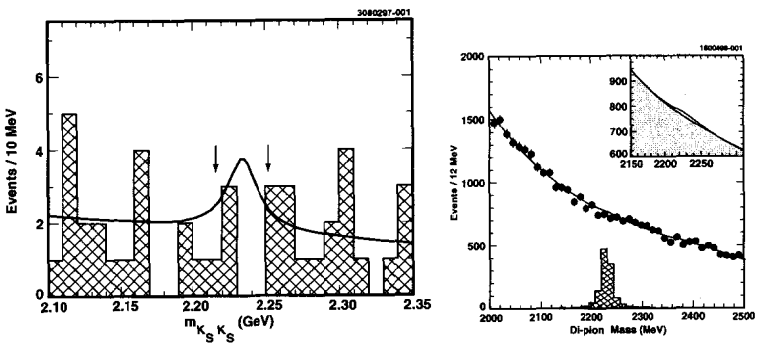


Fig. 24. The $K_S K_S$ (left) and $\pi\pi$ (right) invariant mass spectrum from searches for two-photon production of the $f_J(2220)$.

***B* Physics:**

- Observation of High Momentum η' Production in B Decay
- Search for $B \to \rho^0 K^0$, $B \to K^{*\pm}\pi^{\mp}$
- Distribution in q^2 of the Decay $\bar{B}^0 \to D^{*+}\ell\bar{\nu}$ via Partial Reconstruction
- $\bar{B} \to D\ell\bar{\nu}$ Branching Fractions and Form Factor Parameters

Charm Physics:

- Measurement of the Decay Asymmetry Parameter in $\Xi_c^0 \to \Xi^-\pi^+$ and $\Xi^- \to \Lambda\pi^-$ and a Search for Direct CP Violation in Hyperon Decays
- Improved Measurement of the Pseudo-Scalar Decay Constant f_{Ds}

Tau Physics:

- Resonance Structure of $\tau^- \to K^-\pi^+\pi^-\nu_\tau$ Decays
- First Search for CP Violation in Tau Lepton Decay
- A Limit on the Mass of the ν_τ

Upsilon Physics:

- Υ Dipion Transitions at Energies near the $\Upsilon(4S)$
- Measurement of the Mass Splittings between the $B\bar{B}$ $\chi_{b,J}$ States.

This is just a sampling of CLEO's diverse activities. There are also many other new (i.e., not merely update) analyses underway.

12 Conclusion

Much of CLEO B Physics revolves around te CK CKM elements $|V_{cb}|$, $|V_{ub}|$ will be measured › even way, CLEO is also pioneering important techiques l efforts to explore the CP-violating aspects o the C a comprehensive set of rare B decay searchs; mai including some surprises, such as $B \to \eta' l$. It i decays is needed to get at the interesting physics di diagrams that contribute to these processes. CLEO'; by the time our run ends for CLEOIII instalition; for most B physics topics.

CLEO is also a major force in charm phyics (b physics, Upsilon physics, and two-photon physics. make use of a silicon vertex detector at the Υ4S), a three future B factories. Our own entry in th B fa online in 1999.

Current preprints and other information ai avail URL: http://www.lns.cornell.edu/public/CLEO/ . B no doubt be more exciting new results from CLEO tion concerning the impressive accomplishmats of available at http://www.lns.cornell.edu/publicCESR

13 Acknowledgments

Many thanks to my CLEO colleagues for the hard and especially to those who supplied figures ad oth these proceedings. I apologize for limiting m refere tal work; additional references should be easiy obta

References

[1] CLEO Collaboration, Y. Kubota *et al.*, Nucl. Instrum. Methods Phys. Res., Sec. A **320**, 66 (1992).

[2] CLEO Collaboration, T. S. Hill, Nucl. Instrum. Methods Phys. Res., Sec. A **418**, 32 (1998).

[3] CLEO Collaboration, R. A. Briere, "Tracking in Helium-Based Gases: Present and Future B Factories," to appear in the *Proceedings of the Seventh International Symposium on Heavy Flavor Physics*, Santa Barbara, CA, July 1997.

[4] CLEO Collaboration, J. Bartelt *et al.*, Phys. Rev. Lett. **71**, 4111 (1993).

[5] CLEO Collaboration, J. P. Alexander *et al.*, Phys. Rev. Lett. **77**, 5000 (1996).

[6] CLEO Collaboration, C. P. Jessop *et al.*, CLEO CONF 98-18; ICHEP 855.

[7] CLEO Collaboration, B. Barish *et al.*, Phys. Rev. Lett. **76**, 1570 (1996).

[8] CLEO Collaboration, B. Barish *et al.*, Phys. Rev. D **51**, 1014 (1995).

[9] CLEO Collaboration, J. Bartelt *et al.*, CLEO CONF 98-21; ICHEP 1013.

[10] CLEO Collaboration, C. P. Jessop *et al.*, Phys. Rev. Lett. **79**, 4533 (1997).

[11] CLEO Collaboration, M. Artuso *et al.*, Phys Rev. Lett. **82**, 3020 (1999).

[12] CLEO Collaboration, M. Athanas *et al.*, Phys. Rev. Lett. **80**, 5493 (1998).

[13] CLEO Collaboration, G. Bonvicini *et al.*, CLEO CONF 98-23; ICHEP 852.

[14] CLEO Collaboration, R. Godang *et al.*, Phys. Rev. Lett. **80**, 3456 (1998).

[15] CLEO Collaboration, B. H. Behrens *et al.*, Phys. Rev. Lett. **80**, 3710 (1998).

[16] CLEO Collaboration, T. Bergfeld *et al.*, Phys. Rev. Lett. **81**, 272 (1998).

[17] CLEO Collaboration, D. M. Asner *et al.*, Phys. Rev. **D53**, 1039 (1996).

[18] CLEO Collaboration, B. H. Behrens *et al.*, CLEO CONF 98-09; ICHEP 860.

[19] CLEO Collaboration, M. Artuso *et al.*, CLEO CONF 98-20; ICHEP 858.

[20] CLEO Collaboration, S. Chan *et al.*, CLEO CONF 98-24; ICHEP 1073.

[21] CLEO Collaboration, M. S. Alam *et al.*, Phys. Rev. Lett. **74**, 2885 (1995).

[22] CLEO Collaboration, S. Glenn *et al.*, CLEO CONF 98-17; ICHEP 1011.

[23] K. Chetyrkin, M. Misiak, and M. Munz, Phys. Lett. **B400**, 206 (1997), E: ibid., **B425**, 414 (1998).

[24] CLEO Collaboration, R. Ammar *et al.*, Phys. Rev. Lett., **71**, 674 (1993).

[25] CLEO Collaboration, K. W. Edwards *et al.*, CLEO CONF 95-6; EPS0160.

[26] CLEO Collaboration, S. Glenn *et al.*, Phys. Rev. Lett. **80**, 2289 (1998).

[27] CLEO Collaboration, R. Godang *et al.*, CLEO CONF 98-22; ICHEP 1012.

[28] CLEO Collaboration, Phys. Rev. Lett. **82**, 4586 (1999). An earlier version of this work is detailed in: CLEO Collaboration, S. J. Richichi *et al.*, CLEO CONF 98-15; ICHEP 864.

[29] CLEO Collaboration, D. Cinabro *et al.*, Phys. Rev. Lett. **72**, 1406 (1994).

[30] CLEO Collaboration, C. P. Jessop *et al.*, Phys. Rev. Lett. **82**, 492 (1999).

[31] CLEO Collaboration, M. Athanas *et al.*, CLEO CONF 98-10; ICHEP 866.

[32] CLEO Collaboration, T. E. Coan *et al.*, CLEO CONF 98-19, ICHEP98 976.

[33] CLEO Collaboration, R. Godang *et al.*, Phys. Rev. Lett. **79**, 3829 (1997).

[34] CLEO Collaboration, M. S. Alam *et al.*, Phys. Rev. Lett. **81**, 3328 (1998).

[35] Review of Particle Physics, C. Caso *et al.*, Eur. Phys. J. C **3** 1 (1998).

HEAVY FLAVOR PHYSICS AT THE TEVATRON

Giorgio Apollinari
Fermi National Accelerator Laboratory
Batavia, IL 60510

Representing the DØ and CDF Collaborations

ABSTRACT

We report on the status of top and b quark physics at the Fermilab Tevatron collider. In particular, we summarize the knowledge obtained by CDF and DØ on the top quark mass and production cross section. We also present some new interesting results obtained by CDF with the discovery of the B_c meson, and a first low-statistics measurement of $\sin(2\beta)$ at a hadronic collider.

© 1998 by Giorgio Apollinari.

1 Introduction

In this paper, we present a review of recent results obtained on the subject of heavy flavor physics at the Tevatron $p\bar{p}$ collider at Fermilab. After a brief historical overview in Sec. 2, we summarize the status of top quark physics at CDF and D$\emptyset$ in Sec. 3. In particular, we discuss the measurement of the top production cross section and the world-average direct determination of the top quark mass. Section 4 is devoted to recent B physics results at the Tevatron collider, with emphasis on a preliminary, low-statistics measurement of the time-dependent asymmetry in the decay B^0, $\overline{B^0} \to J/\psi K_s^0$. We will also discuss the observation of the B_c meson. Finally, a brief outlook at the future prospects for top quark physics, as well as CP-violation measurements in the b sector, is given in Sec. 5.

2 Historical Overview

Heavy flavor physics at the Tevatron collider is synonymous with physics of the third quark generation. Since the discovery of the τ lepton[1] in 1976, the third quark generation has been the "most anticipated" quark family. At least part of the anticipation was based on theoretical arguments[2] that explained the observed CP violation in the K^0–$\overline{K^0}$ system[3] through the existence of three quark families, and an appropriate mixing between the mass and the weak interaction eigenstates.

In 1977, the b quark was discovered as a dimuon resonance in 400 GeV proton-nucleus collisions at Fermilab,[4] and soon after its discovery, the DORIS e^+e^- storage ring at DESY[5] confirmed the existence of the Υ resonances at a mass of about 9.5 GeV/c^2. These narrow resonances allowed the assignment of a charge -1/3 to the newly discovered b quark.

The Standard Model (SM) is an anomaly free theory[6] if and only if the sum of all the left-handed fermion charges in a given family is null ($\Sigma_l Q^l + N_c \Sigma_q Q^q = 0$, where N_c is the number of colored quarks, Q^l are the lepton charges, and Q^q are the quark charges in a family). Therefore, the -1/3 charge for the b quark and the existence of the τ lepton implied the existence of a b quark partner with a charge of +2/3. This partner is by definition the top quark.

Similarly, in the SM, the b and top quark are members of a weak isospin doublet, and further evidence for the existence of the top quark was produced by the weak isospin measurement of the b quark, extracted from the forward-backward asymmetry A_{FB} in $e^+e^- \to b\bar{b}$ at PETRA[7] and LEP-SLD,[8] where the measurements supported the assignment of $I_3 = -1/2$ for the weak isospin of the b quark, and therefore, the necessity of a $I_3 = +1/2$ partner of the b quark.

The discovery of the top quark has been one of the goals of the Fermilab Tevatron Collider since its commissioning in 1985. However, the top quark has not been the only heavy quark studied at the Tevatron. As will be shown in Sec. 4, the Tevatron collider, and CDF in particular, have also made major contributions to the study of the b sector of the SM.

2.1 Experimental Tools

The Fermilab $p\bar{p}$ Tevatron Collider, with a center-of-mass energy of 1.8 TeV, is the highest energy collider in the world.

Designed for a peak luminosity of 10^{30} cm^{-2}s^{-1}, it routinely reached luminosities of 2×10^{31} cm^{-2}s^{-1} during the 1992-1993 and 1994-1995 data-taking periods, thanks to a set of electrostatic insertions[9] separating the beams, except at the CDF and D$\emptyset$ intersection regions. These insertions improved the beam lifetime and decreased the beam emittance. The bunches cross every 3.5 μs producing a luminous region with a Gaussian distribution having a width of $\sigma_z = 30$ cm along the beam line, and $\sim 36\ \mu m$ in the direction perpendicular to the beams.

The CDF detector has been extensively described in detail elsewhere.[10] Briefly, it consists of a magnetic spectrometer surrounded by calorimeters and muon chambers. A new low-noise, radiation hard four layer silicon vertex detector (SVX),[11] located just immediately outside the beam pipe, provides precise track reconstruction in the direction perpendicular to the beams ($\sigma_{d_0} \sim 17\ \mu m$) and is used to identify secondary vertices from b and c quark decays. The momentum of charged particles is measured in the central tracking chamber (CTC) which is immersed in a 1.4-T superconducting solenoidal magnet, with an excellent momentum resolution ($\sigma(P_t)/P_t = 0.1\%P_t \oplus 0.6\%$). Outside the solenoid, electromagnetic ($\sigma(E_{em})/E_{em} \sim 17\%/\sqrt{E_{em}}$) and hadronic ($\sigma(E_{had})/E_{had} \sim 50\%/\sqrt{E_{had}}$) calorimeters cover the whole azimuthal angle and extend to pseudorapidities of $|\eta| \sim 4.2$. They are used to identify and measure jets and electron candidates. The calorimeters are also used to measure the missing transverse energy ($\not\!\!E_T$) which can indicate the presence of undetected energetic neutrinos. Outside the calorimeter, drift chambers in the region $|\eta| < 1.0$ provide muon identification. A three level trigger selects inclusive electrons and muons used for the

top search. The detection efficiency for $t\bar{t}$ events is improved by the inclusion of triggers based on $\not\!\!E_T$.

A schematic view of the CDF detector is shown in Fig. 1(a).

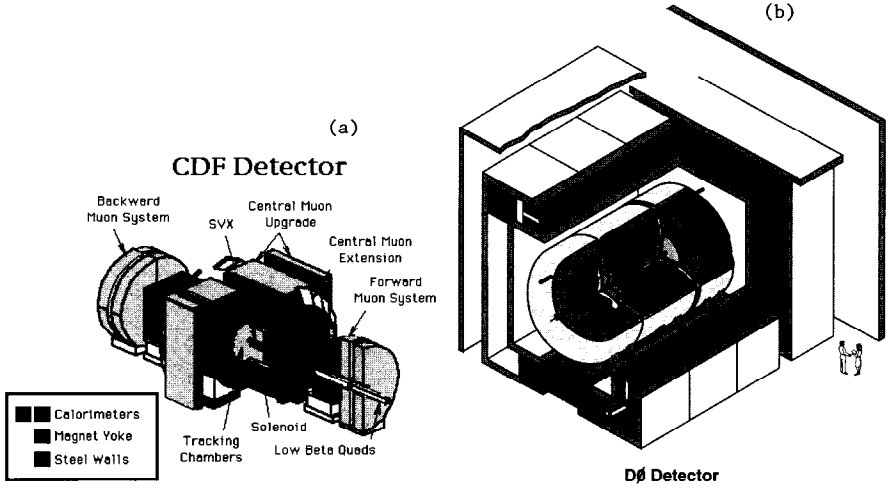


Figure 1: Isometric views of (a) CDF and (b) D$\emptyset$ detectors.

The D$\emptyset$ detector and data collection system are also described elsewhere.[12] The D$\emptyset$ detector has a hermetic, compensating sampling calorimeter with fine longitudinal and transverse segmentation in pseudorapidity and azimuthal angle. The energy resolutions are slightly better than those measured at CDF ($\sigma(E_{em})/E_{em} \sim 15\%/\sqrt{E_{em}}$) and ($\sigma(E_{had})/E_{had} \sim 50\%/\sqrt{E_{had}}$). Since there is no central field, charged particle tracks are reconstructed with a sign degeneracy using drift chambers located between the interaction region and the calorimeter. Electrons are identified by a transition radiation detector. Muons are detected by reconstructing tracks in proportional drift tubes before and behind a set of magnetized iron toroids located outside the calorimeter which provide some momentum measurement with a resolution of $\sigma(P_t)/P_t = 0.3\%P_t \oplus 17.\%$ for pseudorapidities in the range $|\eta| < 3.0$. The good calorimeter hermeticity provides a good missing transverse energy resolution. The CDF transverse energy resolution is approximately 20% worse than D$\emptyset$. A schematic view of the D$\emptyset$ detector is shown in Fig. 1(b).

3 Top Quark at the Tevatron Collider

As mentioned previously, the top quark has been searched since the discovery of its partner, the b quark, in 1977. The first indication that the top quark was a heavy object came from the measurement of a large mixing parameter x_d in the B^0–$\overline{B^0}$ mixing, first observed by the UA1[13] and ARGUS[14] Collaborations.

Precise electroweak fits from LEP have constrained the top quark mass with ever increasing precision since the turn-on of LEP in 1990, as shown in Fig. 2.

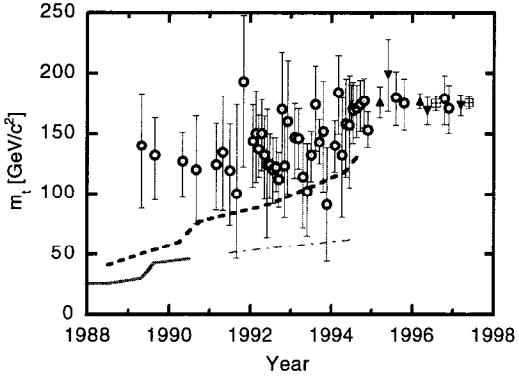


Figure 2: Historical evolution[15] of the indirect fits to the top quark mass from precise electroweak measurements at LEP (circles). The solid and dashed lines indicate the 95% C.L. on the lower M_{top} bound from direct e^+e^- and $p\bar{p}$ searches. The last points are the direct measurements from CDF (triangles) and D$\emptyset$ (inverted triangles).

The first evidence for the top quark's existence was published by CDF[16] in 1994, with a 2.8σ excess of events over the background expectations. Under the assumption of top production, CDF measured a top mass of $M_{top} = 174 \pm 16$ GeV/c^2 and a production cross section $\sigma_{t\bar{t}} = 13.9^{+6.1}_{-4.8}$ pb. Both CDF[17] and D$\emptyset$[18] announced the definitive top discovery in 1995, reporting mass values of:

$$M_{top} = 176 \pm 8 \pm 10 \text{ GeV/c}^2 \quad \text{(CDF)},$$
$$M_{top} = 199^{+19}_{-21} \pm 22 \text{ GeV/c}^2 \quad (D\emptyset),$$

and production cross sections of:

$$\sigma_{t\bar{t}} = 6.8^{+3.6}_{-2.4}\, pb \quad \text{(CDF)},$$
$$\sigma_{t\bar{t}} = 6.4 \pm 2.2\, pb \quad (D\emptyset).$$

At $M_{top} \sim 175\,\text{GeV/c}^2$, the last (expected) quark is the heaviest known elementary particle. The large mass, similar in order of magnitude to the scale for electroweak symmetry breaking, implies a Yukawa coupling for top of the order of 1, while the same coupling for the electron is a million times smaller. Is this telling us that the top quark is the only "normal" fermion, while all the other fermions "lost" part of their mass in some symmetry breaking mechanism?

On the other hand, Fig. 3 shows the rather simple mass pattern of the presently know quarks and leptons. It has been noted[19] that on the logarithmic scale of Fig. 3, the large top mass is not that exceptional and seems to follow an established pattern.

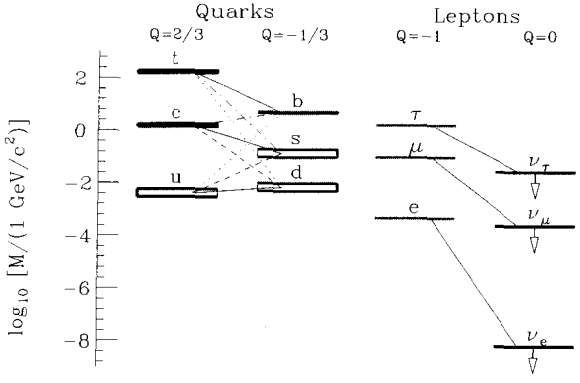


Figure 3: Quark and Lepton mass patterns[19] on a logarithmic scale. The upper-bound on neutrino masses are indicated by the arrow, while the solid and dotted lines correspond to the allowed interquark and interlepton transitions.

In any case, it is clear that due to its large mass the top couples very strongly to the Higgs field, and that the lifetime of the top quark ($\tau_{top} \propto 1/(|V_{tb}|^2 \times M^3) \sim 10^{-24}$ sec) is much shorter than the typical timescale for QCD hadronization ($\tau_{QCD} \propto 1/\Lambda_{QCD} \sim 10^{-23}$ sec). This implies that the top is the only quark that can be studied in a free state, since it decays before hadronizing. Consequently, predictions from perturbative QCD on its production rate are expected to be ratheaccur of the top quark may also imply that the physis in th richness of phenomena observed in the b secor). A to be created mostly from $q\bar{q}$ rather than gg iteract to $\sim$ 90% of the total $\sigma_{t\bar{t}}$. The NLO calculaons[20] where the NLO corrections contribute approximatel contributions of initial and final state radiatioicontri

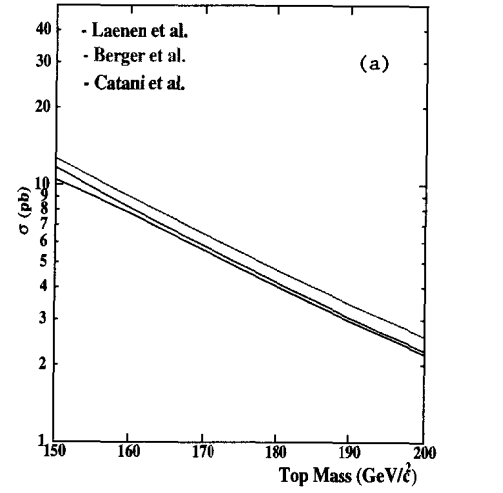


Figure 4: NLO calculation for (a) $\sigma_{t\bar{t}}$ as a fnctior relative magnitude of different production processes

Figure 4(b) gives a simple pictorial view of the p nomena at the Tevatron Collider. As can be sen, the duction is expected to be $\sim$ 3 times lower tha $\sigma_{t\bar{t}}$. subject of study for CDF and D$\emptyset$.

In the SM, assuming V-A coupling and $|V_{tl}| \sim 1$, through the channel $t \to Wb$. The decays $t \to W$ suppressed by factors of $|V_{ts}|^2/|V_{tb}|^2 \sim 10^{-3}$ nd $|V_t$ signature of $t\bar{t}$ events is characterized by the esence the decay modes of the two W's present in thevent. following three channels:

- Dilepton channel (ee, $\mu\mu$, or $e\mu$), corresponding to a branching ratio of $\sim 5\%$ of the total decay.
- Lepton + jets channel (e or μ), corresponding to a branching ratio of $\sim 30\%$.
- Hadronic channel, corresponding to the remaining $\sim 45\%$ of the total decay.

In addition, there are $\sim 21\%$ of the $t\bar{t}$ decays containing a τ lepton in the final state.

3.1 Top Quark Production at the Tevatron Collider

3.1.1 *Dilepton* Channel

Given the process $t\bar{t} \rightarrow W^+W^-b\bar{b} \rightarrow l^+l^-\nu\bar{\nu}b\bar{b}$, the final state for this channel is determined by two oppositely charged leptons with high transverse momenta, large $\not\!\!E_T$, and two b-jets. The dominant backgrounds are WW, $Z \rightarrow \tau^+\tau^-$, and the Drell-Yan production. The dilepton channel is expected to have a very good signal-to-background ratio. However, this decay mode is still limited by statistics, and therefore, not ideal (yet) for a precise measurement of the top mass.

The CDF *dilepton* search starts with the identification of a lepton (e or μ) with $P_T > 20$ GeV/c and satisfying a set of isolation requirements in the cental region ($|\eta| < 1.0$). The second lepton is required to have $P_T > 20$ GeV/c and to satisfy a looser set of isolation requirements. The two leptons must be oppositely charged, and events with ee or $\mu\mu$ candidates where the invariant mass is between 75 and $105\,\mathrm{GeV/c^2}$ are rejected as being consistent with Z^0 candidates.

In order to reject the Drell-Yan events, CDF requires $\not\!\!E_T > 25\,\mathrm{GeV}$. For events with $\not\!\!E_T < 50\,\mathrm{GeV}$, the azimuthal angle between the $\not\!\!E_T$ and the closest lepton or jet must be greater than 20° to reduce the background coming from $Z \rightarrow \tau\tau$ and Drell-Yan events, where a mismeasured jet produces an artificial $\not\!\!E_T$. Finally, all the events are required to have at least two jets with $E_T^{jet} > 10$ GeV and $|\eta| < 2.0$. Background contributions are estimated from a combinations of data and Monte Carlo simulations.

Figure 5(a) shows the distribution of the nine dilepton candidates (one ee, one $\mu\mu$, and seven $e\mu$ events) surviving these cuts in the total integrated luminosity of 110 pb^{-1} at CDF.

The D$\emptyset$ analysis proceeds in a parallel way, the major differences being the cuts on the lepton P_T ($P_T > 15$ (20) GeV/c for the $e\mu$, $\mu\mu$ (ee) channels) and a minimum requirement on H_T ($H_T > 120\,\mathrm{GeV}$ for the "electron" channels, or $H_T > 100\,\mathrm{GeV}$ for the "muon" channel). H_T is defined as the scalar sum of the transverse energies E_T^{jet} of

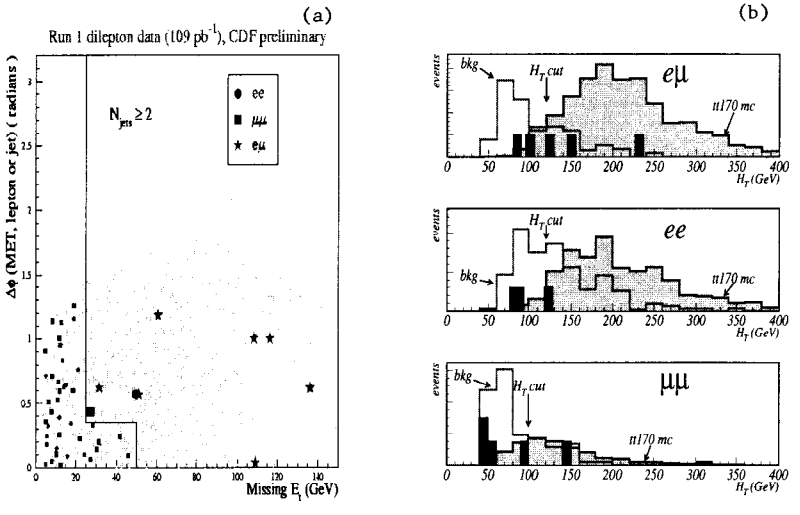


Figure 5: The azimuthal angle $\Delta\phi$ between the $\not\!\!E_T$ vector and the nearest lepton or jet vs $\not\!\!E_T$ in CDF dilepton events. (a) The small points show the distribution expected for the $t\bar{t}$ signal, while the larger symbols represent the data. (b) H_T distributions of dilepton events in the D$\emptyset$ analysis. The observed events are shown by the darker histograms.

Table 1: Summary of the *dilepton* counting experiments.

Sample		DØ	CDF
$e\mu$	Observed	3	7
	Background	0.3 ± 0.1	0.8 ± 0.2
	Expected ($M_{top} \approx 175$ GeV/c^2)	1.7 ± 0.3	2.5 ± 0.2
ee or $\mu\mu$	Observed	2	2
	Background	1.1 ± 0.4	1.6 ± 0.5
	Expected ($M_{top} \approx 175$ GeV/c^2)	1.4 ± 0.1	1.6 ± 0.2
$e\tau$ or $\mu\tau$	Observed	—	4
	Background	—	1.96 ± 0.35
	Expected ($M_{top} \approx 175$ GeV/c^2)	—	0.7 ± 0.1

the jets (for the $\mu\mu$ channels) or the scalar sum of the E_T's of the leading e and the jets (for the $e\mu + jets$ and $ee + jets$ channels). Figure 5(b) shows the discriminating power of the H_T variable for $M_{top} = 200$ GeV/c^2 and the dilepton events surviving the DØ selection.

The tests performed by DØ to understand the behavior of H_T for background events include the comparison between data and Monte Carlo in background-dominated channels such as $e + 2$ jets and $e + 3$ jets, where the H_T distribution agrees well with background calculations based on the VECBOS Monte Carlo and multijet events. DØ observes three $e\mu$, one ee, and one $\mu\mu$ candidate.

In both experiments, the acceptance is much higher for the $e\mu$ than for the ee and $\mu\mu$ channels because of the cut to reject Z^0 candidates. Table 1 summarizes the *dilepton* counting experiments results.

The production cross section is determined through the standard formula:

$$\sigma_{t\bar{t}} = \frac{N - B}{\epsilon \int Ldt}, \quad (1)$$

where N is the observed number of events, B is the expected background (determined from a combination of data and Monte Carlo), ϵ is the total efficiency, and $\int Ldt$ is the integrated luminosity. The efficiency depends on the mass of the top quark, and the two experiments determined the efficiencies for the top quark mass corresponding to the respective direct mass measurement (~ 175 GeV/c^2 for CDF and ~ 172 GeV/c^2 for DØ). The values determined by the two exprimen

$$\sigma_{t\bar{t}} = 8.2^{+4.4}_{-3.4}\, b \quad (($$
$$\sigma_{t\bar{t}} = 5.0 \pm 3.\, pb \quad (1$$

3.1.2 $e\nu$ Channel

A top quark decay channel recently explore by D on the requirement of one isolated electror $\not\!\!E_T$ with transverse momentum greater than 20 GV. A mass of the electron and $\not\!\!E_T$ is designed to rect th $M_T(e\nu) > 150$ GeV/c^2. This inclusive selecon, wl subdivision of the events into *dilepton* and epton increase the top acceptance. With the selectio cuts ple is expected to contain 50% dilepton decys, 3 of $e\tau$ decays. Figure 6 shows the data distriution data distributions for comparison. DØ finds our ev of 1.2 ± 0.4 events, and measures[24] a cross setion c

$$\sigma_{t\bar{t}} = 9.6 \pm 7.5\, pb \quad (D\not{O}$$

When combined with the previous dilepton measure

$$\sigma_{t\bar{t}} = 6.4 \pm 4.4\, pb \quad (L\not{O}\ dil$$

3.1.3 $Lepton + Jets$ Channel

Given the process $t\bar{t} \to W^+W^-b\bar{b} \to l\nu qq\bar{b}$, the l lepton with high transverse momentum, large $\not\!\!E_T$, tw and two b-jets. The set of cuts used in the to exp the signal region of $W+ \geq 3$ jets, the cuts ued by expected $t\bar{t}$ signal (70% in the case of a 175 eV/c^2 QCD $W+$ multijet background.

However, after the removal of Z^0 candidats, the level of 1/6 (for example, CDF is left with ~30 eve events). The background contributions are oming production ($\sim$ 160 expected events) with smll con

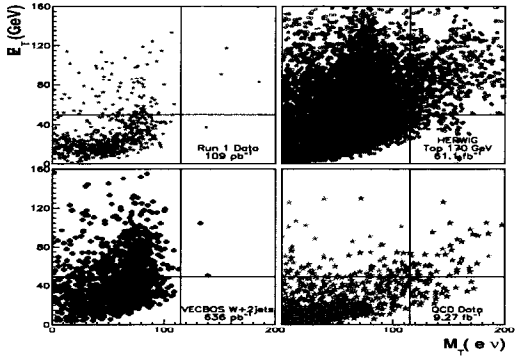


Figure 6: Aplanarity vs H_T for the $e\nu$ analysis in D$\emptyset$ for data, $t\bar{t}$ Monte Carlo events, multijet background, and $W+$ 4 jets Monte Carlo background. The dashed lines indicate the cuts.

$b\bar{b}$ jets ($\sim$ 30 events), $Z \rightarrow ll$ ($\sim$ 26 events), and diboson production ($\sim$ 15 events). Clearly, additional background rejection is required. D$\emptyset$ and CDF solve the problem of the rejection of W+jets QCD backgrounds with three different approaches:

- Event Shape analysis[24] (D$\emptyset$). This approach relies on the fact that, for heavy top, the overall event is different (more *spherical*, and with more energy) than normal QCD $W+$ multijet events. The variables used by D$\emptyset$ to discriminate against background include the event aplanarity $\mathcal{A}$ and the already-mentioned H_T. No attempt is made at the identification of the original flavor of the jets in the event.
- b-quark tagging through the semileptonic decay (CDF[25] and D$\emptyset$[24]). This approach identifies the b-nature of the jets present in the event through the presence of a soft lepton embedded in the jets and originated by a semileptonic decay of the parent b quark (SLT, or Soft Lepton Tagging).
- b-quark tagging by mean of displaced vertices or displaced tracks (CDF[25]). This approach relies on the finite b-quark lifetime and the superb precision of the Silicon Vertex (SVX) detector to identify displaced vertices (SECVTX tagging) or displaced tracks (JPB tagging).

Table 2: Cuts used by CDF and D$\emptyset$ for the *lepton + jets* search.

	CDF	D$\emptyset$	
	All	$e(\mu)$	e + jets/μ
P_T^l	20 GeV	20 GeV	20 GeV
$\not\!\!E_T$	20 GeV	25(20) GeV	20 GeV
Number of Jets	$\geq 3, \mid \eta \mid < 2$	$\geq 4, \mid \eta \mid < 2.5$	$\geq 4, \mid \eta \mid < 2.5$
E_T^{jet}	$\geq$ 15 GeV	$\geq$ 15 GeV	$\geq$ 20 GeV
$\mathcal{A}$	-	> 0.065	> 0.040
H_T	-	$>$ 180 GeV	$>$ 110 GeV

The typical efficiencies per b-quark jet for the three tagging tools are ~23% (SECVTX), ~22% (JPB), and ~7% (SLT). Clearly, due to the different lifetime and decay kinematics, the efficiencies of these tagging tools for c-quark jets are different [~4% (SECVTX), ~9% (JPB), and ~4% (SLT), respectively] and can be used to select samples of data with different c-quark and b-quark compositions.

Figure 7 shows the number of observed $W+$ multijet events after the SVX and SLT tagging algorithms in CDF. The background expectations are based on a mistag rate measured in inclusive jets, while the fraction of $W+$ multijet events due to $Wb\bar{b}$, $Wc\bar{c}$, or Wc are estimated from the Monte Carlo using the measured tagging efficiency. In the $W + 1$ *jet* bin, which is expected to be completely populated by background, the observed data are in good agreement with expectations.

The excess of events in the $W+$ 3- and $W+$ 4-jet bins are attributed completely to top production.

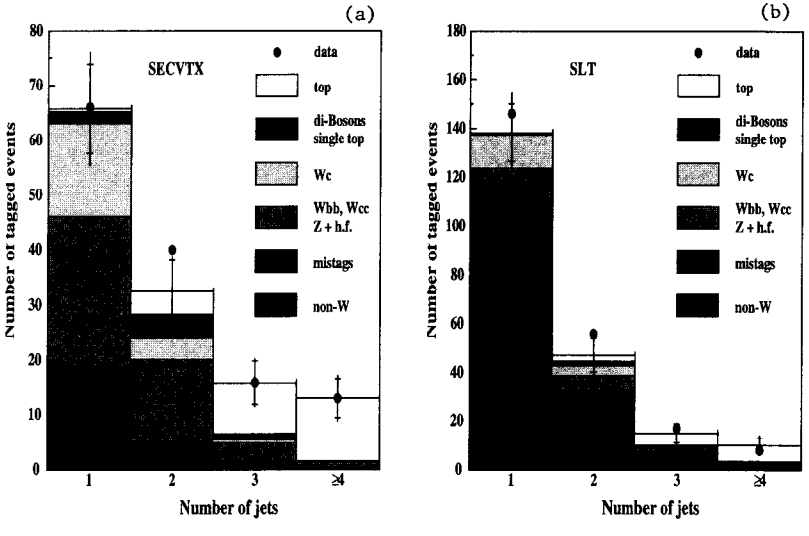


Figure 7: The CDF jet multiplicity distribution of (a) SVX tagged and (b) SLT tagged lepton + jets events.

Figure 8 shows the D$\emptyset$ distribution of events in $\mathcal{A}$ and H_T variables for the data and various Monte Carlo simulations.

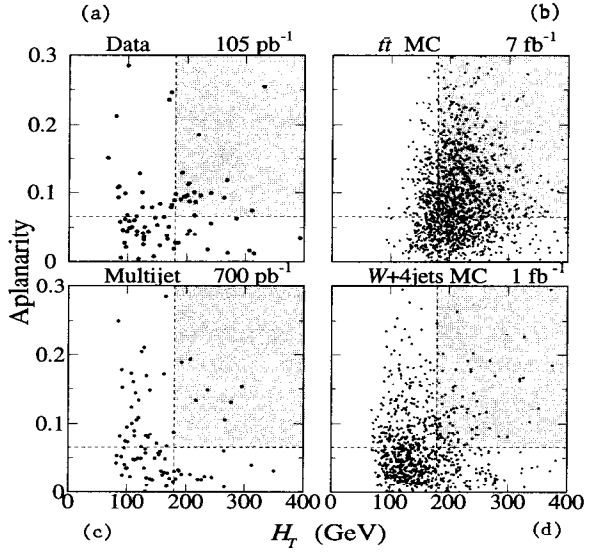


Figure 8: Aplanarity vs H_T for the D$\emptyset$ lepton + jets events for (a) data, (b) $t\bar{t}$ Monte Carlo, (c) multijet background, and (d) W + jets VECBOS Monte Carlo.

Table 3 gives the total number of observed events and the expected background and signals for the two experiments. The cross sections measured by CDF and D$\emptyset$ are:[24,25]

$$\sigma_{t\bar{t}} = 6.2^{+2.1}_{-1.7}\, pb \quad (\mathrm{CDF} - \mathrm{SVX})$$
$$\sigma_{t\bar{t}} = 9.2^{+4.3}_{-3.6}\, pb \quad (\mathrm{CDF} - \mathrm{SLT})$$
$$\sigma_{t\bar{t}} = 4.1 \pm 2.1\, pb \quad (\mathrm{D}\emptyset - \text{Event Shape})$$
$$\sigma_{t\bar{t}} = 8.3 \pm 3.5\, pb \quad (\mathrm{D}\emptyset - \mathrm{SLT}).$$

3.1.4 Hadronic Top Search

The final state for the full hadronic decay of the top consists of six jets, two of them being b-quark jets. The event selection in CDF starts with the requirement of ≥ 5 jets with $E_T > 15$ GeV and $|\eta| < 2.0$ The background, coming predominantly from QCD multi-jet events, is large even after tagging of the b-quarks as shown from the Monte Carlo predictions in Table 4.

In order to further reduce the background, CDF applies a kinematical selection[26]

Table 3: Summary of the $lepton + jets$ counting experiments.

Sample		DØ	CDF
Event Shape	Observed	19	22
	Background	9.7 ± 1.7	7.2 ± 2.1
	Expected ($M_{top} \sim 170$ GeV)	14.1 ± 3.1	—
$b \to lX$	Observed	11	40
	Background	2.4 ± 0.5	24.3 ± 3.5
	Expected ($M_{top} \sim 170$ GeV)	5.8 ± 1.0	9.6
Displaced Vertex	Observed	—	34 Events (42 Tags)
	Background	—	8.4 ± 1.4
	Expected ($M_{top} \sim 170$ GeV)	—	19.8 ± 4.0

Table 4: Predicted signal and backgrounds in the CDF Hadronic Top search.

$\frac{S}{S+B}$	≥ 5 $jets$	≥ 6 $jets$
No tag	1/500	1/200
$1\,b-tag$	1/100	1/30
$2\,b-tag$	1/20	1/10

which includes requiring the $\Sigma E_T \geq 300$ GeV and $C > 0.75$, where the centrality C is defined as $C = H_T/\sqrt{\hat{s}}$ and $\hat{s}$ is the invariant mass of the multijet system. After these kinematical cuts a requirement of one b-tagged jet produces the multiplicity distribution shown in Fig. 9(a) which shows a clear excess of events over the expected background from QCD contributions.

A similar analysis from DØ[27] uses a selection based on the discriminating power of a neural network fed with 14 variables which include kinematical quantities (like aplanarity and total scalar energy) as well as soft lepton tagging of b-quark jets. The output of the neural network is shown in Fig. 9(b).

The cross section values obtained by the two collaborations for the all-hadronic decay modes are:

$$\sigma_{t\bar{t}} = 10.1^{+4.5}_{-3.6}\,pb \quad (\text{CDF})$$
$$\sigma_{t\bar{t}} = 7.1 \pm 3.2\,pb \quad (D\emptyset\,).$$

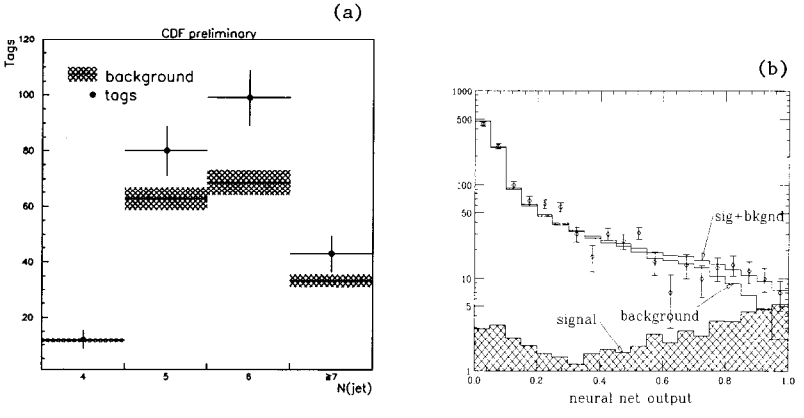


Figure 9: (a) CDF multiplicity distribution for *all hadronic* analysis with the tagged events and the background expectation. (b) Distribution of the final neural network output in the DØ *all hadronic* analysis, showing the results of a fit of the observed tagged distribution to the predictions for signal and background, with $M_{top} = 180$ GeV/c^2.

3.1.5 Summary of the $\sigma_{t\bar{t}}$ Cross-Section Measurements

The summary of the $\sigma_{t\bar{t}}$ cross-section measurements performed at the Tevatron is shown in Fig. 10. The DØ and CDF measurements are in good agreement with the average[28] of $\sigma_{t\bar{t}} = 6.7 \pm 1.3\,pb$, which is slightly larger than, but still in good agreement with, the theoretical SM predictions of Fig. 4.

3.2 Top Mass Determination

A precise measurement of the top quark mass plays a central role in our understanding of the mechanism for the symmetry breaking in the SM. A precise direct measurement can provide a consistency check of the experimental data from different sources, and a combination of the top and W mass measurements provides information on the mass of the Higgs boson.

As for the cross-section measurement, the top mass determination can be performed in any one of the three different decay topologies of the $t\bar{t}$ event (*dilepton*, *lepton + jets*, and *all hadronic*). At the time of the top discovery, the *lepton + jets* channels were studied more extensively given the relative large signal-to-background ratio,

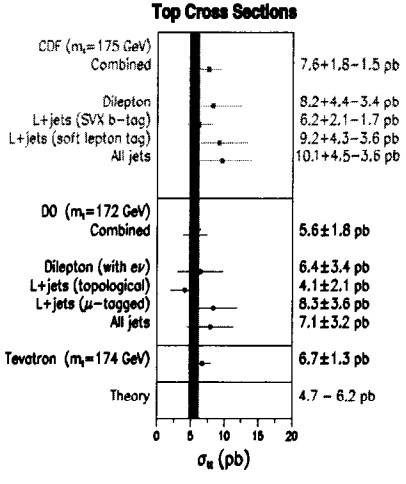


Figure 10: The $t\bar{t}$ production cross section measured in the channels studied by CDF and D$\emptyset$. Also shown is the range of theoretical predictions.

the large branching ratio, and the possibility of performing a constrained fit to the top mass hypothesis. As a result, in 1995 the top mass was known with an uncertainty of $\sim$ 20 GeV/c^2. Since then both collaborations improved their understanding of systematic errors and the analysis of the *dilepton* and *all hadronic* decay channels, developing tools which are allowing the knowledge of the top quark mass with a very good precision ($\sim$ 5 GeV/c^2 i.e., $\sim$ 3%). The following sections will describe the top mass measurements by D$\emptyset$ and CDF, as well as the average mass determined by the two experiments.

3.2.1 Lepton + Jets Channel

The hypothesis under study includes:

- $p\bar{p} \rightarrow t_1\overline{t_2}$ + anything,
- $t_1 \rightarrow Wb \rightarrow l\nu j_1$, and
- $t_2 \rightarrow Wb \rightarrow j_2 j_3 j_4$.

The use of known masses and energy-momentum conservation gives a 2C fit since there are 52 variables, 34 measured (or known) quantities, and 20 equations of conservation (the mass constraints being $M(l\nu) = M_W^{PDG}$, $M(j_2 j_3) = M_W^{PDG}$, and $M_{t_1} = M_{t_2}$).

Clearly, the method is usable only if all four jets are observed in a top decay. In CDF,[29] the event selection starts with the usual lepton ($P_T^l > 20$ GeV/c) and $\not\!\!E_T$ requirements ($\not\!\!E_T > 20$ GeV). At least four jets are required in each event, three of which must have an observed $E_T \geq 15$ GeV and $|\eta| \leq 2$. In order to increase the acceptance, the requirements on the fourth jet is relaxed to be $E_T > 8$ GeV and $|\eta| < 2.4$, provided one of four leading jets is tagged by the SVX or SLT algorithms. SVX tags are only allowed on the three leading jets ($E_T > 15$ GeV), while SLT tags are allowed on all the jets. If no such tag is present, the fourth jet must satisfy the same E_T and η requirements as the first three. The above selection defines a mass sample containing 83 events. Following the 2C fitting scheme described above, CDF requires that SVX and SLT-tagged jets are assigned to b-partons and chooses the configuration with lowest χ^2 out of the many possible combinatorial combinations. Events with $\chi^2 > 10$ are rejected. The precision of the top quark mass measurement is expected to increase with the number of observed events, the signal-to-background ratio, and the narrowness of the reconstructed-mass distribution. Monte Carlo studies demonstrated that the optimal way to partition the sample consists of subdividing the events into the four statistically independent subsamples shown in Fig. 11(a). In each distribution, the inset shows the shape of the log-likelihood function. A combination of the four different samples is shown in Fig. 11(b).

Table 5: Summary of systematic uncertainties in the CDF top mass measurements.

Source	Dilepton	*Lepton + jets*	All Hadronic
Jet E_T	3.8	4.4	5.0
Gluon Radiation	3.1	2.6	1.8
Background	0.3	1.3	1.7
PDF, Monte Carlo	1.1	0.5	1.0
Total	4.8	5.3	5.7

From this, CDF measures $M_{top} = 175.9 \pm 4.8$ GeV/c^2, where the uncertainty corresponds to a half-unit change in the negative log-likelihood with respect to its minimum.

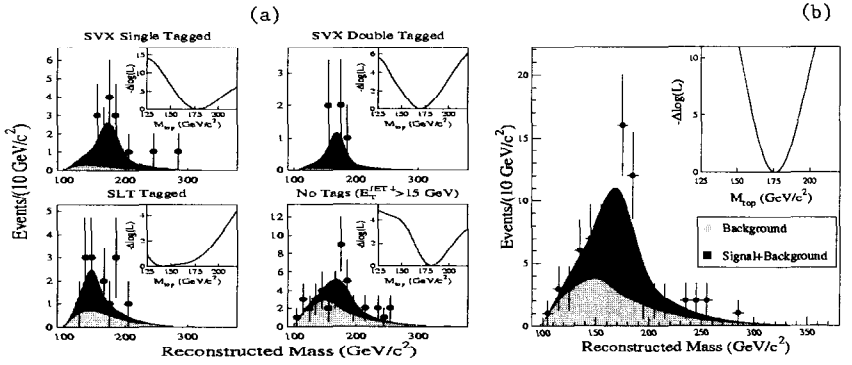


Figure 11: CDF top quark mass distributions for the *lepton + jets* sample. The points are the data, the dark area is the top signal+background resulting from the fit, while the lightly shaded area is the background alone. The plots on (a) show the four independent samples, while (b) shows the combination of the four samples.

The CDF systematic uncertainties for the various $t\bar{t}$ decay modes considered here are listed in Table 5. The final measurement is

$$M_{top} = 175.9 \pm 4.8(stat.) \pm 4.9(syst.)\ \mathrm{GeV/c^2} \quad (\mathrm{CDF} - lepton + jets).$$

D$\emptyset$ performs a two-dimensional likelihood fit to extract the top mass value. One variable in the two-dimensional distribution is the best fit mass obtained by the 2C analysis of the data. The other variable is a top discriminant, which provides a distinct separation between the top signal and the background, without biasing the mass analysis. D$\emptyset$ uses two discriminants[30] based on the following four variable:

- $\not\!\!E_T$
- $\mathcal{A}$
- $H_{T2}/\Sigma|p_z|$, where H_{T2} is defined as the H_T minus the E_T of the leading jet. This variable measures the centrality of the event.
- $(\Delta R_{jj}^{min})E_T^{min}/(E_T^l + \not\!\!E_T)$, where (ΔR_{jj}^{min}) is the minimum ΔR between all pairs of jets and E_T^{min} is the smaller jet E_T from the minimum ΔR pair. This variable measures the extent to which the jets are clustered together.

These variables are combined in a Neural Network (NN) and a Low Bias (LB) discriminant to provide the kind of separation illustrated in Fig. 12(a). D$\emptyset$ then fits the two-dimensional distributions to templates determined from simulated $t\bar{t}$ events and background estimated using a combination of Monte Carlo and data. The experiment obtains a mass measurement of $M_{top} = 173.3 \pm 5.6(stat.)\,\mathrm{GeV/c^2}$ shown in Fig. 12(b). The systematic effects, coming mostly from the jet energy scale and the Monte Carlo modeling, sum to $\pm 5.5\ \mathrm{GeV/c^2}$. The final mass obtained by D$\emptyset$ in the *lepton + jets* channel is:

$$M_{top} = 173.3 \pm 5.6(stat.) \pm 5.5(syst.)\,\mathrm{GeV/c^2} \quad (D\emptyset - lepton + jets).$$

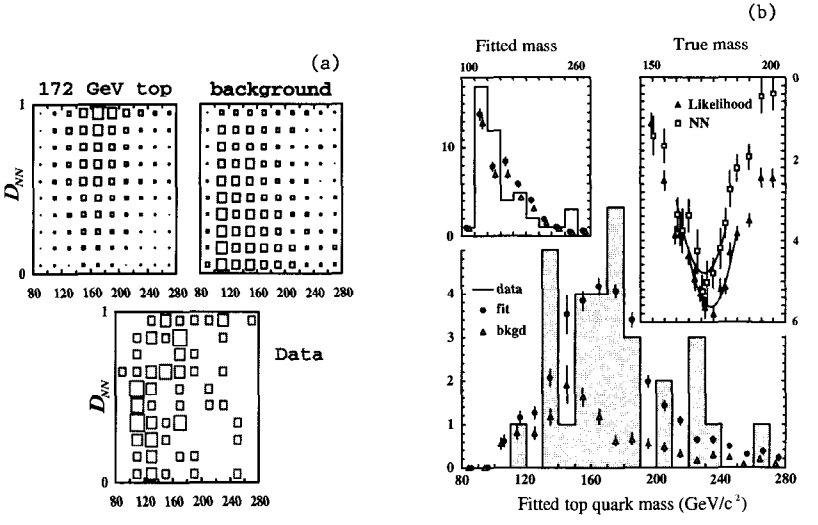


Figure 12: (a) Events per bin in the $\mathcal{D}_{NN}, m_{fit}$ plane for the D$\emptyset$ neural network discriminant analysis, showing the expectation for top, background, and data. (b) Results from the D$\emptyset$ *lepton + jets* mass analysis with the $\mathcal{D}_{LB}$ discriminant for events poor in top signal, rich in top signal, and the final log-likelihood distribution.

3.2.2 Dilepton Top Mass Measurement

Due to the presence of two neutrinos, dilepton events do not contain enough information for a constrained fit. Therefore, to determine the top mass, one must use some other

mass estimators. One possibility is the determination of the top mass through the study of kinematic variables which have a mass dependence, like the b-jet energy spectrum ($< E_b > \propto M_{top}$) or the invariant mass of a lepton and b-jet coming from the same top quark. These methods were originally used by CDF[23] but are limited by a rather large systematic error ($\sigma_{M_{top}} \sim 10\ \mathrm{GeV/c^2}$).

D$\emptyset$ developed a method similar to the one used in the *lepton + jets* analysis. The missing constraint in the dilepton events is provided by assuming a top quark mass and reconstructing the event for every assumed top mass.[31] Then a weight is computed which characterizes the probability for the event to be from a $t\bar{t}$ decay with the assumed mass. D$\emptyset$ developed two algorithms to determine the weight. The matrix element weighting (MTW) uses the proton structure function and the probability density function for the energy of the decay lepton in the rest frame of the top quark. The neutrino weighting method (νWT) assigns the weight based on the available phase space for the neutrinos, consistent with the measured $\not\!\!E_T$. A maximum likelihood fit is then performed to shape the weight distribution, using the Monte Carlo derived probability density function for signal and background, as shown in Fig. 13.

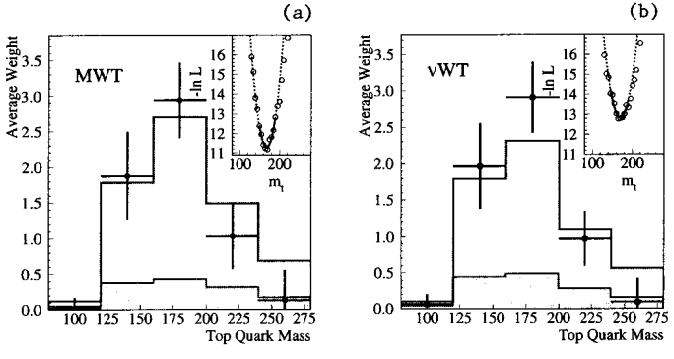


Figure 13: Sum of the normalized candidate likelihoods for (a) MWT and (b) νWT analyses, compared to the results of the likelihood fit. The solid histogram is the best-fit $t\bar{t}$ signal plus background, and the dotted histogram is the background alone.

By using the full information of the event, the total systematic error on the measured mass is lower ($\sim 4\ \mathrm{GeV/c^2}$) and the precision of the measurement is limited by the available statistics. D$\emptyset$ determined

$$M_{top} = 168.4 \pm 12.3(stat.) \pm 3.6(syst.)\ \mathrm{GeV/c^2} \quad (D\emptyset - dilepton).$$

CDF applied a similar technique, assuming atop qu rapidity (η) values, to solve for the neutrino momenta to each solution by comparing the sum of the neutrino solution to the measured $\not\!\!E_T$. For each choice o the to (100) pairs of (η_1, η_2) combinations are summed to gen particular mass. The mass is then varied, to generate th Fig. 14(a) for each event. From each distribution, an a the averages are then fit together with the Monte Carl and top signal as shown in Fig. 14(b). Includingthe sy Table 5, the top quark mass is measured[29] to be:

$$M_{top} = 167.4 \pm 10.3(stat.) \pm 4.8(syst GeV/$$

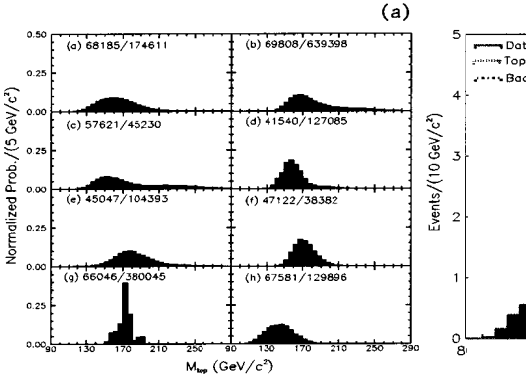


Figure 14: Weight distributions normalized to unity as eight CDF dilepton candidates, and (b) reconstucted t events. The data are shown by the solid line, the backgr area and the top Monte Carlo plus the background conti

3.2.3 Hadronic Top Mass

CDF performed a mass analysis of the $t\bar{t}$ event in wh quark-antiquark pairs. The study of this channe, with complements the leptonic modes, and the mass measur reconstructed final states, but it suffers from very large

reduce this background, events with at least one identified SVX b-jet are required to pass strict kinematic criteria that favor $t\bar{t}$ production and decay. To determine the top quark mass, full kinematic reconstruction is applied to the sample of events with six or more jets. All combinations are tried, with the constraint that an SVX-tagged jet must be assigned to a b-parton. The data sample consists of 136 events, of which 108 ± 9 events are expected to come from background. The reconstructed three-jet mass distribution is shown in Fig. 15. The inset shows the shape of the difference log-likelihood as a function of top mass. With the systematic uncertainty shown in Table 5, the CDF measurement[29] is:

$$M_{top} = 186.0 \pm 10.0(stat.) \pm 5.7(syst.)\ \mathrm{GeV/c^2} \quad (\mathrm{CDF} - \text{all hadronic}).$$

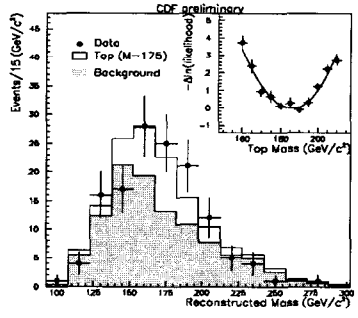


Figure 15: Reconstructed mass distribution for *all hadronic* events with at least one b-tag. Also shown are the background distribution (shaded) and the $t\bar{t}$ Monte Carlo events added to background (hollow). The inset shows the log-likelihood and the fit used to determine the top mass.

3.2.4 Top Mass Summary

Figure 16(a) shows the summary of the direct measurements of the top quark mass at the Tevatron collider. When the appropriate correlations are taken into account between the mass systematic errors in CDF and D$\emptyset$, the world average is determined to be

$$M_{top} = 174.3 \pm 5.1\ \mathrm{GeV/c^2} \quad (\text{Tevatron average}).$$

Figure 16(b) shows the relative weight of the various mass determinations on the final Tevatron average. When the top and W boson masses are interpreted in the framework of the SM, Fig. 17 shows the correlation between the top and the W masses as a function of the Higgs mass. With the latest values of the W and top masses and the LEP measurements[32] the data seems to favor a light Higgs, with $m_H < 260\ \mathrm{GeV/c^2}$ at 95% C.L.

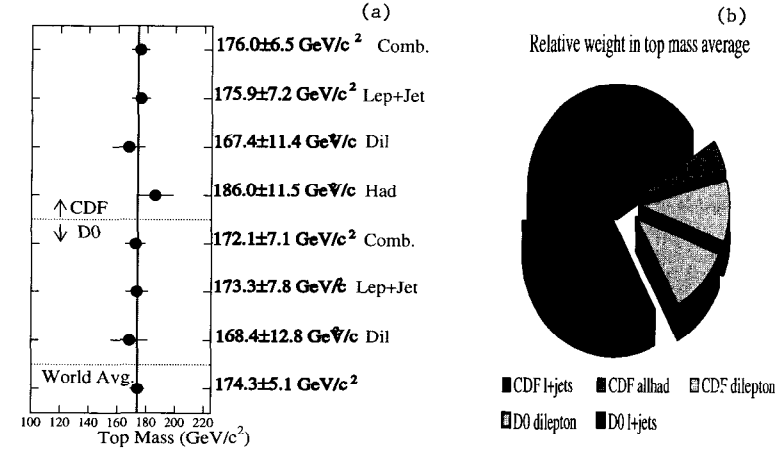


Figure 16: Summary of all the top quark mass measurements from (a) CDF and D$\emptyset$, and (b) relative weight of the various single measurements in the overall average.

3.3 Other Top Quark Measurements

3.3.1 Measurement of $|V_{tb}|$

In the previous discussions, $|V_{tb}|$ has been assumed to be approximately equal to 1. From the knowledge of the tagging efficiencies, the number of dilepton, and *lepton + jets* events with one, two, or no jets tagged as b quark, the following ratio can be derived:

$$R = \frac{B(t\to Wb)}{B(t\to Wq)} = \frac{|V_{tb}|^2}{|V_{tb}|^2+|V_{td}|^2+|V_{ts}|^2}.$$

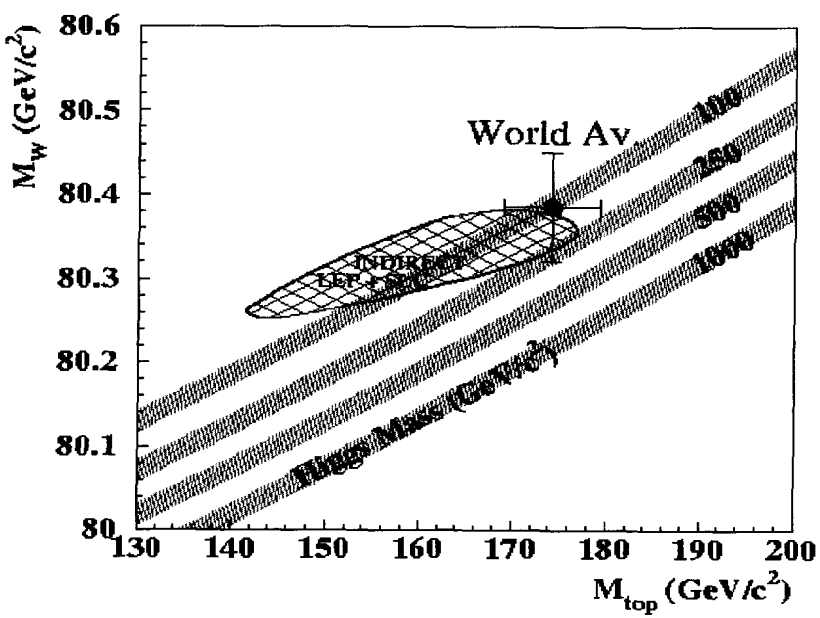


Figure 17: Relation between the top quark mass and the W boson mass. The dependence on the Higgs mass is shown through the different bands for several assumptions of m_H.

CDF has determined[33] $R = 0.99 \pm 0.29$, where the uncertainty is dominated by the statistical component. If three-generation unitarity is assumed, then $|V_{tb}| = 0.99 \pm 0.15$ and $|V_{tb}| \geq 0.76$ at 95% C.L. If the unitarity condition is relaxed, then by setting $|V_{td}| = 0.009$ and $|V_{ts}| = 0.04$ (midpoint of their 90 % C.L. determined with the unitarity assumption) CDF determines $|V_{tb}| \geq 0.048$ at 95 % C.L.

3.3.2 Flavor-Changing Neutral Current (FCNC) Decays of the Top Quark

FCNC decays can be used to probe for new physics at mass scales which are otherwise not accessible to present-day experiments. A typical historical example is the absence of the FCNC decay $K_L^0 \to \mu^+\mu^-$ which was indicative of the existence of the charm quark, a state much heavier than the kaon.[34] The SM expectations for FCNC decays of the top quark are extremely small, and therefore, any observation of these modes would be evidence for physics beyond the SM.

CDF searched for $t \to q\gamma$ and $t \to qZ$. A single event is observed in the $q\gamma$ mode. Without any subtraction of the expected background, CDF finds

$$B(t \to u\gamma) + B(t \to c\gamma) < 0.032.$$

Similarly, in the $t \to qZ$ channels, with $Z \to \mu^+\mu^-$ or $Z \to e^+e^-$, CDF finds one candidate, and with no background subtraction the branching fraction limit is

$$B(t \to uZ) + B(t \to cZ) < 0.33.$$

This search is less sensitive than the $t \to q\gamma$ search because of the small branching fraction of the Z boson into charged leptons.

3.3.3 W Boson Helicity in Top Quark Decays

The W boson from the top decay can be polarized either transversely or longitudinally. The SM prediction for the fraction of longitudinally polarized bosons is $F_0 = M_{top}^2/(2M_W^2 + M_{top}^2) \sim 0.70$ for $M_{top} = 174$ GeV/c^2. A measurement of F_0 can be extracted from the P_T distribution of the observed leptons, since transversely polarized W bosons emit the charged lepton in a direction preferentially antiparallel to the direction of the boost from the top quark rest frame to the W boson rest frame, while longitudinally polarized W's emit the charged lepton perpendicular to the boost direction. Fitting simultaneously the lepton P_T spectrum in the dilepton and lepton+jets

events (Fig. 18), CDF[35] finds

$$F_0 = 0.55 \pm 0.32(stat.) \pm 0.12(syst.).$$

4 b Quark at the Tevatron

The principal motivation for studying b-quark physics in the context of the SM arises from the possibility of gathering valuable information on the CKM matrix elements. In fact, a study of b decays allows access to five of the nine CKM elements (V_{cb}, V_{ub}, V_{td}, V_{ts}, and V_{tb}), some of which (V_{td}, V_{ts}) would be very hard to study in decays of the top quark system.

Traditionally, b-quark physics has been the domain of e^+e^- machines. However, UA1 already demonstrated the possibility of studying b physics at a hadron collider. CDF, with a superb mass resolution and vertex detection capabilities, has really expanded the b-physics program achievable at a hadron collider. The D$\emptyset$ experiment has also published several b-physics results,[36] but due to the lack of precision momentum measurement of charged particles and the absence of a precision microvertex detector, D$\emptyset$ is not ideally suited to study the b sector with the same broad coverage.

A hadron machine has several advantages (and some disadvantages) compared to an e^+e^- machine at the $\Upsilon(4S)$. All species of B hadrons can be produced at the Tevatron Collider (B^+, B^0, B_s^0, B_c, and Λ_b), with a large production cross section ($\sigma_b \sim 50\ \mu b$, while $\sigma_{\Upsilon(4S)} \sim 1$ nb and $\sigma_{Z^0 \to b\bar{b}} \sim 7$ nb at LEP). This very large cross section results in about $5 \times 10^9\ b\bar{b}$ pairs produced during Run I at the Tevatron detectors. Unfortunately, the inelastic cross section is three orders of magnitude larger, which puts very specific requirements on the trigger system designed to recognize b-hadrons for further processing. Moreover, the b-quark production cross section drops almost exponentially with the transverse momentum of the produced b quark. This puts the trigger threshold for b-physics events on a collision course with the experiment's data acquisition (DAQ) bandwidth.

All b physics triggers at CDF and D$\emptyset$ are based on leptons, with the possibility of requiring both single leptons and dileptons events. As an example, CDF dilepton trigger consists of a dimuon trigger with $P_T > 2$ GeV/c for both muon legs, and an $e\mu$ trigger with $P_T^\mu > 3$ GeV/c and $E_t^e > 5$ GeV. The dimuon trigger is the source of the J/ψ sample, and both dilepton triggers are used for b mixing analysis. The thresholds for single-lepton triggers are higher with $P_T > 7.5$ GeV/c for muons and $E_T > 8$ GeV

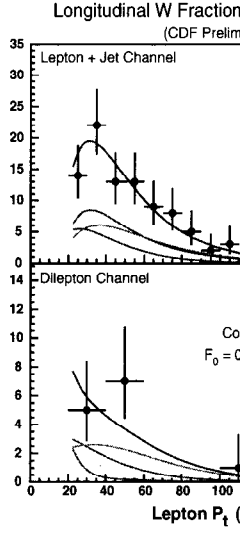


Figure 18: The CDF lepton P_T spectra for the sum of longitudinal W boson decays, transverse background.

for electrons. Analyses involving semileptonic decays are based on these single lepton triggers.

The publications from CDF and DØ on the field of b-physics have been numerous. In the last year alone (1998), many papers have documented the results from the two experiments on b-quark production,[37] lifetimes,[38] rare decays,[39,40] and B^0–$\overline{B^0}$ mixing.[41]

In Sec. 4.1, we will concentrate on two recent results from CDF; the discovery of the B_c meson, and the first measurement at a hadronic collider of the time-dependent asymmetry in the decay $B^0, \overline{B^0} \to J/\psi K^0_S$ with a low-statistics determination of the CP-violation parameter $\sin(2\beta)$.

4.1 Discovery of the B_c

During recent decades, quark spectroscopy has evolved in the same manner as atom and isotope spectroscopy did during the first half of this century. Just as the periodic table was built, one can build a quark periodic table also, as shown in Fig. 19(a).

One of the last missing items in the quark periodic table was the B_c, a very tight bound state of the two heaviest quarks (b and c) with a lifetime long enough to allow hadronization into a bound system. Nonrelativistic QCD potential models are expected to give a reliable description of an interesting spectroscopy with many states below the level for direct B–D production, as shown in Fig. 19(b). The same models predict a B_c mass between 6.2 and 6.3 GeV/c^2, and the lifetime is between 0.4 and 1.4 ps. Although significantly shorter than those of other B mesons, the B_c lifetime is expected to be measurable with the CDF SVX detector. Fragmentation models predict that the B_c production is suppressed by $\simeq 10^{-3}$ with respect to the production of B_d and B_u mesons.

Many exclusive B_c decay modes have been explored by the CDF and LEP experiments.[42] Although a few candidates were found, their number was so low that no clear claim could be put forward. For this search, CDF[43] investigated the semileptonic decay with a J/ψ in the final state, as shown in Fig. 20(a). The J/ψ is reconstructed through the decay $J/\psi \to \mu^-\mu^+$ which is required at trigger level.

After the J/ψ candidate's reconstruction, shown in Fig. 20(a), the two muons and the third lepton are required to come from a common displaced secondary vertex. Due to the presence of the undetected ν in the final state, the only measurable quantities are the mass and the pseudolifetime of the trilepton system. The pseudolifetime ct^* is

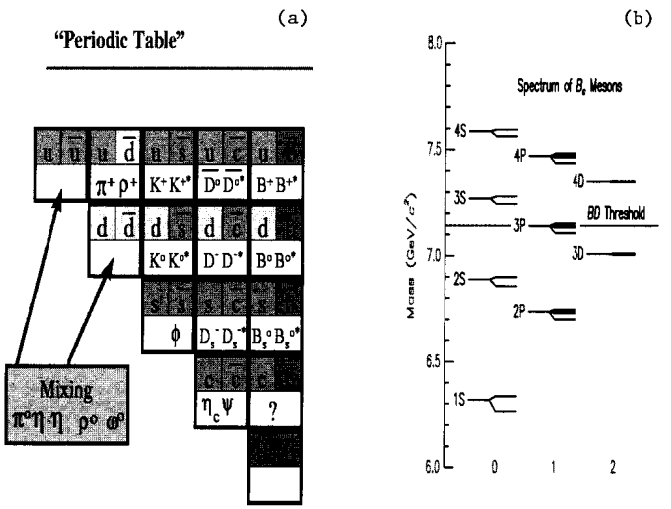


Figure 19: (a) Quark Periodic Table, and (b) level expectations for the bound states of b and c quarks.

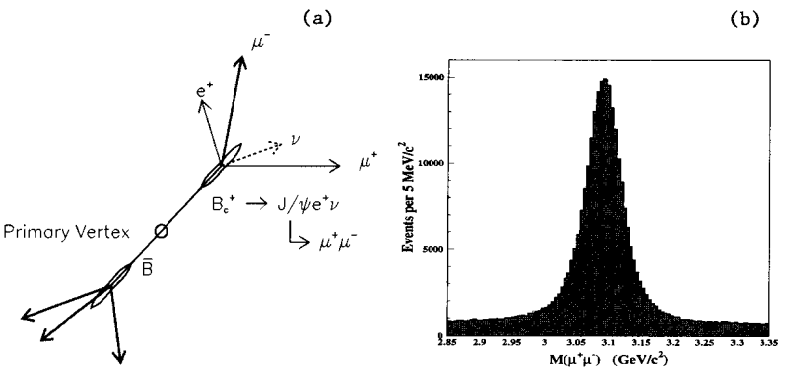


Figure 20: (a) Decay pattern for $B_c \to J/\psi l X$, and (b) invariant mass of $\mu^+\mu^-$ pairs showing the J/ψ candidates.

defined as

$$ct^* = \frac{L_{xy} \times M(J/\psi l)}{P_T(J/\psi l)},$$

where L_{xy} is the distance between the reconstructed decay vertex and the average beam position in the transverse plane.

From Monte Carlo studies (where M_{B_c} is set to 6.27 GeV/c^2) most signal events are expected to have $4 \leq M(J/\psi l) \leq 6$ GeV/c^2. To select possible decays of long lived particles, CDF requires $ct^* > 60\mu$m. This cut is removed later for the lifetime measurement.

Starting with a sample of 196,000 J/ψ reconstructed in the SVX and then rejecting candidates compatible with being $B^+ \to J/\psi K^+$ decays, and events where the electron was identified as coming from a photon conversion, CDF finds a sample of 31 $J/\psi l$ candidates (19 $J/\psi e$, and 12 $J/\psi \mu$). The main sources of background are expected to be due to real J/ψ which form a good displaced vertex when paired with a hadron misidentified as a third lepton, and to $b\bar{b}$ events with one B hadron decaying to a J/ψ and the other B hadron decaying semileptonically, with a topology compatible with having the J/ψ and the lepton in a common vertex. Hadrons misidentified as the third lepton are found to be the main source of background. For muons, this is due to light hadrons (pions or kaons) which punch through the calorimeter and are then detected in the muon chambers, or decays in flight producing a muon with a kink small enough to be well-linked to the track of the hadron. For electrons, this happens when the shower of a hadron in the electromagnetic calorimeter is indistinguishable from that of an electron. The contribution of these sources is estimated from a J/ψ + track sample obtained by releasing the lepton identification criteria on the third track. This sample is then weighted with the probability, estimated from real data as a function of P_T, that a hadron is misidentified as a lepton. With this method, CDF also obtains the mass shape of the background. Real $J/\psi l$ background from $b\bar{b}$ events is estimated from a Monte Carlo simulation.

A summary of all the background sources and the estimated signals in both channels is given in Table 6.

The number of B_c mesons and the statistical significance of the excess is also estimated from a likelihood fit of the $J/\psi l$ mass distribution, as shown in Fig. 21(a). The mass shapes for the signal and background are constrained to the results of signal simulation and background measurement, respectively. The only free parameter returned by the fit is the number of B_c mesons, $N(B_c) = 20.4^{+6.2}_{-5.5}$. The null hypothesis (i.e., the

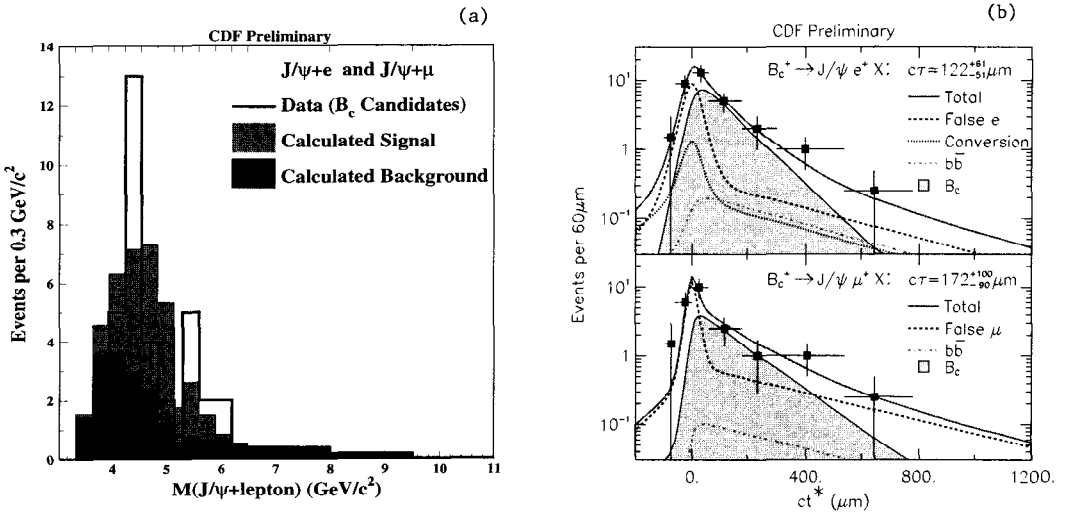


Figure 21: (a) Mass distribution of B_c candidates. The result of the fit for the B_c signal and the measured background are superimposed. (b) Pseudolifetime distribution for data (crosses) with the result of the fit for signal (shaded histogram) and background (dashed line).

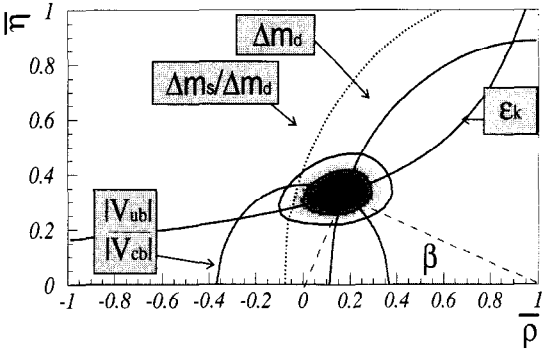


Figure 22: Indirect constraints[44] on the unitary triangle.

Table 6: Summary of the background sources and signal events estimates for a $J/\psi l$ mass between 4 and 6 GeV/c^2.

	J/ψ e results	J/ψ μ results
False Electrons	$2.6 \pm 0.05 \pm 0.3$	
Conversions	$1.2 \pm 0.8 \pm 0.4$	
Total False Muons		$6.4 \pm 0.5 \pm 1.3$
Punch-Through		$0.88 \pm 0.13 \pm 0.33$
Decay-in-Flight		$5.5 \pm 0.5 \pm 1.3$
$B\overline{B}$ Background	1.2 ± 0.5	0.7 ± 0.3
Total Background	5.0 ± 1.1	7.1 ± 1.5
Events Observed	19	12
Net Signal	14.0	4.9
$P_{counting}$(Null)	2.1×10^{-5}	0.084

probability for the background to generate the observed distribution) is rejected at the 4.8 σ level.

To determine the B_c mass, the observed $J/\psi l$ mass distribution is fitted to different mass templates generated assuming a B_c mass between 5.5 and 7.5 GeV/c^2. CDF finds

$$M_{B_c} = 6.40 \pm 0.39(stat.) \pm 0.13(syst.)\ \mathrm{GeV/c^2}.$$

The B_c lifetime is determined by a fit to the pseudolifetime ct^* distribution shown in Fig. 21(b), where the background is parameterized by a prompt contribution plus negative and positive exponentials, while the signal is parameterized by a single positive exponential on which a statistical correction for the missing neutrino P_T is applied. Both background and signal distributions are convoluted with the experimental resolution on the decay length. The fit returns:

$$c\tau_{B_c} = 137^{+53}_{-49}\ \mu\mathrm{m}.$$

From this, CDF infers a lifetime of 0.46 $\pm$ 0.16 $\pm$ 0.03 ps, thus favoring the hypothesis that the c quark decays first and almost independently of the b quark.

The production rate for the B_c meson is determined by measuring the ratio

$$\frac{\sigma(B_c^+) \cdot Br(B_c^+ \to J/\psi l\nu)}{\sigma(B_u^+) \cdot Br(B_u^+ \to J/\psi K^+)} = 0.132^{+0.041}_{-0.037}(stat.) \pm 0.031(syst.)^{+0.032}_{-0.020}(lifet.).$$

In this ratio, the uncertainties in the J/ψ trigg
well as the luminosity determination, cancel. Sin
the theoretical uncertainty on the b-quark productio
lider. The CDF data are consistent with the theoretic
obtained by the LEP experiments.

4.2 Measurement of the CP Asymmetr

Since its discovery[3] over thirty years ago in K^0 dec
eluded further experimental testing. The fact that th
the matter-antimatter asymmetry of the universe r i
priority of many high-energy physics programs. Al
CP violation lies in the relationship between the
different generations of quarks if three or more ger
SM, this relationship is parameterized by the unitar

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = \begin{pmatrix} V_{ud} & V_{us} & V_{ul} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix}$$

where d, s, and b are the mass eigenstates and d', s', and b' are the weak eigenstates. With three generations, this matrix cont
capable of accommodating CP violation. As mentio
1973 proposal of a third-quark generation to explain
unexpected discovery of the charmonium states wh
generation.

The CKM matrix is a unitary matrix, and one
expressed as:

$$V_{ud}V_{ub}^* + V_{cd}V_{cb}^* + V_{td}V_{tb}^*.$$

This equation can be represented as a triangle in th

By measuring enough quantities in the triangles
constrain and cross-check the basic parameters of th
parameters, the angle β, is shown in Fig. 22 and exp

$$\beta \equiv arg(-\frac{V_{cd}V_{cb}^*}{V_{td}V_{tb}^*}).$$

β can be measured by comparing the relative decay rates of B^0 and $\overline{B^0}$ to the common CP eigenstate mode $J/\psi K_S^0$. By exploiting the interference between the direct decay path ($B^0 \to J/\psi K_S^0$) and the mixed decay path ($B^0 \to \overline{B^0} \to J/\psi K_s^0$), β can be measured through the time-dependent asymmetry

$$A_{CP} = \frac{\overline{B^0}(t) - B^0(t)}{\overline{B^0}(t) + B^0(t)} = \sin(2\beta)\sin(\Delta m_d t),$$

where $B^0(t)$ and $\overline{B^0}(t)$ are the number of decays to $J/\psi K_S^0$, at the time t, assuming that the meson produced at $t = 0$ was a B^0 or a $\overline{B^0}$, respectively. The effect of the mixing between B^0 and $\overline{B^0}$ appears through the mass difference Δm_d, while the CP-phase difference between the two decay amplitudes appears via the factor $\sin(2\beta)$. Indirect evidence shown in Fig. 22 implies $0.30 < \sin(2\beta) < 0.88$ at 95% C.L., while OPAL recently reported[45] $\sin(2\beta) = 3.2^{+1.8}_{-2.0} \pm 0.5$ using the same decay channel.

During Run I, CDF collected approximately 200 $B^0, \overline{B^0} \to J/\psi K_S^0$ decays. Although this sample is not sufficient to allow a precise measurement of $\sin(2\beta)$, it is the largest reconstructed sample of $B^0, \overline{B^0} \to J/\psi K_S^0$ decays in the world, and it can be studied[46] to determine the feasibility of this measurement in a hadron collider. As for the B_c discovery, the selection of $B^0, \overline{B^0} \to J/\psi K_S^0$ candidates starts from the $J/\psi \to \mu^+\mu^-$ reconstruction. The muon tracks are required to be measured in the SVX detector, thereby obtaining a precise determination of the J/ψ vertex. The other pairs of oppositely charged tracks in the event are then searched for those consistent with the $K_S^0 \to \pi^+\pi^-$ decay hypothesis, where the K_S^0 decay point is significantly displaced from the J/ψ vertex. Each K_S^0 candidate is then combined with the J/ψ candidate in a four-particle fit which requires the K_S^0 to point back to the J/ψ vertex, and the combined $J/\psi K_S^0$ system to point back at the primary vertex. The mass calculated by the fit has a typical resolution of $\sigma_M \sim 9$ MeV/c^2. The proper decay length has a typical resolution of $\sim 50\ \mu m$. Figure 23(a) shows the distribution of positive-lifetime candidate events as a function of the normalized mass $M_N = (M_{FIT} - M_0)/\sigma_{FIT}$, where M_0 is the central value of the B^0 mass peak (5.277 GeV/c^2). A maximum likelihood fit yields 198 ± 17 mesons. Since the CP asymmetry varies in time as $\sin(\Delta m_d t)$, it reaches its maximum close to a proper decay length of $\sim 1000\ \mu m$, which is a region in which the background is strongly suppressed, as shown in Fig. 23(b).

4.2.1 Same-Side Flavor Tagging

Once the sample of B's is obtained, the next step is to determine ("tag") whether they were B^0's or $\overline{B^0}$ when they were produced.

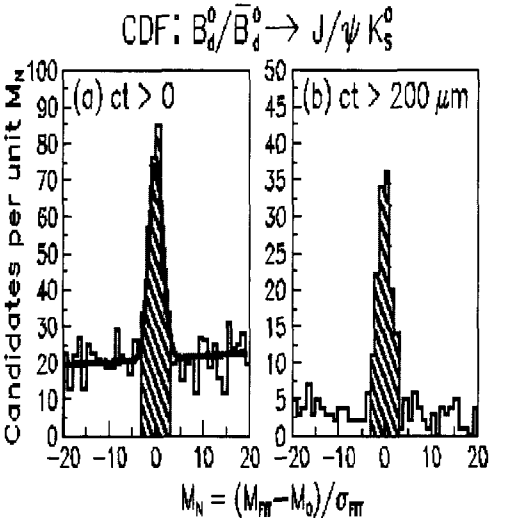


Figure 23: Normalized mass distribution for (a) $B^0 \to J/\psi K_S^0$ candidates with $ct \geq 0$, and (b) $ct \geq 200\ \mu m$. The curve is the Gaussian signal plus background from the full likelihood fit.

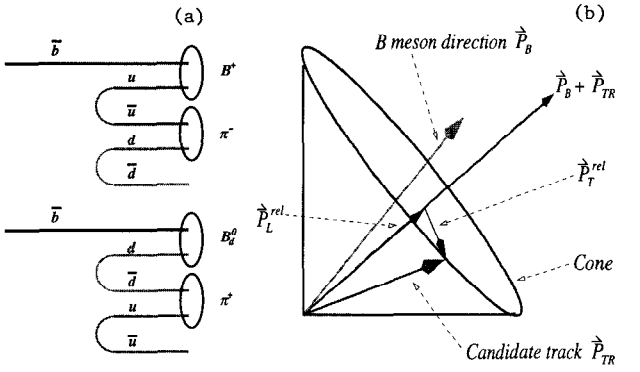


Figure 24: (a) A simple picture of b quarks hadronizing into B mesons, and (b) definition of tagging candidate based on P_T^{rel}.

Different methods have been suggested for this tagging. Most of them rely on the "opposite-side tagging," i.e., the identification of the flavor of the second b in the events through its semileptonic decay ($b \to e^-$, while $\overline{b} \to e^+$), or its decay into charged Kaons. Similarly, the Opposite Side Jet-Charge method reconstructs the away-side b-hadron and uses statistical methods to determine its flavor. Alternatively, one can consider the "same-side tagging" (SST), i.e., the determination of the b quark flavor through the examination of the particles produced in association with the reconstructed B. In particular, the method used by CDF relies on the correlation between the flavor of the reconstructed B and the charge of a nearby particle. This idea was first proposed in order to take advantage of the fact that the b quark may first hadronize to a B^{**} state, whose decay products would be the B^0 as well as a "tagging" pion:[47]

$$B^{**-} \to \overline{B^0}\pi^-,$$
$$B^{**+} \to B^0\pi^+.$$

In this scenario, a $\overline{B^0}$ would always be associated with a π^-, and a B^0 with a π^+. The same correlation is expected to exist between the B meson and the "leading" pion from fragmentation, as shown in Fig. 24(a), and the CDF analysis utilizes both sources of correlation. The nearby track is selected according to the following criteria:

- the track must lie in an $\eta - \phi$ cone with a half-angle of 0.7, around the B meson direction,
- $P_T^{min} \geq 400\,\mathrm{MeV/c}$, and
- the track must consistently be coming from the primary vertex.

When there is more than one candidate track, the one with the smallest P_T^{rel} is selected as the tagging track, as shown in Fig. 24(b). P_T^{rel} is defined as the component of the particle momentum transverse to the momentum of the combined $B + particle$ system.

The tagging algorithm, based upon physical processes that happen before the B meson decays, is applicable to other decay modes, and indeed, it has been applied successfully to the observation of the B^0–$\overline{B^0}$ mixing and the measurement of Δm_d using the $B \to lD^{(*)}X$ decay,[48] as shown in Fig. 25.

The same algorithm has been applied for a measurement of the mixing parameter Δm_d in a lower-statistic sample of $B^0 \to J/\psi K^*$ decays (which kinematically are very similar to the $J/\psi K_S^0$ decays used for the CP measurement) yielding a measurement of Δm_d consistent with the higher precision determination of Fig. 25.

If the tagging algorithms were perfect, the time-dependent mixing measurement

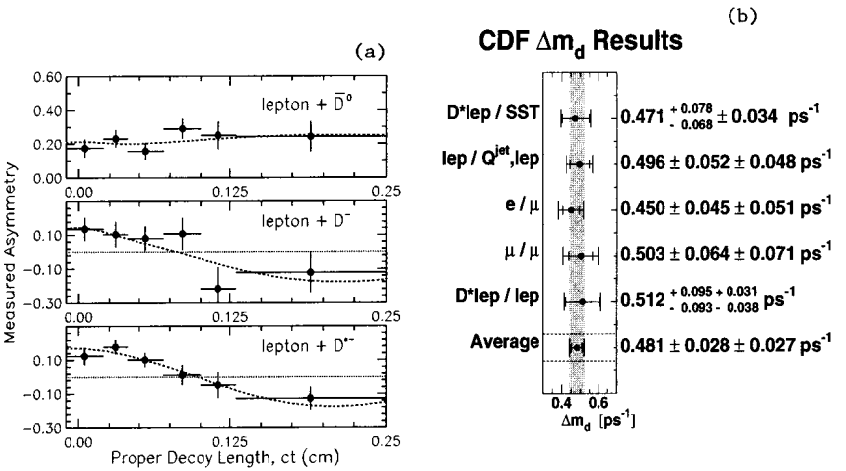


Figure 25: (a) Measured tagging asymmetries as a function of ct for $B \to l + D^0$ candidates (coming mostly from B^+), where no mixing is expected, and $B \to l + D^{(*)-}$ (coming mostly from B^0) for which mixing is expected. (b) Compilation of Δm_d measurements from CDF, showing that the SST method is competitive with the other standard methods.

shown in Fig. 25(a) would be a cosine curve of amplitude one. An amplitude smaller than one is an indication of a "dilution" of the measurement. The dilution, $\mathcal{D}_0$, is related to the mistag probability and to the observed asymmetry by:

$$\begin{aligned} A_{CP}^{obs.} &= \mathcal{D}_0 \sin(2\beta)\sin(\Delta m_d t), \\ \mathcal{D}_0 \equiv \frac{N_{RS} - N_{WS}}{N_{RS} + N_{WS}} &= 1 - 2\mathcal{P}_{mistag}, \end{aligned}$$

where N_{RS} are events with the correct-sign correlation, and N_{WS} events with the wrong-sign correlation. For this same-side tagging method, $\mathcal{D}_0 \sim 20\%$. This dilution determination, measured on the data, is necessary for the extraction of $\sin(2\beta)$.

4.2.2 Tagging Asymmetry

The SST technique tags approximately 65% of the $B^0, \overline{B^0} \to J/\psi K_S^0$ decays. Figure 26 shows the sideband-subtracted asymmetry in bins of the proper decay-time, where the asymmetry is calculated by counting the sideband-subtracted number of positive tags, N^+ , and negative tags, N^-, in each proper decay-time bin:

$$\mathcal{A}(ct) = \frac{N^-(ct) - N^+(ct)}{N^-(ct) + N^+(ct)}.$$

The signal and background samples were defined according to the dashed region shown in Fig. 23. The events in the signal region generally prefer negative tags (i.e., a positive asymmetry), whereas events in the sideband regions favor positive tags (negative asymmetry). As noted before, however, the signal purity is high at large ct, and the sideband subtraction is, correspondingly, a small effect.

Two fits are shown in Fig. 26. The dashed curve gives the results of a simple χ^2 fit of the function $\mathcal{A}_0 \sin(\Delta m_d t)$ to the binned asymmetries, where Δm_d has been fixed to its 1996 world-averaged[49] value of 0.474 ps^{-1}. The solid curve is the result of an unbinned maximum likelihood fit which incorporates both signal and background distributions in mass and proper decay-time. Sideband and negative-lifetime events are included to help constrain the background distributions. The likelihood function also incorporates resolution effects and corrections for systematic biases, such as the small inherent charge asymmetry favoring positive tracks resulting from the wire plane orientation in the main CDF drift chamber. Clearly, the result is dominated by the sample size and not by the particular fitting procedure applied to the data. Also shown

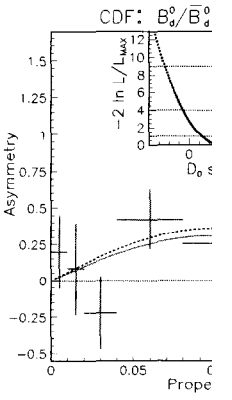


Figure 26: The sideband-subtracted taggir
structed $J/\psi K_S^0$ proper decay length (poin
discussed in the text, while the inset shows t

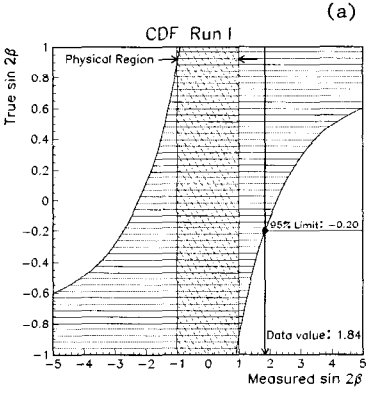


Figure 27: 95% confidence belts for the si
the actual measurement; (a) intersection wit
interval, and (b) the world summary of sin(2

in the Fig. 26 inset is the relative log-likelihood. It is very close to a parabola, indicating Gaussian errors.

Before entirely ascribing the above asymmetry to CP Violation, all other sources of charge asymmetry must be eliminated. The small charge asymmetry of the main drift chamber has been measured in an independent sample of inclusive $B \to J/\psi X$ decays and corrected for in the maximum likelihood fit. Backgrounds from other B decays, such as $B^0 \to J/\psi K^{*0}$, $K^{*0} \to K^0_S\pi^0$, and $B^0 \to J/\psi K^+$, where the π^0 has not been reconstructed, have been considered and found to be negligible. The high signal purity at large decay times also limits contributions to the asymmetry from backgrounds which are present in the sidebands. CDF determines the systematic uncertainty on $\mathcal{D}_0 \sin(2\beta)$ by shifting the central value of each fixed input parameter of the likelihood fit by $\pm 1\sigma$ and refitting it to find the shift in $\mathcal{D}_0 \sin(2\beta)$.

The following systematic effects were investigated:

- B^0 lifetime,
- parametrization of the intrinsic charge asymmetry of the detector, and
- Δm_d.

The systematic effects are added in quadrature, giving

$$\mathcal{D}_0 \sin(2\beta) = 0.31 \pm 0.18(stat.) \pm 0.03(syst.).$$

4.2.3 Extracting $\sin(2\beta)$

As mentioned above, the dilution $\mathcal{D}_0$, which reduces the amplitude of the CP asymmetry, can be measured in other data samples, including the $B^0 \to J/\psi K^{*0}$ decays and the $B^0 \to lD^{(*)}X$ samples. These different dilution measurements can be extrapolated to the kinematic range appropriate for the $J/\psi K^0_S$ data. The extrapolation is done using a Monte Carlo simulation based upon a version of the PYTHIA event generator tuned to CDF data;[46] the necessary adjustments to the different dilutions are at the level of 10% at most, and CDF finds that the appropriate dilution for the $B^0\overline{B^0} \to J/\psi K^0_S$ data is $\mathcal{D}_0 = 0.166 \pm 0.018 \pm 0.013$, where the first uncertainty is due to the uncertainties in the contributing dilution measurements, and the second uncertainty is due to the Monte Carlo extrapolation, which is calculated by varying the parameters of the Monte Carlo model. Using this value of $\mathcal{D}_0$, CDF finds

$$\sin(2\beta) \quad = 1.8 \pm 1.1(stat.) \pm 0.3(syst.),$$

where the dilution uncertainty has been added to the systematic uncertainty. The central value is unphysical since the amplitude of the measured asymmetry exceeds the measured dilution. The result may be phrased in terms of confidence intervals. The CDF analysis follows the frequentist construction of Feldman and Cousins,[50] which gives proper confidence intervals even for measurements in the unphysical region, as is the case here. The confidence interval is shown in Fig. 27(a). It is found that the measurement corresponds to an exclusion of $\sin(2\beta)$ values below -0.20 at 95% C.L.

It is also interesting to note that as long as $\mathcal{D}_0 \neq 0$, then the exclusion of negative $\sin(2\beta)$ from this result is independent of the actual value of $\mathcal{D}_0$. Given that $\mathcal{D}_0 > 0$, the same prescription for calculating a limit yields an exclusion of negative values of $\sin(2\beta)$ at 90% C.L. Figure 27(b) shows the world summary of the direct measurements of $\sin(2\beta)$. Clearly, one of the priorities for future experiments is to bring $\sin(2\beta)$ into the physical region!

5 Prospects for Run II

After the termination of Run I in February 1996, the accelerator and both detectors at Fermilab started a program of major upgrades for the next data taking period (Run II), which is expected to start in the year 2000. The major addition to the accelerator complex is the Main Injector, a 150 GeV machine that will be used to inject protons into the Tevatron and will replace the Main Ring in the creation of antiprotons. The Main Injector is expected to increase the Tevatron peak luminosity from $\sim 2 \times 10^{31}\, cm^{-2}s^{-1}$ to $\sim 8 \times 10^{31}\, cm^{-2}s^{-1}$. A second machine, the Recycler, will capture unused antiprotons at the end of a collider run and reuse them, boosting the peak luminosity to $\sim 2 \times 10^{32}\, cm^{-2}s^{-1}$. In addition, the Tevatron energy will increase from $\sqrt{s} = 1.8$ TeV to $\sqrt{s} = 2$ TeV, thus boosting the $t\bar{t}$ cross section by 40%. The total integrated luminosity expected to be delivered to the experiments during Run II is ~ 2 fb^{-1}. Both collider detectors are being upgraded[51] for Run II. The major upgrades include: D$\emptyset$ acquiring a tracking system with a central solenoidal magnetic field and a silicon vertex detector, and CDF expanding the coverage of its vertex detector, its calorimetry, and its muon detectors.

5.1 Potentials for Top Physics in Run II

The reach of top physics at the Tevatron during Run II has been studied.[52] Each detector will record $\sim$ 160 dilepton and $\sim$ 1200 lepton+jets events, with $\sim$ 500 of them double-tagged. For the cross-section measurement, the limiting factor will probably be the error on the luminosity, which can be measured to $\sim$ 5% using the $W \rightarrow l\nu$ rate. Therefore, the error on the production cross section is expected to be around 8–10%.

The error on the mass measurement is presently dominated by statistics, but the uncertainty on the jet energy scale is a close second. As it is usual in the study of systematic errors, the understanding of the energy scale will likely improve with improving statistics (by studying, for instance, the $Z + multijet$ events). The total uncertainty on M_{top} in Run II should be at the level of 2 GeV/c^2 (i.e., at the 1% level).

The precision on the R measurement should improve to 2%, corresponding to a 95% C.L. of $|V_{tb}| \geq 0.20$, while limits of $B(t \rightarrow q\gamma) < 3 \times 10^{-3}$ and $B(t \rightarrow qZ) < 0.02$ should be achievable.

From single top production, which is directly proportional to $|V_{tb}|^2$, the Tevatron experiments will be able to determine $|V_{tb}|^2$ to $\sim$ 10%. Other fields of study may arise from the large top quark mass and the *expected* (or, rather, *hoped for*) connection with the symmetry breaking mechanism of the Standard Model.

5.2 Potentials for B Physics in Run II

The goal of the B physics community in the next decade is the measurement of the unitary triangle parameters. Therefore, we will concentrate on the expectations for this kind of physics, even though it is expected that the Tevatron experiments will continue exploring the "standard" b physics avenues, like lifetimes, B^0 and B_s mixing, and heavy states searches.

Studies of CP violation will concentrate on the measurement of $\sin(2\beta)$ and $\sin(2\alpha)$ in $B^0 \rightarrow J/\psi K^0_s$ and $B^0 \rightarrow \pi^+\pi^-$ decays. CDF has demonstrated that the Tevatron Collider environment allows these kind of measurements. Moreover, CDF plans to also look for CP violation in $B_s \rightarrow D_s K$ and $B \rightarrow DK$, probing $\sin(2\gamma)$.

As shown in the discussion of the $\sin(2\beta)$ measurements, the important ingredients for the CP measurements are the efficiency ϵ to tag the original b flavor and the dilution $\mathcal{D}$ in the tagging. These ingredients are combined in a figure of merit ($\epsilon\mathcal{D}^2$) that compares different tagging methods, since the error on the determination of a given CP asymmetry is proportional to $1/\sqrt{N\epsilon\mathcal{D}^2}$, where N is the number of reconstructed decays. The $\epsilon\mathcal{D}^2$ for the three tagging methods discussed in the CP asymmetry measurements at CDF are shown in Table 7.

Table 7: Summary of the *dilepton* counting experiments.

Tagger	$\epsilon\mathcal{D}^2$ (Run I)	$\epsilon\mathcal{D}^2$ (Run II)	Relevant Upgrade
Soft Lepton (SLT)	0.9 ± 0.1	1.7	Extended Coverage
Same Side π (SST)	1.7 ± 0.2	2.0	New Tracking
Jet Charge (JetQ)	1.0 ± 0.4	3.0	New Tracking
All Combined	2.9	5.4	

Based on the experience gathered in Run I, CDF expects to reconstruct $\sim$10,000 $B^0 \rightarrow J/\psi K^0_S$ decays in Run II. This takes into account the increased luminosity, the extended silicon vertex detector, and the upgraded muon detector. This expectation is conservative because it does not include the possibility of using the $J/\psi \rightarrow e^+e^-$ decay channel, which will be possible in Run II with a contribution of $\sim$ 5000 more reconstructed B^0 decays. The asymmetry and its error are defined as:

$$\begin{aligned} \mathcal{A} &= \mathcal{D}\sin 2\beta, \\ \sigma(\sin 2\beta) &= \frac{\sigma(\mathcal{A})}{\mathcal{D}} \oplus (\sin 2\beta)\frac{\sigma(\mathcal{D})}{\mathcal{D}}, \end{aligned}$$

where $\oplus$ means addition in quadrature. Using the values determined in the Run I analysis and extrapolating according to Table 7, CDF expects to measure $\sin(2\beta)$ with an error of $\sigma(\sin 2\beta) = 0.071 \oplus 0.044 = 0.08$.

The identification and reconstruction of $B \rightarrow \pi^+\pi^-$, for the determination of $\sin(2\alpha)$, is very challenging at hadronic colliders because of the very low branching ratio ($\sim 10^{-5}$) and the huge combinatorial background from the low P_T charged tracks produced in $p\bar{p}$ collisions. CDF made an effort to study the trigger and the background rejection. A fully dedicated trigger, based on the idea of selecting tracks with large impact parameter ($d_0 \geq 100\mu m$) is thought to be sufficient to limit the event rate. Assuming 10,000 events in 2 fb^{-1}, $\epsilon\mathcal{D}^2 = 5.4\%$ and a signal-to-background ratio of one to four, CDF expects to measure the asymmetry with an error of $\sigma_{\mathcal{A}} = 0.12$.

References

[1] M. Perl *et al.*, Phys. Rev. Lett. **35**, 1489 (1975).

[2] N. Cabibbo, Phys. Rev. Lett. **10**, 531 (1963).
M. Kobayashi and T. Maskawa, Prog. Theor. Phys. **49**, 652 (1973).

[3] J. Christenson *et al.*, Phys. Rev. Lett. **13**, 138 (1964).

[4] S. W. Herb *et al.*, Phys. Rev. Lett. **39**, 252 (1977).

[5] C. Berger *et al.*, (PLUTO), Phys. Lett. B **76**, 243 (1978); C. W. Darden *et al.*, (DASP), Phys. Lett. B **76**, 246 (1978).

[6] C. Itzykson, J.-B. Zuber, *Quantum Field Theory* (McGraw-Hill, New York 1980).

[7] W. Behrends *et al.*, (JADE), Phys. Lett. B **146**, 437 (1984).

[8] The LEP Collaborations ALEPH, DELPHI, L3 OPAL, the LEP Electroweak Working Group, and the SLD Heavy Flavor Group:
D. Abbaneo *et al.*, LEPEWWG/97-02.

[9] V. Bharadwa *et al.*, 10^{th} *Topical Workshop on* $p\bar{p}$ *Collider Physics,* Fermilab, May 1995.

[10] F. Abe *et al.*, Nucl. Instr. Methods A **271**, 387 (1988).

[11] P. Azzi *et al.*, Nucl. Instr. Methods A **360**, 137 (1995); FERMILAB-CONF-94/205-E;
D. Amidei *et al.*, Nucl. Instr. Methods A **350**, 73 (1994).

[12] S. Abachi *et al.*, Nucl. Instr. Methods A **338**, 185 (1994).

[13] C. Albajar *et al.*, Phys. Lett. B **186**, 247 (1987).

[14] H. Albrecht *et al.*, Phys. Lett. **192**, 245 (1987).

[15] C. Quigg, presented at the *Advanced School on Electroweak Theory,* Mao, Menorca, 1996, FERMILAB-CONF-97/157-T, hep-ph/9707508.

[16] F. Abe *et al.*, Phys. Rev. D **50**, 2966 (1994);
F. Abe *et al.*, Phys. Rev. Lett. **73**, 225 (1994).

[17] F. Abe *et al.*, Phys. Rev. Lett. **74**, 2626 (1995).

[18] S. Abachi *et al.*, Phys. Rev. Lett. **74**, 2632 (1995).

[19] J. Rosner, "Top Quark Mass," presented at the *NATO Advanced Study Institute on Masses of Fundamental Particles*, Cargese, France, 1996. CERN-TH-96-245, hep-ph/9610222.

[20] E. Laenen *et al.*, Phys. Rev. Lett. B **321**, 2 (19

[21] E. Berger and H. Contopanagos, Phys. ReD **54**

[22] S. Catani *et al*, Phys. Lett. B **378**, 329 (19).

[23] F. Abe *et al.*, Phys. Rev. Lett. **80**, 2779 (18).

[24] S. Abachi *et al.*, Phys. Rev. Lett. **79**, 1203997

[25] F. Abe *et al.*, Phys. Rev. Lett. **80**, 2773 (18).

[26] F. Abe *et al.*, Phys. Rev. Lett. **79**, 1992 (17).

[27] B. Abbott *et al.*, FERMILAB-PUB-98/133 (19 to Phys. Rev. D.

[28] B. Klima, presented at the *12th Les Rencre d Results and Perspectives in Particle Physi* La FERMILAB-CONF-98/137-E, hep-ex/98017.

[29] F. Abe *et al.*, Phys. Rev. Lett. **80**, 2767 (18); F. Abe *et al.*, Phys. Rev. Lett. **82**, 271 (19, *Er*

[30] B. Abachi *et al.*, Phys. Rev. Lett. **79**, 1197997

[31] B. Abbott *et al.*, Phys. Rev. Lett. **80**, 2063998)

[32] G. Quast, *29th International Conferencen Hi* Vancouver, BC, CANADA; see also http:/ww.

[33] F. Abe *et al.*, in *Proceedings of Internatial Ei Energy Physics (HEP 97)*, Jerusalem, Israd 997

[34] S. Glashow *et al.*, Phys. Rev. D **2**, 1285 (10).

[35] F. Abe *et al.*, in *Proceedings of 33rd Rencres ergy Hadronic Interactions,* Les Arcs, Frar, 19 E.

[36] B. Abbott *et al.*, Phys. Rev. Lett. **74**, 2548995) B. Abbott *et al.*, Phys. Lett. B **370**, 239 (16).

[37] B. Abbott. *et al.*, Phys. Rev. Lett. **82**, 35 (99).

[38] F. Abe *et al.*, Phys. Rev. D **57**, 5382 (1998

[39] F. Abe *et al.*, Phys. Rev. D **57**, R3811 (19.

[40] B. Abbott *et al.*, Phys. Lett. B **423**, (1998)

[41] F. Abe *et al.*, Phys. Rev. Lett. **80**, 2057 (1998).

[42] P. Abreu *et al.*, Phys. Lett. B **398**, 207 (1997);
K. Ackerstaff *et al.*, Phys. Lett. B **420**, 157 (1998);
R. Barate *et al.*, Phys. Lett. B **402**, 213 (1997).

[43] F. Abe *et al.*, Phys. Rev. Lett. **81**, 2432 (1998).

[44] F. Parodi *et al.*, LAL-98-49, hep-ph/9802289 (1998).

[45] K. Ackerstaft *et al.*, Eur. Phys. J. C **5**, 379 (1998).

[46] F. Abe *et al.*, Phys. Rev. D **59**, 32001 (1999).

[47] M. Gronau, A. Nippe, and J. L. Rosner, Phys. Rev. D **47**, 1988 (1993);
M. Gronau and J. L. Rosner, Phys. Rev. D **49**, 254 (1994).

[48] F. Abe *et al.*, Phys. Rev. Lett. **80**, 2057 (1998);
F. Abe *et al.*, Phys. Rev. D **59** 32001 (1999).

[49] Particle Data Group; R. Barnett *et al.*, Phys. Rev. D **54**, 1 (1996).

[50] G. Feldman and R. Cousins, Phys. Rev. D **57**, 3873 (1998).

[51] CDF Collaboration, FERMILAB-PUB-96/390-E, 1996;
D$\emptyset$ Collaboration, FERMILAB-PUB-96/357-E, 1996.

[52] The TeV-2000 Group Report, FERMILAB-PUB-96/082, 1996; R. Frey *et al.*,
FERMILAB-CONF-97/085 1997.

NEW HERA RESULTS ON DEEP–INELASTIC e^+p SCATTERING AT VERY HIGH Q^2

Ulrich F. Katz (ZEUS)

Physikalisches Institut, Universität Bonn
Nußallee 12, 53115 Bonn, Germany
Email: katz@physik.uni-bonn.de

Representing the H1 and ZEUS Collaborations

ABSTRACT

Until the end of 1997, the H1 and ZEUS experiments at the HERA collider at DESY both have collected an integrated luminosity of roughly $40\,\mathrm{pb}^{-1}$ of e^+p collisions. These data allow high–precision studies of deep–inelastic scattering of positrons on protons at large four–momentum transfer. Measurements of the differential cross–sections $d\sigma/dQ^2$ and $d\sigma/dx$ are reported and compared to predictions derived from measurements performed in fixed–target experiments at lower energies, extrapolated to the HERA kinematic regime using Standard Model calculations. In general, the HERA results agree well with the predictions. An excess of events at highest Q^2 in the 1994–1996 data, which was reported by both experiments, has not been corroborated by the new data but is still present in the combined data samples. Searches for signatures of production of positron–quark resonances or $(eeqq)$ contact interactions have found no evidence for physics processes beyond the Standard Model, and new limits on the parameters of these scenarios are set.

© 1998 by Ulrich F. Katz.

1 Introduction

The electron[a]–proton accelerator HERA at DESY in Hamburg, Germany, is providing ep collisions to the experiments H1[1] and ZEUS[2] since 1992. Between mid–1994 and the end of 1997 HERA has been operated with positrons (e^+) at an energy of $E_e = 27.5\,\text{GeV}$ and protons (p) of $E_p = 820\,\text{GeV}$, yielding a center–of–mass energy of $\sqrt{s} = 300\,\text{GeV}$. In this time, ZEUS and H1 have collected e^+p data samples corresponding to integrated luminosities of $46.6\,\text{pb}^{-1}$ and $37.0\,\text{pb}^{-1}$, respectively. The analyses reported in this paper are based on these data, and are hence sensitive to processes with cross–sections as low as about 50 fb.

In the Standard Model (SM), deep–inelastic lepton–proton scattering proceeds to leading order via the two diagrams shown in Fig. 1, i.e. by the t–channel exchange of a γ or Z boson (neutral current, NC) or of a W boson (charged current, CC). Reaction products are the scattered lepton and a multi–particle hadronic system which evolves by fragmentation from the partonic final state and reflects its configuration in terms of energy flow and jet topology.

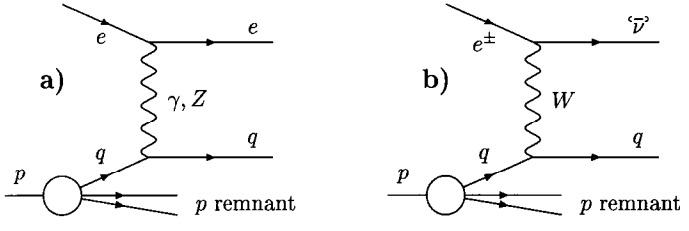


Figure 1: *Leading order Feynman graphs for deep-inelastic ep scattering (a) for NC and (b) for CC reactions. Note that scattering on antiquarks, produced in the proton by virtual processes, is also possible.*

For both reaction types, the relevant kinematic variables are given by

$$Q^2 = -q^2 = -(k - k')^2 \quad (1)$$

$$x = Q^2/(2\,q \cdot P) \quad (2)$$

$$y = (q \cdot P)/(k \cdot P)\,, \quad (3)$$

[a] Here and in the following, the term *electron* will denote both electrons and positrons, unless explicitly stated otherwise.

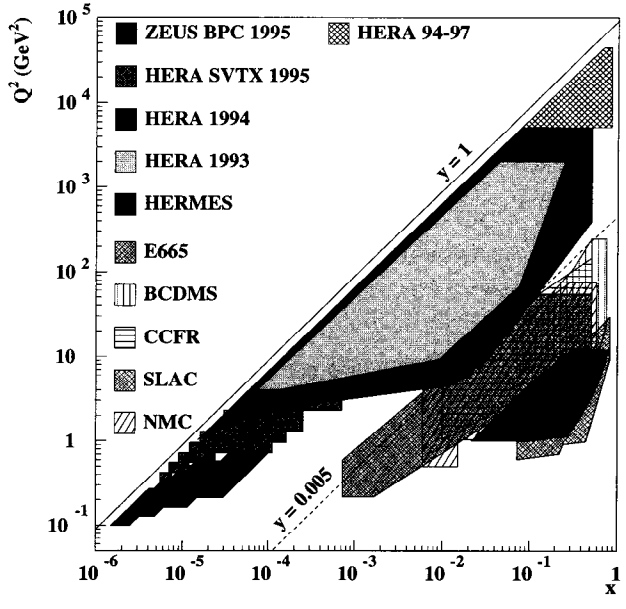


Figure 2: *Regions in the HERA kinematic (x, Q^2)–plane covered by structure function and cross–section measurements of ZEUS and H1 and of the fixed–target experiments with charged lepton beams (BCDMS, NMC, E665, SLAC) and with neutrino beams (CCFR).*

where k, k' and P are the four–momenta of the initial and the final–state lepton[b] and of the incoming proton, respectively (see Fig. 1). Q^2 measures the virtuality of the gauge boson and reaches values of above $10^4\,\text{GeV}^2$ at HERA, corresponding to a spatial resolution of $\mathcal{O}(10^{-16}\,\text{cm})$ with which structures in the proton can be resolved. The Bjorken scaling variable x indicates the fraction of the proton momentum carried by the struck quark, and y is related to the eq center–of–mass scattering angle θ^* by $y = (1 - \cos\theta^*)/2$ (where $\theta^* = 0$ for forward scattering). For $k_\gamma = 0$, these variables are related by $Q^2 = x \cdot y \cdot s$. The invariant mass of the electron–quark system is given by $M = \sqrt{xs}$.

The kinematic regime covered by the HERA and the fixed–target experiments is indicated in Fig. 2. This report will focus on Q^2 values exceeding a few $100\,\text{GeV}^2$,

[b] If a photon with four–momentum k_γ is radiated off the initial or final–state lepton, q has to be replaced by $q = k - k' - k_\gamma$ in eqs. (1-3).

and in particular on the region of highest Q^2 indicated by the cross–hatched region, where the differential cross–sections are small and where detailed studies only have become possible with the high–statistics data collected in 1996 and 1997. The high–Q^2 event topologies are shown in Fig. 3. Note that for Q^2 above $\mathcal{O}(1000\,\mathrm{GeV}^2)$ the scattered electron is deflected under large angles w.r.t. its original direction and has energies up to several 100 GeV.

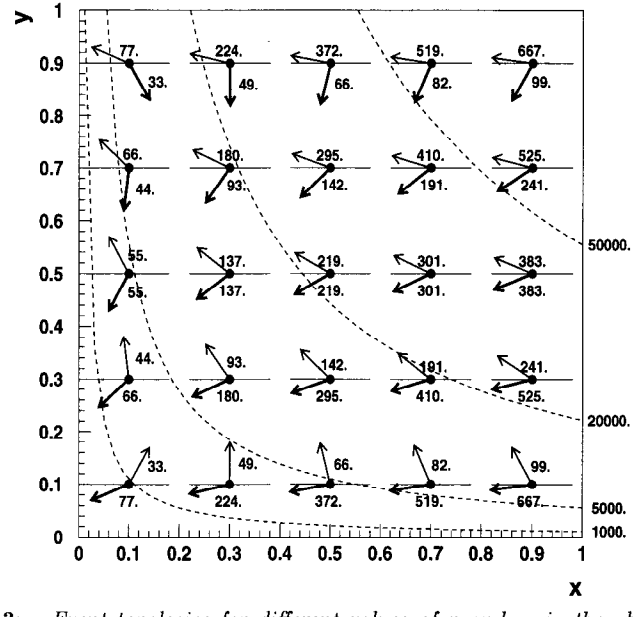


Figure 3: *Event topologies for different values of x and y, in the absence of QED radiation (i.e. $k_\gamma = 0$). Each event pictogram shows the directions in which lepton (upper arrow) and quark (lower arrow) are emitted in reactions with x and y being given by the position of the central dot. Quasi–free, massless quarks have been assumed for relating x and y with the scattering angles. The numbers next to the arrows represent the electron and quark energies in GeV. The dashed lines indicate constant values of Q^2 (in GeV^2).*

Particular interest in DIS at highest Q^2 was stimulated when both the ZEUS[3] and the H1[4] collaborations reported an excess of NC events over the SM expectation at $Q^2 \gtrsim 2 \cdot 10^4\,\mathrm{GeV}^2$ in their 1994–1996 data. Figure 4 shows the NC cross–section as a function of a lower cutoff in Q^2 as measured by ZEUS and H1 using the data taken until mid–1997. Several scenarios beyond the SM have been suggested to account for this high–Q^2 excess. The main objective of this report is to summarize the latest developments of cross–section measurements and searches for new physics in the high–Q^2 e^+p DIS data taken at HERA.

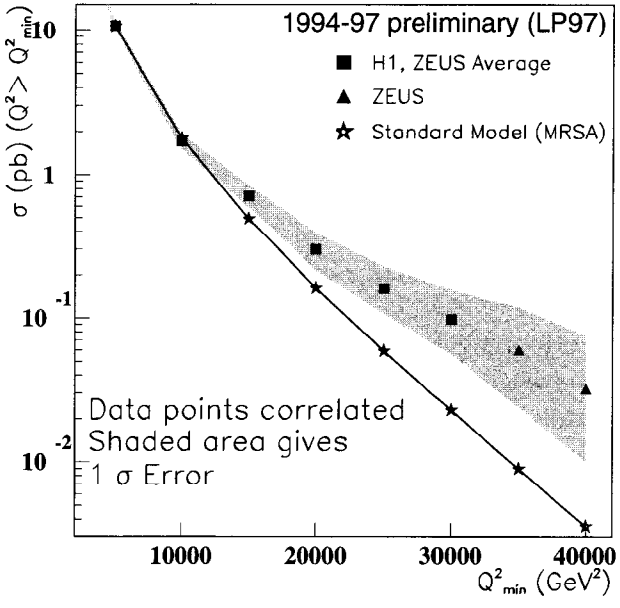


Figure 4: *NC cross sections for $Q^2 > Q^2_{\mathrm{min}}$ with $Q^2_{\mathrm{min}} > 5000\ GeV^2$. Shown are the combined ZEUS and H1 results (squares) and the corresponding one–sigma error band (shaded region). Note that the triangles in the domain $Q^2 \geq 35000\ GeV^2$ are the ZEUS cross–sections and contain no averaging with the zero events observed by H1 in this region.*

After a short synopsis of experimental issues (section 2), we summarize some essential formulae describing the DIS cross–sections in the SM (section 3), review new NC and CC cross–section measurements at high Q^2 reported recently by ZEUS[5,6] and H1[7] (section 4) and finally report on searches for resonance production[8,9] and contact interactions[10,11] in the high–Q^2 DIS data (section 5). A short summary and outlook concludes this report (section 6). A review of QCD analyses based on the HERA data which also refers to the high–Q^2 cross–sections can be found in a separate contribution to these proceedings.[12]

2 Experimental Issues

The calorimeters are the principal components for both the ZEUS and the H1 high-Q^2 analyses. ZEUS uses a compensating uranium–scintillator sandwich calorimeter with energy resolutions of $\Delta E/E = 18\%/\sqrt{E\,[\mathrm{GeV}]}$ for electrons and $\Delta E/E = 35\%/\sqrt{E\,[\mathrm{GeV}]}$ for hadrons (under test beam conditions). H1 employs a liquid–argon (LAr) calorimeter covering the central and forward[c] region, and a lead scintillating–fiber calorimeter in the rear direction. The LAr resolutions (relevant for high-Q^2 analyses) are $\Delta E/E = 12\%/\sqrt{E\,[\mathrm{GeV}]}$ for electrons and $\Delta E/E = 50\%/\sqrt{E\,[\mathrm{GeV}]}$ for hadrons. The absolute energy scales are known to a precision of about 1.5% (electrons) and $3-4\%$ (hadrons) in both experiments.

Momenta of charged particles and event vertices are reconstructed using the central tracking devices which are complemented by forward and rear planar chambers and operate in a longitudinal magnetic field (1.43 T in ZEUS, 1.15 T in H1).

In both experiments the luminosity is measured to a precision of about 1.5% using external electromagnetic calorimeters placed down–stream in e–beam direction which are designed to measure the rates of electrons and photons from Bethe–Heitler processes ($ep \to ep\gamma$).

A schematic overview over the general detector layout of both experiments, together with typical high-Q^2 event signatures, is shown in Fig. 5.

The trigger decision for NC candidates is based mainly on energies deposited in the calorimeter, specifically on the electromagnetic energy and on various energy sums like the total transverse energy (E_t). For CC events the net transverse momentum constitutes the main signature.

In the offline event reconstruction an algorithm is applied to identify scattered electrons using the topology of their calorimeter signal and the tracking information. NC events are identified mainly by the presence of an isolated electron candidate with energy above typically 10 GeV, and by longitudinal energy–momentum containment. The kinematic variables for NC candidate events are calculated either from the scattering angle of the electron and an overall hadronic angle (double–angle method, ZEUS), or from a combination of energy and angle of the scattered

[c] Both experiments define their Z axis parallel to the proton beam direction, with the nominal interaction point being the coordinate system origin. The polar angle, θ, and the forward and rear directions are defined w.r.t. the Z axis. The pseudorapidity is given by $\eta = -\ln\tan(\theta/2)$.

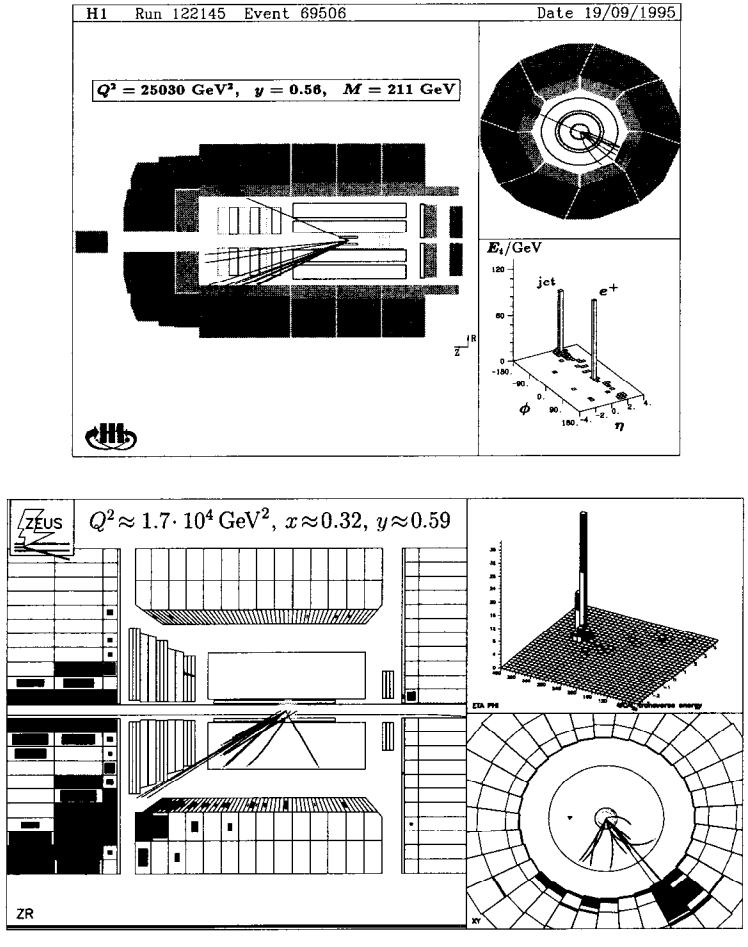


Figure 5: *Event displays of high-Q^2 e^+p reactions: A NC event in the H1 detector (top) and a CC event in ZEUS (bottom). The left parts of both pictures show vertical cuts through the detectors along the beam direction, with the positron entering from the left and the proton from the right. The tracks reconstructed from the hits in the inner tracking chambers and the energy deposits in the surrounding calorimeters are indicated. The two smaller displays at the right show transverse cuts through the central detectors and the flow of transverse energy in the pseudorapidity–azimuth plane, respectively. The reconstructed kinematic variables of each event are indicated on top of the displays.*

electron and hadronic variables ($e\Sigma$–method, H1). The resolutions of x, y and Q^2 for NC events are typically of $\mathcal{O}(5\%)$ and generally better than 10%.

CC events are selected by requiring a net transverse momentum above a threshold and the absence of a scattered electron. The kinematic variables have to be determined from hadronic measurements only, resulting in resolutions of about $15\% - 25\%$ except at very high y, where they are worse.

3 DIS in the Standard Model

In the naïve quark–parton model (QPM), e^+p DIS proceeds via lepton–quark scattering according to the Feynman graphs shown in Fig. 1, where the quark is assumed to be quasi–free, massless, and collinear with the incoming proton. In this picture the DIS cross section is given by the $e^{\leftrightarrow}\bar{q}^{\rightarrow}$ matrix elements (which are fully calculable within the SM) and the probability densities $q(x)$ ($\bar{q}(x)$) to find a quark q (antiquark $\bar{q}$) of a given flavor in the proton which carries a fraction x of its momentum. The effects of the strong interaction can be taken into account by applying the QCD–based DGLAP evolution prescription[13] to the parton density functions (PDF's) which thereby acquire an effective Q^2 dependence.

The NC cross–section incorporating leading–order QCD corrections is given by

$$\frac{d^2\sigma^{e^\pm p\to e^\pm X}}{dx\,dQ^2}(x,Q^2) = \frac{2\pi\alpha^2}{Q^4}\left[\left(1+(1-y)^2\right)\mathcal{F}_2^{\rm NC} \mp \left(1-(1-y)^2\right)\mathcal{F}_3^{\rm NC}\right] \quad (4)$$

$$\mathcal{F}_2^{\rm NC}(x,Q^2) = \sum_{q=d,u,s,c,b} A_q(Q^2)\cdot\left[q(x,Q^2)+\bar{q}(x,Q^2)\right] \quad (5)$$

$$\mathcal{F}_3^{\rm NC}(x,Q^2) = \sum_{q=d,u,s,c,b} B_q(Q^2)\cdot\left[q(x,Q^2)-\bar{q}(x,Q^2)\right] \quad (6)$$

$$A_q(Q^2) = \frac{1}{2}\left[(V_{\rm L}^q)^2+(V_{\rm R}^q)^2+(A_{\rm L}^q)^2+(A_{\rm R}^q)^2\right] \quad (7)$$

$$B_q(Q^2) = \left[(V_{\rm L}^q)(A_{\rm L}^q)-(V_{\rm R}^q)(A_{\rm R}^q)\right] \quad (8)$$

with

$$\begin{aligned} V_{L,R}^q &= Q_q - (v_e \pm a_e)\,v_q\,P_Z \\ A_{L,R}^q &= \quad -(v_e \pm a_e)\,a_q\,P_Z \\ v_f &= \frac{I_f^3 - 2\sin^2\theta_W\,Q_f}{2\sin\theta_W\cos\theta_W}, \quad a_f = \frac{I_f^3}{2\sin\theta_W\cos\theta_W} \\ P_Z &= \frac{Q^2}{Q^2+M_Z^2}\,. \end{aligned} \quad (9)$$

In eq. (9), the plus sign applies to the left–handed coefficients (index L), v_f and a_f are the SM vector and axial–vector coupling constants of a fermion f, Q_f and I_f^3 denote its charge and third component of the weak isospin, and M_Z and θ_w are the Z mass and the Weinberg angle. The top quark contribution to the NC ep cross–section is completely suppressed at HERA energies.

For CC scattering, the corresponding cross–section is

$$\frac{d^2\sigma^{e^\pm p\to{\overset{\leftrightarrow}{\nu}}^{\rightarrow}X}}{dx\,dQ^2} = \frac{\pi\alpha^2}{8\sin^4\theta_W\,Q^4}\left[\left(1+(1-y)^2\right)\mathcal{F}_2^{\rm CC\pm} \mp \left(1-(1-y)^2\right)\mathcal{F}_3^{\rm CC\pm}\right] \quad (10)$$

$$\mathcal{F}_{2,3}^{\rm CC+} = \left(\frac{Q^2}{Q^2+M_W^2}\right)^2\cdot\left[\sum_{q=d,s} q(x,Q^2) \pm \sum_{q=u,c}\bar{q}(x,Q^2)\right] \quad (11)$$

$$\mathcal{F}_{2,3}^{\rm CC-} = \left(\frac{Q^2}{Q^2+M_W^2}\right)^2\cdot\left[\sum_{q=u,c} q(x,Q^2) \pm \sum_{q=d,s}\bar{q}(x,Q^2)\right] \quad (12)$$

with M_W being the W mass. Note that electrons and positrons couple to quarks and antiquarks of different flavor, as required by charge conservation: $Q_e + Q_q = Q_{q'}$. Quark mass effects are neglected in eqs. (11) and (12). The tiny contributions of bottom and top quarks to the cross–section are ignored.

In order to evaluate the above cross–sections, two sets of ingredients are needed: the electroweak parameters like $\sin^2\theta_w$, M_Z etc., which are known to high precision from the e^+e^- experiments at LEP and SLAC, and the PDF's. The latter have been parameterized by several theory groups (see e.g. ref. 14–16) using fits to the relevant data (mainly measurements of DIS structure functions and cross–sections) in which the next–to–leading order QCD evolution is imposed. A typical set of PDF's is shown in Fig. 6. It becomes obvious that at $x \gtrsim 0.1$ the valence quark distributions (i.e. $q_v \equiv q - \bar{q}$) are dominant, and also that the d/u ratio falls far below the naïve expectation of 0.5 at high x.

From eqs. (4-9) it can be seen that at $Q^2 \ll M_Z^2$, the NC cross–section is dominated by the photon exchange, and the five "light" quark flavors contribute in proportion to their charge squares. In particular, u and d enter in the combination $4u+d$. For higher Q^2, this relation becomes y–dependent, but the u contribution remains to be the largest one. For CC scattering, (see eqs. (4-12)) $e^+(e^-)$ beams "see" d–type quarks and u-type antiquarks (u–type quarks and d-type antiquarks). In particular, e^+p CC scattering at high x is dominated by the d distribution. For $Q^2 \gtrsim M_W^2, M_Z^2$ the contributions of Z and W exchange

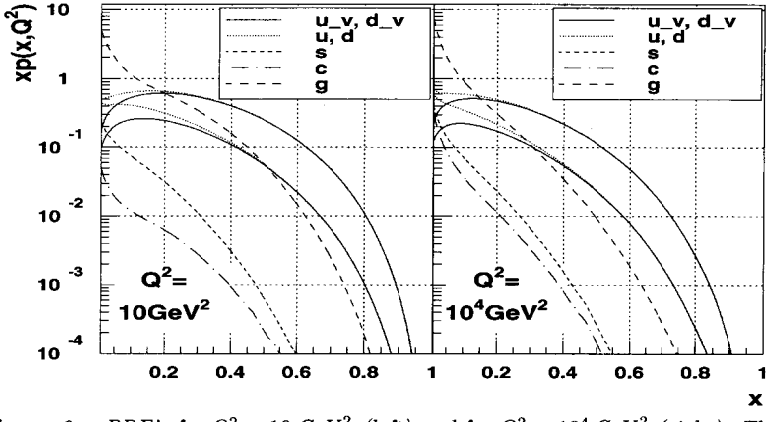


Figure 6: *PDF's for $Q^2 = 10\ GeV^2$ (left) and for $Q^2 = 10^4\ GeV^2$ (right). The larger of the two valence distributions represents u_v, the differences between the dotted and the solid lines indicate the u and d sea quark distributions.*

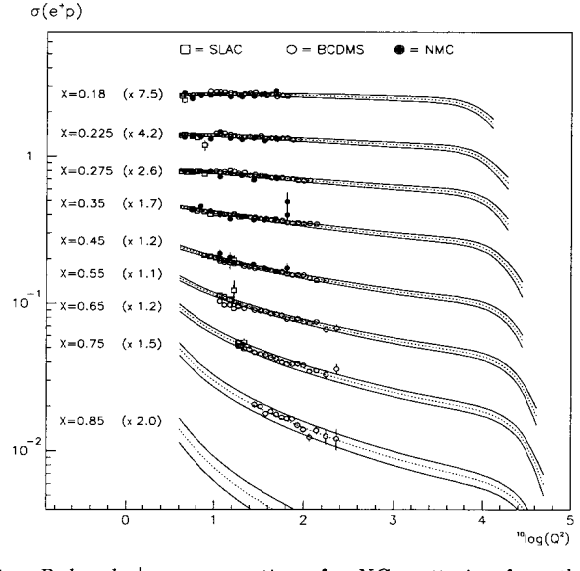


Figure 7: *Reduced e^+p cross-sections for NC scattering from the PDF fit by M.Botje (ZEUS). The error bands indicate the overall cross-section uncertainties resulting from propageting the errors on the input data and on α_s through the fit procedure.*

to the DIS cross–sections become of similar magnitude as that of the photon exchange. The HERA experiments are the first having the possibility to observe this manifestation of electroweak unification in the space–like regime.

The uncertainties of the cross–sections calculated from eqs. (4) and (10) (or rather from their equivalents incorporating the next–to–leading order QCD corrections and the full one–loop electroweak corrections) are dominated by the PDF's, more specifically by the experimental errors of the input data to the PDF fits, and by theoretical uncertainties related to higher–order corrections, to the treatment of the heavy quark PDF's, and to the value of α_s. In order to obtain a realistic PDF error estimate, M.Botje from the ZEUS collaboration has performed a QCD fit[17] to the relevant fixed–target data and propagated the sources of uncertainties into errors on the reduced NC cross–section,

$$\tilde{\sigma}^{e^\pm p \to e^\pm X}(x, Q^2) = \frac{xQ^4}{2\pi\alpha^2(1+(1-y)^2)} \cdot \frac{d^2\sigma^{\rm NC}}{dx\, dQ^2} \qquad (13)$$

(see Fig. 7). The resulting overall cross–section uncertainty amounts to about $\pm 7\%$ at high x and Q^2. A similar study for e^+p CC reactions yielded larger uncertainties which rapidly grow with Q^2 and exceed 20% for $Q^2 \gtrsim 2 \cdot 10^4\,\mathrm{GeV}^2$ or $x \gtrsim 0.5$.

4 Cross–Section Measurements

Both ZEUS[5,6] and H1[7] reported new preliminary cross–section measurements for high–Q^2 NC and CC DIS at the ICHEP98 conference in Vancouver. These analyses are based on the full e^+p data sets collected until end of 1997 (integrated luminosities of $46.6\,\mathrm{pb}^{-1}$ for ZEUS and $37.0\,\mathrm{pb}^{-1}$ for H1), and include substantial improvements of data statistics, detector calibration, analysis methods, and Monte Carlo simulation as compared to the status of summer 1997.[18]

NC Cross–Sections:

For the NC analyses, ZEUS considers reactions with $Q^2 > 400\,\mathrm{GeV}^2$ and $y_e < 0.95$ (y_e being the value of y as reconstructed from the measured momentum of the scattered electron). H1 applies the cuts $Q^2 > 200\,\mathrm{GeV}^2$ and $y_e < 0.9$. The final event samples consist of 38000 events (ZEUS) and 75000 events (H1).

On the next pages, the latest HERA NC e^+p cross–section measurements are presented: $d\sigma/dQ^2$ from ZEUS (Fig. 8) and H1 (Fig. 9), $d\sigma/dx$ from ZEUS (Fig. 10) and the reduced cross–section $\tilde{\sigma}(x, Q^2)$ from H1 (Fig. 11). The cross–sections are radiatively corrected to the Born–level. The systematic errors are added in quadrature to the statistical ones and are mainly caused by uncertainties of the energy scale calibration[d], the detector simulation, the trigger and vertex–reconstruction efficiencies, and the luminosity measurement. None of these sources is dominant, and the combined systematic uncertainties amount to a few percent. For $Q^2 \gtrsim 1000\,\mathrm{GeV}^2$, the precision of the analyses is statistics–limited.

In general, very good agreement is observed between measured and expected cross–sections. Possible exceptions are the region $Q^2 \gtrsim 20000\,\mathrm{GeV}^2$, where both ZEUS and H1 observe a higher cross–section than expected (see Figs. 8 and 9), and the x–interval around $x = 0.45$, where the reduced cross–section of H1 (Fig. 11) is above the expectation[e]. No anomaly around $x = 0.45$ is apparent in the ZEUS data of Fig. 10. Clearly, these observations are related to the high–Q^2 excess reported in refs. 3 and 4. Without further discussion, we quote the relevant event numbers: for $Q^2 > 35000\,\mathrm{GeV}^2$, ZEUS observes 2 events and expects 0.29 ± 0.02 (was 2 events vs. 0.15 ± 0.01 expected in 1994–96 data). For $|M - 200\,\mathrm{GeV}| < 12.5\,\mathrm{GeV}$ (corresponding to $0.39 < x < 0.50$), H1 now observes 8 events and expects 3.0 ± 0.5 (was 7 events vs. 0.95 ± 0.18 expected in the 1994–96 data).

CC Cross–Sections:

For the measurements of the CC cross–sections ZEUS uses the event sample with $p_t > 10\,\mathrm{GeV}$ (p_t being the net transverse momentum of the event), $Q^2 > 400\,\mathrm{GeV}$

[d] This affects H1 more than ZEUS who use the energy–scale independent double–angle method, which in turn suffers from systematic effects related to the hadronic measurements.

[e] Note that the statistical errors in the cross–section plots are calculated from the numbers of observed events and are therefore not suited to assess the significance of a possible excess, which has to be evaluated using confidence intervals based on the numbers of expected events.

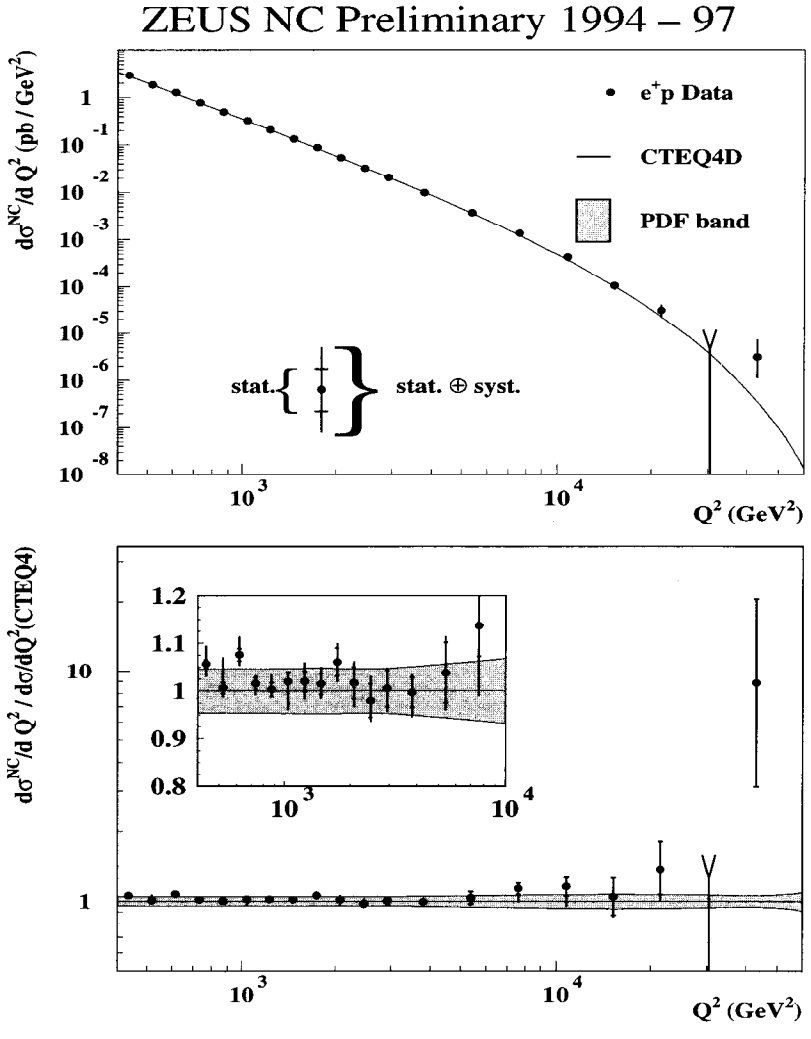


Figure 8: *In the top plot, the points with error bars indicate the ZEUS results for the radiatively corrected differential NC e^+p cross–section $d\sigma/dQ^2$. The line is the theoretical prediction, calculated by integrating eq. (4) over x. The bottom plot shows the ratio of measured to expected cross–sections, with the shaded band indicating the PDF uncertainty (see Fig. 7).*

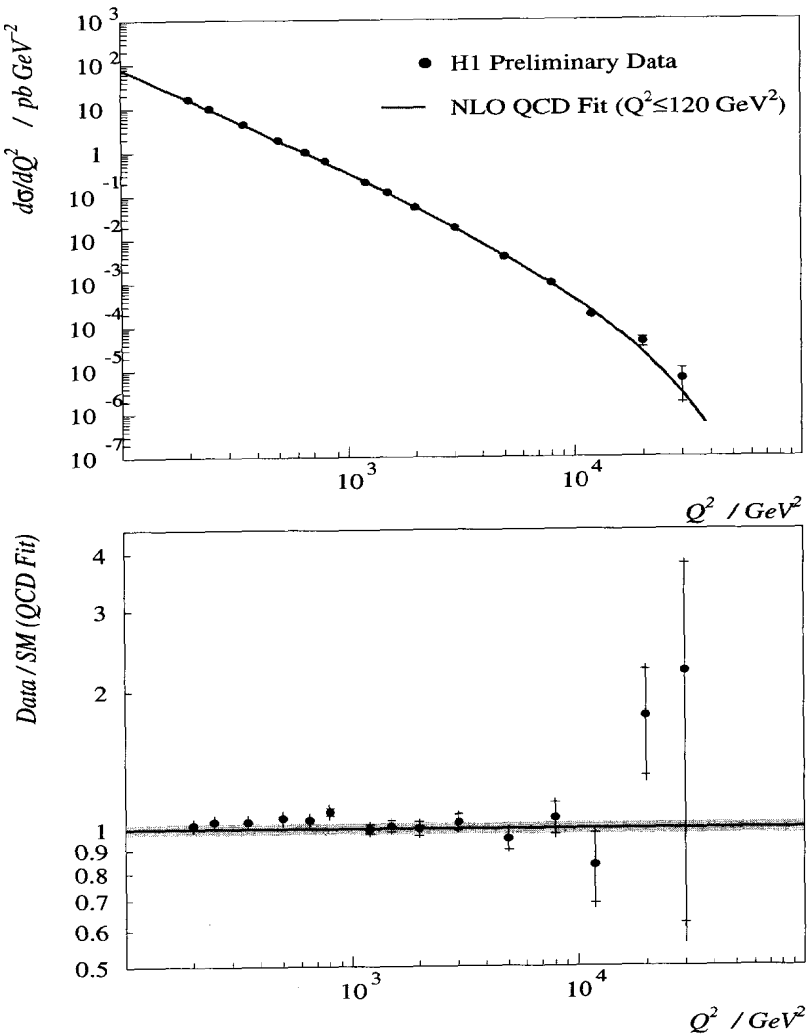


Figure 9: *In the top plot, the points with error bars indicate the H1 results for the radiatively corrected differential NC e^+p cross-section $d\sigma/dQ^2$. The line is the theoretical prediction, calculated by integrating eq. (4) over x. The bottom plot shows the ratio of measured to expected cross-sections, with the shaded band indicating the H1 luminosity uncertainty.*

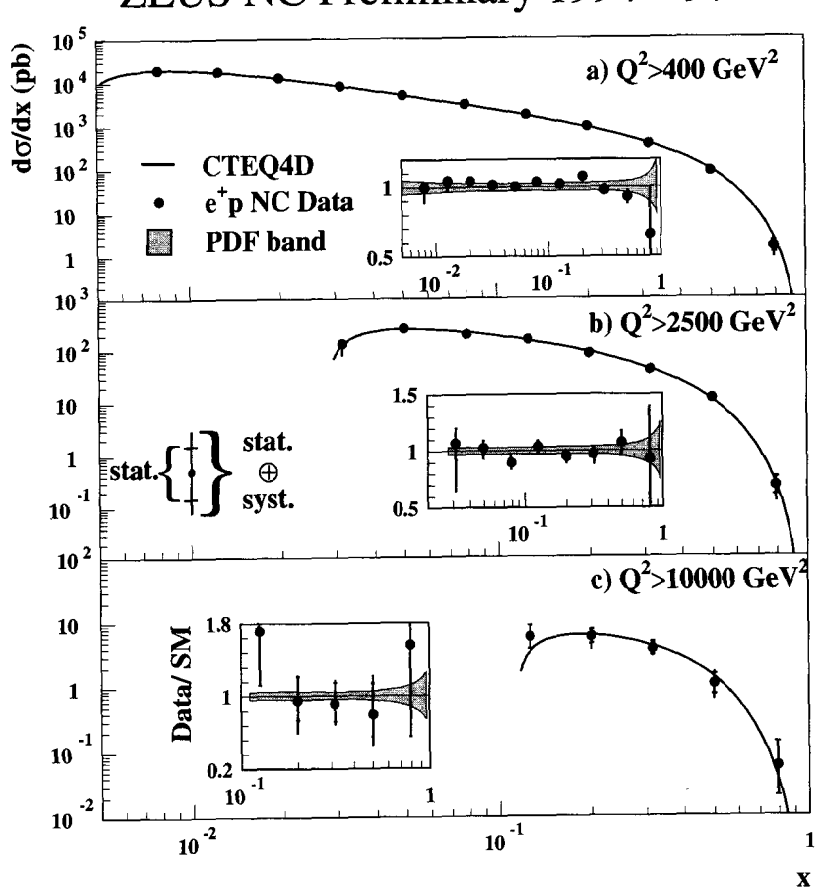


Figure 10: *Radiatively corrected differential NC e^+p cross-section $d\sigma/dx$ as measured by ZEUS, for three different lower Q^2-cutoffs. The points with error bars are the experimental results, the lines show the theoretical predictions. The insert plots show the ratio of measured to expected cross-sections, with the shaded band indicating the PDF uncertainty (see Fig. 7).*

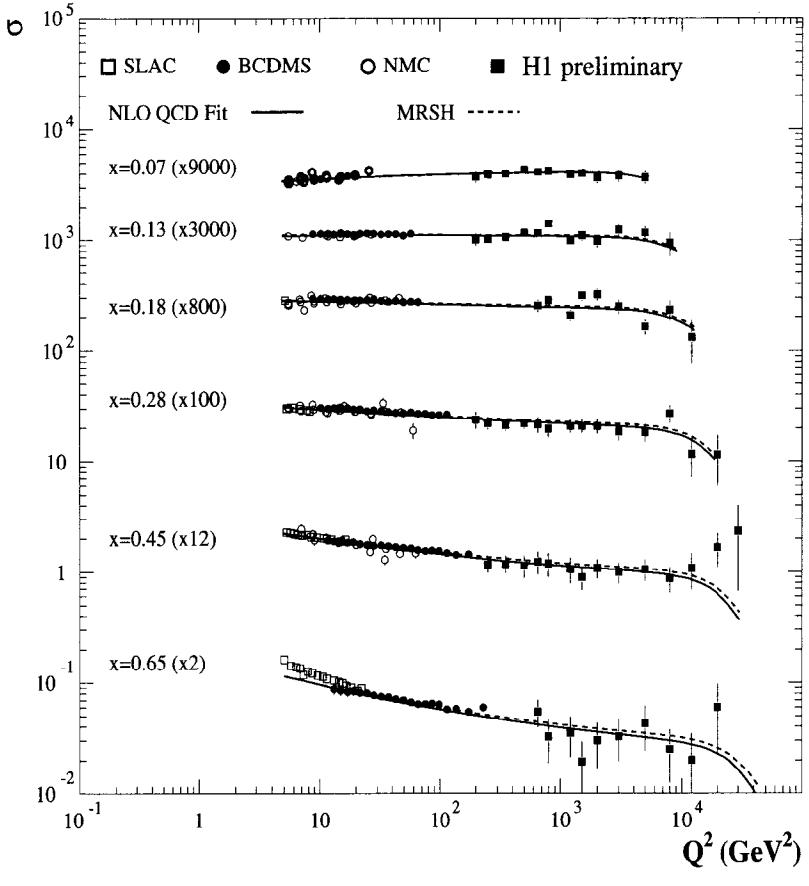


Figure 11: *Radiatively corrected reduced NC e^+p cross-section $\tilde{\sigma}(x, Q^2)$ as measured by H1, as a function of Q^2 for different fixed values of x (big dots with error bars). Also indicated are corresponding results of fixed-target experiments (small symbols). The lines indicate the theoretical predictions, derived (i) from the PDF set MRSH[19] (dashed), and (ii) from a QCD-fit performed by the H1 collaboration (solid).*

and $y < 0.9$ (altogether 869 events). H1 requires $p_t > 12\,\text{GeV}$, $x > 0.01$ and $y < 0.9$ (656 events). For both experiments, the systematic uncertainty of the cross-section measurement is dominated by the hadronic energy scale uncertainty.

On the following pages, the differential cross-sections $d\sigma/dQ^2$ from H1 (Fig. 12) and ZEUS (Fig. 13) as well as $d\sigma/dx$ (Fig. 14) and $d\sigma/dy$ (Fig. 15) from ZEUS are shown. Again, there seems to be a weak indication for an excess of measured over expected cross-sections at highest Q^2 (see Figs. 12 and 13) and highest x (Fig. 14). Figure 15 may indicate that this excess is related to a y-behavior which is slightly softer than expected. Indeed one may speculate whether these (admittedly not very significant) observations hint to an underestimation of the d distribution at large x, which could both explain the tendencies in the experimental data and be in accordance with recent phenomenological and theoretical work.[20,21]

The Q^2 dependence of the CC cross-section is proportional to the square of the W propagator (see eq. (10)) and can hence be used to infer the W mass. Although the resulting precision is inferior to that of the Tevatron and LEP results, it constitutes an important check of the SM consistency in the regime of highly virtual, space-like W propagators. The values obtained by H1 ($M_W = 81.2 \pm 3.3\,(\text{stat.}) \pm 4.3\,(\text{syst.})$ GeV) and by ZEUS ($M_W = 78.6^{+2.5}_{-2.4}\,(\text{stat.})^{+3.3}_{-3.0}\,(\text{syst.})$ GeV) are in perfect agreement with the current world average.

Comparison of NC and CC Cross-Sections:

A comparison of $d\sigma/dQ^2$ for NC and CC e^+p reactions as measured by both experiments is shown in Fig. 16. Note that the ZEUS and the H1 results agree well over the full cross-section range of more than five decades. For $Q^2 \gtrsim 10000\,\text{GeV}^2$, the NC and the CC differential cross-sections are indeed similar in magnitude and Q^2-slope, indicating the restoration of full electroweak symmetry at this scale.

5 Scenarios Beyond the Standard Model

In this section, we will constrain ourselves to hypothesized processes beyond the SM producing event signatures which are indistinguishable from NC scattering on the event-by-event basis, and which have matrix elements which in fact interfere with the SM ones. There are two such scenarios which have been tested by both collaborations: contact interactions and the on-shell production of a heavy

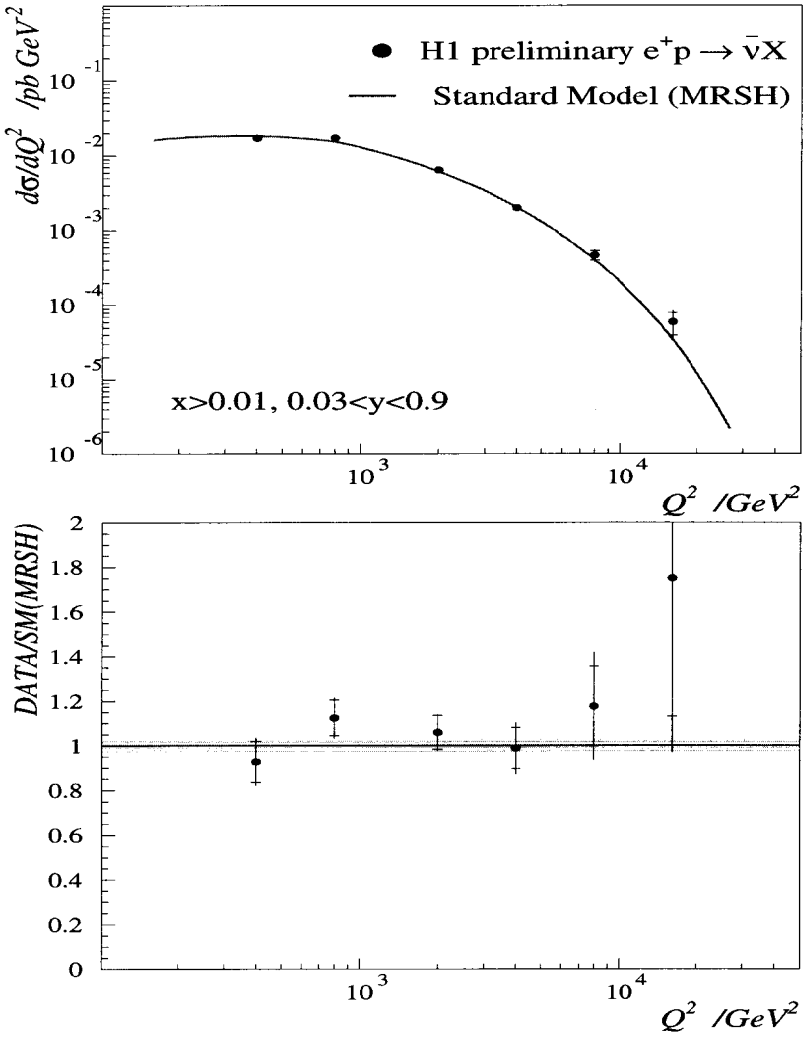


Figure 12: *In the top plot, the points with error bars indicate the H1 results for the radiatively corrected differential CC e^+p cross-section $d\sigma/dQ^2$, restricted to the kinematic region $x > 0.01$ and $y < 0.9$. The line is the theoretical prediction, calculated by integrating eq. (10) over x. The bottom plot shows the ratio of measured to expected cross-sections, with the shaded band indicating the H1 luminosity uncertainty.*

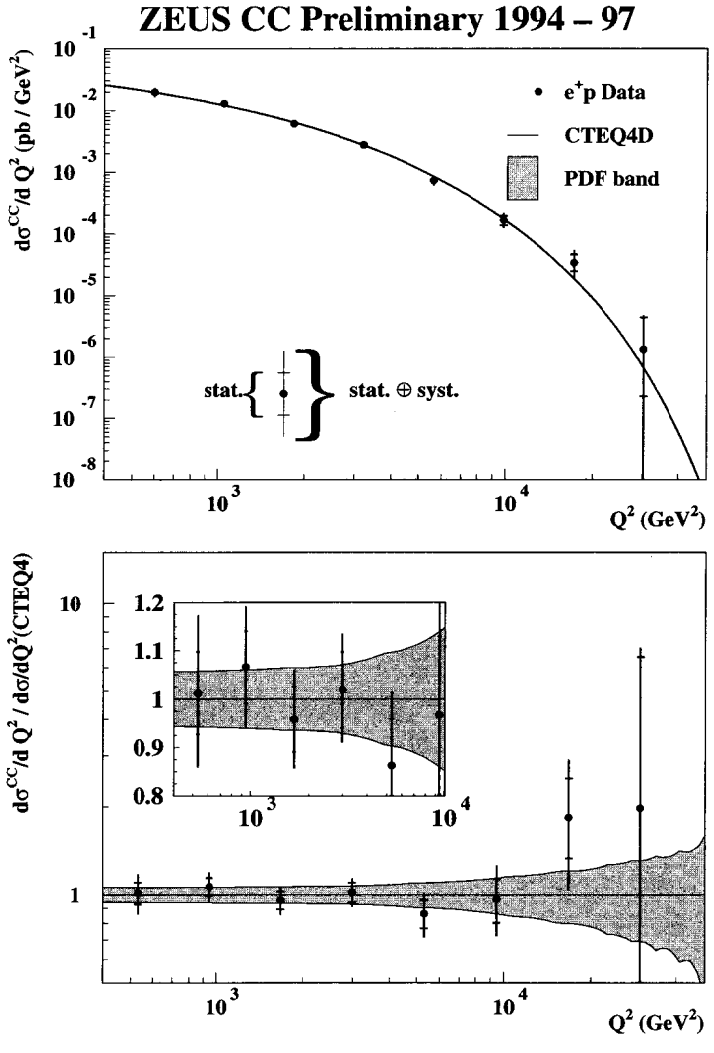


Figure 13: *In the top plot, the points with error bars indicate the ZEUS results for the radiatively corrected differential CC e^+p cross-section $d\sigma/dQ^2$. The line is the theoretical prediction, calculated by integrating eq. (10) over x. The bottom plot shows the ratio of measured to expected cross-sections, with the shaded band indicating the PDF uncertainty.*

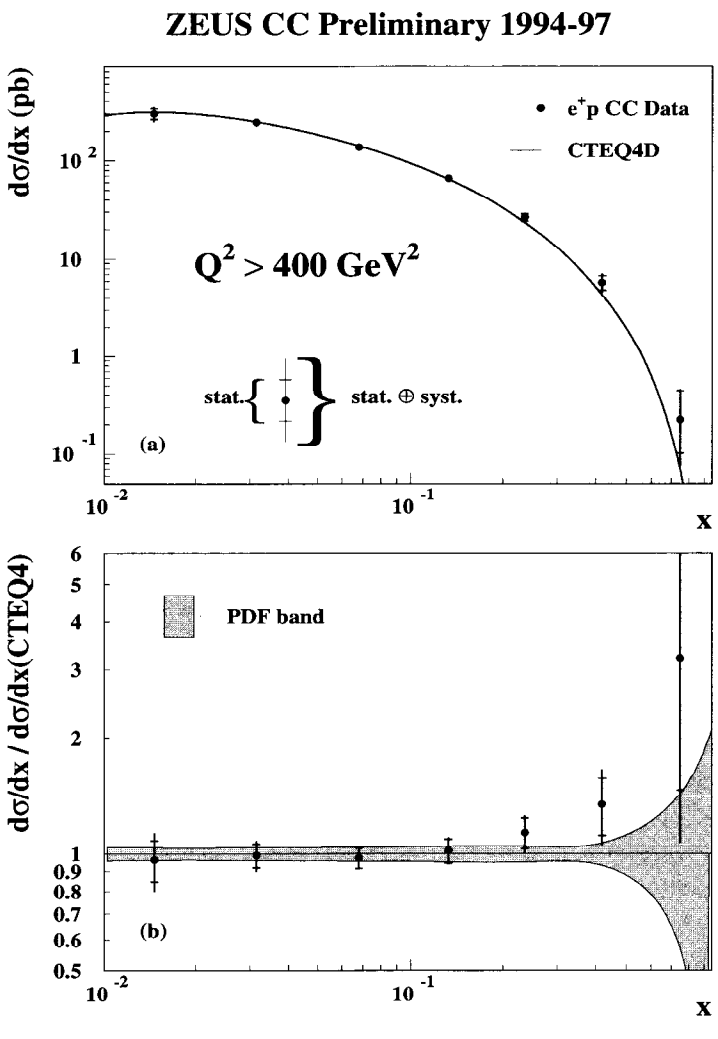


Figure 14: *In the top plot, the points with error bars indicate the ZEUS results for the radiatively corrected differential CC e^+p cross-section $d\sigma/dx$. The line is the theoretical prediction, calculated by integrating eq. (10) over Q^2. The bottom plot shows the ratio of measured to expected cross-sections, with the shaded band indicating the PDF uncertainty.*

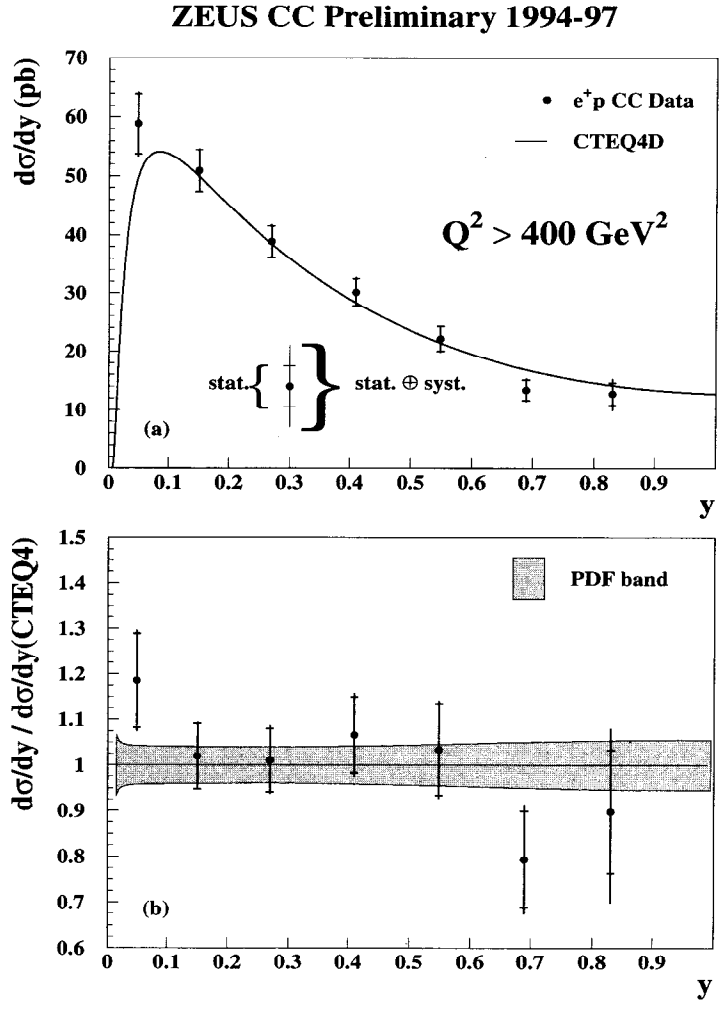


Figure 15: *In the top plot, the points with error bars indicate the ZEUS results for the radiatively corrected differential CC e^+p cross-section $d\sigma/dy$. The line is the theoretical prediction. The bottom plot shows the ratio of measured to expected cross-sections, with the shaded band indicating the PDF uncertainty.*

e^+p resonance. We will not consider searches for pair production of supersymmetric MSSM particles, for excited fermions, or for new physics inducing CC–like reactions or final states with muons having high transverse momentum.

Search for Contact Interactions:

A wide class of hypothesized new interactions at characteristic mass scales in the TeV range would modify the differential HERA DIS cross–sections in a way which can be parameterized by effective four–fermion contact interactions (CI) coupling electrons to quarks ($eeqq$), in much the same way as four–fermion interactions were once used to describe weak interactions.

The Lagrangian for a ($eeqq$) CI is given by

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\rm SM} + \frac{g^2}{\Lambda^2}\Big[& \eta_s^q(\bar{e}_L e_R)(\bar{q}_L q_R) + \eta_s^{q\prime}(\bar{e}_L e_R)(\bar{q}_R q_L) + \text{h.c.} && \text{scalar} \\ & + \eta_{LL}^q(\bar{e}_L\gamma^\mu e_L)(\bar{q}_L\gamma_\mu q_L) + \eta_{LR}^q(\bar{e}_L\gamma^\mu e_L)(\bar{q}_R\gamma_\mu q_R) \\ & + \eta_{RL}^q(\bar{e}_R\gamma^\mu e_R)(\bar{q}_L\gamma_\mu q_L) + \eta_{LR}^q(\bar{e}_R\gamma^\mu e_R)(\bar{q}_R\gamma_\mu q_R) && \text{vector} \\ & + \eta_T^q(\bar{e}_L\sigma^{\mu\nu} e_R)(\bar{q}_L\sigma_{\mu\nu} q_R) + \text{h.c.}\Big] , && \text{tensor}\end{aligned} \quad (14)$$

where g is the coupling, Λ is the effective mass scale, the η_i^q determine the relative size and sign of the individual terms, and the indices L, R denote the helicity of the respective fermions. ZEUS and H1 only consider the vector terms because strong limits beyond the HERA sensitivity have already been placed on the scalar and on the tensor terms.[22] Since g and Λ always enter in the combination g^2/Λ^2, we use the convention $g^2 = 4\pi$ to relate CI strengths to the corresponding effective mass scales. Furthermore, each of the $\eta_{\alpha\beta}^q$ ($\alpha, \beta = L, R$) is assumed to be either zero or ± 1, and to be flavor–symmetric (i.e. $\eta^d = \eta^s = \eta^b$ and $\eta^u = \eta^c = \eta^t$).

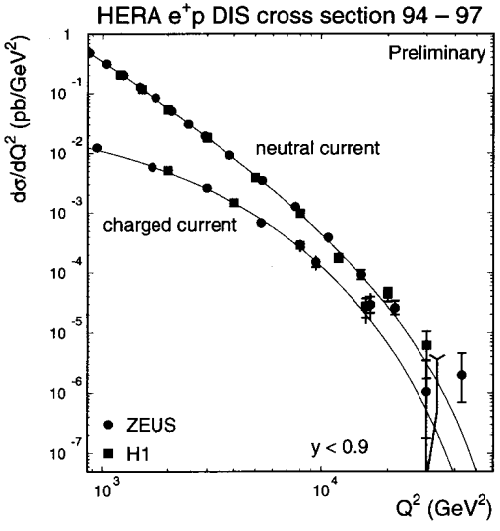


Figure 16: *Comparison of the NC and CC cross–sections $d\sigma/dQ^2$ for $y < 0.9$. The circles indicate the ZEUS measurements, the squares those of H1, the lines represent the theoretical predictions. All cross–sections are radiatively corrected.*

CI	η_{LL}^u	η_{LR}^u	η_{RL}^u	η_{RR}^u	η_{LL}^d	η_{LR}^d	η_{RL}^d	η_{RR}^d	studied by
VV	+	+	+	+	+	+	+	+	ZEUS, H1
AA	+	−	−	+	+	−	−	+	
VA	+	−	+	−	+	−	+	−	
X1	+	−	0	0	+	−	0	0	ZEUS
X2	+	0	+	0	+	0	+	0	
X3	+	0	0	+	+	0	0	+	
X4	0	+	+	0	0	+	+	0	
X5	0	+	0	+	0	+	0	+	
X6	0	0	+	−	0	0	+	−	
U1	+	−	0	0	0	0	0	0	ZEUS
U4	0	+	+	0	0	0	0	0	
U5	0	+	0	+	0	0	0	0	
LL	+	0	0	0	+	0	0	0	H1
LR	0	+	0	0	0	+	0	0	
RL	0	0	+	0	0	0	+	0	
RR	0	0	0	+	0	0	0	+	

Table 1: *CI scenarios studied by ZEUS and H1. Each line describes two different η patterns, corresponding to an overall sign $+1$ or -1.*

Different patterns of the $\eta_{\alpha\beta}^q$ define different CI scenarios, corresponding to distinct physics models. The CI scenarios considered in the HERA analyses are summarized in Tab. 1. Except for the purely chiral terms, all scenarios obey $\eta_{LL}^q + \eta_{LR}^q - \eta_{RL}^q - \eta_{RR}^q = 0$ in order to avoid strong CI limits[23–25] implied by recent measurements of parity–violating transition amplitudes in cesium atoms.[26]

The cross–section modification in presence of CI affects mainly $d\sigma/dQ^2$, which acquires two extra terms: one proportional to $\pm Q^2/\Lambda^2$ (from the SM–CI interference), and one proportional to $+Q^4/\Lambda^4$ (pure CI). In order to search for such signatures, ZEUS[10] used log–likelihood techniques based on the comparison of the measured Q^2 and (x, y) distributions to the corresponding Monte Carlo (MC) simulations, in which CI's were taken into account by reweighting the events. No

Type	95% C.L. lower limits on Λ [TeV]						
	ZEUS,H1		CDF	OPAL,ALEPH,L3			APV (u/d)
	ref. 10	ref. 11	ref. 27	ref. 28	ref. 29	ref. 30	ref. 24
$VV+$	4.9	4.5	3.5	4.1	6.7	3.2	—
$VV-$	4.6	2.5	5.2	5.7	7.4	3.9	—
$AA+$	2.0	2.0	3.8	6.3	7.4	4.3	—
$AA-$	4.0	3.8	4.8	3.8	8.2	2.9	—
$VA+$	2.8	2.6	—	—	—	—	—
$VA-$	2.8	2.8	—	—	—	—	—
$X1+$	1.8	—	—	—	—	—	—
$X1-$	3.0	—	—	—	—	—	—
$X2+$	3.9	—	—	—	—	—	—
$X2-$	1.9	—	—	—	—	—	—
$X3+$	2.8	—	—	4.4	6.9	3.2	—
$X3-$	1.5	—	—	3.8	7.7	2.8	—
$X4+$	4.5	—	—	3.1	2.9	2.4	—
$X4-$	4.1	—	—	5.5	4.5	3.7	—
$X5+$	3.8	—	—	—	—	—	—
$X5-$	3.0	—	—	—	—	—	—
$X6+$	3.0	—	—	—	—	—	—
$X6-$	1.9	—	—	—	—	—	—
$U1+$	2.9	—	—	—	—	—	—
$U1-$	1.6	—	—	—	—	—	—
$U4+$	4.4	—	2.3	—	—	1.8	—
$U4-$	4.6	—	3.2	—	—	2.2	—
$U5+$	3.6	—	—	—	—	—	—
$U5-$	4.2	—	—	—	—	—	—
$LL+$	—	2.4	2.5	4.4	5.6	3.0	7.4/7.9
$LL-$	—	1.2	3.7	2.8	6.4	2.1	11.7/12.3
$LR+$	—	3.0	2.8	3.3	3.0	2.4	7.4/7.9
$LR-$	—	1.5	3.3	3.6	3.2	2.6	11.7/12.3
$RL+$	—	3.0	2.9	2.5	2.3	2.0	11.7/12.3
$RL-$	—	1.6	3.2	4.9	4.0	3.2	7.4/7.9
$RR+$	—	2.4	2.6	3.0	4.1	2.3	11.7/12.3
$RR-$	—	1.2	3.6	3.9	4.5	2.7	7.9/7.4

Table 2: *Lower limits on Λ at 95% C.L. from the HERA experiments, compared to corresponding limits reported recently by CDF and by the LEP experiments, and to limits derived from measurements of parity-violating atomic transitions. The ZEUS, H1 and ALEPH results are preliminary.*

significant indications for CI's were found. One–sided 95% C.L. lower limits on Λ were set using MC techniques. H1 performed a χ^2–fit of the predicted to the measured $d\sigma/dQ^2$ and inferred two–sided 95% C.L. limits from the resulting changes in χ^2 with respect to the fit to the SM, which describes the data well.

In Tab. 2, the Λ limits are summarized and compared to corresponding results recently reported by CDF[27] ($q\bar{q} \to e^+e^-$) and the LEP experiments[28–30] ($e^+e^- \to q\bar{q}$), and to the limits derived from the atomic parity violation results (see ref. 24 and references therein). All LEP and CDF limits except those for for the U4 scenario assume flavor–symmetry. The HERA and the other experiments are all sensitive to CI at mass scales of a few TeV. Limits for the X1, X2, X5, X6, U1 and U5 scenarios are derived only by ZEUS. The relative sensitivity to different CI scenarios depends on the SM–CI interference sign which is opposite for e^+p and for e^+e^- and $p\bar{p}$ reactions. Where available, the CERN limits mostly exceed those of HERA and CDF. However, this conclusion depends on the assumption of flavor–symmetric CI; CI coupling e.g. only to the first–generation quarks would give almost the same results for ep or $p\bar{p}$ scattering but significantly different limits in the case of LEP.

Search for e^+p resonance production:

The HERA kinematic range allows to access eq reactions at a center–of–mass energy of 200 GeV and beyond, where new particles coupling to electrons and quarks (leptoquarks, LQ) might be resonantly produced. If the LQ decayed back to eq, the events would be NC–like. The experimental signature would be a peak in $M = \sqrt{xs}$ and a y distribution which is either flat (scalar LQ) or proportional to $(1-y)^2$ (vector LQ). The LQ cross–section depends on the LQ–e–q coupling (λ), on the LQ mass ($M_{\rm LQ}$), and on the branching ratio of the decay LQ$\to eq$ (β). Generally, two types of LQ's are distinguished: (i) the "classical" LQ's[31] which only couple to eq or $\nu q'$ and have $\beta = 0, 0.5$ or 1; (ii) squarks in R_P–violating extensions of Supersymmetry, which have additional R_P–conserving decay modes and therefore can have $\beta \ll 1$ (for details see e.g. ref. 32). The type–(ii) LQ's avoid the mass limits of the DØ collaboration at the Tevatron who have searched for LQ pair production and exclude at 95% C.L. scalar[33]/vector[34] LQ's with $M_{\rm LQ} <$

225/298 GeV ($M_{LQ} < 204/270$ GeV) for $\beta = 1$ ($\beta = 0.5$).[f]

Both ZEUS[8] and H1[9,36] have reported searches for LQ signals. Figure 17 shows the M–distribution of H1, for two different cuts in y_e. The excess of events around $x = 0.45$ seen in Fig. 11 is again evident at $M \approx 200$ GeV. From the comparison of the measured and expected M distributions, H1 sets limits on β as a function of M_{LQ} for fixed values of λ. As can be seen from Fig. 18, the HERA experiments still have a discovery window for a LQ with $\beta \lesssim 0.3$ and $M_{LQ} \gtrsim 200$ GeV.

The M distribution as measured by ZEUS[8] is shown in Fig. 19. Again, some excess around $M \approx 220$ GeV is visible. However, the kinematic properties of the events in this region do not exhibit significant indications for LQ production, and ZEUS hence sets limits on the corresponding cross–section $\sigma(e^+p \to \text{LQ})$. The 95% C.L. limit on $\beta \cdot \sigma(e^+p \to \text{LQ})$ resulting from a comparison of the measured and the predicted M distributions is displayed in Fig. 20.

6 Conclusion and Outlook

The HERA experiments ZEUS and H1 have both collected a total of approximately 40 pb^{-1} of e^+p collision data in the years 1994–1997. These data allow for the first time to study deep–inelastic scattering in the region of high Q^2 above several 100 GeV with good statistical precision.

Particular interest in e^+p scattering at $Q^2 \gtrsim 15000$ GeV^2 was stimulated in early 1997, when both experiments reported an excess of events in their 1994–1996 data samples above the expectations in this kinematic regime. Meanwhile the data samples have about doubled, and careful measurements of the high–Q^2 differential cross–sections, both for NC and for CC reactions are available. Very good agreement between measured and expected cross–sections is generally observed, with the possible exception of the highest x and Q^2 intervals, where the data are above the prediction. The data allow for the first time to observe the restoration of full electroweak symmetry at a scale of $Q^2 \sim M_W^2, M_Z^2$.

It has been hypothesized in a series of papers that the excess at high Q^2 may be a first glimpse at new physics beyond the Standard Model. Dedicated searches for

[f] A recent combined analysis[35] of the CDF and DØ data yields a mass limit of $M_{LQ} > 242$ GeV for scalar LQ's with $\beta = 1$.

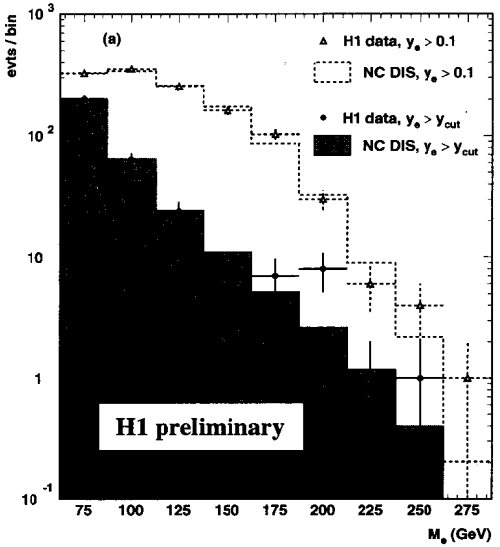


Figure 17: *Distributions of M from H1 for two different cuts on y_e (error bars). The dashed (solid) histograms represent the expected distributions for $y_e > 0.1$ ($y_e > y_{cut}$, where y_{cut} is a function of M).*

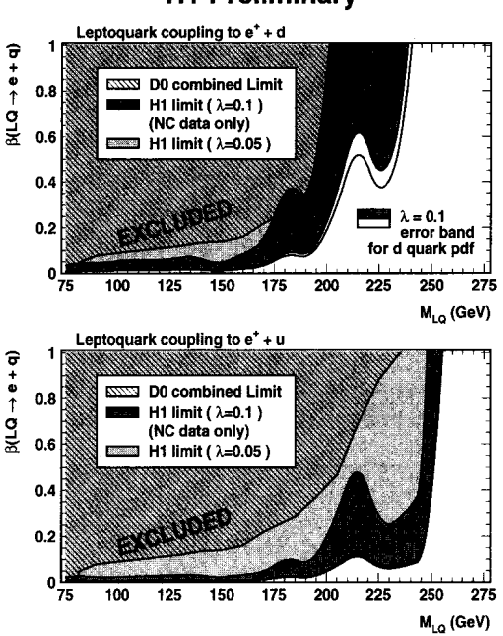


Figure 18: *Limits on LQ production from H1, for LQ's coupling to e^+d (top) and to e^+u (bottom). For different assumptions on the coupling λ, the shaded regions indicate as a function of M_{LQ} the values of the branching ratio β which are excluded at 95% C.L. The DØ exclusion limits are marked by the hatched area.*

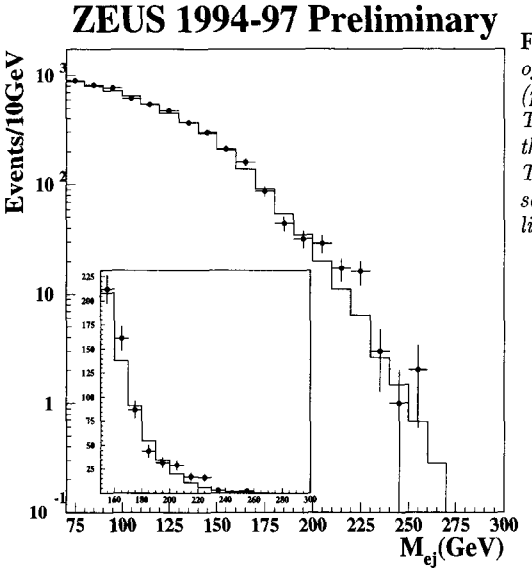


Figure 19: *Distribution of M measured by ZEUS (points with error bars). The histogram represents the expected distribution. The insert plot shows the same distributions with a linear vertical scale.*

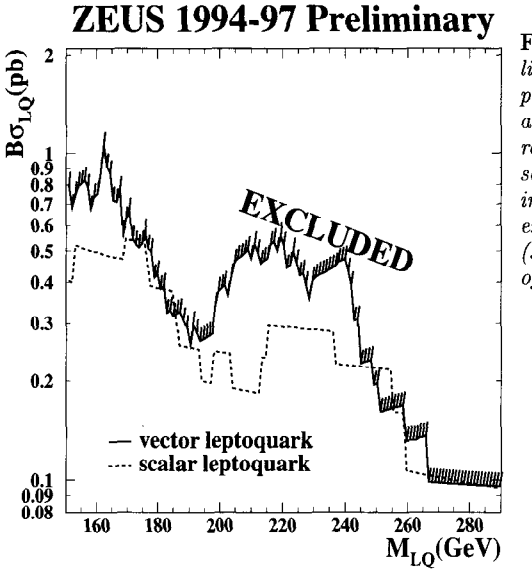


Figure 20: 95% *C.L. limits on the product of LQ production cross-section and β (here denoted B) reported by ZEUS. The solid (dashed) histogram indicate the borders of the excluded regions for vector (scalar) LQ's as function of M_{LQ}.*

signatures of different such scenarios have been performed. No evidence for new processes has been found, and new limits on contact interactions and positron-quark resonance production are reported. It remains to be studied with increased data statistics whether the observed excess is a statistical fluctuation or a sign of something new.

Currently HERA is operated with electrons (e^-), and these data together with the existing e^+p sets will allow a detailed study of high-Q^2 structure functions and parton distributions and will help to reduce the uncertainties of the theoretical cross-section predictions. In a shutdown of about nine months in 1999/2000, HERA will be upgraded to deliver significantly higher luminosity and also longitudinally polarized e beams to the experiments. From 2001 onwards, the e polarization and the expected integrated luminosities of more than $100\,\mathrm{pb}^{-1}$ per experiment and year will allow to study the region of highest Q^2 with yet much higher precision and with an additional degree of freedom. We await with excitement the measurements and possible discoveries which will be made possible by these data.

Acknowledgments. None of the research reported here would have been possible without the dedicated efforts of the HERA machine physicists and technicians, and of all those who contributed to the design, construction, maintenance and operation of the detectors. I'm indebted to Martin Erdmann and Stefan Schlenstedt for the careful reading of the manuscript. Last but not least, I would like to thank the organizers of the SLAC Topical Conference for a very inspiring meeting and a pleasant stay in California.

References

[1] H1 Coll., I.Abt et al., Report DESY-93-103, DESY (1993);
H1 Coll., I.Abt et al., Nucl. Instrum. Methods **A 386** (1997) 310.

[2] ZEUS Coll., ed. U. Holm, Status Report, 1993.

[3] ZEUS Coll., J.Breitweg et al., Z. Phys. **C 74** (1997) 207.

[4] H1 Coll., C.Adloff et al., Z. Phys. **C 74** (1997) 191.

[5] ZEUS Coll., Subm. paper 753 to ICHEP98, Vancouver, July 23–29, 1998.

[6] ZEUS Coll., Subm. paper 751 to ICHEP98, Vancouver, July 23–29, 1998.

[7] H1 Coll., Subm. paper 533 to ICHEP98, Vancouver, July 23–29, 1998.

[8] ZEUS Coll., Subm. paper 754 to ICHEP98, Vancouver, July 23–29, 1998.

[9] H1 Coll., Subm. paper 579 to ICHEP98, Vancouver, July 23–29, 1998.

[10] ZEUS Coll., Subm. paper 753 to ICHEP98, Vancouver, July 23–29, 1998.

[11] H1 Coll., Subm. paper 584 to ICHEP98, Vancouver, July 23–29, 1998.

[12] G.Bernardi, Contribution to these proceedings.

[13] V.N.Gribov, L.N.Lipatov, Sov. J. Nucl. Phys. **15** (1972) 438;
L.N.Lipatov, Yad. Fiz. **20** (1974) 181, Sov. J. Nucl. Phys **20** (1975) 94;
G.Altarelli, G.Parisi, Nucl. Phys. **B 126** (1977) 298;
Yu.L.Dokshitser, Sov. Phys. JETP **46** (1977) 641.

[14] A.D.Martin, et al., Eur. Phys. J. **C 4** (1998) 463.

[15] CTEQ Coll., H.L.Lai et al., Phys. Rev. **D 55** (1997) 1280.

[16] M.Glück, E.Reya, A.Vogt, Z. Phys. **C 67** (1995) 433.

[17] M.A.J.Botje, Preprint NIKHEF–97–028, NIKHEF (1997), hep–ph/9707289.

[18] B.Straub, for the ZEUS and H1 Collaborations, in *Proc.* 18^{th} *Int. Symp. on Lepton-Photon Interactions (LP97), Hamburg*, 1997, p. 3.

[19] A.D.Martin, W.J.Stirling, R.G.Roberts, Preprint DTP 93–86/RAL–93–077, Univ. Durham/Rutherford Lab. (1993).

[20] W.Melnitchouk, A.Thomas, Acta Phys. Pol. **B 27** (1996) 1407;
W.Melnitchouk, A.W.Thomas, Phys. Lett. **B 377** (1996) 11;
W.Melnitchouk, J.C.Peng, Phys. Lett. **B 400** (1997) 220.

[21] U.K.Yang, A.Bodek, Q.Fan, Preprint UR–1519, Univ. Rochester (1998), hep–ph/9806457.

[22] P.Haberl, F.Schrempp, H.U.Martyn, in *Proc. Workshop Physics at HERA*, ed. W.Buchmüller, G.Ingelman, Hamburg, Germany, 1991, p. 1133.

[23] V.Barger et al., Phys. Rev. **D 57** (1998) 391.

[24] A.Deandrea, Phys. Lett. **B 409** (1997) 277.

[25] L.Giusti, A.Strumia, Phys. Lett. **B 410** (1997) 229.

[26] C.S.Wood et al., Science **275** (1997) 1759.

[27] CDF Coll., F.Abe et al., Phys. Rev. Lett. **79** (1997) 2198.

[28] OPAL Coll., G.Abbiendi et al., Eur. Phys. J. **C 6** (1999) 1.

[29] ALEPH Coll., ALEPH 98–060/CONF 98–031, subm. paper 906 to ICHEP98, Vancouver, July 23–29, 1998.

[30] L3 Coll., M.Acciarri et al., Phys. Lett. **B 433** (1998) 163.

[31] W.Buchmüller, R.Rückl, D.Wyler, Phys. Lett. **191 B** (1987) 442.

[32] G.Altarelli et al., Nucl. Phys. **B 506** (1997) 3.

[33] DØ Coll., B.Abbott et al., Phys. Rev. Lett. **80** (1998) 2051.

[34] J.A.Valls, in *Proc. 32. Recontres de Moriond: QCD and High–Energy Hadronic Interactions, Les Arcs, France*, 1997.

[35] CDF and DØ Coll., C.Grosso–Pilcher et al., hep–ex/9810015, 1998.

[36] H1 Coll., Subm. paper 580 to ICHEP98, Vancouver, July 23–29, 1998.

Search for New Particles at LEP

S. Rosier-Lees
LAPP (IN2P3-CNRS)
Annecy-Le-Vieux - France

Abstract

The LEP energy upgrade up to $\sqrt{s}$ =189 GeV has allowed us to extend substantially the potential of searches for new physics. Results on searches for Higgs bosons and supersymmetric particles obtained by the ALEPH, DELPHI, L3, and OPAL experiments are reported. No evidence of any signal is observed. Therefore, new limits on the Higgs boson masses as well as on the masses of the various supersymmetric particles are derived. They significantly improve those obtained either at LEP1 or LEP1.5. The LEP200 discovery potential for the neutral Higgs bosons is also shown.

© 1998 by S. Rosier-Lees.

1 Introduction

A wide range of searches for new particles has been performed at LEP and no deviation with respect to the Standard Model predictions was observed yet, unfortunately. Therefore, new limits on the masses of these particles, assuming different models, are set by the various experiments. In what follows, all the presented limits are derived at 95% C.L.

We present first the results of Higgs searches, which are interpreted in the framework of the Standard and Minimal Supersymmetric Models. Then the searches for supersymmetric particles are reviewed. In the latter case, results are interpreted within Supersymmetric Models like gravity SUSY breaking models or gauge mediated SUSY breaking models.

After collecting $\sim$ 150 pb^{-1} at LEP1 (1989 $\rightarrow$ 1995), each experiment (ALEPH, DELPHI, L3, OPAL) has accumulated data at LEP2: $\sim$ 5.5 pb^{-1} at $\sqrt{s}$ = 133 GeV (1995), $\sim$ 10 pb^{-1} at $\sqrt{s}$ = 161 GeV (1996), $\sim$ 10 pb^{-1} at $\sqrt{s}$ = 172 GeV (1996), and $\sim$ 55 pb^{-1} at $\sqrt{s}$ = 183 GeV (1997). This year, an amount of $\sim$ 175 pb^{-1} at $\sqrt{s}$ = 189 GeV (November 1998) is already recorded by each experiment, but the results including all the statistics for 1998 are not yet available; a fraction of the overall integrated luminosity is only used corresponding to the statistics collected just before the summer conference time (up to July 1998: $30 - 40$ pb^{-1}).

2 The Higgs Bosons

2.1 The Standard Model Higgs Boson

The search for the Standard Model Higgs boson is presently only performed at LEP. The current experimental lower limit for its mass m_H, obtained from large samples of Z boson decays at LEP1, is 63.9 GeV/c^{2}as reported in Ref. [1]. The accurate electroweak measurements, sensitive to the Higgs mass via the radiative corrections, exclude at 95% C.L. a Higgs boson heavier than 280 GeV/c^2 [2].

At LEP, the Higgs boson is expected to be produced through the "Higgs-strahlung" process $e^+e^- \rightarrow ZH$, the $e^+e^- \rightarrow WW$, and $e^+e^- \rightarrow ZZ$ contributions being still negligible at those center-of-mass energies. For a Higgs mass of 85 GeV, an amount of $\sim$ 20 events is expected per experiment at $\sqrt{s}$=183 GeV.

The background sources are distributed into three subsamples: the $\gamma\gamma$ interactions; the four fermion processes including in particular the WW, $We\nu$, ZZ final states; and the two fermion processes. The environments at LEP1 and LEP2 are different as summarized below:

- The fraction Signal(HZ)/background increases from 10^{-6} at LEP1 to $\sim 10^{-2}$ at LEP2 as both the Higgs boson production cross section increases and, above all, the main background from $e^+e^- \rightarrow Z \rightarrow q\bar{q}$ decreases by a factor of three in magnitude, allowing us to search for Higgs bosons through all the possible channels including the four jets channel.
- For $M_H < \sqrt{s} - M_Z$, on shell Z bosons are produced. This is an additional constraint very useful for the background rejection.
- However, compared to LEP1, background sources arising from the four fermion processes become more important, bringing in new irreducible background contaminations to the HZ search.

In this new context, the strategy for the Higgs boson searches changed compared to what was performed at LEP1. This search consists of four different channels, determined by the Higgs decay channels together with the Z decay channels. They are presented hereafter sorted according to their branching ratio.

- "The four jet channel": $e^+e^- \rightarrow HZ \rightarrow b\bar{b}b\bar{b}$ is produced with the highest rate ($\sim$ 20%). The signal in this channel consists of four hadronic jets, b flavored. The key point for this topology is the b-tagging technique which allows us to eliminate most of the background arising from $e^+e^- \rightarrow q\bar{q}$, WW, and ZZ. In addition, the invariant mass of two of the four jets has to be consistent with the Z boson mass.
- The "missing energy channel" due to the reaction $e^+e^- \rightarrow HZ \rightarrow b\bar{b}\nu\bar{\nu}$ is characterized by a large missing energy and acoplanar b jets. The best hermeticity and an effective b-tagging are required in order to reduce the background arising from $We\nu$ or WW (through semileptonic decay of the W pair). This channel is useful due to the relatively high branching fraction ($\sim$ 17%) and the good sensitivity achieved.
- The "tau channel": $e^+e^- \rightarrow HZ \rightarrow q\bar{q}\tau^+\tau^-$ in which two isolated taus and two jets are required in the final state. It represents $\sim$ 9% of the overall decays.
- The "electron and muon channels" from $e^+e^- \rightarrow HZ \rightarrow q\bar{q}e^+e^-, q\bar{q}\mu^+\mu^-$ provide a clean signal consisting of two isolated leptons with an invariant mass equal to m_Z. With this channel, the best resolution on the Higgs mass is achieved ($\sim$ 2%) as well as the best sensitivity. Unfortunately, the branching fraction is low and is equal to 6%.

The typical overall signal efficiency reached is of the order of 30% for an expectation of 6 to 11 events from standard processes; no excess in any channel for any experiment is observed, as illustrated on the mass distribution of the candidate events observed by the OPAL experiment (Fig. 1). The agreement between the observation and the Standard Model expectations allowed us to extract a limit on the Higgs boson mass. The ranges of the limits obtained by each experiment are listed in Table 1. Since the limit results do not reach the kinematic limit, more luminosity will be useful. Therefore, in the framework of the LEP Higgs Working Group [3], an effort has been made to combine the results of the four LEP experiments. The combination leads to a gain of about 3–4 GeV with respect to the individual experiments. The 95% confidence level upper limit obtained for each statistical method is shown in Fig. 2 as a function of the Higgs mass. The four different combinations all agree within a spread of $\pm$ 0.15 GeV; conservatively, the lowest limit (method C) is quoted in Table 1; it has to be noticed that the limits have been improved, for each LEP experiment, by almost 4 GeV/c^2 with the recent data collected at 189 GeV [4], reaching therefore the level of the combined limit at 183 GeV. In Ref. [5], prospect studies show that a Higgs boson with a mass up to 107 GeV/c^2 could

be discovered at LEP200 when running at $\sqrt{s}$=200 GeV and assuming an integrated luminosity of 200 pb^{-1} collected by each experiment.

	$M_{H_{SM}}$ Lower Limit (GeV/c^2)	
	Range of Individual Limit (ADLO)	Lower LEP Combined
H (Obs.)	**85.7–88.3**	**89.8**
(Exp.)	85.0–86.5	90.4

Table 1: Individual and LEP combined observed and expected mass limits for the Standard Model Higgs boson [3], up to $\sqrt{s} = 183$ GeV.

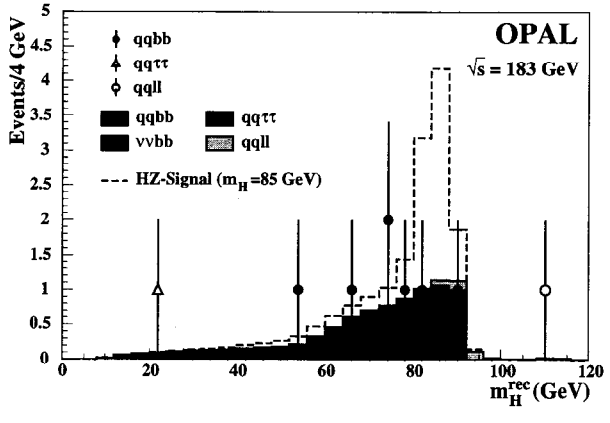


Figure 1: Mass distribution for the candidate events selected by the OPAL experiment in the searches for $e^+e^- \rightarrow HZ$ at center-of-mass energies up to 183 GeV [3].

2.2 The MSSM Higgs Bosons

In the MSSM, all SUSY particle masses, their couplings, and their production cross sections and decay widths are predicted in terms of only six free independent parameters: $m_{1/2}$ (the common gaugino mass parameter at the GUT scale), m_0 (the common mass for scalar fermions at the GUT scale), μ (the higgsino mixing parameter), $\tan\beta$ (the ratio of the vacuum expectation values of the two Higgs doublets), m_A (the mass of the neutral CP-odd Higgs boson A), and A (the trilinear coupling in the Higgs sector at the GUT scale); the renormalization group equations are used to determine the parameters at low

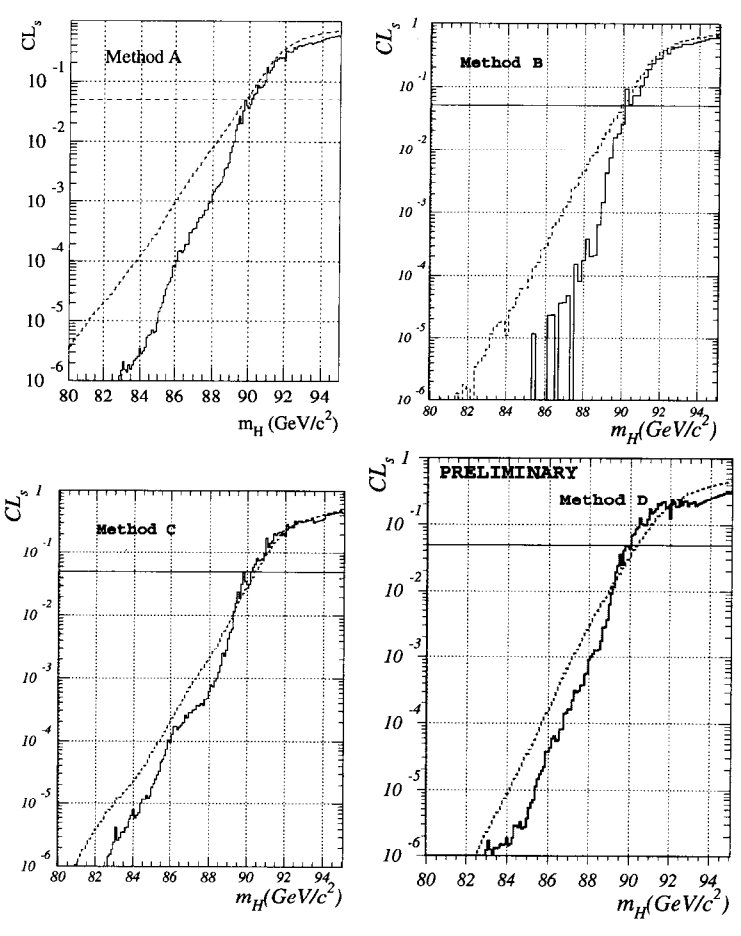


Figure 2: *Average expected (dashed lines) and observed (solid lines) confidence levels, CL_s, obtained from combining the results of the four LEP Collaborations using the four statistical methods. The intersections of the curves with the horizontal lines at 0.05 define the 95% confidence level lower bounds, expected and observed, for the mass of the SM Higgs boson.*

energies. Instead of $m_{1/2}$, the LEP experiments usually derive their results as a function of the parameter M_2, the SUSY breaking mass term associated with the $SU(2)_L$ gauge group ($m_{1/2} \approx M_2/0.81$).

Presently, only the lightest neutral Higgs bosons h and A can be discovered at LEP2 [1], since the CP–even Higgs boson H and the charged Higgs bosons $H^{\pm}$ are expected to be too heavy. However, searches for charged Higgs bosons are also performed at LEP2 in the framework of a non–SUSY two–Higgs doublet model.

The lightest neutral Higgs bosons

The lightest neutral Higgs bosons (h and A) could be observed at LEP through the two following processes:

- The Standard Model-like Higgs–strahlung process, $e^+e^- \rightarrow hZ$; the cross section for this process is equal to its SM analogue, reduced by a factor $\sin^2(\alpha - \beta)$ which means that this process occurs mainly at low $\tan\beta$ values.
- The associated pair production process, $e^+e^- \rightarrow hA$, with a cross section proportional to $\cos^2(\beta - \alpha)$; this process is dominant at large $\tan\beta$ values.

In the first channel, searches are similar to those performed for the SM Higgs boson. For $\tan\beta > 1$ the main decay modes of the h and A bosons are into $b\bar{b}$, and to a lesser extent $\tau^+\tau^-$. Most of the experimental analyses required therefore at least two b quarks in the final states, as detailed in Table 2. Clearly, b–tagging plays a crucial role in all the Higgs searches at LEP; it allows us to reach very high sensitivities by rejecting most of the WW background. Peculiar decays such as $h \rightarrow AA$ or $h \rightarrow \chi_1^0\chi_1^0$ have been also considered; the sensitivities achieved in the latter case are better compared to those obtained with the standard decay channels at LEP1 but worse at LEP2. Typical efficiencies and background expectations are listed in Table 3, where observation and expectation from SM processes agree very well.

Therefore, limits have been extracted [6]- [11], for two "benchmark" sets of the MSSM parameters where M_{SUSY}, representing the stop mass mean value, is fixed to 1 TeV and $m_t = 175$ GeV, as shown in Fig. 3; M_A varies up to 2 TeV and $\tan\beta$ from 0.5 to 50, and either minimal or no stop mixing [$A = 0$ TeV, $\mu = -0.1$ TeV] or maximal squark mixing [$A = \sqrt{6}$ TeV, $\mu = -0.1$ TeV] is assumed.

The limits obtained by each experiment within these assumptions are listed in Table 4. The combined limit, obtained as described in Ref. [3], leads to a gain of about 3–4 GeV with respect to the individual experiments. All the methods used for the combination agree within a spread of $\pm$ 1.9 GeV; conservatively, the lowest limit is quoted in Table 4 (method C of Ref. [3]), while the highest limits (with method D) are given in Fig. 3. These limits are valid for $\tan\beta$ greater than 0.8, irrespectively of the mixing value or of the method used for the combination. For the no-mixing scenario, the range of $\tan\beta$ between 0.8 and 2.1 is excluded independently of the method. It has to be noticed that the limits have been improved, for each LEP experiment, by almost 4 GeV/c^2 with the recent data collected at 189 GeV [4], reaching therefore the level of the combined limit at 183 GeV.

$\sqrt{s}$ (GeV)	hA	hZ
161,172,183	hA$\rightarrow$ bbbb	Zh$\rightarrow$ $\nu\bar{\nu}$(h $\rightarrow$ all)
	hA$\rightarrow$ b$\bar{\text{b}}\tau^-\tau^+$, $\tau^-\tau^+$ b$\bar{\text{b}}$	Zh$\rightarrow$ l^-l^+(h $\rightarrow$ all)
	AAA$\rightarrow$ b$\bar{\text{b}}$b$\bar{\text{b}}$b$\bar{\text{b}}$	Zh$\rightarrow$ $\tau^-\tau^+$(h $\rightarrow$ all)
		Zh$\rightarrow$ (h $\rightarrow$ all) $\tau^-\tau^+$
		Zh$\rightarrow$ q$\bar{\text{q}}$b$\bar{\text{b}}$
		Zh$\rightarrow$ q$\bar{\text{q}}$ or $l^-l^+(h \rightarrow \chi_1^0\,\chi_1^0)$

Table 2: Search channels for the neutral Higgs bosons at LEP.

It has been pointed out recently in Refs. [11], [9], [7], that selected sets of the MSSM parameters lead to less stringent exclusions, either because of an unexpectedly small bremsstrahlung cross section or because of a reduced decay branching ratio into $b\bar{b}$. The theoretical status of these pathological parameter sets is currently under investigation. Meanwhile, the limits obtained in the benchmark cases can be regarded as sufficiently robust for practical purposes as demonstrated in Ref. [7].

Selection	ϵ(%) for hZ M_h=80 GeV/c^2	N^{exp}_{signal}	N^{exp}_{bkg}	N^{obs}
hl$^+$l$^-$	78.3	1.1	2	3
h$\nu\bar{\nu}$	20.9	0.8	0.16	0
b$\bar{\text{b}}$q$\bar{\text{q}}$	23.6	3.6	1.4	1
b$\bar{\text{b}}\tau^+\tau^-$	22.7	0.14	0.17	0
$\tau^+\tau^-$q$\bar{\text{q}}$	9.8	0.16	0.16	0
Selection	ϵ(%) for hA $M_h = M_A$=75 GeV/c^2	N^{exp}_{signal}	N^{exp}_{bkg}	N^{obs}
b$\bar{\text{b}}$b$\bar{\text{b}}$	60.5	3.1	2.4	2
b$\bar{\text{b}}\tau^+\tau^-$	28.6	0.27	0.07	0

Table 3: Typical efficiencies and expected/observed number of events for the neutral Higgs boson searches performed by the ALEPH experiment [6] at $\sqrt{s}$=183 GeV.

The charged Higgs boson

The $H^{\pm}$ bosons can be produced in e^+e^- interactions via the process $e^+e^- \rightarrow \gamma^*, Z^* \rightarrow H^+H^-$ and are expected to decay mainly into the heaviest lepton kinematically allowed and its associated neutrino, or into the heaviest kinematically allowed quark pair whose decay widths are not Cabibbo suppressed; the relative branching ratios are model dependent. Therefore searches are performed in the three possible decay modes: $e^+e^- \rightarrow H^+H^- \rightarrow \tau^-\bar{\nu}_\tau\tau^+\nu_\tau$, $\tau^+\nu_\tau s\bar{c}$, and $c\bar{s}s\bar{c}$. For each experiment [12]- [15] and each decay channel, the number of selected events in the data is consistent, with the number of expected events from SM processes, as can be seen in Table 5. Compared to the neutral Higgs boson searches, the sensitivities achieved are lower; this is due to the high WW

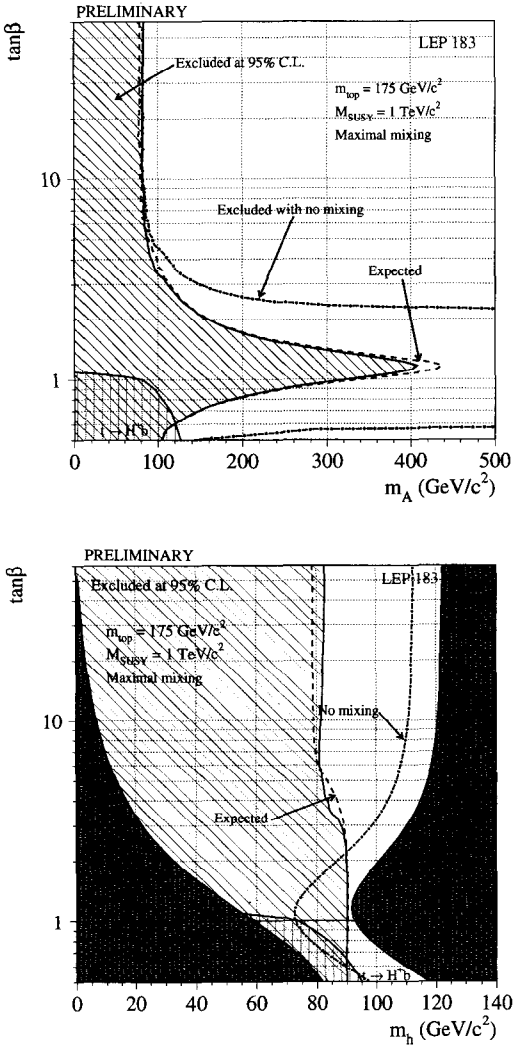


Figure 3: LEP combined exclusion domains (method D) in the plane $(M_A, \tan\beta)$ (top) or $(M_h, \tan\beta)$ (bottom) by the LEP direct searches for $e^+e^- \to hZ$ and hA at center-of-mass energies up to 183 GeV assuming either a maximal mixing or no mixing [3] and [16].

	$M_{h,A}$ Lower Limit (GeV/c^2)	
	Range of Individual Limit (ADLO)	Lower LEP Combined
h (Obs.)	**70.7–74.4**	**78.8**
(Exp.)	67.4–70.3	76.3
A (Obs.)	**71.0–76.1**	**79.1**
(Exp.)	68.4–72.0	76.3

Table 4: Individual and LEP combined mass limits for the neutral h and A bosons [3].

background contamination. The 95% C.L. individual limits on the charged Higgs boson

$M_{H^\pm}$=60 GeV/c^2	$\tau^-\bar{\nu}_\tau\tau^+\nu_\tau$	$\tau^+\nu_\tau s\bar{c}$	$c\bar{s}s\bar{c}$
Efficiency	24%	42%	40%
Expected events	9.2	30.1	99.4
Observed events	6	28	93

Table 5: Typical efficiencies and number of expected and observed events obtained by the L3 experiment in the charged Higgs boson searches at a c.m. energy of 183 GeV [14].

	$M_{H^\pm}$ Lower Limit (GeV/c^2)	
	Range of Individual Limit (ADLO)	Lower LEP Combined
$\mathbf{H^\pm}$ (Obs.)	**56.6–59.0**	**68.0**
(Exp.)	56.0–62.0	69.0

Table 6: Individual and LEP combined mass limits for the charged Higgs bosons.

mass is shown as a function of the branching ratio $\mathrm{Br}(H^\pm \to \tau^\pm\nu_\tau)$ in Fig. 4. Combining all LEP results, a lower limit on the charged Higgs boson mass of 68 GeV is established at the 95% C.L., assuming that the sum $\mathrm{BR}(H^+ \to \tau^+\nu_\tau) + \mathrm{Br}(H^+ \to c\bar{s})$ is equal to one. It represents a gain of almost 10 GeV with respect to the individual limits, Table 6.

3 Other SUSY Particles

The supersymmetric particles searched for at LEP2, not counting the Higgs bosons, are sleptons, stops, sbottoms, charginos, and neutralinos as detailed in Refs. [17] and [18]. The most popular SUSY breaking mechanisms are the gravity and the gauge mediated SUSY breaking models (GMSB). The results of SUSY searches at LEP are interpreted in the framework of these two models. In general, to avoid lepton and baryon number

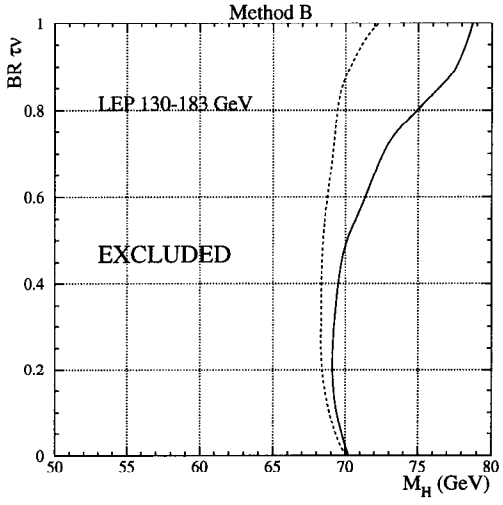


Figure 4: LEP combined exclusion domains in the plane $[M_{H^\pm}, Br(H^\pm \to \tau\nu_\tau)]$ by the LEP direct searches for $e^+e^- \to H^+H^-$ at center-of-mass energies up to 183 GeV [3].

violations induced by SUSY loops, the conservation of a new quantum number is invoked: $R = (-1)^{L+3B+2J}$, $= 1$ for standard particles and $= -1$ for SUSY particles. Within this assumption, the lightest SUSY particle (LSP) is stable and escapes detection since it is supposed to be neutral and weakly interacting. In particular in gravity SUSY breaking models, the LSP is the lightest neutralino ($\tilde{\chi}_1^0$), while in GMSB models it is the gravitino ($\tilde{G}$, with $M_{\tilde{G}} \sim 1$ eV).

3.1 $\tilde{\chi}_1^0$ LSP Scenario

The various pair-produced SUSY particles decay to SM particles and LSPs. SUSY signatures consist of some combination of jets or/and leptons and missing energy carried out by the LSP. The signal topology and the background conditions are in practice affected by the SUSY particle and the LSP mass difference ($\Delta M = m_{\rm SUSY} - m_{\chi_1^0}$) which controls the visible energy. In the low $\Delta M (= 5 - 10$ GeV) range, the expected topologies for the signals are characterized by a low multiplicity and a low visible energy, and the background is dominated by two-photon interactions. For large $\Delta M (= 50 - 60$ GeV) values, the signal signatures are very similar to those of W-pair production. Since all the background sources are due to well-calculable processes with reasonable production cross sections compared to the signal, most of the decay channels are studied at LEP2.

Scalar leptons

LEP experiments excluded scalar neutrinos with masses up to 43 GeV, from the comparison of the measured Z widths with the SM expectations [19].

In e^+e^- collisions, the production of scalar muons $\tilde{\mu}$ and scalar taus $\tilde{\tau}$ proceeds via γ or Z exchange in the s-channel only, whereas scalar electrons $\tilde{e}$ can also be produced by exchanging neutralinos in the t-channel.

The scalar leptons dominantly decay into their SM partners and the lightest neutralino χ_1^0. If the scalar lepton is not the next-to-lightest SUSY particle (NLSP), cascade decays via χ_2^0 or via the lightest chargino χ_1^+ (for $\tilde{l}_L$ only for a pure gaugino-type chargino) may be possible. The topologies arising from scalar lepton production are then usually acoplanar leptons plus missing energy. As already explained in the introduction, the sensitivities for these searches depend strongly on the mass difference, ΔM, between the scalar lepton and the χ_1^0. Typical efficiencies and expected events from SM processes for different ΔM ranges are given in Table 7.

$m_{\tilde{l}^\pm} = 75$ GeV									
	$\tilde{e}_R$			$\tilde{\mu}_R$			$\tilde{\tau}_R$		
ΔM (GeV)	ϵ (%)	$N_{\rm sign}^{\rm exp}$	$N_{\rm back}^{\rm exp}$	ϵ (%)	$N_{\rm sign}^{\rm exp}$	$N_{\rm back}^{\rm exp}$	ϵ (%)	$N_{\rm sign}^{\rm exp}$	$N_{\rm back}^{\rm exp}$
75	65	33.8	7.6	66	4.5	6.9	43	2.9	5.9
15	57	7.4	0.5	45	4.3	0.15	38	2.4	3.2
5	11	0.9	0.7	13	0.9	0.9	7	0.5	1.5

Table 7: Scalar electron, scalar muon, and scalar tau sensitivities. Results are obtained at $\sqrt{s} = 183$ GeV by the ALEPH experiment [20] as a function of ΔM for $m_{\tilde{l}^\pm} = 75$ GeV.

No excess with respect to the number of events expected from SM processes was observed at LEP up to $\sqrt{s}$=183 GeV [20]- [24]. Since in general, the cross section for the pair production of $\tilde{l}_R$ is smaller than for $\tilde{l}_L$, limits on masses are given by default taking into account $\tilde{l}_L\tilde{l}_R$ production only. The expected background from W-pair production is subtracted for all LEP analyses. All limits are derived in the plane $(m_{\chi_1^0}, m_{\tilde{l}_R})$. These limits depend only slightly on μ or $\tan\beta$ for $\tilde{\mu}_R$ and $\tilde{\tau}_R$, since these parameters may just change their couplings to χ_2^0 (relevant only for low masses of χ_1^0 in the region covered anyway by the χ_1^+ searches) but do not affect their cross section for pair-production. For $\tilde{e}_R$, the production cross section in the χ_1^0 t-channel exchange depends on these parameters. The combined exclusion contours of all LEP analyses [25] up to 183 GeV are shown in Fig. 5. The individual limits are derived for different $\tan\beta$ values ranging from 1.4 to 2, and for $\mu = -200$ GeV.

Since we are far from the kinematical limit, more luminosity helps and a gain ranging from 2 to 10 GeV, depending on the scalar lepton flavor, is obtained in the combination as shown in Table 8. We should mention also that the individual limits obtained recently at $\sqrt{s}$=189 GeV with an integrated luminosity of 30-40 pb^{-1} for each experiment reach the level of the LEP combined (up to 183 GeV) limit [25].

Assuming a common scalar lepton mass at the GUT scale, the relation between the

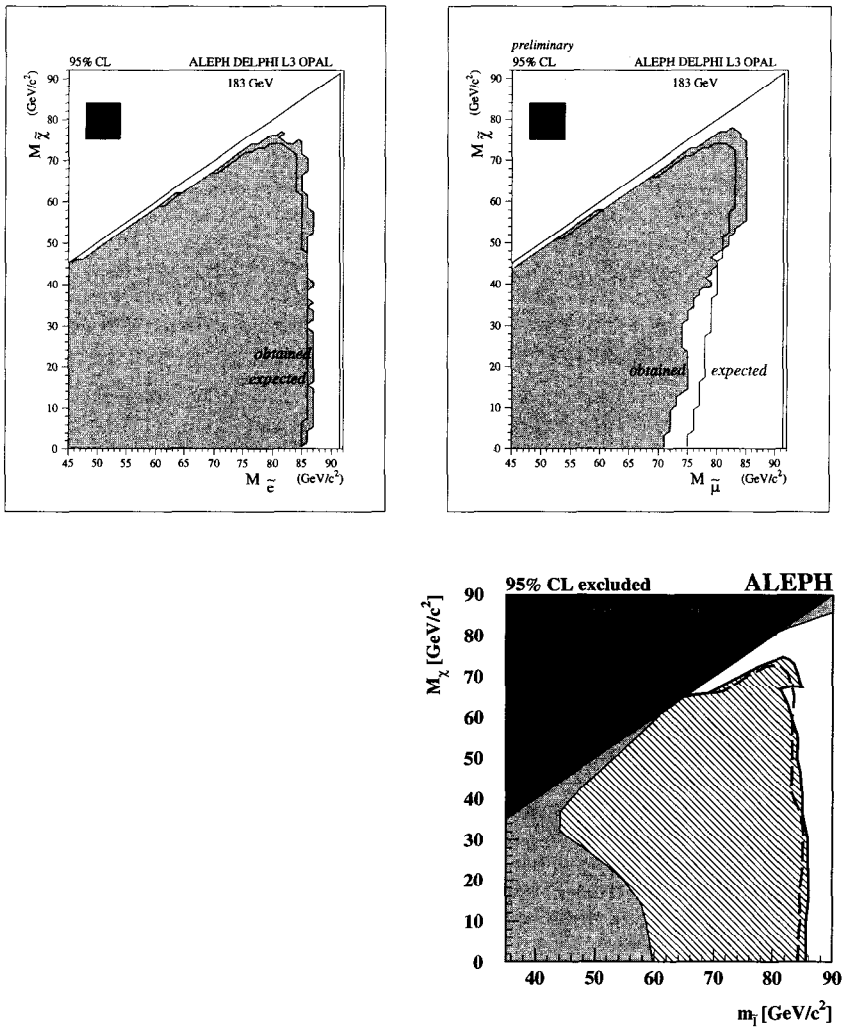


Figure 5: Combined exclusion contours for scalar electrons, muons, and taus [25] as a function of the LSP mass and the limit for sleptons obtained by the ALEPH experiment [20] when combining all selectron and smuon searches up to 183 GeV (bottom–right).

	$m_{\tilde{l}_R}$ Lower Limit (GeV/c^2)	
	Individual Limit (ADLO)	LEP Combined
$\tilde{e}_R$ (Obs.)	**79–83**	**85**
(Exp.)	78–83	85
$\tilde{\mu}_R$ (Obs.)	**55–62**	**71**
(Exp.)	56–67	75
$\tilde{\tau}_R$ (Obs.)	**45–63**	**75**
(Exp.)	45–56	67

Table 8: Individual and LEP combined mass limits for $\tilde{e}_R$, $\tilde{\mu}_R$, and $\tilde{\tau}_R$ up to $\sqrt{s} = 183$ GeV for ΔM greater than 20 GeV, and assuming $\tan\beta = 2$ and $\mu = -200$ GeV for $\tilde{e}_R$.

masses of right– and left–handed sleptons can be used to combine results of the searches for acoplanar leptons (muons, electrons together) [20] coming from the $\tilde{e}_R\tilde{e}_L$ process when the $\tilde{e}_R$ and the χ_1^0 are mass degenerate (ΔM below 3 GeV). The result is shown in Fig. 5 (bottom right), for $\tan\beta = 2$ and $\mu = -100$ GeV; this value of μ minimizes the product of cross section times branching fraction for the process $e^+e^- \to \tilde{e}_R\tilde{e}_L$ with $\tilde{e}_L \to e\chi_1^0$ for vanishing ΔM. In this case, a scalar lepton mass limit of 65 GeV/c^2 is set independently of ΔM; this limit holds for higher values of $\tan\beta$.

Scalar quarks

The squarks of the two first generations should be abundantly produced at the Tevatron which already places limits far above the kinematical reach of LEP2. Due to a large Yukawa coupling, the mass of one of the top squarks can be significantly smaller compared to those of the other scalar quarks; the large mixing between the left and right stops, proportional to m_t, leads to a large splitting of the two mass eigenstates. The lighter stop $\tilde{t}_1$ could be within the discovery range of LEP2. In addition, for large $\tan\beta$ values ($\gtrsim 30$), the mixing angle in the sbottom sector can also be large and the lighter sbottom $\tilde{b}_1$ could also be within the reach of LEP2. Stop and sbottom production at LEP proceed via Z/γ exchange in the s–channel; the production cross sections depend on the squark mass and the squark mixing angle $\cos\theta_q$, and at $\cos\theta_q \sim 0.57(0.39)$ the stop (sbottom) decouples from the Z–boson and the cross section is minimal; the dominant decay modes of the $\tilde{t}_1$ are expected to be $\tilde{t}_1 \to c\chi_1^0$ or $\tilde{t}_1 \to b\tilde{\nu}l^+$; both of these decay modes have been searched for at LEP. The dominant decay mode of $\tilde{b}_1$ is expected to be $\tilde{b}_1 \to b\chi_1^0$.

Under the assumption of R–parity conservation, the χ_1^0 and the $\tilde{\nu}$ are invisible in the detector; thus stop or sbottom pair events are characterized by two acoplanar jets or two acoplanar jets plus two leptons, with missing energy. When the $\tilde{t}_1 \to c\chi_1^0$ decay mode is dominant, the corresponding decay width is small enough for the stop to hadronize into a colorless "stop hadron" before decaying; this feature has been implemented by the LEP experiments in their Monte Carlo programs. No excess of events was observed by any of the four LEP experiments [26]– [29] in the searches for stop and sbottom. Typical sensitivities obtained by the OPAL [29] experiment in the various channels are listed in Table 9. All LEP results have been combined [25] and are summarized in Table 10, for ΔM greater

$m_{\tilde{t}_1,\tilde{b}_1} = 80$ GeV						
	$\tilde{t}_1 \to c\chi_1^0$		$\tilde{b}_1 \to b\chi_1^0$		$\tilde{t}_1 \to bl\tilde{\nu}$	
ΔM (GeV)	ϵ (%)	$N_{\rm back}^{\rm exp}$	ϵ (%)	$N_{\rm back}^{\rm exp}$	ϵ (%)	$N_{\rm back}^{\rm exp}$
7	23	1.97	37	1.97	11.8	1.07
15	58	1.97	61	1.97	64	2.05
80	37	1.97	31	1.97	58	2.05

Table 9: $\tilde{t}_1$ and $\tilde{b}_1$ efficiencies (ϵ) and the number of events expected from the SM ($N_{back}^{\rm exp}$) obtained at $\sqrt{s} = 183$ GeV by OPAL [29] as a function of ΔM for $m_{\tilde{q}} = 80$ GeV.

	$m_{\tilde{t}_1}, m_{\tilde{b}_1}$ Lower Limits (GeV/c^2)	
	Individual Limit (ADLO) or AO only: *	LEP Combined **obs** [exp]
$\tilde{t}_1 \to c\chi_1^0$, $\cos\theta_t = 1$	**80–85**	**86** (85)
$\tilde{t}_1 \to c\chi_1^0$, $\cos\theta_t = 0.57$	**72–81**	**83** (80)
$\tilde{t}_1 \to b\tilde{\nu}l$, $\cos\theta_t = 1$*	**84–85**	**87** (86)
$\tilde{t}_1 \to b\tilde{\nu}l$, $\cos\theta_t = 0.57$*	**80–82**	**85** (84)
$\tilde{b}_1 \to b\chi_1^0$, $\cos\theta_b = 1$	**78–84**	**86** (83)
$\tilde{b}_1 \to b\chi_1^0$, $\cos\theta_b = 0.39$	**45–68**	**75** (60)

Table 10: Individual and LEP combined mass limits for stop and sbottom up to $\sqrt{s} = 183$ GeV for $\Delta M > 15$ GeV, assuming maximal or minimal production cross sections.

than 15 GeV/c^2. Since a slight lack of events appears for sbottom searches, conservatively, the combination in this channel is done without any background subtraction. The LEP combined excluded regions in the plane $(m_{\chi_1^0}, m_{\tilde{q}})$ are shown in Fig. 6.

Charginos and neutralinos

We now turn to the search for charginos and neutralinos at LEP2. At e^+e^- colliders, charginos $\chi_1^+\chi_1^-$ (neutralinos $\chi_i^0\chi_j^0$ with $i,j = 1,\ldots,4$ ordered by their masses) are pair-produced via s-channel γ/Z (Z) boson and t-channel sneutrino $\tilde{\nu}$ (selectron $\tilde{e}$) exchange. When the masses of the sfermions are very large, the $\chi_1^\pm$ (χ_j^0 with $j \geq 2$) states decay via an exchange of a virtual W (Z) boson as follows: $\chi_1^\pm \to \chi_1^0 W^* \to \chi_1^0 f\bar{f}'$ ($\chi_j^0 \to \chi_k^0 Z^* \to \chi_k^0 f\bar{f}$ with $k < j$). If the slepton and sneutrino masses are comparable to M_W (M_Z), the charginos (next-to-lightest neutralinos) decay via virtual slepton or sneutrino exchange and the leptonic branching fraction is enhanced. Finally, for a slepton or sneutrino lighter than the chargino (next to lightest neutralino), the decay modes $\chi_1^\pm \to \tilde{l}^\pm\nu$ or $l^\pm\tilde{\nu}$ ($\chi_j^0 \to \tilde{l}^\pm l^\mp$ or $\tilde{\nu}\nu$) become dominant. The radiative decays $\chi_j^0 \to \chi_k^0\gamma$ are also possible via higher-order diagrams. For all LEP2 studies, the gluino is supposed to be heavy, otherwise the chargino decay patterns would be different.

The charginos and neutralinos are searched for at LEP2 [30]– [36], for the first time

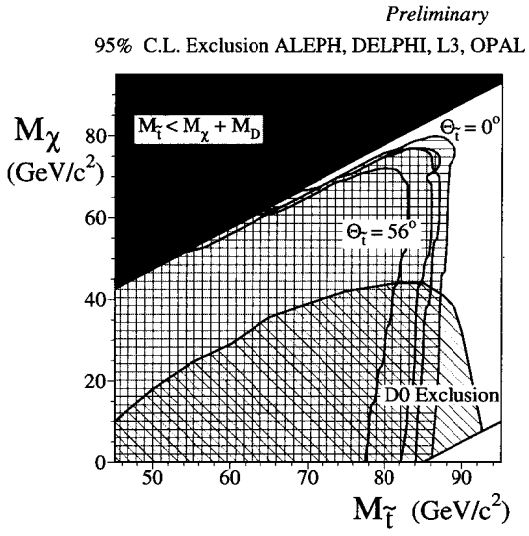


Figure 6: Combined LEP2 [25] exclusion domains in the plane $(m_{\chi_1^0}, m_{\tilde{t}_1})$ from stop decay $\tilde{t}_1 \to c\chi_1^0$ searches. The domain excluded by D0 is superimposed.

within the large m_0 scenario, *i.e.*, large scalar fermion masses. In this case, charginos and neutralinos decay into the LSP accompanied by virtual W and Z bosons. The three possible final state topologies arising from the W and Z decays are: hadrons plus missing energy, acoplanar lepton pairs plus missing energy, and mixed final state (hadrons plus leptons) plus missing energy. As explained before, signal topologies and the associated background sources depend on ΔM. Therefore, for each topology, selections were optimized for four different ΔM ranges: the very low (3–5 GeV), the low (5–10 GeV), the medium (20–40 GeV), and the large ($\geq$ 50 GeV). In addition, the DELPHI experiment searched for topologies arising in the very small ΔM regime, with decays in flight and ISR tags [33]. Altogether a total of 27 events was observed by the four experiments with 25$\pm$5 expected from standard processes with efficiencies ranging from 5% to 65% depending on the analysis and the ΔM ranges as illustrated in Table 11. Depending on the neutralino–chargino field content, one distinguishes the following cases for the determination of the lower limits on neutralino–chargino masses:

– Higgsino–like χ_2^0 and $\chi_1^\pm$ ($M_2 \gg |\mu|$): in this case, the production cross sections do not depend on the scalar lepton masses, and ΔM is low and decreases with increasing M_2. Consequently, the limits on the masses of the next-to-lightest neutralino and the lightest chargino decrease with M_2. The present LEP2 combined [25] limit on the chargino mass is shown in Fig. 7 (top left) turning into a limit on the chargino mass of 63 GeV/c^2 for

$M_{\chi_1^\pm}$=91 GeV/c^2		$\chi_1^\pm \to \chi_1^0 W^*$	
ΔM (GeV)	ϵ (%)	N_{back}^{exp}	N^{obs}
5	15	10.3 ± 1.1	8
20	55	17.4 ± 1.1	18
60	14	17.4 ± 1.1	18

Table 11: Typical efficiencies and number of expected and observed events obtained by the DELPHI experiment in the chargino searches at a c.m. energy of 183 GeV [32].

$\Delta M \geq 3$ GeV/c^2. The DELPHI experiment has excluded regions for lower ΔM ranges as depicted in Fig. 7 (bottom). Masses lower than 80 GeV/c^2 are excluded by all LEP experiments for the next-to-lightest neutralino χ_2^0 assuming M_2 less than 1500 GeV/c^2. This limit holds only for higgsino-like χ_2^0, since the coupling to the Z-boson vanishes for gaugino-like neutralinos.

– Gaugino-like chargino ($|\mu| \gg M_2$): the cross section depends strongly on the scalar neutrino mass. For $50 \leq m_{\tilde{\nu}} \leq 80$ GeV/c^2, the destructive interference term reduces the cross section by one order of magnitude compared to what is expected for $m_{\tilde{\nu}} \geq 300$ GeV/c^2. When the two-body decay $\chi_1^\pm \to l^\pm \tilde{\nu}$ is dominant, the relevant ΔM becomes equal to $m_{\chi_1^\pm} - m_{\tilde{\nu}}$. Therefore, the limit on the chargino mass will clearly depend on the $m_{\tilde{\nu}}$ value. For large $m_{\tilde{\nu}}$ values (≥ 300 GeV/c^2), the situation is the ideal one: we benefit there from the best detection sensitivity and the highest cross section production. This is the reason why the kinematical limit is reached with only a few pb^{-1}. Up to now, including the recent 189 GeV data, LEP experiments exclude a chargino mass below 94.3 GeV/c^2, irrespective of tanβ. Extending the study to all $m_{\tilde{\nu}}$ values where the chargino leptonic decay is enhanced for low slepton masses, dedicated analyses were then developed by the LEP experiments. However, it appears that the chargino cannot be directly detected when it is degenerate in mass with the sneutrino since the soft leptons produced in the final state escape detection. In this case, the LEP1 limit of 45 GeV/c^2 remains, that is deduced from the Z-boson total decay width measurement which is insensitive to the chargino decay pattern. Recently, this limit has been improved in two independent ways: (i) the ALEPH experiment excludes $m_{\chi_1^\pm}$ below 51 GeV/c^2 by measuring the invisible W-boson width, within some model assumptions [31]; and (ii) the L3 experiment [34] exploits the complementarity of the scalar lepton searches, excluding chargino masses below 57.2 GeV/c^2 irrespective of tanβ and m_0, as depicted in Fig. 7 (top right).

Indirect limits on the $\tilde{\chi}_1^0$ mass

Finally, let us discuss the neutralino–LSP mass limits. Since the neutralino χ_1^0 is the LSP and escapes detection, direct searches for the LSP neutralino cannot be performed at LEP2. However, indirect limits have been derived from the constraints on chargino–neutralino and slepton searches. As detailed previously, the limits depend on the slepton masses, in particular on $m_{\tilde{\nu}}$.

– Large sneutrino mass: For high tanβ values, the chargino search provides the most

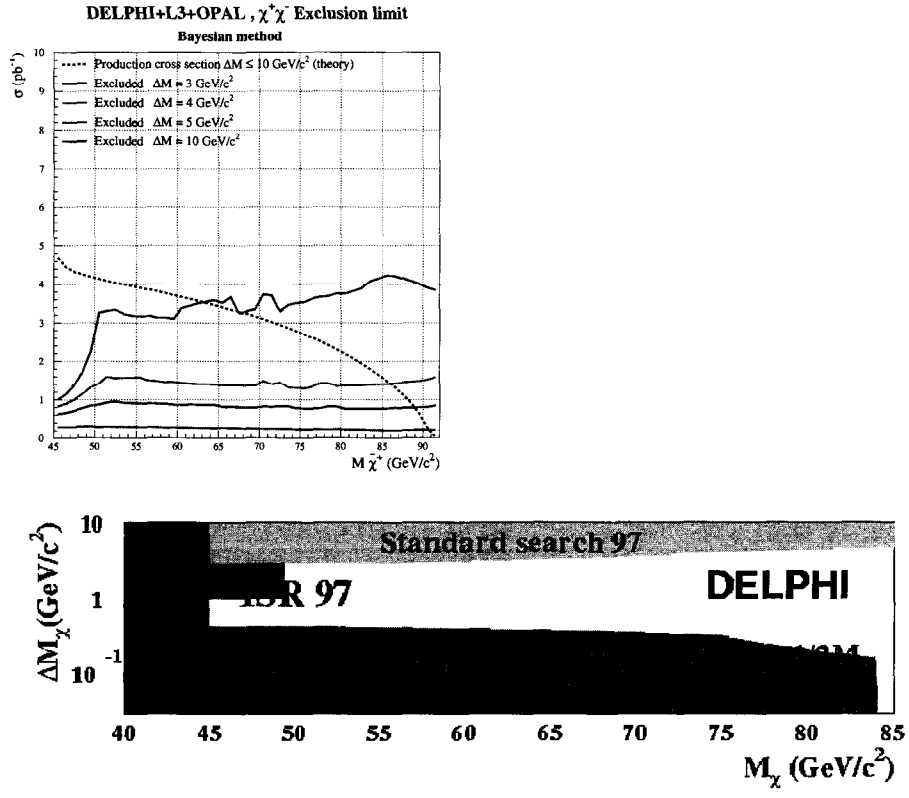


Figure 7: Top left: LEP combined [25] lower limit on the chargino mass for different ΔM values at c.m. energies up to 183 GeV. Top right: lower limit on the chargino mass for different field content, as a function of tanβ, by the L3 experiment [34]. Bottom: excluded domains in the plane ($m_{\chi_1^\pm}$, ΔM) in the very low ΔM range obtained by DELPHI [33] at a c.m. energy up to 183 GeV.

	$m_{\chi_1^0} > (\mathrm{GeV/c^2})$	
	Up to 183 Gev	Up to 189 GeV
Large $m_0, \tan\beta \geq 1$	~ 30.5 (ADLO)	~ 32.5 (ADLO)
Any $m_0, \tan\beta \geq 1$		
A μ up to -2000 GeV/c^2	26*	28
D μ up to -200 GeV/c^2	23.4	
L μ up to -500 GeV/c^2	25.9	28.2
O μ up to -500 GeV/c^2	25.4	

Table 12: Indirect individual mass limits for the LSP χ_1^0 up to $\sqrt{s} = 183$, 189 GeV for large m_0 or any m_0. In the latter case, the minimum value (*) for the ALEPH experiment is found in the deep–gaugino region (up to 183 GeV data) while other experiments found the minimum in the mixed region.

stringent bound on $m_{\tilde{\chi}_1^0}$. For $\tan\beta$ values below 2, the neutralino searches, including the processes $e^+e^- \to \tilde{\chi}_1^0\tilde{\chi}_3^0, \tilde{\chi}_1^0\tilde{\chi}_4^0, \tilde{\chi}_2^0\tilde{\chi}_3^0$ but also the radiative decays $\tilde{\chi}_2^0 \to \tilde{\chi}_1^0\gamma$, exclude additional regions of the parameter space (mixed region). The LEP1 exclusion limits still play some role, due to the fact that the neutralino search is limited in the region near the point where the χ_1^0 mass limit is set by the value of the coupling to the Z–boson rather than by kinematics. The lower limit obtained for $m_{\chi_1^0}$ as a function of $\tan\beta$ and for heavy sleptons is displayed in Fig. 8 for LEP center–of–mass energies increasing from 91.5 up to 183 GeV. A summary of all limits [30]– [36] obtained both up to 183 GeV and more recently to 189 GeV is listed in Table 12.

– Any sneutrino mass: For lower slepton masses, as already discussed, the constraints in the chargino–neutralino sectors are weaker and the limits on $m_{\chi_1^0}$ are therefore degraded. Constraints on $m_{\chi_1^0}$ have been obtained by all LEP experiments with a systematic scan of the MSSM parameter space, as a function of $\tan\beta$ for all m_0 values, as depicted in Fig. 8(bottom). For low $\tan\beta$ values (≤ 1.6), the minimum allowed value for $m_{\chi_1^0}$ is found in the parameter space where the production cross section for charginos is minimal [$m_{\tilde{\nu}} \sim$ 90 GeV/c^2 in the mixed region $\mu \sim -70$ GeV/c^2] [32]– [36] and the heavy neutralino $\chi_{3,4}^0$ are mostly invisible while for higher $\tan\beta$ values (≥ 2) the lower χ_1^0 mass limit is found in the parameter space region where the chargino and sneutrino mass difference is small and where the next–to–lightest neutralinos χ_2^0 decay invisibly. The exclusion limit came only from slepton searches; it is localized in the deep gaugino–like region for charginos ($\mu \leq -400$ GeV/c^2) [30]. The limits obtained by each experiment are listed in Table 12 as well as the MSSM parameter ranges used in the scan. The most recent limit on the $\tilde{e}$ mass obtained with 189 GeV data by ALEPH [30] and L3 [35] allows us to set the best absolute limit on the LSP mass of ~ 28 GeV/c^2, which can be turned into an absolute limit on M_2.

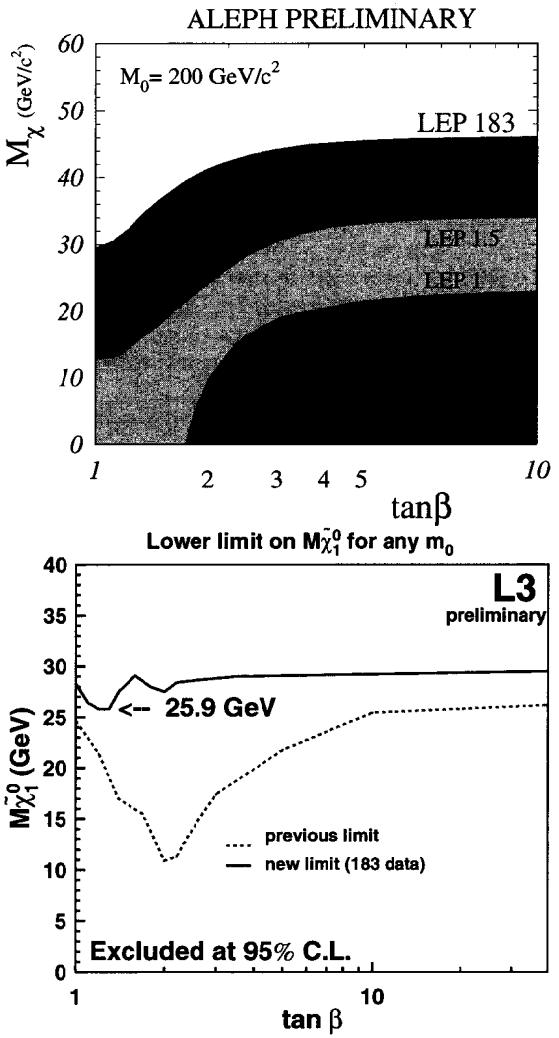


Figure 8: Top: ALEPH [30] lower limit on the LSP mass as a function of $\tan\beta$, for large slepton masses, at c.m. energies up to 183 GeV. Bottom: Lower limit on the LSP mass as a function of $\tan\beta$, for any slepton mass obtained by L3 [34] at c.m. energies up to 183 GeV.

3.2 Search for Charginos and Neutralinos Assuming No R–Parity Conservation

In supersymmetry, R–parity violation can occur through the following Yukawa coupling terms in the superpotential

$$\lambda_{ijk} L_i L_j \bar{E}_k + \lambda'_{ijk} L_i Q_j \bar{D}_k + \lambda''_{ijk} \bar{U}_i \bar{D}_j \bar{D}_k$$

where L and Q are the SU(2) doublet lepton and quark superfields and E, U, D are the singlet superfields with i, j, k as the generation indices. The simultaneous presence of the last two types of terms would induce fast proton decay, which is excluded experimentally. Therefore, it is generally assumed that among the 45 R–parity violating coupling terms only one (or more generally only one type) dominates. The three types (λ, λ', λ'') have been studied at LEP2.

The main consequence of R–parity violation is that the LSP is no longer stable. The R–parity violating couplings under consideration are strong enough so that the LSP decays within the detector (it requires a strength of the coupling greater than 10^{-3} which is within the existing upper bound obtained from low energy experiments). The L-violating λ type coupling terms will give rise to multilepton final states, for instance, a λ_{ijk} term may induce the decay of a neutralino LSP to final states such as $\bar{\nu}_i l_j^+ l_k^-$ (via virtual slepton or sneutrino exchange); the L-violating λ' type coupling terms involve multijets with leptons or with some missing energy. Finally, the B-violating λ'' type coupling terms lead to events with multijet signatures and without missing energy.

In what follows the process involving chargino pair production as well as neutralino pair production ($\tilde{\chi}_1^0\tilde{\chi}_1^0$ but also $\tilde{\chi}_1^0\tilde{\chi}_2^0$) have been studied in which R–parity violation manifests itself only in the decay of the LSP. No signal above the expected Standard Model backgrounds was detected at LEP [37]– [40] in any of the topologies investigated, resulting in mass limits or constraints on the parameters of the MSSM. As shown by ALEPH in Fig. 9 [37], the kinematic limit for the chargino is reached everywhere, in particular in the deep-higgsino region in contrast to the case of R parity conservation. Assuming a dominant λ_{133} coupling (channel providing the worst sensitivity for this type of coupling), the LEP experiments [38]– [40] have derived a lower limit on the LSP mass equal to 26.8 GeV/c^2 irrespective of tanβ.

3.3 The Light Gravitino Scenario

A class of supersymmetric models in which the gravitino is the lightest supersymmetric particle has recently received renewed attention, since the observation by the CDF experiment [41] of one event containing two electrons, two photons, and a large transverse missing energy. Within this scenario ($\tilde{G}$ is the LSP), the phenomenology will depend on the nature of the next-to-lightest supersymmetric particle. At LEP, one distinguishes two cases: the NLSP is either the lightest neutralino (χ_1^0) or a slepton ($\tilde{l}$).

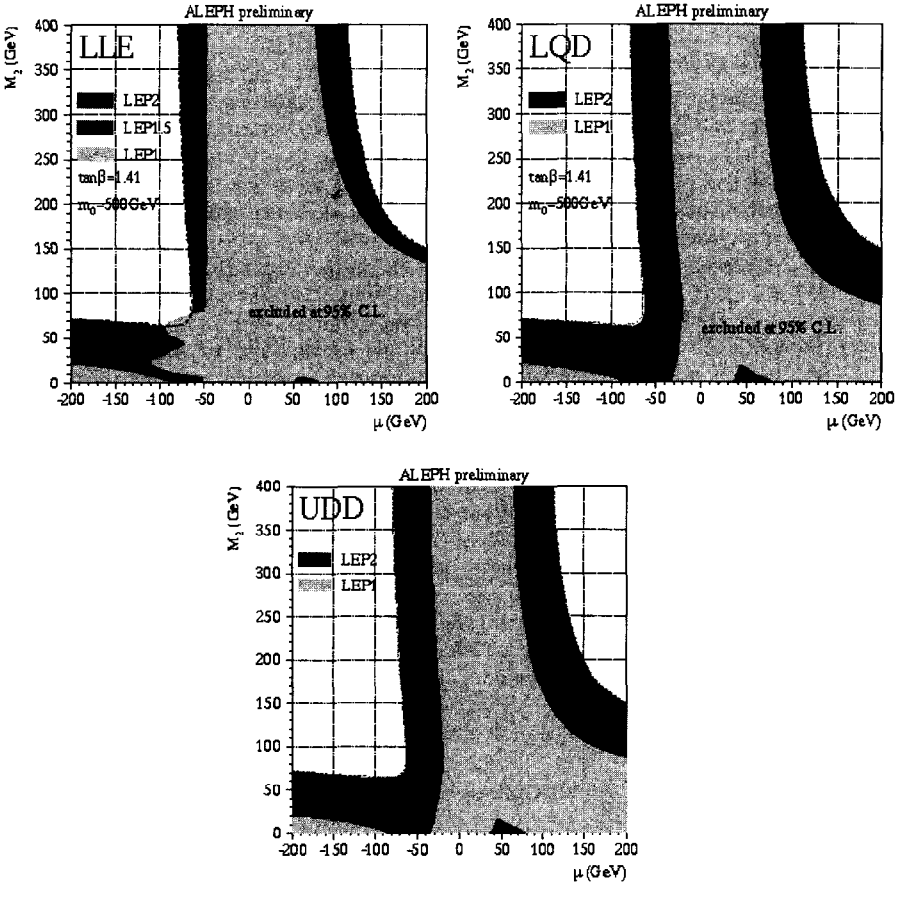


Figure 9: Regions excluded in the (μ,M_2) plane for the three λ type couplings. The dotted line is the chargino kinematic limit.

The lightest neutralino (χ_1^0) as the next-to-lightest supersymmetric particle

In such a scenario, the lightest neutralino is expected to decay into a photon and a gravitino, $\tilde{\chi}_1^0 \to \gamma\tilde{G}$. For practical purposes, the neutralino lifetime is negligibly small as soon as the gravitino mass is smaller than a few eV/c^2. In the light gravitino scenario, large production cross sections are predicted at LEP2 ($\sim$ 1 pb) for the two processes: $e^+e^- \to \tilde{\chi}_1^0\tilde{\chi}_1^0 \to \gamma\gamma\tilde{G}\tilde{G}$ and $e^+e^- \to \tilde{\chi}_1^0\tilde{G} \to \gamma\tilde{G}\tilde{G}$ (the latter only for very light $\tilde{G}$, $M_{\tilde{G}} \sim 10^{-4}$ eV/c^2). These two reactions lead respectively to the two photon and single photon plus missing energy signatures. For these kinds of signatures, there is however an irreducible background due to the reaction $e^+e^- \to \nu\bar{\nu}\gamma(\gamma)$, which should be kept under control.

The combined result obtained by the four LEP experiments is shown in Fig. 10 (left), where the single or two photon recoil mass for data is displayed which is in perfect agreement with the SM prediction.

In the single photon channel, upper limits on the production cross section for the process $e^+e^- \to \tilde{\chi}_1^0\tilde{G}$ have been derived, ranging from 0.1 to 0.7 pb, depending on the analysis and on the mass of the lightest neutralino. In models with superlight gravitinos, the process $e^+e^- \to \tilde{G}\tilde{G}$ can also be relevant. When accompanied by initial state radiation, this reaction leads to single (or multi-) photon signature. If $M_{\tilde{\chi}_1^0} \geq \sqrt{s}$, this is the only accessible reaction to produce SUSY particles [42]. In the absence of a signal, lower limits on the gravitino mass ranging from 6.6 to 8.8 $\times 10^{-6}$ eV/c^2 have been set [43]- [46]; these can be translated to an upper limit on the SUSY breaking scale $\sqrt{F} \geq 182.5$ GeV.

The two-photon final state refers to the pair production of the lightest neutralinos, $e^+e^- \to \tilde{\chi}_1^0\tilde{\chi}_1^0$. The main background from the process $e^+e^- \to \nu\bar{\nu}\gamma\gamma$ is reduced by the requirement that the mass recoiling against the two photons should not be close to the Z boson mass; in addition, a minimum energy for both photons is required. In the absence of a signal, the LEP combined upper limit on the production cross section of $\tilde{\chi}_1^0$ pairs reaches the level of 0.02-0.04 pb [25]. To turn this constraint into a $\tilde{\chi}_1^0$ mass limit, further model requirements are needed (right or left selectron masses, $\tilde{\chi}_1^0$ field content). Values below 85 GeV/c^{2}are typically excluded. Finally, the resulting combined upper limit is confronted with the kinematic region defined by the CDF event. This is shown in Fig. 10 (right) where LEP2 data exclude most of the allowed region.

$\tilde{l}$ as the next-to-lightest supersymmetric particle

For a part of the parameter space, the slepton and in particular the stau can be the NLSP, when the gravitino is the LSP; leading to the process:

$$e^+e^- \to \tilde{l}^+\tilde{l}^- \to l^+l^-\tilde{G}\tilde{G}.$$

The slepton may decay inside or outside the detector, since the decay length depends on the $\tilde{l}$ and the $\tilde{G}$ masses. Three strategies are developed by the experiments depending on this decay length: First, if the decay is prompt at the interaction vertex, the search is similar to the slepton search in the conventional MSSM, in the limiting case of a vanishing neutralino mass. Then for an increasing decay length, one looks for particles decaying

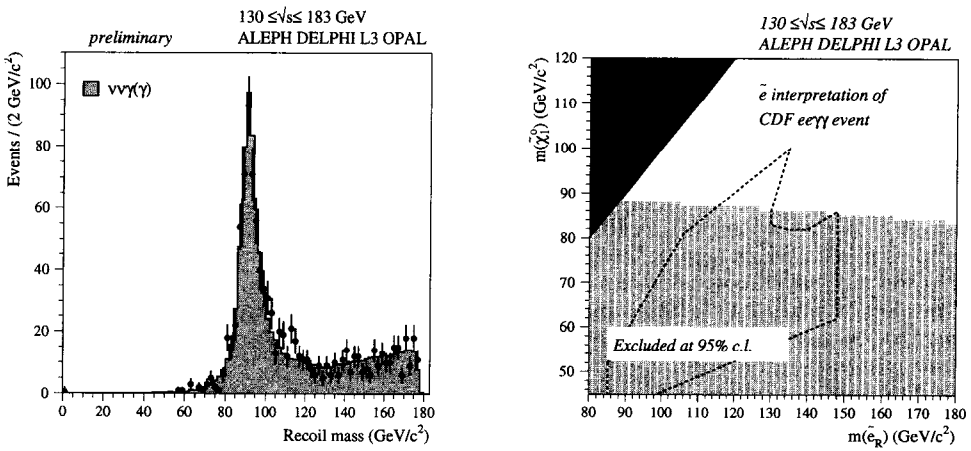


Figure 10: Left: LEP combined recoil mass for single and two-photon events [25]. Right: The excluded region in the neutralino, selectron mass plane. Overlaid is the region determined from the properties of the event observed by CDF [41].

into the detector, characterized by events containing tracks with large impact parameters and secondary vertices (kinks). Finally, if the particles decay outside the detector, the particles are seen as heavy stable ionizing particles (HIP). The four LEP experiments have searched for such HIPs; since no excess of events has been observed (two events expected and one candidate event observed), the LEP combined results, shown in Fig. 11 (left), lead to a lower limit on the mass of the $\tilde{\mu}$ (valid also for the $\tilde{\tau}$) equal to 86.5 GeV/c^2(87 GeV/c^2), depending on if the $\tilde{\mu}$ is the supersymmetric partner of the right- (left-) handed μ, respectively.

The DELPHI and ALEPH experiments have developed specific analyses for all the different signatures for each lepton flavor. Since no excess of events with respect to the Standard Model expectation has been observed, all the analyses have been combined to derive cross section upper limits which can be turned out in an exclusion contour in the plane $(m_{\tilde{l}}, m_{\tilde{G}})$. For instance, the DELPHI experiment [47] derives a lower limit on the $\tilde{\tau}$ mass of 68 GeV/c^2 irrespective of the mass of the gravitino, as illustrated in Fig. 11 (right).

4 Conclusions and Outlook

A wide variety of searches has been carried out at LEP2 and despite all the efforts, no indication for new physics has been observed yet up to 189 GeV.

Higgs boson searches considerably improve previous results obtained at LEP1 and ex-

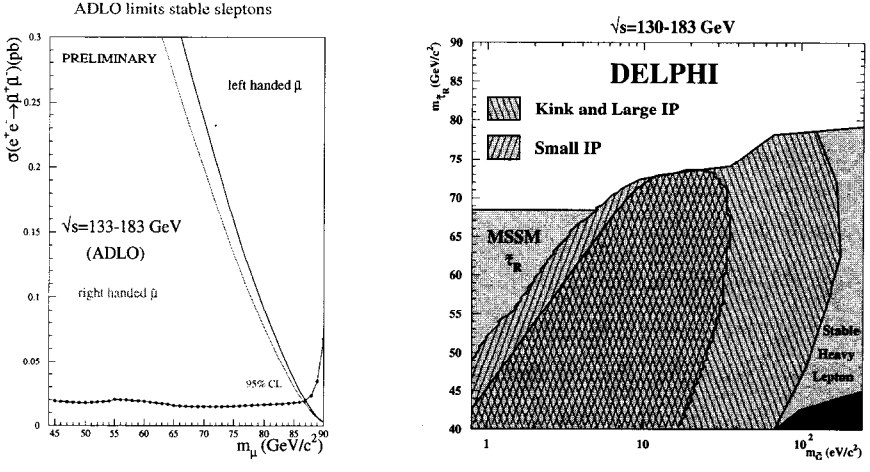


Figure 11: Left: LEP combined upper limit on the production cross section for stable sleptons as a function of the mass of the sleptons. Right: Exclusion region in the $(m_{\tilde{\tau_R}}, m_{\tilde{G}})$ plane by DELPHI. The positive slope hatched area shows the region excluded by the small impact parameter search. The negative slope hatched area shows the region excluded by the combination of the large impact parameter and kink searches.

clude Higgs masses below 89.8 GeV/c^2. New limits in the SUSY sector have been derived, in particular for the neutral Higgs bosons h and A , masses below 77 ,78 GeV/c^2(respectively) are excluded for tanβ values greater than 0.8. With the coming data foreseen at $\sqrt{s} \sim 200$ GeV, the SM Higgs with a mass up to 100-110 GeV/c^2 can be discovered at LEP2.

LEP2 is also a unique place for SUSY searches: Due to the good detector performances and the relatively good knowledge of the background reactions, a large domain is covered and all channels are exploited. With the absence of any signal, LEP2 excluded chargino masses up to the kinematic limit (94.3 GeV/c^2) for most of the SUSY parameter space and derived an absolute limit equal to 57.2 GeV/c^2. In addition, an indirect and absolute lower limit, equal to 28 GeV/c^2, on the Lightest SUSY Particle has been derived in the framework of the CMSSM. With future LEP runs, chargino masses up to 100 GeV/c^2will be explored and a lower mass limit on the LSP of the order of $\sim$ 35 GeV/c^2will be reached.

Acknowledgments

Preparing such a review needs a lot of discussion and interaction with many colleagues. For this reason I would like to thank Wim deBoer, Laurent Duflot, Alvise Favara, Marta Felcini, Pascal Gay, Paolo Giacomelli, Fabiola Gianotti, Eilam Gross, Peter Igo Kemenes, Patrick Janot, Luc Pape, and Pierpaolo Rebecchi from the ALEPH, DELPHI, L3, and OPAL experiments as well as all the members of the Higgs and SUSY LEP working groups for their great help in preparing this talk. I am especially grateful to G. Coignet and J. F. Grivaz for the numerous interesting discussions and for a careful reading of the report. It is a pleasure for me to thank the organizers of the SSI for the very stimulating session and for the very pleasant atmosphere at the conference.

References

[1] P. Janot, "Searching for Higgs bosons at LEP1 and LEP2," *Perspectives on Higgs Physics II*, World Scientific Publishing Company, ed. G. L. Kane.

[2] G. Quast, *Electroweak physics*, talk given on behalf of the LEP Electroweak working group, LEPC, 15 Sept. 1998.

[3] The LEP HIGGS Working Group, Abstracts 577-582, ICHEP98, Vancouver, July 1998. F. Di Lodovico, *Report from the LEP Higgs working group,* talk given on behalf of the LEP HIGGS working group, LEPC, 15 Sept. 1998.

[4] ALEPH collaboration, Abstract 903, ICHEP98, Vancouver, July 1998; DELPHI collaboration, Abstract 219, ICHEP98, Vancouver, July 1998; L3 collaboration, Abstract 485, ICHEP98, Vancouver, July 1998; OPAL collaboration, Abstract 1069, ICHEP98, Vancouver, July 1998.

[5] E. Gross *et al.*, *Prospects for the Higgs Boson Search in Electron-Positron Collisions at LEP 200*, CERN-EP/98-094.

[6] ALEPH collaboration, Phys. Lett. B **440**, 403 (1998); Phys. Lett. B **447**, 336 (1999).

[7] ALEPH collaboration, Abstract 897, ICHEP98, Vancouver, July 1998.

[8] DELPHI collaboration, Abstract 200, ICHEP98, Vancouver, July 1998.

[9] DELPHI collaboration, Abstract 209, ICHEP98, Vancouver, July 1998.

[10] L3 collaboration, Phys. Lett. B **436**, 389 (1998).

[11] OPAL collaboration, Abstract 355, ICHEP98, Vancouver, July 1998.

[12] ALEPH collaboration, Abstract 900, ICHEP98, Vancouver, July 1998.

[13] DELPHI collaboration, Abstract 214, ICHEP98, Vancouver, July 1998.

[14] L3 collaboration, Phys. Lett. B **444**, 516 (1998).

[15] OPAL collaboration, Abstract 1069, ICHEP98, Vancouver, July 1998.

[16] P. Janot, *New particles searches*, talk given at the E.P.S conference in Jerusalem, August 1997.

[17] J. F. Grivaz, "Supersymmetric particle searches at LEP," *Perpectives on Supersymmetry*, World Scientific Publishing Company, ed. G. L. Kane.

[18] A. Djouadi, S. Rosier-Lees et al., *The Minimal Supersymmetric Standard Group Summary Report*, "`http://www.lpm.univ-montp2.fr:7082/djouadi/GDR/gdr.html`."

[19] Particle Data Group, C. Caso *et al.*, "The review of particle physics," The European Physical Journal, C3 (1998).

[20] ALEPH collaboration, Phys. Lett. B **433**, 176 (1998).

[21] ALEPH collaboration, Abstract 942, ICHEP98, Vancouver, July 1998.

[22] DELPHI collaboration, Abstract 204, ICHEP98, Vancouver, July 1998.

[23] L3 collaboration, Abstract 492, ICHEP98, Vancouver, July 1998.

[24] OPAL collaboration, Eur. Phys. J. C **12**, 551 (2000).

[25] F. Cerutti, *Report from the LEP SUSY working group*, talk given on behalf of the LEP SUSY working group, LEPC, 15 Sept. 1998.

[26] ALEPH collaboration, Phys. Lett. B **434**, 189 (1998), and Abstract 942, ICHEP98, Vancouver, July 1998.

[27] DELPHI collaboration, Abstract 204, ICHEP98, Vancouver, July 1998.

[28] L3 collaboration, Phys. Lett. B **445**, 428 (1999).

[29] OPAL collaboration, Eur. Phys. J. C **6**, 225 (1999), and Abstract 276, ICHEP98, Vancouver, July 1998.

[30] ALEPH collaboration, Abstracts 942 and 952, ICHEP98, Vancouver, July 1998.

[31] ALEPH collaboration, Abstract 915, ICHEP98, Vancouver, July 1998.

[32] DELPHI collaboration, Abstracts 201 and 219, ICHEP98, Vancouver, July 1998.

[33] DELPHI collaboration, Abstract 202, ICHEP98, Vancouver, July 1998.

[34] L3 collaboration, Abstracts 485 and 493, ICHEP98, Vancouver, July 1998.

[35] R. Clare, L3 report to LEPC, November 1998.

[36] OPAL collaboration, Eur. Phys. J. C **8**, 255 (1999), and Abstract 273, ICHEP98, Vancouver, July 1998.

[37] ALEPH collaboration, Abstract 949, ICHEP98, Vancouver, July 1998.

[38] DELPHI collaboration, Abstract 210, ICHEP98, Vancouver, July 1998.

[39] L3 collaboration, Abstract 496, ICHEP98, Vancouver, July 1998.

[40] OPAL collaboration, Abstracts 278 and 279, ICHEP98, Vancouver, July 1998.

[41] CDF collaboration, F. Abe *et al.*, FERMILAB-PUB-98-024-E, submitted to Phys. Rev. Lett.

[42] A. Brignole *et al.*, Nucl. Phys., B516 (1998) 13.

[43] ALEPH collaboration, Phys. Lett. B **429**, 201 (1998).

[44] DELPHI collaboration, Abstracts 206 and 207, ICHEP98, Vancouver, July 1998.

[45] L3 collaboration, Phys. Lett. B **444**, 503 (1998).

[46] OPAL collaboration, Eur. Phys. J. C **8**, 23 (1999).

[47] DELPHI collaboration, Abstract 205, ICHEP98, Vancouver, July 1998.

SPIN STRUCTURE MEASUREMENTS FROM E155 AT SLAC

Lee Sorrell*
American University
Department of Physics
4400 Massachusetts Avenue
Washington, DC 20016

Representing the E155 Collaboration

ABSTRACT

The E155 Collaboration at SLAC has collected precise polarized deep-inelastic scattering data which will provide further insights into the structure of the nucleon. The SLAC 48.3 GeV electron beam was used to obtain data from both proton and deuteron targets covering $0.01 < x < 0.9$ and $1(\mathrm{GeV/c})^2 < Q^2 < 40(\mathrm{GeV/c})^2$. E155 is the first high dose electron scattering experiment to use lithium deuteride ($^6\mathrm{Li}^2\mathrm{H}$) as a target material.

*Supported by DOE Contract DE–AC03–76F00515.

© 1998 by Lee Sorrell.

1 Introduction

Polarized deep-inelastic scattering is a powerful tool for studying the internal spin structure of the proton and neutron (collectively known as nucleons).[1] Early results obtained by experiments at SLAC[2,3] and CERN[4] indicated that the quarks contribute very little to the nucleon's spin. This contradicted expectations based on the naive quark-parton model (QPM), in which one expects the valence quarks to carry the nucleon spin, and led to the "spin crisis." This motivated additional theoretical and experimental studies of the spin structure of the nucleon which have yielded a richer view of the nucleon. Experiments at CERN,[5] SLAC,[6–9] and DESY[10] have provided increasing levels of precision and enabled the determination of polarized parton distribution functions (PDF's) using next-to-leading-order perturbative QCD calculations and the (Altarelli-Parisi or DGLAP) evolution equations such as the fits performed by E154[8] and SMC.[11] The main goal of the E155 experiment was to obtain large samples for both the proton and deuteron covering the widest possible kinematic range in a single experiment in order to constrain the polarized PDF's.

1.1 Inclusive Deep-Inelastic Scattering

The SLAC experiments have focused on inclusive deep-inelastic scattering, in which only the scattered electron is detected (see Fig. 1). The kinematics of the scattering event are determined by the incident electron energy (E), and the energy (E′) and angle (θ) of the scattered electron with respect to the incident beam direction. From this information we can determine the four-momentum transfer squared Q^2, and the Bjorken scaling variable $x = \frac{Q^2}{2M\nu} = \frac{Q^2}{2M(E-E')}$. The Bjorken scaling variable gives the fraction of the nucleon momentum carried by the struck quark.

In the unpolarized case, the cross section obtained from the inclusive measurements can be related to the unpolarized structure functions F_1 and F_2 as shown in Eq. (1).

$$\frac{d\sigma}{d\Omega dE'} = \frac{4\alpha^2 E'^2 \cos^2(\theta/2)}{Q^4}\left[\frac{F_2(x,Q^2)}{\nu} + \frac{2F_1(x,Q^2)}{M}\tan^2(\theta/2)\right] \quad (1)$$

In this equation, M is the nucleon mass and θ is the scattering angle with respect to the original beam direction. In leading order, the structure function F_1 is given by a sum over the spin-averaged parton distributions

$$F_1(x,Q^2) = \frac{1}{2}\sum_i e_i^2[q_i^{\uparrow}(x,Q^2) + q_i^{\downarrow}(x,Q^2)]. \quad (2)$$

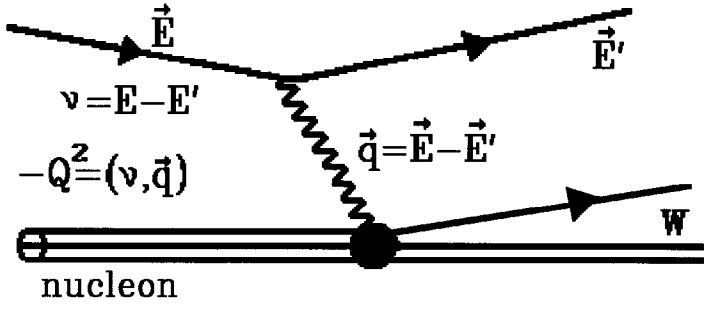


Fig. 1. Inclusive deep-inelastic scattering.

The functions $q_i^{\uparrow}(x,Q^2)$ and $q_i^{\downarrow}(x,Q^2)$ are the probability distributions for the ith quark flavor to be aligned and anti-aligned with the nucleon spin (respectively). The index i in the sum runs over the kinematically allowed quark flavors and e_i is the corresponding charge.

1.2 Virtual Photon-Nucleon Asymmetries

In the polarized case, cross-section asymmetries are measured and the polarized structure functions g_1 and g_2 are determined from these asymmetries. A simple way to understand how these asymmetries provide information on the net quark polarization is to first consider the case of a polarized virtual photon incident on a polarized nucleon. If the nucleon is polarized along the direction of the photon motion, then we can measure the longitudinal virtual photon asymmetry,

$$A_1(x,Q^2) = \frac{\sigma_T^{1/2} - \sigma_T^{3/2}}{\sigma_T^{1/2} + \sigma_T^{3/2}} = \frac{g_1(x,Q^2) - \gamma^2 g_2(x,Q^2)}{F_1(x,Q^2)}, \quad (3)$$

in which $\gamma^2 = Q^2/\nu^2$. The transverse virtual photoabsorption cross sections $\sigma_T^{1/2}$ and $\sigma_T^{3/2}$ are measured with the virtual photon polarization anti-aligned and aligned with the nucleon spin. In leading order, a spin-one virtual photon with its spin aligned with the nucleon will couple to the quarks with spins anti-aligned with the nucleon spin and flip them from $s_z = -1/2$ to $+1/2$ (where s_z indicates the spin component along the beam axis). Similarly, a photon with its spin anti-aligned with the nucleon spin will couple with the quarks that have their spins aligned with the nucleon spin. In this naive picture we will have $\sigma_T^{1/2} \sim \frac{1}{2}\sum_i e_i^2 q_i^{\uparrow}$ and $\sigma_T^{3/2} \sim \frac{1}{2}\sum_i e_i^2 q_i^{\downarrow}$.

The other virtual photon asymmetry A_2 is related to the transverse-longitudinal interference term σ_{TL}

$$A_2(x,Q^2) = \frac{2\sigma_{TL}}{\sigma_T^{1/2} + \sigma_T^{3/2}} = \gamma \frac{g_1(x,Q^2) + g_2(x,Q^2)}{F_1(x,Q^2)}. \quad (4)$$

This asymmetry does not lend itself to the naive interpretations given above, but does play a role in the extraction of the structure functions. In the past, $A_2 = 0$ has been used in the determination of the structure functions, but more recent data suggest that while A_2 is small, it is not consistent with zero.

1.3 Polarized DIS

At SLAC, cross-section asymmetries are measured using polarized electrons, which interact with the polarized nucleons through the exchange of virtual photons (see Fig. 2). The "depolarization" factor, which reflects the difference between the electron polarization and virtual photon polarization, can be calculated and used in relating the experimental asymmetries to the desired structure functions. The cross section asymmetry $A_{\|}$ is obtained by polarizing the target parallel to the direction of the beam and measuring the difference in scattering rates for left and right electron helicity:

$$A_{\|} = \frac{\sigma^{\downarrow\uparrow} - \sigma^{\uparrow\uparrow}}{\sigma^{\downarrow\uparrow} + \sigma^{\uparrow\uparrow}} = f_k \left\{ g_1(x,Q^2)[E + E'\cos\theta] - \frac{Q^2}{\nu} g_2(x,Q^2) \right\}. \quad (5)$$

The factor f_k includes the unpolarized structure function F_1 and the ratio of the transverse and longitudinal virtual photon cross sections R

$$f_k = \frac{1-\epsilon}{\nu F_1(x,Q^2)[1 + \epsilon R(x,Q^2)]}, \quad (6)$$

where

$$\epsilon = \left[1 + 2(1 + \frac{\nu^2}{Q^2}) \tan^2(\theta/2) \right]^{-1}. \quad (7)$$

In the expression for the cross-section asymmetry, the first arrow indicates the electron helicity and the second arrow indicates the target helicity. The SLAC experiments have also performed the same measurement with the target polarization reversed to minimize electroweak contributions and check for systematic effects. This experimental asymmetry can be related to the virtual photon asymmetries using

$$A_{\|} = D(A_1 + \eta A_2), \quad (8)$$

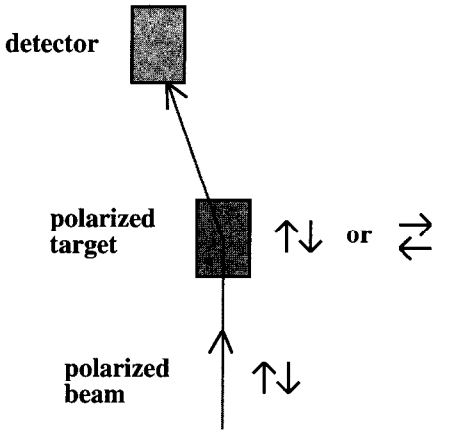


Fig. 2. Polarized inclusive deep-inelastic scattering.

where $\eta = \epsilon\sqrt{Q^2}/(E - E'\epsilon)$ and the depolarization factor is given by $D = (1 - E'\epsilon/E)/(1 + \epsilon R)$. Since A_2 is typically small, we see that $A_{\|} \approx DA_1$, i.e., the asymmetry obtained using electrons is proportional to the virtual photon asymmetry, but is reduced by the depolarization factor.

Another experimental cross-section asymmetry $A_{\perp}$ is obtained by polarizing the target in a direction transverse to the beam and measuring the rates for incident electrons with left and right helicity:

$$A_{\perp} = \frac{\sigma^{\downarrow\leftarrow} - \sigma^{\uparrow\leftarrow}}{\sigma^{\downarrow\leftarrow} + \sigma^{\uparrow\leftarrow}} = f_k E' \sin\theta \left\{ g_1(x,Q^2) + \frac{2E}{\nu} g_2(x,Q^2) \right\}. \quad (9)$$

As in Eq. (5), the factor f_k is given by Eq. (6). For the recent SLAC experiments, θ is small, so $A_{\perp}$ is primarily sensitive to g_2, whereas $A_{\|}$ is primarily sensitive to g_1.

1.4 Polarized Structure Functions

Using Eqs. 2 and 3, one can see that in leading order we expect the longitudinal structure function g_1 to be given by

$$g_1(x,Q^2) = \frac{1}{2}\sum_i e_i^2 \Delta q_i = \frac{1}{2}\sum_i e_i^2 [q_i^{\uparrow}(x,Q^2) - q_i^{\downarrow}(x,Q^2)], \quad (10)$$

where Δq_i is the polarized quark distribution for the ith quark flavor. In leading order, g_1 is only sensitive to the net quark polarizations. In particular, the data from the

isoscalar deuteron target provides a particularly strong constraint on the net quark polarization, $\Delta\Sigma = \Delta u + \Delta d + \Delta s$. However, in next-to-leading order (NLO) perturbative QCD (PQCD) g_1 is also sensitive to the polarized gluon distribution ΔG through the Q^2 evolution of the parton distributions and we can use the factorization theorem[12,13] to express g_1 as

$$g_1(x,Q^2) = \frac{1}{2}\sum_q e_q^2 \left[C_q \bigotimes \Delta q + \frac{1}{N_f} C_G \bigotimes \Delta G \right]. \quad (11)$$

In this equation, N_f is the number of accessible quark flavors and the sum runs over the quark and anti-quark flavors. The coefficients C_q and C_g are the Wilson coefficients and correspond to the photon-quark and photon-gluon hard scattering cross sections. The convolution $\bigotimes$ is given by

$$(C \bigotimes q)(x,Q^2) = \int_x^1 \frac{dz}{z} C\left(\frac{x}{z}, \alpha_s\right) q(z,Q^2), \quad (12)$$

where $\alpha_s(Q^2)$ is the running coupling for the strong interaction. The Q^2 evolution of the parton distributions is governed by the Altarelli-Parisi (or DGLAP) evolution equations.[14] This evolution reflects the increasing resolution of our "microscope" as we increase Q^2. One important consequence of this evolution is to link the quark distributions to the gluon distributions through q $\rightarrow$ gq splitting. This makes it possible to indirectly measure the gluon distributions by measuring the quark distributions over a wide range in Q^2. Given sufficient statistical power over a wide range in x and Q^2, an NLO PQCD analysis of the Q^2 dependence of g_1 can be used to constrain the net polarization of the gluon.

Although the naive picture of the g_2 structure function is somewhat unclear, twist-two calculations from Wandzura and Wilczek provide a relation between g_2 to g_1 via[15]

$$g_2(x,Q^2) = -g_1(x,Q^2) + \int_x^1 g_1(\xi,Q^2) d\xi/\xi. \quad (13)$$

More generally, we expect

$$g_2(x,Q^2) = g_2^{WW}(x,Q^2) - \int_x^1 \frac{\partial}{\partial y}\left(\frac{m}{M} h_T(y,Q^2) + \xi(y,Q^2)\right)\frac{dy}{y}, \quad (14)$$

where the second and third terms are twist-two and -three (respectively). The contribution from the quark transverse spin distribution h_T is suppressed by the ratio of the quark mass to the nucleon mass $\frac{m}{M}$ and should be negligible. The remaining twist-three term reflects the quark-gluon correlations in the nucleon. This term can be obtained from the deviations from g_2^{WW}.

2 The Experiment

The E155 experiment ran in early 1997 and collected $\sim$ 180 million events. Keys to the success of E155 included the high efficiency of beam delivery, the use of lithium deuteride as the deuteron target material, and the addition of a third spectrometer. These factors enabled E155 to collect the data in 2/3 of the originally proposed time and cover almost twice the range in Q^2 as planned in the original proposal. The major components of the experiment are described below.

2.1 Polarized Electron Beam

Longitudinally polarized 48.3 GeV electron beam pulses[16] of up to 400 ns duration[17] were produced at 120 Hz by a circularly polarized laser beam illuminating a strained GaAs photocathode. The helicity for each pulse was selected by a 32-bit pseudo-random number generator to minimize instrumental asymmetries. The beam polarization P_b was determined periodically during the data run using Møller scattering from 20–154 μm thick Fe-Co-V polarized foils. Results from independent single arm[18] and coincidence detector systems agreed within 1%. No statistically significant deviations from the average value of $P_b = 0.813 \pm 0.20$ were seen during the experiment.

2.2 Target

Figure 3 shows a schematic of the polarized target used for E143 and E155.[19] The target material was contained in two copper-coated aluminum cups which were 3 cm long and 2.5 cm in diameter and could be moved into and out of the beam. Polarization of the material was achieved using the technique of dynamic nuclear polarization (DNP).[20] In this technique, microwaves and a high magnetic field are used to polarize paramagnetic electrons, and transfer their polarization to the nucleons. In E155, 140 GHz microwaves were delivered to each of the aluminum cups by separate waveguides. A superconducting Helmholtz pair provided a 5.004 T field with 10^{-4} uniformity in the region occupied by the material. DNP also requires the presence of paramagnetic centers (with a free or quasi-free electron). These centers were created by pre-irradiating the materials at relatively warm temperatures (80–90 K for ammonia and 180 K for lithium deuteride) and preserved by storing the material at liquid nitrogen temperatures. During normal running, a ^{4}He evaporation refrigerator kept the material at 1 K and the electron beam was rastered across the face of the target cup to minimize local polarization loss due to beam

heating. Nuclear magnetic resonance (NMR) measurements were made using coils embedded in the target materials to determine the polarization at regular intervals. These measurements were calibrated to the signal measured at thermal equilibrium (TE) near 1.6 K. These TE polarizations are small (typically $\leq 0.5\%$), but can be compared with the predictions of the Boltzmann distribution to obtain an accurate calibration.[21]

As in E143, ammonia ($^{15}NH_3$) was used as the proton target material. Polarization values of $\sim 90\%$ were typically achieved after 10–20 minutes. The polarization slowly decreased due to radiation damage and was restored by periodic annealing at about 80 K. Unlike E143, which used deuterated ammonia ($^{15}ND_3$), E155 used lithium deuteride[22] because it provides a larger ratio of polarized (effective) deuterons to the total number of nucleons and higher radiation resistance. The lithium (^{6}Li) can, to the first order, be treated as a polarized deuteron plus an unpolarized alpha particle, and therefore half of the nucleons in ^{6}LiD are the desired polarizable species.[23] Lithium deuteride is also five times as radiation resistant as $^{15}ND_3$. Pre-irradiation doses of 1.3-4.5$\times 10^{17} e^-/cm^2$ were used to create the paramagnetic centers necessary for dynamic nuclear polarization. An average in-beam free deuteron polarization $\langle P_t \rangle$ of 22% was achieved during E155. The maximum value was obtained after an exposure of approximately $5\times 10^{15} e^-/cm^2$ in the 48.3 GeV beam (see Fig. 4). An overall relative uncertainty of 4% on P_t was achieved for the deuteron.

2.3 Spectrometers and Detectors

Scattered electrons were detected in three independent magnetic spectrometers at central angles of 2.75°, 5.5°, and 10.5° with respect to the incident beam (see Figs. 5 and 6). The spectrometers at 2.75° and 5.5° were used previously in the E154 experiment.[8] These spectrometers used two dipole magnets with opposing field directions to perform momentum selection, and the 2.75° line included a quadrupole magnet to optimize the distribution of particles at the shower counter. Electrons were identified by two threshold gas Cherenkov counters and an electromagnetic calorimeter that consisted of a 20 by 10 stack of lead glass blocks 24 radiation lengths (X_0) thick. Particle momenta and scattering angles were measured with two sets of scintillator hodoscopes. The 10.5° spectrometer was added for E155 to double the Q^2 range of the experiment (see Fig. 7). It consisted of a single dipole magnet between two quadrupole magnets followed by a scintillator hodoscope, a threshold gas Cherenkov counter, and an electromagnetic calorimeter with a preradiator (PR) section and a total absorption (TA) section. The PR

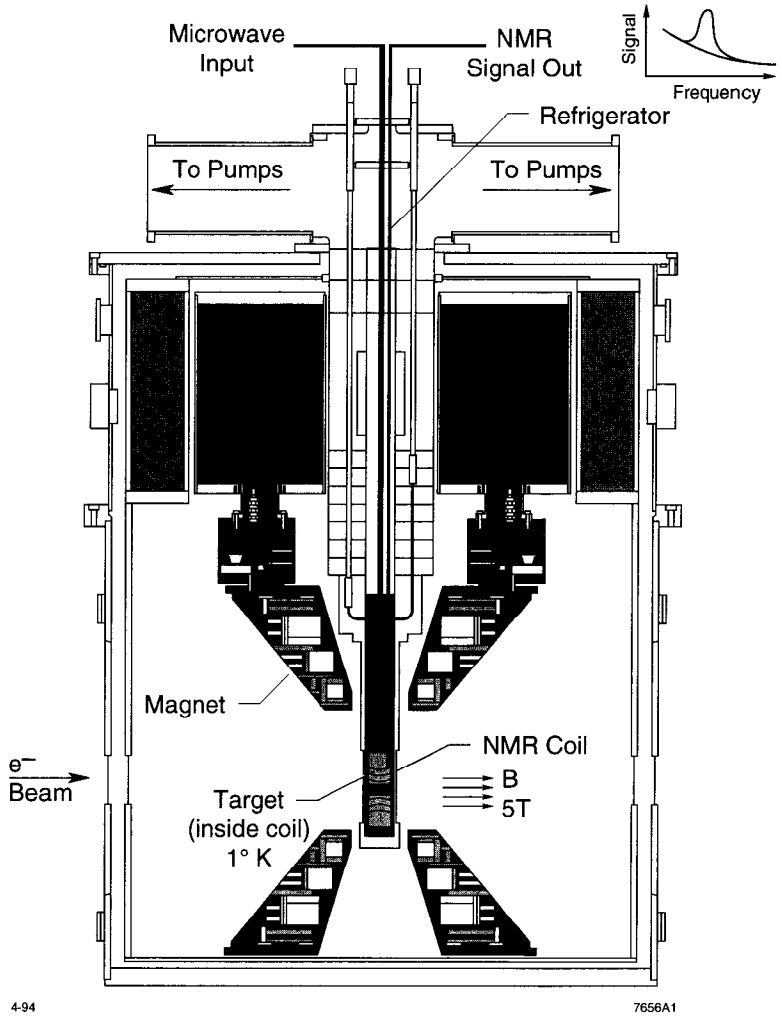


Fig. 3. The E155 target assembly.

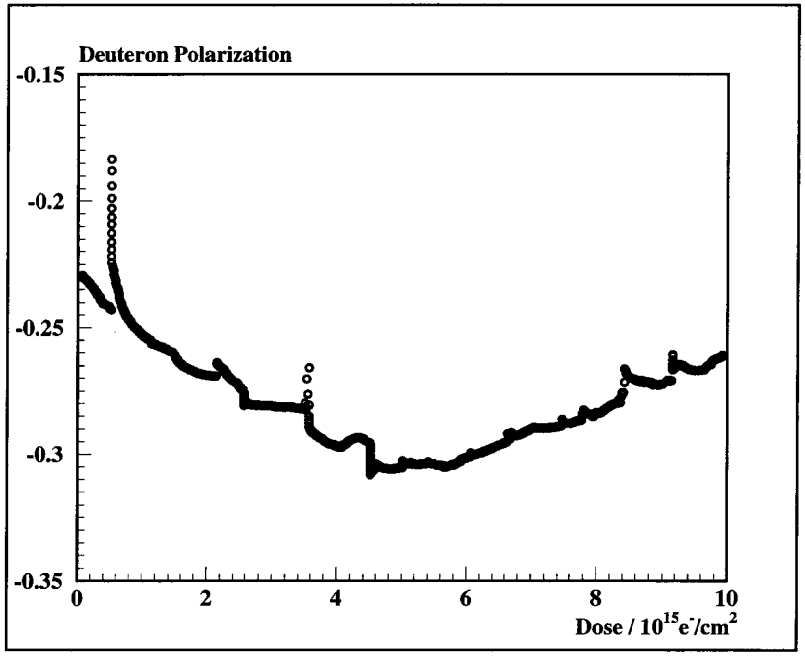


Fig. 4. Lithium Deuteride polarization as a function of accumulated dose. Deuterated ammonia shows the same behavior but reaches a maximum polarization after only 10^{15} e^-/cm^2 exposure.

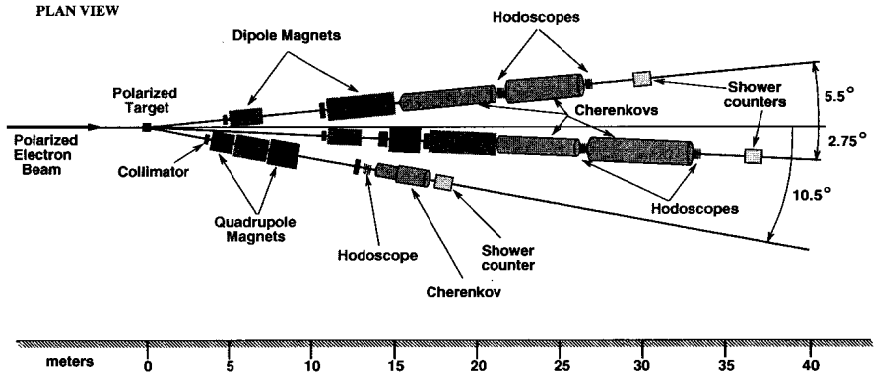


Fig. 5. Plan view of E155 spectrometers.

section consisted of ten $2X_0$ thick lead glass bars. These bars sampled the initial portion of the particle showers to separate electrons from hadrons. They were arranged to measure y position information which was combined with the hodoscope information to determine particle momenta. The TA section consisted of a five by six stack of lead glass blocks $16X_0$ thick.

3 Experimental Asymmetries

The experimental longitudinal ($A_{||}$) asymmetries were determined from the numbers of scattered electrons per incident beam charge for negative (N_-) and positive (N_+) beam helicity for each target using

$$A_{||}^d = \left(\frac{N_- - N_+}{N_- + N_+} \frac{1}{f P_b P_t C_1} - C_2 A_{||}^p \right) \frac{1}{f_{RC}} + A_{RC} \tag{15}$$

for the deuteron and

$$A_{||}^p = \left(\frac{N_- - N_+}{N_- + N_+} \frac{1}{f P_b P_t C_1} \right) \frac{1}{f_{RC}} + A_{RC} \tag{16}$$

for the proton. The rates N_- and N_+ were corrected for contributions from charge symmetric background processes, which were measured by reversing the spectrometer polarity. The asymmetry for the charge symmetric background is consistent with zero,

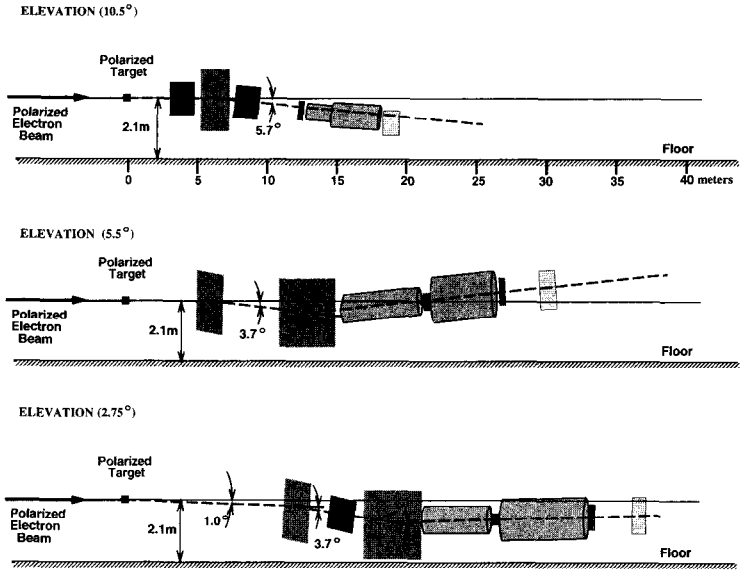


Fig. 6. Elevation view of E155 spectrometers.

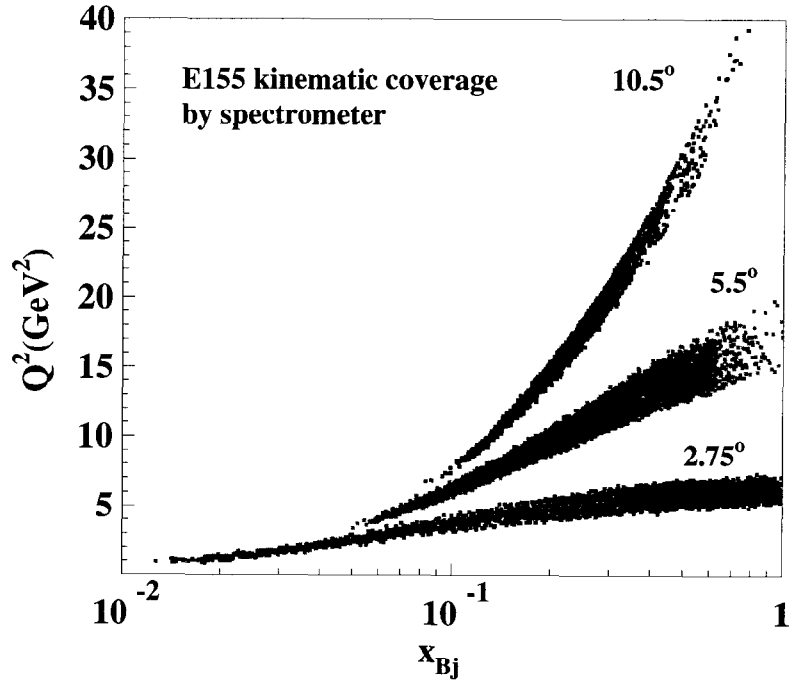


Fig. 7. x and Q^2 range for the E155 spectrometers (note the log scale for x). The data correspond to $\sim 10^6$ beam spills (a few hours of data).

leading to dilution corrections of 10%−15% at the lowest x bin in each spectrometer and decreasing rapidly at higher x. The rates N_- and N_+ were also corrected for mis-identified hadrons, which were typically about 2% or less of the electron candidates. Corrections for the rate dependence of the electron detection and reconstruction efficiency were a few percent for each spectrometer.

3.1 Dilution Factor

The dilution factor f is the fraction of events originating from the polarizable protons or free deuterons. It was determined from the composition of the target materials, containers, and NMR coils in the beam. The x dependence of the cross sections for these nuclei caused f to vary from 0.13 at low x to 0.18 at high x for the proton and from 0.18 at low x to 0.20 at high x for the deuteron. The relative systematic uncertainty on f of $\sim 2\%$ was dominated by the uncertainties of the density and packing fraction of the granules.

3.2 Nuclear Corrections

The factors C_1 and C_2 account for the presence of several polarizable nuclear species in the target. In the case of the proton, the only nuclear correction comes from the polarization of the nitrogen nuclei. This leads to a small correction of $C_1 \approx 1.015$ and no C_2 term. For the deuteron, C_1 includes the contributions from the free deuterons and effective deuterons in ^{6}Li. Calculations based on isospin conservation, shell model, solutions of Fadeev equations and a Green's function Monte Carlo provide consistent values for the valence nucleon polarization[23] in ^{6}Li. From these analyses, we conclude that the effective deuteron in ^{6}Li has a net polarization of 87% of the Li polarization. From this information we determined that C_1 ranges from 1.77 to 1.85, with an uncertainty of ± 0.05. $C_1 f$ then gives an effective dilution factor of ~ 0.36 for ^{6}LiD, as compared with ~ 0.22 for $^{15}ND_3$. The C_2 term accounts for contributions from polarized protons in Li^1H and ^{7}Li. This correction is the product of the fraction of the species present in the granules and a factor for the effective polarization of the protons. The spin-3/2 ^{7}Li nucleus is described well in terms of two clusters: an α particle and a triton. The neutrons in the triton are predominantly anti-aligned, so the only significant polarized contribution comes from the proton.[23] Values for C_2 range from -0.023 to -0.030 with an uncertainty of ± 0.003. The proton asymmetry $A_{||}^p$ was obtained from a fit to world data.

3.3 Radiative Corrections

Both internal[24] and external[25] radiative corrections were obtained using an iterative global fit of all data,[2–10] including the E155 data. Previous SLAC data were recorrected in a manner consistent with the E155 corrections. The radiative multiplicative factor f_{RC} and additive correction A_{RC} were determined in a manner similar to that used in E154.[8] This method leads to somewhat larger corrections than those in E143,[7] which only took into account the quasielastic radiative tails. For the proton, the net corrections range from 0.02 to 0.01. For the deuteron, $A_{||}$ is changed by less than 2% of its value for x above ~ 0.3, and is decreased by ~ -0.005 at low x. The x-and Q^2-dependent systematic uncertainties on $A_{||}$ due to the radiative corrections were typically 0.001, 0.002, and 0.004 for the 2.75°, 5.5°, and 10.5° spectrometers, respectively.

4 Results

4.1 Longitudinal Structure Function

Preliminary results for g_1^p and g_1^d are shown in Figs. 8 and 9. A simple parameterization was used to evolve the data from the experimental values of Q^2 to $Q_0^2 = 5$ (GeV/c)^2, which is roughly the average Q^2 for the experiment. The world data from proton, neutron, and deuteron targets,[2–10] including E155, with $Q^2 > 1$ (GeV/c)^2 were used to constrain separate proton and neutron functions of the form,

$$(g_1/F_1) = x^\alpha(a + bx + cx^2)(1 + \beta/Q^2). \tag{17}$$

The deuteron data were included in the fit using the relation, $g_1^d = \frac{1}{2}(g_1^p + g_1^n)(1 - 1.5\omega_D)$, where the deuteron D-state probability is $\omega_D = 0.05 \pm 0.01$. The variation[26] of ω_D with x is comparable to the uncertainty in ω_D for $x < 0.75$ and negligible compared to the statistical uncertainties for $x > 0.75$. The β values of -0.17 ± 0.05 and -0.14 ± 0.30 obtained from the fit are small and indicate that the Q^2 dependence of g_1 is very similar to the Q^2 dependence of the unpolarized structure function. The parameterization in Eq. (17) and F_1 determined from the recent SLAC fit for R[27] and the NMC fit for F_2[28] were used to evolve the measured g_1 data to $Q_0^2 = 5$ (GeV/c)^2 as shown in Figs. 8 and 9.

The E155 results for g_1^d/F_1^d are shown in Fig. 10 as a function of Q^2. They are in good agreement with world data[2–10] as demonstrated by representative data from SMC

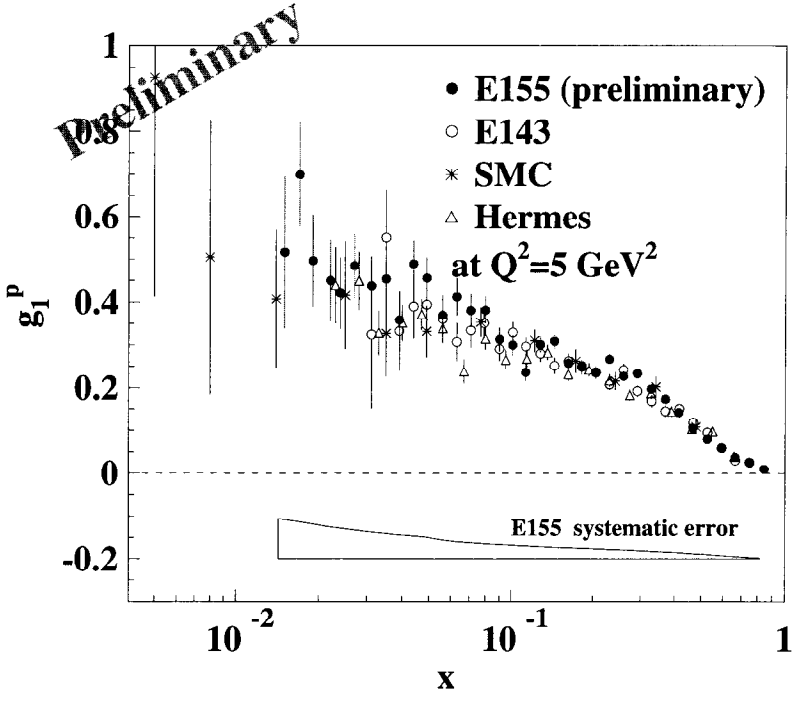


Fig. 8. Proton results for g_1^p evolved to $Q^2 = 5$ (GeV/c)^2 from E155 as a function of x (filled circles). The uncertainties on the points are statistical. The shaded band indicates the systematic uncertainties. Results from E143 (open circles), SMC (stars), and HERMES (triangles) are also shown.

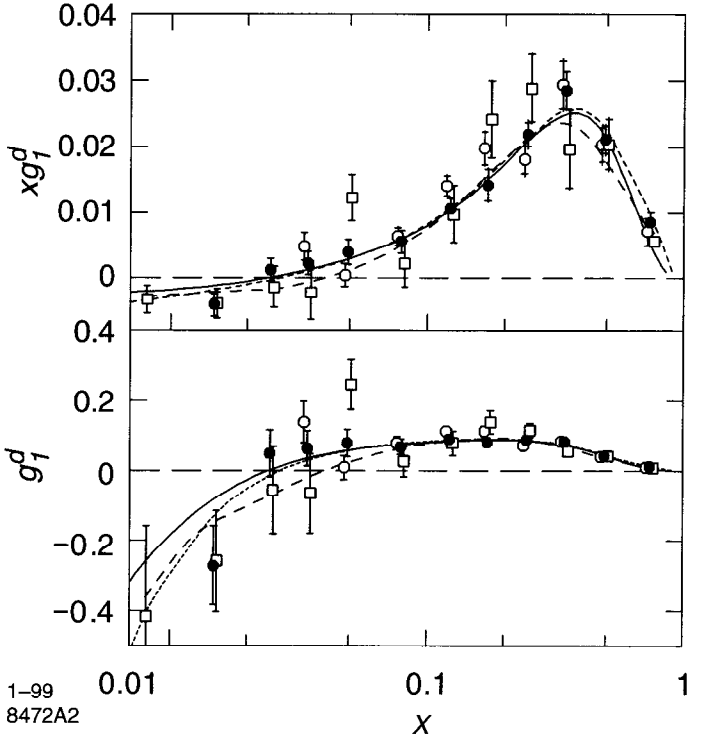


Fig. 9. Deuteron results for g_1^d (bottom) from this experiment (E155) as a function of x (filled circles). The inner bars indicate the statistical uncertainties, while the outer bars include the systematic errors added in quadrature. The top section shows xg_1^d for E155. Results from E143 (open circles) and SMC (open boxes) are also shown.

and E143. The data from the three spectrometers provide both large Q^2 coverage and good statistical resolution in the mid-x region. There is no significant Q^2 dependence for g_1^d/F_1^d, indicating that the polarized and unpolarized structure functions evolve similarly.

Integrating g_1^d over the data range yields $\int_{0.01}^{0.9} g_1^d(x,Q_0^2)dx = 0.0408 \pm 0.0025(\text{stat}) \pm 0.0024(\text{syst})$. The extrapolation for the high x contribution was found to be negligible. However, the contribution from the low x region from 0 to 0.01 does not converge for Eq. (17), reinforcing the need for additional data at very low x. Using the E154[8] perturbative QCD fit for the low x extrapolation gives a value of $\int_0^1 g_1^d(x,Q^2)\,dx = \Gamma_1^d = 0.0288 \pm 0.0025(\text{stat}) \pm 0.0071(\text{syst})$. Alternatively, using the SMC[11] perturbative QCD fit for the low x extrapolation gives a value of $\Gamma_1^d = 0.0299$ with comparable uncertainties.

Using the expression obtained from the operator product expansion,[7] and information on the weak hyperon decay constants from the particle data group,[29] $\Delta\Sigma$ can be extracted from Γ_1^d. Using the E154 and SMC fits for the low x region, values of $\Delta\Sigma = 0.17 \pm 0.03 \pm 0.08$ and 0.19 (respectively) are obtained in the $\overline{MS}$ scheme. These values are well below the Ellis-Jaffe prediction of 0.58,[30] but are in agreement with previous experimental values. The precision of the $\Delta\Sigma$ result obtained using only the E155 deuteron data is comparable with results obtained from NLO fits to all previous world data, such as the value of $\Delta\Sigma = 0.20 \pm 0.06 \pm 0.05$ obtained by E154.

4.2 "Transverse" Structure Function

E155 ran for a short period during the 1997 run at 38.8 GeV beam energy with the target polarization transverse to the beam direction. Results for the g_2 structure function obtained from this data are shown in Figs. 11 and 12. The data from the three spectrometers show no Q^2 dependence within the uncertainties, so the data from the spectrometers have been combined. The data correspond to a fairly short running period with the target polarization transverse to the beam, but the results are comparable with those from E143. The data are generally consistent with g_2^{WW}. However, the current results lack the power to strongly differentiate among g_2^{WW}, bag model predictions[31–33] and zero. This will be addressed by the much larger data set planned for the E155 extension run, which will focus on measurements with the target polarization perpendicular to the beam direction.

Additional information can be obtained from the moments of g_2 using the operator

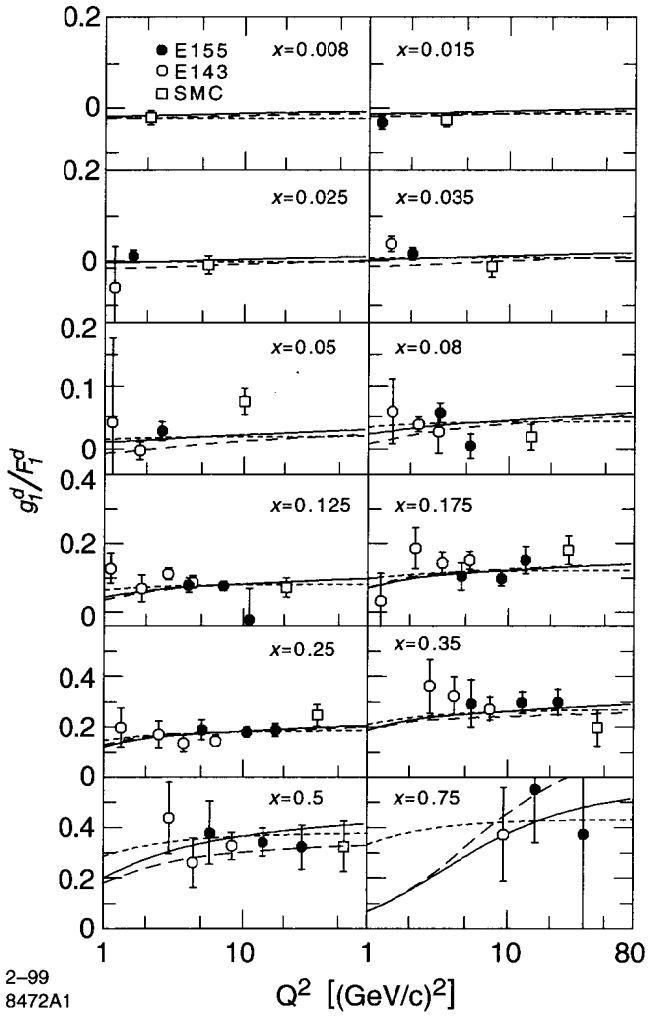


Fig. 10. Deuteron structure function ratios $\frac{g_1^d}{F_1^d}$ from E155 as a function of Q^2 for various x bins (filled circles). The inner bars indicate the statistical uncertainties, while the outer bars include the systematic errors added in quadrature. Results from E143 (open circles) and SMC (open boxes) are also shown.

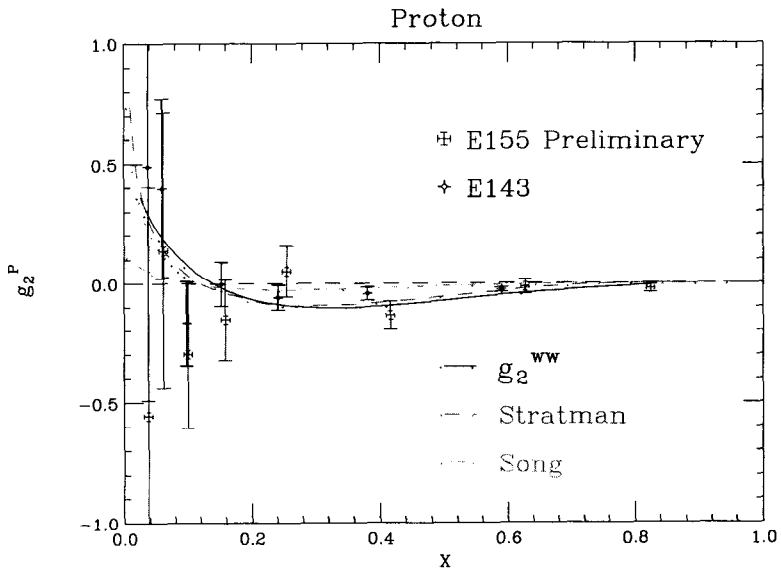


Fig. 11. Preliminary E155 results for the proton structure function g_2^p versus x (crosses). The uncertainties on the points are statistical. The E143 data have also been shown for comparison (diamonds). The solid line indicates the prediction from Wandzura and Wilczek[15] using a fit to world g_1 data. Also shown are calculations from Song[31] (dot-dashed line) and Stratmann[32] (dashed line).

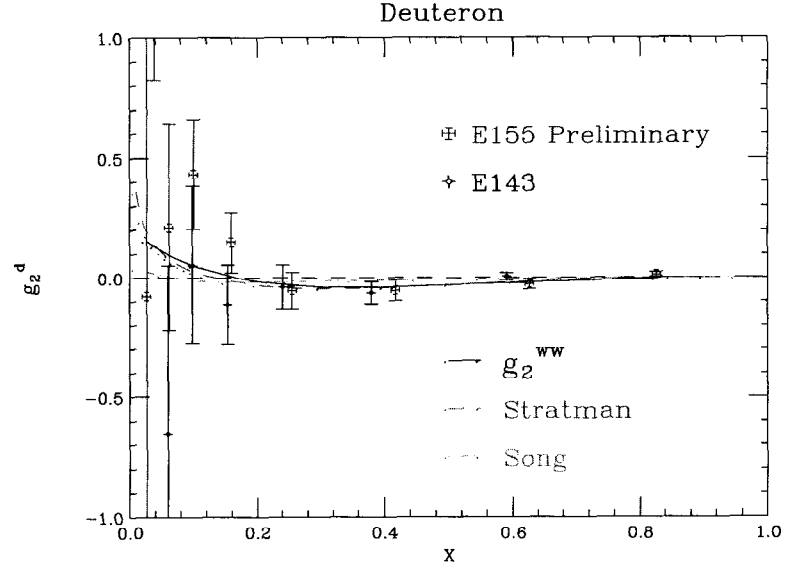


Fig. 12. Preliminary E155 results for the deuteron structure function g_2^d versus x (crosses). The uncertainties on the points are statistical. The E143 data have also been shown for comparison (diamonds). The solid line indicates the prediction from Wandzura and Wilczek[15] using a fit to world g_1 data. Also shown are calculations from Song[31] (dot-dashed line) and Stratmann[32] (dashed line).

product expansion (OPE). The OPE allows us to write the hadronic matrix elements in deep-inelastic scattering in terms of a series of renormalized operators of increasing twist.[34,35] The moments of g_1 and g_2 at fixed Q^2 can be related to the twist-two and twist-three reduced matrix elements, a_n and d_n,[35] and higher twist terms which are suppressed by powers of $1/Q$. Neglecting the quark mass terms, the relations are given by

$$\Gamma_1^{(n)} = \int_0^1 x^n g_1(x,Q^2)dx = \frac{a_n}{2} + \mathcal{O}(M^2/Q^2) \quad \text{for} \quad n = 0,2,4,..., \tag{18}$$

$$\Gamma_2^{(n)} = \int_0^1 x^n g_2(x,Q^2)dx = \frac{1}{2}\frac{n}{n+1}(d_n - a_n) + \mathcal{O}(M^2/Q^2) \quad \text{for} \quad n = 2,4,.... \tag{19}$$

Neglecting higher order terms and rearranging, we see that d_n measures deviations from the g_2^{WW} twist-two prediction,

$$d_n = \frac{2(n+1)}{n}\int_0^1 x^n \left[g_2(x,Q^2) - g_2^{WW}(x,Q^2)\right] dx. \tag{20}$$

Evaluating this expression for $n = 2$ using the E155 data, we obtain values of $d_2^p = 0.007 \pm 0.007$ and $d_2^d = 0.008 + 0.005$ at an average Q^2 of 5 (GeV/c)^2. Combining these results with those from E142,[6] E143,[7] and E154[8] we obtain $d_p = 0.007 \pm 0.004$ for the proton and $d_n = 0.003 \pm 0.010$ for the neutron. Figure 13 shows these results compared with several bag model predictions, QCD sum rule calculations, and lattice QCD. In general, the results seem to be in reasonable agreement with the bag model predictions but do not favor the QCD sum rule predictions as strongly. The proton lattice QCD prediction is currently being redone. Additional data are desirable in order to differentiate between the various predictions.

5 The Future

E155 is currently completing the analysis of data taken using ammonia as the target material. When this analysis has been completed, the data from both target materials will be used along with the existing world data to perform an NLO global fit. The addition of the E155 data will improve the constraints on the parton distributions, especially the poorly understood polarized gluon distribution.

The E155 extension run, scheduled for early 1999, will be a dedicated high precision measurement of the g_2 structure function. The data from this run will significantly improve our knowledge of g_2 and will definitively distinguish between $g_2 = 0$ and g_2^{WW}. The uncertainty in d_2 will be reduced by a factor of three.

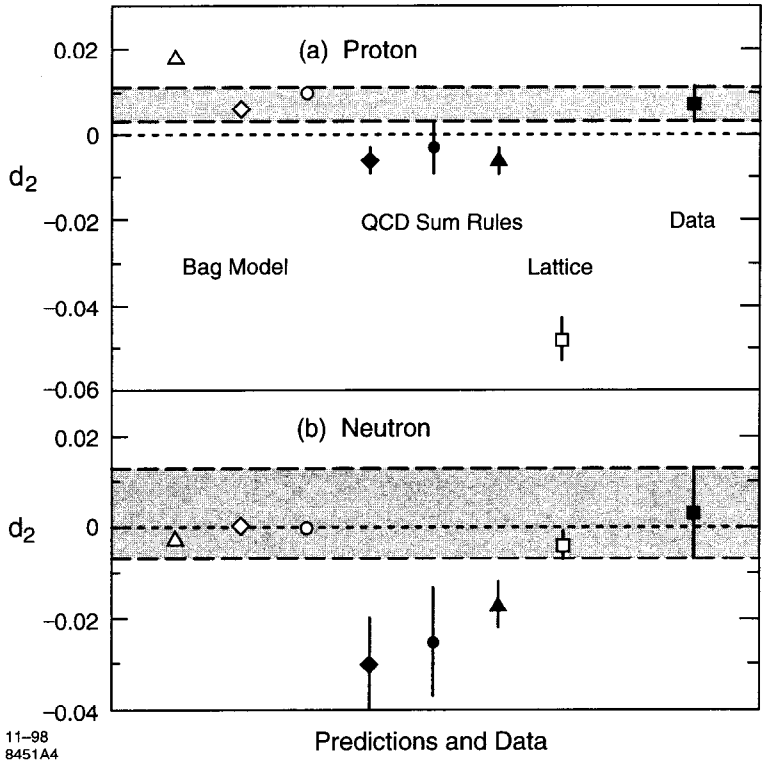


Fig. 13. SLAC results including E155 for the twist-three element d_2. The results are in reasonable agreement with bag model calculations from Song,[31] Stratmann,[32] and Ji and Unrau.[33] The agreement with predictions from QCD sum rules from Stein *et al.*,[36] Balitsky *et al.*,[37] and Ehrnsperger *et al.*[38] is not as strong.

References

[1] R. G. Roberts, *The Structure of the Proton*, Cambridge Univ. Press (1990).

[2] SLAC E80, M. J. Alguard *et al.*, Phys. Rev. Lett. **37**, 1261 (1976).

[3] SLAC E130, G. Baum *et al.*, Phys. Rev. Lett. **51**, 1135 (1983).

[4] EMC, J. Ashman *et al.*, Nucl. Phys. B **328**, 1 (1989).

[5] SMC, B. Adeva *et al.*, Phys. Rev. D **58**, 112001 (1998).

[6] SLAC E142, P. L. Anthony *et al.*, Phys. Rev. D **54**, 6620 (1996);

[7] SLAC E143, K. Abe *et al.*, Phys. Rev. D **58**, 112003 (1998).

[8] SLAC E154, K. Abe *et al.*, Phys. Lett. B **405**, 180 (1997); K. Abe *et al.*, Phys. Rev. Lett. **79**, 26 (1997).

[9] SLAC E155, P. Anthony *et al.* Phys. Lett. B **458**, 529 (1999); SLAC-PUB-8041, `hep-ex/9904002`.

[10] HERMES Collaboration, A. Airapetian *et al.*, Phys. Lett. B **442**, 484 (1998); K. Ackerstaff *et al.*, Phys. Lett. B **404**, 383 (1997).

[11] SMC Collaboration, B. Adeva *et al.*, Phys. Rev. D **58**, 112002 (1998).

[12] H. Georgi and H. D. Politzer, Phys. Rev. D **9**, 416 (1974); D. J. Gross and F. Wilczek, Phys. Rev. D **9**, 980 (1974).

[13] "CTEQ Handbook of Perturbative QCD," Rev. Mod. Phys. **67**, Number 1, 157 (1995).

[14] G. Altarelli and G. Parisi, Nucl. Phys. B **126**, 298 (1977); V. N. Gribov and L. N. Lipatov, Sov. J. Nucl. Phys. **15**, 438, 675 (1972); Yu. L. Dokshitzer, Sov. Phys. JETP **46**, 461 (1977).

[15] S. Wandzura and F. Wilczek, Phys. Lett. B **72**, 195 (1977).

[16] R. Alley *et al.*, Nucl. Instrum. Methods A **365**, 1 (1995); J. E. Clendenin *et al.*, Report No. SLAC-PUB-7619 (1997).

[17] F. J. Decker *et al.*, Report No. SLAC-PUB-7214 (1996).

[18] H. R. Band et al., Nucl. Instrum. Methods Phys. Res., Sect. A **400**, 24 (1997).

[19] D. G. Crabb and D. B. Day, Nucl. Instrum. Methods Phys. Res., Sect. A **356**, 9 (1995).

[20] A. Abragam, *Principles of Nuclear Magnetism*, Oxford University Press (1961).

[21] D. G. Crabb and W. Meyer, Annu. Rev. Nucl. Part. Sci **47**, 67 (1997).

[22] S. Bültmann *et al.*, Nucl. Instr. Meth. A **425**, 23 (1999).

[23] O. A. Rondon, Phys. Rev. C **60**, 035201, (1999).

[24] T. V. Kukhto and N. M. Shumeiko, Nucl. Phys. B **219**, 412 (1983); I. V. Akusevich and N. M. Shumeiko, J. Phys. G **20**, 513 (1994).

[25] Y. S. Tsai, Report No. SLAC-PUB-848 (1971); Y. S. Tsai, Rev. Mod. Phys. **46**, 815 (1974).

[26] W. Melnitchouk, G. Piller, and A. W. Thomas, Phys. Lett. B **346**, 165 (1995).

[27] K. Abe *et al.*, Phys. Lett. B**452**, 194 (1999).

[28] NMC Collaboration, P. Arneodo *et al.*, Phys. Lett. B **364**, 107 (1995).

[29] Particle Data Group, Eur. Phys. J. C **3**, (1998).

[30] J. Ellis and R. Jaffe, Phys. Rev. D **9**, 1444 (1974); *ibid.* D **10**, 1669 (1974).

[31] X. Song, Phys. Rev. D **54**, 1955 (1996).

[32] M. Stratmann, Z. Phys. C **60**, 763 (1993).

[33] X. Ji and P. Unrau, Phys. Lett B **333**, 228 (1994).

[34] E. Shuryak and A. Vainshtein, Nuc. Phys. B **201**, 141 (1982).

[35] R. Jaffe and X. Ji, Phys. Rev. D **43**, 724 (1991).

[36] E. Stein *et al.*, Phys. Lett B **343**, 369 (1995).

[37] I. Balitsky, V. Braun, and A. Kolesnichenko, Phys. Lett. B **242**, 245 (1990); *ibid.* B **318**, 648 (1993) (Erratum).

[38] B. Ehrnsperger and A. Schafer, Phys. Rev. D **52**, 2709 (1995).

RESULTS FROM HERMES

Wolfgang Lorenzon*
Randall Laboratory of Physics
University of Michigan, Ann Arbor, Michigan 48109-1120
and
Deutsches Elektronen-Synchrotron, DESY, D-22603 Hamburg, Germany

Representing the HERMES Collaboration

ABSTRACT

HERMES, HERA Measurement of Spin, is an experiment to study the spin structure of the nucleon. The experiment was installed and commissioned in 1995 in the HERA storage ring at the DESY laboratory in Hamburg, Germany. Inclusive and semi-inclusive spin-dependent deep inelastic scattering on polarized internal gas targets are measured simultaneously. Data were taken on polarized ^{3}He and ^{1}H targets as well as on unpolarized targets of hydrogen, deuterium, ^{3}He, and nitrogen. The experiment is described briefly and some recent results from the 1995 to 1997 running periods are presented.

*Supported by NSF Contract PHY-9724838.

© 1998 by Wolfgang Lorenzon.

1 Introduction

Understanding how the spin of the nucleon is constructed from its elementary constituents has proven to be a significant challenge both theoretically and experimentally. In a simple parton model the spin of the nucleon can be decomposed as[1]

$$\frac{1}{2} = \frac{1}{2}\Delta\Sigma + \Delta G + \Delta L_q + \Delta L_G, \quad (1)$$

where $\Delta\Sigma$ and ΔG are the intrinsic spins carried by the quarks and gluons, and ΔL_q and ΔL_G the orbital angular momenta of the quarks and gluons. Whereas the left hand side of Eq. (1) is very well known, the right hand side has been investigated for more than twenty years, starting at SLAC in 1975. Scattering polarized leptons off polarized nucleons has been the method of choice in the past to understand the contributions to the nucleon spin from quarks. Using an analysis based on the quark parton model with QCD corrections, the main conclusion of these polarized deep inelastic scattering experiments is that the quarks contribute only at the 20% level to the nucleon's spin,[2] much lower than expected from the relativistic quark parton model. No direct measurements of the other components are available at present.

Many open questions remain: what are the contributions of the other components; what are the individual quark contributions to $\Delta\Sigma$, i.e., what is the contribution of the valence quarks, the sea quarks—in particular the strange quarks, and what is the role of the glue?

2 HERMES Experiment

HERMES has been designed[3] to address many of these open questions using a new experimental technique to measure spin-dependent deep inelastic scattering. Inclusive and semi-inclusive spin-dependent measurements are done simultaneously on undiluted polarized ^{1}H, ^{2}D, and ^{3}He targets. Semi-inclusive spin asymmetries resulting from hadron particle identification provide the means to investigate the spin structure of the nucleon in a new way. HERMES provides data not only on polarized structure functions but also on unpolarized structure function ratios, investigates the light sea flavor asymmetry, extracts information on fragmentation functions, and studies hadronization in nuclei. A sizeable program exists also in diffractive vector meson production of ρ, ϕ, ω, and J/Ψ mesons to extract cross sections, decay angular distributions, and to study nuclear transparency. HERMES also studies transversity and azimuthal distributions in polarized Λ production, etc.

The HERMES experiment is located in the east section of the HERA electron-proton collider. The experiment is arranged such that the 27.5 GeV longitudinally polarized electron beam interacts with internal polarized gas targets without interference from the proton beam. The beam is self-polarized transverse to the beam direction by emission of synchrotron radiation in the curved sections (Sokolov-Ternov effect[4]). Spin rotators located at the entrance and exit of the east straight section[5] precess the spin direction from vertical to longitudinal at the target position. The polarization builds up with a rise time of about 25 minutes and is measured with two polarimeters. Figure 1 shows a typical polarization rise time recorded by both polarimeters.

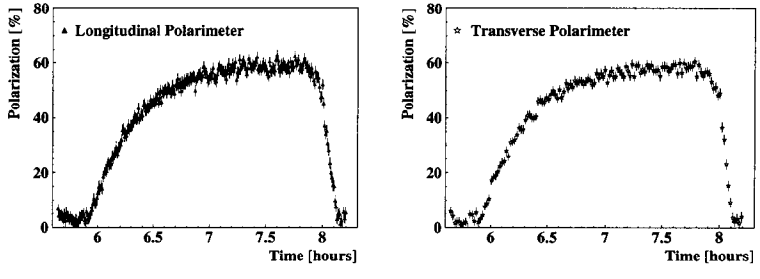


Figure 1. Polarization rise time recorded simultaneously with the longitudinal and transverse polarimeters to compare build-up time and general performance of both polarimeters. At the beginning and the end of the spectrum, the beam was depolarized on purpose.

In 1995 data were taken with a polarized ^{3}He target, and in 1996/97 HERMES was running with a polarized hydrogen target. The HERMES target[3,6] is contained in an open-ended thin walled storage cell through which the circulating electron beam of the HERA accelerator passes. The target region surrounding the storage cell is displayed schematically in Fig. 2. A magnetic holding field provides a quantization axis for the target polarization. Polarized gas enters the target cell through a central feed tube and then leaves the cell at the ends, where high speed pumps prevent the gas from flowing into the electron beam pipe. In 1995, a typical ^{3}He polarization of 50% was obtained. In 1996 and 1997, the atomic beam source had achieved hydrogen polarization of 80–95%. Unpolarized hydrogen, deuterium, ^{3}He, and nitrogen gases up to 10^{15} atoms/cm^2 have also been used to check the detectors in the spectrometer and to take

unpolarized data. The maximum target thickness is determined by the impact on the stored electron lifetime.

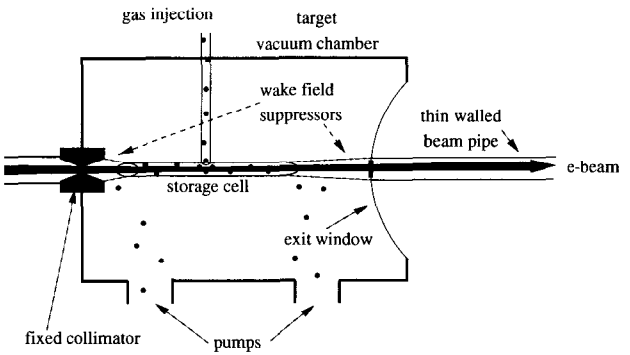


Figure 2. Schematic of the HERMES target region. See text for more explanation.

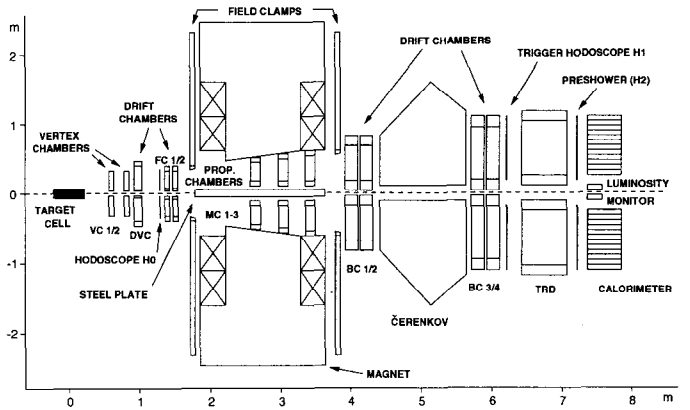


Figure 3. Schematic side view of the HERMES spectrometer.

The HERMES spectrometer, as shown in Fig. 3, was designed[3,7] to detect scattered electrons and hadrons from deep inelastic scattering. The spectrometer consists of two identical halves above and below the electron ring plane. This provides two independent measurements of observables and thus, a cross check on systematic uncertainties. The HERMES spectrometer contains a bending magnet to measure particle momenta and to reject background. Tracking chambers before the magnet, inside the magnetic field, and behind the magnet provide charged particle tracking. The Cherenkov, transition radiation detector (TRD), hodoscope (H1), and calorimeter are used for particle identification, and the magnet and calorimeter for momentum and energy analysis.

3 The Spin Structure of the Nucleon

Experimentally, the spin structure function g_1 is extracted from the measured asymmetry, $A_{\parallel}$, of the scattering cross section as the beam or target spin is flipped. These asymmetries are measured with longitudinally polarized beams and longitudinally polarized targets. Ignoring contributions from transverse asymmetries for the purpose of simplicity here, this can be written as $A_{\parallel} \approx D(x,y)A_1$, where $D(x,y)$ is the virtual photon depolarization factor and A_1 the longitudinal virtual photon asymmetry. Finally, g_1 is determined using

$$g_1(x,Q^2) = \frac{A_{\parallel}}{D}\frac{F_2(x,Q^2)}{2x(1+R(x,Q^2))}\,, \qquad (2)$$

with parametrizations of the unpolarized structure function[9] $F_2(x,Q^2)$ and $R(x,Q^2)$ (Ref. 10), the ratio of longitudinal to transverse virtual photon absorption cross sections.

The first HERMES results for the proton spin structure function[11] taken on polarized ^{1}H in 1997 are presented in Fig. 4 and compared to the published data from SLAC[12] and SMC.[8,13] The structure function ratio g_1^p/F_1^p, which is approximately equal to A_1, is given for the measured Q^2 at each value of x. There is good agreement between these three sets of data, although the Q^2 values of E-143 and HERMES differ from those of SMC by a factor between five and ten. This demonstrates that there is no statistically significant Q^2 dependence of the ratio g_1^p/F_1^p in the Q^2 range of these experiments.

In Fig. 5 the spin structure function g_1^p on the proton is shown as a function of x, evolved to a common Q^2 of 2 and 10 GeV^2 for comparison with the SLAC and SMC results. The evolution was done under the assumption that the ratio g_1^p/F_1^p does not depend on Q^2. The evolved HERMES data are in excellent agreement with both data sets. The apparent Q^2 dependence of g_1^p, seen when comparing the left and right plots, entirely originates from the Q^2 dependence of F_2^p and R. In summary, the HERMES results were obtained using an entirely different technique from that used in all previous experiments, indicating that the systematic uncertainties are well understood for both techniques over the entire measured x range.

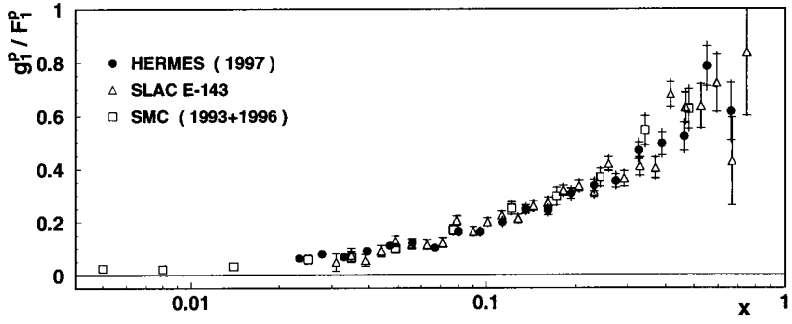


Figure 4. The structure function ratio g_1^p/F_1^p of the proton as a function of x from HERMES is compared with recent results for g_1^p/F_1^p measured at SLAC (E143) and A_1^p at CERN (SMC). The inner error bars show the statistical uncertainties and the outer ones the quadratic sum of the statistical and total systematic uncertainties.

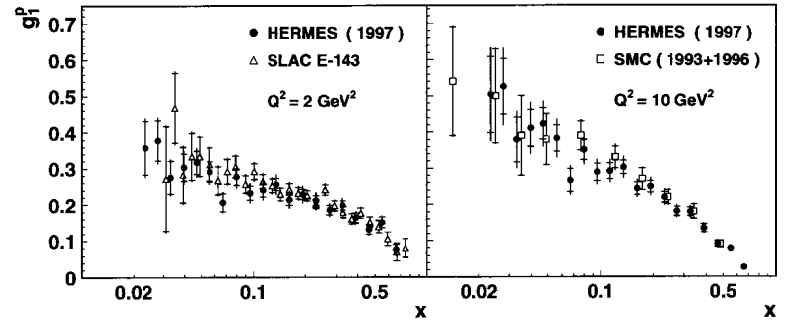


Figure 5. The spin structure function g_1^p on the proton is shown as a function of x. In the left panel, the HERMES data are evolved to $Q^2 = 2$ GeV^2 and compared to recent results for $Q^2 > 1$ GeV^2 from E143 (Ref. 12). In the right panel, the HERMES data are evolved to $Q^2 = 10$ GeV^2 and compared to recent results for $Q^2 > 1$ GeV^2 and $x > 0.01$ from SMC (Ref. 13). The error bars are defined as in Fig. 4.

A summary of spin structure function measurements is shown in Table 1, where the beams, targets, energies, and measured quantities are listed for each experiment. The experiments complement each other in their kinematic coverage and their sensitivities to possible systematic errors associated with the measured quantities. The pioneering

Table 1. Summary of experiments at SLAC, CERN, and DESY to measure spin-dependent deep inelastic scattering.

Lab	Experiment	Year	Beam	Target	Measured Quantities
SLAC	E80	75	23 GeV e^-	H-butanol	A_1^p
	E130	80	23 GeV e^-	H-butanol	A_1^p
	E142	92	25 GeV e^-	^{3}He	A_1^n
	E143	93	29 GeV e^-	NH_3/ND_3	A_1^p,A_1^d,$A_2^{p,d}$
	E154	95	49 GeV e^-	^{3}He	A_1^n, A_2^n
	E155	97	49 GeV e^-	NH_3/LiD	A_1^p,A_1^d
CERN	EMC	85	100-200 GeV μ^-	NH_3	A_1^p
	SMC	92	100 GeV μ^+	D-butanol	A_1^d,$A_{1,h}^d$
	SMC	93	190 GeV μ^+	H-butanol	A_1^p,$A_{1,h}^p$
	SMC	94/95	190 GeV μ^+	D-butanol	A_1^d,$A_{1,h}^d$
	SMC	96	190 GeV μ^+	NH_3	A_1^p,$A_{1,h}^p$
DESY	HERMES	95	28 GeV e^+	^{3}He	A_1^n, $A_{1,h}^n$, $A_{\perp,\pi}^n$
	HERMES	96/97	28 GeV e^+	^{1}H	A_1^p, $A_{1,h}^p$, $A_{1,\pi}^p$

experiments on the proton were performed at SLAC[14] with modest polarizations of beam and target. They were followed by more precise proton data from EMC[15] which, while consistent with the SLAC data, showed a large deviation from the Ellis-Jaffe sum rule[16] and a corresponding small value for the implied quark contribution to the proton's spin. This result spurred a new generation of experiments at SLAC, CERN, and DESY with significant improvements in precision.

The new experiments at SLAC, CERN, and DESY over the last six years have focused on high precision measurements of the structure function g_1 over a large kinematic range to extract information on the spin structure. They appear to be consistent if they are evolved to the same Q^2. The experiments confirmed EMC's initial observation of the violation of the Ellis-Jaffe sum rule. The Bjorken sum rule,[17] on the other hand, is verified at the 1 σ (8%) level and is considered a successful test of QCD on spin observables. They however indicate that only about 20% of the spin comes from the spin of the quarks. Therefore, the question of how the spin of the nucleon is related to its internal structure is not resolved. The need for measurements of other observables using more sophisticated experimental techniques is evident: g_1 needs to be measured at very small x; the polarization of the sea, in particular the polarization of the strange

quarks using flavor decomposition, needs to be measured; and a direct measurement of the gluon helicity distribution ΔG is important to make definitive statements about the nucleon's spin structure.

4 Semi-Inclusive Physics

To this end, HERMES is currently running a broad scope of semi-inclusive measurements. This is done by measuring the hadronic reaction products in coincidence with the scattered lepton. As shown schematically in Fig. 6, it should be possible in appropriate kinematics to tag the flavor of the struck quark[18] which is reflected by the type of hadron in the current jet. In this way the right hand side of Eq. (1) may be decomposed.

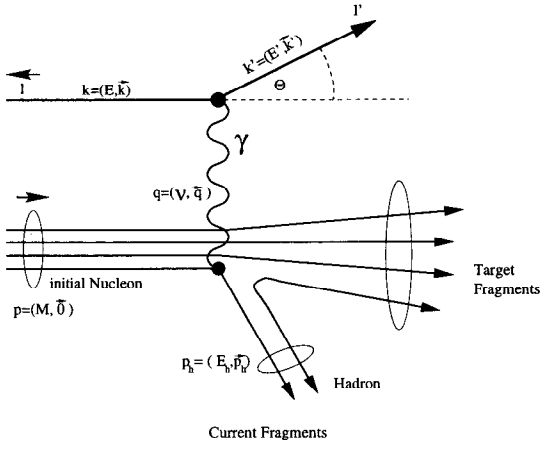


Figure 6. A schematic illustration of the hadronization process.

4.1 Flavor Tagging

Assuming factorization between the hard scattering process and the hadronization of the struck quark, the hadron asymmetry can be expressed as

$$A_1^h(x,z) \simeq \frac{A_{||}^h(x,z)}{D(x,Q^2)} \simeq \frac{\sum e_f^2 \Delta q_f(x) \cdot D_{q_f}^h(z)}{\sum e_f^2 q_f(x) \cdot D_{q_f}^h(z)}. \quad (3)$$

This equation differs from the inclusive formula by the factor $D_{q_f}^h(z)$, which represents the probability for a struck quark of flavor f to fragment into a hadron of type h with energy fraction of the virtual photon $z = E_h/\nu$. A measurement of the hadron asymmetry will then give access to new linear combinations of polarized quark distributions Δq.

Of course, the hadronization is a statistical process, and a quark with flavor f may produce either a leading π^+ or π^-. However, because of the valence quark structure of mesons, the probabilities are not equal and one can define both *favored* and *disfavored* fragmentation functions, D^+ and D^-, respectively. These pion fragmentation functions, as extracted from the HERMES data using the EMC extraction method, are shown in Fig. 7. They are compared to the EMC data.[19] There is good agreement between these two data sets.

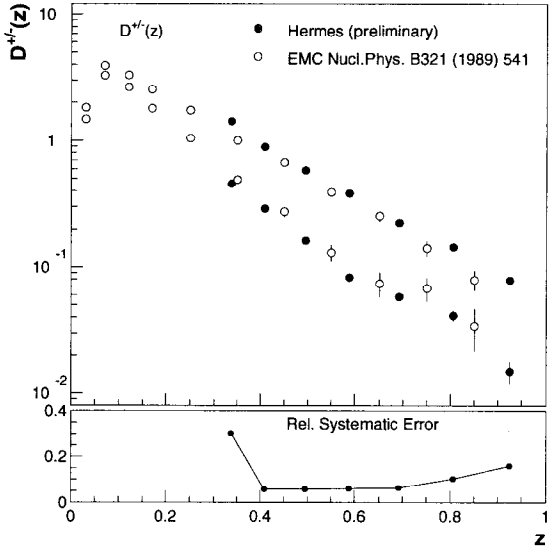


Figure 7. Preliminary data on the favored D^+ and disfavored D^- pion fragmentation functions from HERMES are compared to data from EMC. They are shown in the top panel. The same extraction method has been used for both data sets. The relative systematic errors are displayed in the lower panel.

Figure 8 displays the hadron asymmetries measured in 1995 on polarized ^{3}He and in 1996 on polarized ^{1}H targets. In the top panel the inclusive asymmetry A_1^p, and the semi-inclusive asymmetries for positive $A_1^p(h^+)$ and negative $A_1^p(h^-)$ hadrons on proton targets are shown. The corresponding asymmetries on ^{3}He are shown in the

bottom panel. The semi-inclusive asymmetries are compared to the available SMC semi-inclusive data.[20] There is generally good agreement between the two data sets.

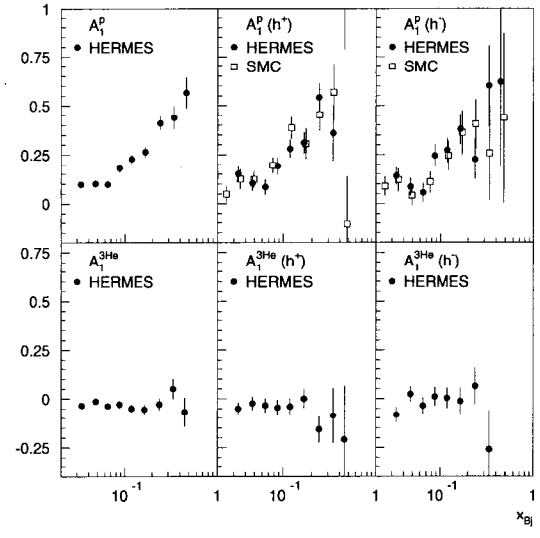


Figure 8. Preliminary data for positive and negative charged hadron asymmetries on the proton and ^{3}He from HERMES are compared to available data from the SMC experiment. Also included are the inclusive asymmetries on the proton and ^{3}He from HERMES for comparison.

In order to extract the polarized quark distributions, several quantities are needed[21] as input: the measured hadron asymmetries $A_1^h(x,z)$, the unpolarized quark distributions $q(x)$, and the fragmentation functions $D_{q_f}^h(z)$. It is further necessary to rewrite the photon-nucleon asymmetry (Eq. (3)) in such a way that one can isolate the quark polarizations, and end up with a sum over quark purities times the quark polarizations:

$$A_1^h(x,z) = \sum_q \frac{e_q^2 q(x) D_q^h(z)}{\sum_{q'} e_{q'}^2 q'(x) D_{q'}^h(z)} \cdot \frac{\Delta q(x)}{q(x)} = \sum_q P_q^h(x,z) \cdot \frac{\Delta q(x)}{q(x)}. \quad (4)$$

The quark purity P_q^h is the conditional probability that a quark q was struck in an event, if a hadron of type h was produced. The purities are spin-independent, unpolarized quantities. They are produced by Monte Carlo simulations tuned using unpolarized HERMES hadron multiplicities. A matrix equation including various hadron asymmetries and quark polarizations can then be defined as

$$\vec{A} = \mathcal{P}\vec{\mathcal{Q}}. \quad (5)$$

To extract the quark polarizations this matrix equation has to be solved.

Since the sea purities are too small for decomposition with the statistics that is currently available, it is assumed that the relative polarization of the sea quarks is flavor symmetric,

$$\frac{\Delta u_{sea}}{u_{sea}} = \frac{\Delta d_{sea}}{d_{sea}} = \frac{\Delta s}{s}, \quad (6)$$

to reduce the number of degrees of freedom. It is then possible to extract the polarizations of the u-valence, d-valence, and sea quarks. In order to compare with the data from SMC,[20] x times the polarized quark distributions are plotted in Fig. 9.

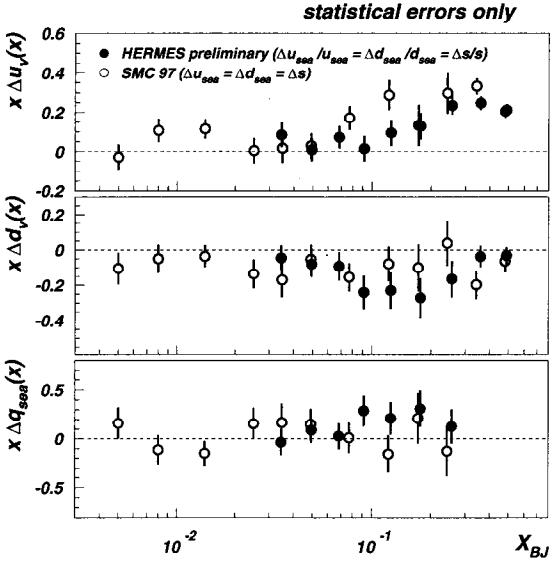


Figure 9. Preliminary data on the quark spin distributions at the measured Q^2 are shown separately for the valence Δu_v, Δd_v, and the sea $\Delta\bar{u}$ as a function of x. The error bars shown are statistical only. The distributions are compared to results from SMC evolved to $Q^2 = 10$ GeV^2. The error bars of the SMC result correspond to its total uncertainty.

Note, that even though different sea quark assumptions are used for HERMES (equal polarizations) and SMC (SU(3) flavor symmetric sea), the results change by less than 10%. The SMC results, evolved to a fixed Q^2 of 10 GeV^2, are in rough agreement with the polarized quark distributions at HERMES, which are given at their measured Q^2.

4.2 Unpolarized Semi-Inclusive Results

As with polarized scattering, flavor tagging with unpolarized targets allows flavor dependent separation of valence and sea quark probability densities. By forming the ratio of charged pion yields on ^{1}H and ^{2}D targets,

$$r(x,z) = \frac{N_p^{\pi^-}(x,z) - N_n^{\pi^-}(x,z)}{N_p^{\pi^+}(x,z) - N_n^{\pi^+}(x,z)}, \tag{7}$$

one can extract the flavor content of the nucleon sea. Here $N^\pi(x,z)$ is the yield of pions coming from deep inelastic scattering off nucleons. Until recently, there has been little experimental constraint on the flavor asymmetry in the distributions of the light sea quarks in the nucleon. The first one was an integral test, based on a comparison of inclusive deep inelastic scattering on the proton and the neutron to determine the Gottfried sum[22] defined as $S_G = \int_0^1 \frac{dx}{x}(F_2^p(x) - F_2^n(x))$, where $F_2^p(x)$ and $F_2^n(x)$ are the structure functions of the proton and neutron, respectively. The assumption of isospin symmetry in the quark parton model allows the sum to be written as $S_G = \frac{1}{3} - \frac{2}{3}\int_0^1(\bar{d}(x) - \bar{u}(x))dx$. A flavor symmetric sea, $\bar{d}(x) = \bar{u}(x)$, leads to the Gottfried Sum Rule $S_G = 1/3$. A measurement by NMC[23] resulted in $S_G = 0.235 \pm 0.026$. If isospin symmetry holds, a global flavor asymmetry $\int_0^1(\bar{d}(x) - \bar{u}(x))dx \approx 0.15$ would account for the NMC result. Two methods have been proposed to measure its x dependence: The Drell-Yan process[24] and semi-inclusive deep inelastic scattering.[25] HERMES can measure the ratio $(\bar{d} - \bar{u})/(u - d)$ from the ratio of charged pion yields defined in Eq. (7), in semi-inclusive DIS on ^{1}H and ^{2}D.

Results[26] are presented for the x dependence of $(\bar{d} - \bar{u})/(u - d)$ in Fig. 10. Within the measured x region, the values are nonzero and positive everywhere, clearly showing an excess of $\bar{d}$ quarks over $\bar{u}$ quarks in the proton. Note that no higher order QCD corrections have been made and so only leading order parametrizations should be compared to these results. At low Q^2, only GRV 94 LO[27] is available at leading order. The distribution $\bar{d} - \bar{u}$ as a function of x is thus derived from $(\bar{d} - \bar{u})/(u - d)$ using the GRV 94 LO parametrization of $u_v(x) - d_v(x)$. The alternative approach to the measurement of the flavor asymmetry is the comparison of the Drell-Yan process on protons and deuterons, which is sensitive to the quantity $\bar{u}/\bar{d}$. The E866 collaboration at Fermilab recently reported a measurement of the flavor asymmetry[31] using the CTEQ 4M parametrization of $\bar{u} + \bar{d}$ to extract $(\bar{d} - \bar{u})$ from the measured quantity $\bar{u}/\bar{d}$. These results at Q^2 of 54 GeV^2 are included in the lower plot. This figure demonstrates that the sea asymmetry measured in deep inelastic scattering and in Drell-Yan experiments agree, even though the Q^2 of the two experiments differ by a factor of about 20.

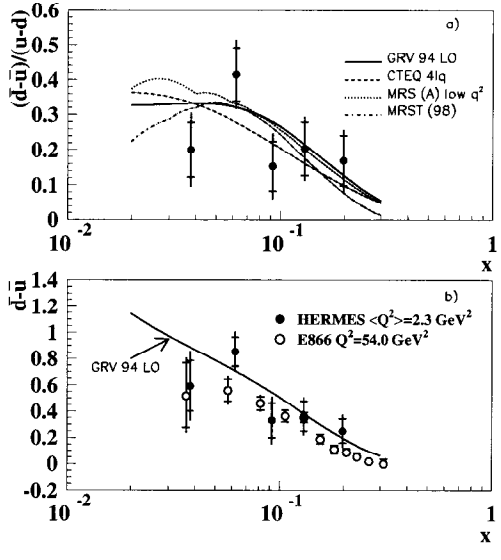


Figure 10. In the top panel, $(\bar{d} - \bar{u})/(u - d)$ as a function of x is shown. Also included are GRV 94 LO (Ref. 27), CTEQ 4lq (Ref. 28), MRS (A) low Q^2 (Ref. 29), and MRST (98) (Ref. 30) parametrizations calculated at the appropriate Q^2 for each x-bin. In the bottom panel $(\bar{d}$-$\bar{u})$ as a function of x is shown. The curve is the GRV 94 LO parameterization. The E866 data determined from $(\bar{d}$-$\bar{u})$ are represented by open circles.

5 Future

At present, no direct measurement of the gluon contribution to the nucelon's spin is available. While NLO analyses of the inclusive data suggest that the gluon carries a

significant fraction of the nucleon spin[32-36] to compensate the deficit due to the intrinsic quark distribution, it is clear that a direct measurement of the gluon spin contribution is very important. Several experimental efforts are underway to measure spin-dependent charm production which should probe the gluons via the photon-gluon fusion reaction. These include the COMPASS experiment at CERN (Ref. 37), the SLAC proposal E156 (Ref. 38), and the HERMES experiment at DESY. The HERMES collaboration has just finished an upgrade of the spectrometer to enhance kaon identification using a Ring Imaging Cherenkov (RICH) counter to probe the strange sea, and to enhance charm identification to probe the gluon spin (Ref. 39). In addition, the RHIC (Ref.40) experiments at BNL plan to probe the gluon spin by direct photon production. Further, it has been shown that HERA, with polarized protons in collider mode and with significantly increased luminosity, can probe gluon spin at low x (Ref. 41). Table 2 summarizes the likely sensitivities of these experiments. The upgrade of the HERMES spectrometer could provide direct information on the sign of the gluon polarization by about 2000. The planned experiments at CERN, RHIC, and SLAC should be able to provide significantly more precise information in the early years of the next decade.

Table 2. Summary of experiments at BNL, CERN, SLAC, and DESY to directly probe gluon spin.

Experiment	Results expected by	Kinematic range	$\delta\left(\frac{\Delta G}{G}\right)$
HERMES	1999	$x_g \sim 0.3$	~ 0.5 /year
RHIC	$\sim 2001/2$	$x_g \sim 0.05 - 0.3$	$\sim 0.01 - 0.3$
COMPASS	$\sim 2002/3$	$x_g \sim 0.15$	~ 0.1
SLAC E-156	deferred	$x_g \sim 0.1 - 0.5$	~ 0.02
pol. HERA	not approved	$x_g \sim 0.01 - 0.1$	~ 0.1

6 Conclusions

The HERMES experiment has successfully finished its third year of data taking. Inclusive and semi-inclusive data are collected simultaneously using polarized internal gas targets. This novel target technology has proven to work well and has provided competitive results with different systematic errors than the previous experiments. The recent upgrades to the spectrometer have opened new possibilities for HERMES. It is expected that by the end of 1999, HERMES will provide precise knowledge of the valence and sea quark polarizations, report on the first direct measurement of the strange quark polarization, and maybe make the first direct measurement of the gluon polarization.

7 Acknowledgements

I wish to thank my colleagues in the HERMES collaboration. I acknowledge C. A. Miller for critical reading of the manuscript. The author's research is supported in part by the U. S. National Science Foundation, Nuclear Physics Division under grant No. PHY-9724838. I thank the University of Michigan and DESY for support while on research leave.

References

[1] R. L. Jaffe and A. Manohar, Nucl. Phys. B **337**, 509 (1990).

[2] D. Adams *et al.*, Phys. Lett. B **357**, 248 (1995).

[3] HERMES Collaboration, Technical Design Report, DESY-PRC 93/06 (1993).

[4] A. A. Sokolov and I. M. Ternov, Sov. Phys. Dokl. **8**, 1203 (1964).

[5] J. Buon and K. Steffen, Nucl. Instrum. Methods A **245**, 248 (1986).

[6] F. Stock *et al.*, Nucl. Instrum. Methods A **343**, 334 (1994); L. H. Kramer *et al.*, Nucl. Instrum. Methods A **365**, 49 (1995).

[7] K. Ackerstaff *et al.*, Nucl. Instrum. Methods A **417**, 230 (1998).

[8] D. Adams *et al.*, Phys. Lett. B **336**, 125 (1994).

[9] M. Arneodo *et al.*, Phys. Lett. B **364**, 107 (1995).

[10] L. W. Whitlow *et al.*, Phys. Lett. B **250**, 193 (1990).

[11] A. Airapetian *et al.*, Phys. Lett. B **442**, 484 (1998).

[12] K. Abe *et al.*, Phys. Rev. Lett. **74**, 346 (1995); Phys. Lett. B **364**, 61 (1995); Phys. Rev. D **58**, 112003 (1998).

[13] D. Adams *et al.*, Phys. Rev. D **56**, 5330 (1997); B. Adeva *et al.*, Phys. Lett. B **412**, 414 (1997).

[14] M. J. Alguard *et al.*, Phys. Rev. Lett. **37**, 1261 (1976); *ibid.* **41**, 70 (1978).

[15] J. Ashman *et al.*, Phys. Lett. B **206**, 364 (1988); Nucl. Phys. B **328**, 1 (1989).

[16] J. Ellis and R. L. Jaffe, Phys. Rev. D **9**, 1444 (1974); *ibid.* **10**, 1669 (1974).

[17] J. D. Bjorken, Phys. Rev. **148**, 1467 (1966); Phys. Rev. D **1**, 1376 (1970).

[18] J. M. Niczyporuk and E. E. W. Bruins, Phys. Rev. D **58**, 091501 (1998).

[19] M. Arneodo *et al.*, Nucl. Phys. B **321**, 541 (1989).

[20] B. Adeva *et al.*, Phys. Lett. B **420**, 180 (1998).

[21] M. A. Funk, "A Measurement of the Polarized Parton Densities of the Nucleon in Deep-Inelastic Scattering at HERMES," Ph.D. thesis, University of Hamburg, DESY-Thesis-1998-017, (1998) unpublished.

[22] K. Gottfried, Phys. Rev. Lett. **18**, 1174 (1967).

[23] M. Arneodo *et al.*, Phys. Rev. D **50**, 1 (1994).

[24] S. D. Ellis and W. J. Stirling, Phys. Lett. B **256**, 258 (1991).

[25] J. Levelt *et al.*, Phys. Lett. B **263**, 498 (1991).

[26] K. Ackerstaff *et al.*, Phys. Rev. Lett. **81**, 5519 (1998).

[27] M. Glück *et al.*, Z. Phys. C **67**, 433 (1995).

[28] H. L. Lai *et al.*, Phys. Rev. D **55**, 1280 (1997).

[29] A. D. Martin *et al.*, Phys. Rev. D **51**, 4756 (1995).

[30] A. D. Martin *et al.*, Eur. Phys. J. C **4**, 463 (1998).

[31] E. A. Hawker *et al.*, Phys. Rev. Lett. **80**, 3715 (1998); J. C. Peng *et al.*, Phys. Rev. D **58**, 092004 (1998).

[32] J. Ellis and M. Karliner, Phys. Lett. B **341**, 397 (1995).

[33] R. D. Ball, S. Forte, and G. Ridolfi, Phys. Lett. B **378**, 255 (1996).

[34] M. Glück, E. Reya, M. Stratmann, and W. Vogelsang, Phys. Rev. D **53**, 4775 (1996).

[35] T. Gehrmann and W. S. Stirling, Phys. Rev. D **53**, 6100 (1996).

[36] G. Altarelli *et al.*, Nucl. Phys. B **496**, 337 (1997).

[37] COMPASS Proposal, CERN/SPSLC-96-14 (1996).

[38] SLAC Proposal E156, March 15, 1997.

[39] M. Amarian *et al.*, HERMES Report 97-004, "The HERMES Charm Upgrade Program," January 24, 1997.

[40] G. Bunce *et al.*, Particle World **3**, 1 (1992).

[41] A. De Roeck *et al., Proceedings of the Workshop on Future Physics at HERA* 1995/96, Volume 2, page 803, September 1996.

RESULTS FROM THE CHORUS EXPERIMENT

Maria-Gabriella Catanesi
CERN
CH1211, Geneva, Switzerland

Representing the CHORUS Collaboration

ABSTRACT

Recent results from the CHORUS (**C**ern **H**ybrid **O**scillation **R**esearch apparat**US**) experiment are summarized. CHORUS is an experiment designed to study neutrino oscillations using the pure and intense Wide Band ν_μ beam at SPS (CERN). A fraction of the neutrino interactions collected during the years 1994–1996 by the CHORUS experiment has been analyzed, searching for ν_τ charged current interactions, followed by the τ lepton decaying into a negative hadron or into a muon. Within the applied cuts, no ν_τ candidate has been found. This result leads to a 90% C.L. limit $P(\nu_\mu \rightarrow \nu_\tau) < 6.0 \cdot 10^{-4}$ for the mixing probability. We also report on the first direct observation of a neutrino-induced charged current interaction with two subsequent decays of short-lived particles, which a complete analysis allows us to interpret as D_s^{+*} production, followed by the decay chain $D_s^{+*} \rightarrow D_s^+\gamma$, $D_s^+ \rightarrow \tau^+\nu_\tau$, and $\tau^+ \rightarrow \mu^+\nu_\mu\bar{\nu}_\tau$.

© 1998 by Maria-Gabriella Catanesi.

1 Introduction

The question of neutrino mass and stability remains one of the most interesting in the actual physics scenario. This question is particularly relevant for the ν_τ, a member of the third generation of fermions, and one of the most promising candidates for Dark Matter in the universe.

Stimulated by these considerations, a new generation of accelerator experiments was built in order to search, with improved sensitivities, for $\nu_\mu \rightarrow \nu_\tau$ oscillations.[1,2] The approach adopted by the CHORUS experiment was to use a visual technique that allows a direct observation of the τ decay topology. The experiment took data in the reoptimized and rebuilt CERN Wide Band ν_μ beam from April '94 to September '97 (see Ref. 3). The new beam presents a higher intensity ($> 2 \times 10^{13}$ protons/cycle) and a higher mean neutrino energy ($\langle E_{\nu_\mu} \rangle \approx 27$ GeV). The mean distance of the neutrino source from the detector is 600 m. This, together with the event statistics and backgrounds, fixes the sensitivity of the experiment. The aim is to explore mixing angles down to $\sin^2(2\theta) \sim 10^{-4}$ putting, in the case of the absence of a signal, a new limit of a factor of 20 better than the existing one (see Ref. 5).

2 The Detector

CHORUS is a "classic" appearance experiment. In fact, it was conceived to identify ν_τ by detecting the τ path and decay vertex. In particular, the aim is to isolate a "few" signal events coming from the charge current interactions

$$\nu_\tau N \rightarrow \tau^- X, \tag{1}$$

followed by one of the decay topologies:

$$\begin{aligned} \tau^- &\hookrightarrow \mu^- \nu\nu && 18\%, \\ &\hookrightarrow h^- \nu + n\pi^0 && 50\%, \\ &\hookrightarrow h^- h^- h^+ + n\pi^0 && 14\%, \end{aligned}$$

and to separate these events from the large background of charged and neutral current interactions:

$$\begin{aligned} \nu_\mu N &\rightarrow \mu^- X, \\ \nu_\mu N &\rightarrow \nu_\mu X. \end{aligned}$$

Because of the very short τ^- decay length ($c\tau = 90$ μm), we need a precise device such as a nuclear emulsion to provide the required high-spatial resolution at the micron range.

The CHORUS target was made with 800 kg emulsion gel, segmented into four stacks of one radiation length each. Each stack was subdivided into 36 sheets. This configuration allows us fast, automatic, and semiautomatic scanning techniques [5] (developed at Nagoya University) which permit us to speed up the search time. Further improvements can be obtained by reducing the scanning area. This goal is reached by introducing, between the emulsion stacks, a scintillating fiber tracking detector which predicts the track's position at the stack exit.[6]

The detection of long-lived particles is done in CHORUS by using two magnetic spectrometers and a calorimeter (see Fig. 1). The first spectrometer, placed downstream from the target, consists of an array of three diamond-shaped fiber planes in a magnetic field of 0.12 T generated by a hexagonal air-core magnet[7] operating in pulsed mode, that allows determination of the sign and momentum of low-energy particles (below 10 GeV).

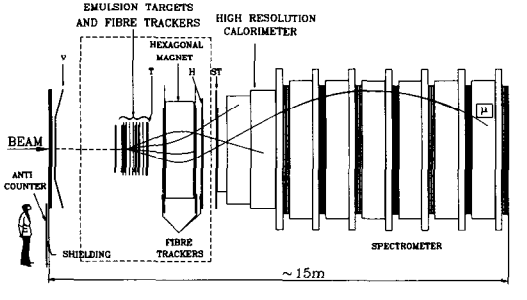


Figure 1: The CHORUS detector.

The second one, placed downstream from the calorimeter, is a "classic" muon spectrometer that allows measurement of the momentum of muons up to 100 GeV/c, with

a resolution of 20% at 75 GeV/c. It consists of six circularly magnetized iron modules with a magnetic field of 1.8 T, interleaved with a tracking section composed of drift chambers and streamer tube planes. The CHORUS sampling calorimeter is the first large-scale application of the so-called "spaghetti technique."[8] It consists of a lead and scintillating fiber matrix optimized for compensation, with an equal response to electromagnetic and hadronic showers. A full description of the CHORUS detector can be found in Ref. 9.

3 Data Selection and Analysis

3.1 Data Collection and Selection Criteria

In the 1994–1997 period, CHORUS collected 2,271,000 triggers corresponding to 5.06×10^{19} protons on target. Of these, 458,601 have a muon identified in the final state (the so-called 1μ events) and 116,049 do not (the so-called 0μ events), and a vertex position compatible with one of the four emulsion target stacks.

All tracks that are associated with the interaction vertex, with an angle of less than 0.4 rad from the beam axis, and (in view of the large emulsion background of muons originating from a nearby secondary target) more than 0.05 rad from the direction of this target, are extrapolated downstream and selected for further analysis. These tracks are searched for in the emulsion if their charge is negative and their momentum is within the range of $0 \leq p_h \leq 20$ GeV/c and $0 \leq p_\mu \leq 30$ GeV/c for hadrons and muons, respectively.

3.2 Vertex Location in Emulsion

The various steps for finding the plate containing the vertex by means of fully automated microscopes are identical to those described in Refs. 10 and 11. They are independently applied to all of the selected tracks in the event, namely the muon for the 1μ events, and all the negative tracks for the 0μ events. A track which has been found in the interface emulsion sheets is followed upstream in the target emulsion using track segments reconstructed in the most upstream 100 μm of each plate until the track disappears. This plate is referred to as the vertex plate. The mean efficiency of this scan-back procedure is found to be $\sim 32\%$ and $\sim 42\%$ for 0μ and 1μ events, respectively. The scanning results are summarized in Table 1.

Table 1: Current status of the CHORUS analysis.

	1994	1995	1996
Protons on target	$0.81 \cdot 10^{19}$	$1.20 \cdot 10^{19}$	$1.38 \cdot 10^{19}$
Emulsion triggers	422,000	547,000	617,000
1μ to be scanned	66,911	110,916	129,669
0μ to be scanned	17,731	27,841	32,548
1μ scanned so far	42,154	49,912	72,615
0μ scanned so far	8,908	12,635	-
1μ vertex located	18,286	20,642	30,128
0μ vertex located	3,401	3,805	0

3.3 Decay Search

Once the vertex plate is defined, automatic microscopic measurements are performed to select events potentially containing a decay topology (kink). As a result of the progress made in scanning procedures and improved performance in speed of the scanning devices, different algorithms have been applied. They are described in Refs. 10 and 11, and briefly recalled here.

In the first procedure, the event is selected either when the scan-back track has a significant impact parameter with respect to the other predicted tracks, or when the change in the scan-back track direction between the vertex plate and the exit from the emulsion corresponds to an apparent transverse momentum, p_T, which is larger than 250 MeV/c. For selected events, and for those with only one predicted track, digital images of the vertex plate are recorded and analyzed offline for the presence of a kink.

The second procedure is restricted to the search of *long* decay paths. In this case, the vertex plate is assumed to contain the decay vertex of a charged parent particle produced in a more upstream plate. With this procedure, only kink angles larger than 0.025 rad are detected.

For the events selected by either one of these procedures, a computer assisted eye-scan is performed to assess the presence of a secondary vertex and to accurately measure its topology. A τ^- decay candidate must satisfy the following criteria:

1. the secondary vertex appears as a kink without black prongs, nuclear recoils, blobs, or Auger electrons;
2. the transverse momentum of the decay muon (hadron), with respect to the parent direction, is larger than 250 MeV/c (to eliminate the decays of strange particles);
3. the kink in the 0μ channel occurs within three plates downstream from the neutrino interaction vertex plate. Because of the lower background, the kink search in the muonic decay channel was extended to five plates, with an efficiency gain of about 8%.

No τ^- decay candidate has been found which satisfies this selection criteria.

3.4 Background Estimates

In this section, we discuss the expected background from known sources for both hadronic and muonic τ decay channels.

The sources of background for the hadronic τ decay channel are:

- the production of negative charmed particles from the antineutrino components of the beam. These events constitute background if the primary μ^+ or e^+ remains unidentified. We expect ~ 0.02 events from these sources;
- the production of positive charmed mesons in charged current interactions, if the primary lepton is not identified and the charge of the charmed particle daughter is incorrectly measured. We expect ~ 0.03 events from this source in the present sample;
- the associated charm production both in charged current (when the primary muon is lost) and neutral current interactions, when one of the charmed particles is not detected. In the present sample, the estimated background from this process, taking into account the total charged current cross section[12] and the upper limit for associated charm production in charged current interactions, is < 0.01 events;
- the main potential background to the hadronic τ^- decays is due to the so-called hadronic "white kinks," defined as one-prong nuclear interactions with no heavily ionizing tracks (*black* and *grey* tracks in emulsion terminology), and no evidence for nuclear break up (evaporation tracks, recoils, blobs, or Auger electrons). Published data, which allows the determination of the white kink interaction cross section, is scarce.[13,14] The main source of data is from a dedicated experiment with 4 GeV pions at KEK.[13]

  Since the experimental information on the p_T dependence is statistically poor at large values, a Monte Carlo simulation based on a modified version of FLUKA[15,16] has been performed. The results of this simulation are in good agreement with the KEK p_T-dependence measurement. We estimate with the current statistics a background of 0.5 events within three plates downstream from the primary vertex plate.

The main source of potential background in the muonic τ channel is charm production. We expect less than 0.1 events in the current sample from the antineutrino components of the beam

$$\bar{\nu}_\mu(\bar{\nu}_e)M \rightarrow \mu^+(e^+)D^-X,$$

followed by

$$D^- \rightarrow \mu^- X^0,$$

in which the $\mu^+(e^+)$ escapes detection, or is not identified.

The prompt ν_τ contamination of the beam[17] is a background common to both the hadronic and muonic decay channels. For the present sample, the expected background is much less than 0.1 events.

4 Results

4.1 Oscillation Sensitivity

In the usual approximation of a two-flavor mixing scheme, the probability of ν_τ appearance in an initially pure ν_μ beam can be expressed as

$$P_{\mu\tau}(E) = \sin 2t \cdot \int \Psi(E,L) \cdot \cdot \sin^2\left(\frac{1.27 \cdot \Delta m^2_{\mu\tau}(\mathrm{eV}^2) \cdot L(\mathrm{km})}{E(\mathrm{GeV})}\right) \cdot dL,$$

where E is the incident neutrino energy, L is the neutrino flight length to the detector, $\theta_{\mu\tau}$ is the effective $\nu_\mu - \nu_\tau$ mixing angle, $\Delta m^2_{\mu\tau}$ is the difference of the squared masses

of the two assumed mass eigenstates, and $\Psi(E,L)$ is the fraction of ν_μ with energy E originating at a distance between L and $L+dL$ from the emulsion target.

The τ^- channels considered in the present $\nu_\mu \to \nu_\tau$ oscillation search are (1) $\tau \to \mu$, (2) $\tau \to h$, (3) $\tau \to e$, and (4) $\tau \to \bar{\mu}$ (where the μ is not identified).

The expected number $N_{\tau i}$ ($i = 1,2,3,4$) of observed τ^- decays into a channel of branching ratio BR_i is then given by

$$N_{\tau i} = BR_i \cdot \int \Phi_{\nu_\mu} \cdot P_{\mu\tau} \cdot \sigma_\tau \cdot A_{\tau i} \cdot \epsilon_{\tau i} \cdot dE, \tag{2}$$

with

- $BR_{(1\ or\ 4)} = BR(\tau \to \nu_\tau \bar{\nu}_\mu \mu^-) = (17.35 \pm 0.10)\%$ (see Ref. 18).
- $BR_2 = BR(\tau \to \nu_\tau h^- n h^0) = (49.78 \pm 0.17)\%$ (see Ref. 18);
- $BR_3 = BR(\tau \to \nu_\tau \bar{\nu}_e e^-) = (17.83 \pm 0.08)\%$ (see Ref. 18);
- Φ_{ν_μ} the incident ν_μ flux spectrum;
- σ_τ the charged current ν_τ interaction cross section;
- $A_{\tau i}$ the acceptance and reconstruction efficiency for the considered channel (up to the vertex plate location); and
- $\epsilon_{\tau i}$ the corresponding efficiency of the decay search procedure.

With proper averaging (denoted by $\langle\rangle$), $N_{\tau i}$ can also be written as a function of n_i

$$N_{\tau i} = BR_i \cdot n_i \cdot \langle P_{\mu\tau} \rangle \cdot \frac{\langle \sigma_\tau \rangle}{\langle \sigma_\mu \rangle} \cdot \frac{\langle A_{\tau i} \rangle}{\langle A_\mu \rangle} \cdot \langle \epsilon_{\tau i} \rangle, \tag{3}$$

where

- $n_1 = N_\mu$ (the number of located charged current ν_μ interactions corresponding to the considered event sample) and $n_2 = n_3 = n_4 = (N_\mu)_{0-\mu}$ (the product of N_μ and the relative fraction of the 0-μ sample for which the analysis has been completed);
- $\langle \sigma_{\mu(\tau)} \rangle = \int \frac{d\sigma_{\mu(\tau)}}{dE} \cdot \Phi_{\nu_\mu} \cdot dE$. It takes into account quasi-elastic interactions, resonance production, and deep-inelastic interactions $[\sigma(\frac{\langle \sigma_\tau \rangle}{\langle \sigma_\mu \rangle})_{syst.} \sim 7\%]$;
- $\langle A_{\mu(\tau i)} \rangle = \int \frac{d\sigma_{\mu(\tau)}}{dE} \cdot A_{\mu(\tau i)} \cdot \Phi_{\nu_\mu} \cdot dE$ $[\sigma(\frac{\langle A_{\tau i} \rangle}{\langle A_\mu \rangle})_{syst.} \sim 7\%]$;
- $\langle \epsilon_{\tau i} \rangle$ is the average efficiency of the decay search procedure for the accepted events $[\sigma(\langle \epsilon_{\tau i} \rangle)_{syst.} \sim 10\%]$;

Table 2: Quantities used in the estimation of the sensitivity.

	1994	1995	1996
N_μ	18,286	20,642	30,128
r_σ	1.89	1.89	1.89
r_A	0.93	0.93	0.93
$\langle A_{\tau\mu} \rangle$	0.39	0.39	0.39
$\langle A_{\tau h} \rangle$	0.17	0.17	–
$\langle A_{\tau e} \rangle$	0.093	0.093	–
$\langle A_{\tau\bar{\mu}} \rangle$	0.026	0.026	–
$\langle \epsilon_{\tau\mu} \rangle$	0.53	0.35	0.37
$\langle \epsilon_{\tau h} \rangle$	0.24	0.25	–
$\langle \epsilon_{\tau e} \rangle$	0.12	0.13	–
$\langle \epsilon_{\tau\bar{\mu}} \rangle$	0.22	0.23	–
N_μ^{eq}	11,987	12,743	–

To allow an easy combination of the results from 1-μ and 0-μ event samples, it is useful to define the "equivalent number of muonic events" of the 0-μ sample by

$$N_\mu^{eq} = (N_\mu)_{0-\mu} \cdot \sum_{i=2}^{4} \frac{\langle A_{\tau i} \rangle}{\langle A_{\tau\mu} \rangle} \cdot \frac{\langle \epsilon_{\tau i} \rangle}{\langle \epsilon_{\tau\mu} \rangle} \cdot \frac{BR_i}{BR_\mu}. \tag{4}$$

The 90% C.L. upper limit on the oscillation probability then simplifies to

$$P_{\mu\tau} \leq \frac{2.38 \cdot r_\sigma \cdot r_A}{BR_\mu \cdot \langle \epsilon_{\tau\mu} \rangle \cdot [N_\mu + N_\mu^{eq}]}, \tag{5}$$

where $r_\sigma = \langle \sigma_\mu \rangle / \langle \sigma_\tau \rangle$ and $r_A = \langle A_\mu \rangle / \langle A_{\tau\mu} \rangle$.

In the above formula, the numerical factor of 2.38 takes into account the total systematic error (17%), following the description given in Ref. 19. The systematic error is mainly due to the reliability of the Monte Carlo simulation on scanning procedures.

The estimated values of the quantities appearing in this expression are given in Table 2. No statistical errors are quoted since they are much smaller than the systematic uncertainties.

Using the present sample, the following 90% C.L. limit is obtained

$$P_{\mu\tau} \leq 6.0 \cdot 10^{-4}. \tag{6}$$

In a two-flavor mixing scheme, the 90% C.L. excluded region in the ($\sin^2 2\theta_{\mu\tau}$, $\Delta m^2_{\mu\tau}$) parameter space is shown in Fig. 2. The maximum mixing between ν_μ and ν_τ is excluded at 90% C.L. for $\Delta\ m^2_{\mu\tau} > 0.9\ \mathrm{eV}^2$; and large Δm^2 values are excluded at 90% C.L. for $\sin^2 2\theta_{\mu\tau} > 1.2 \cdot 10^{-3}$.

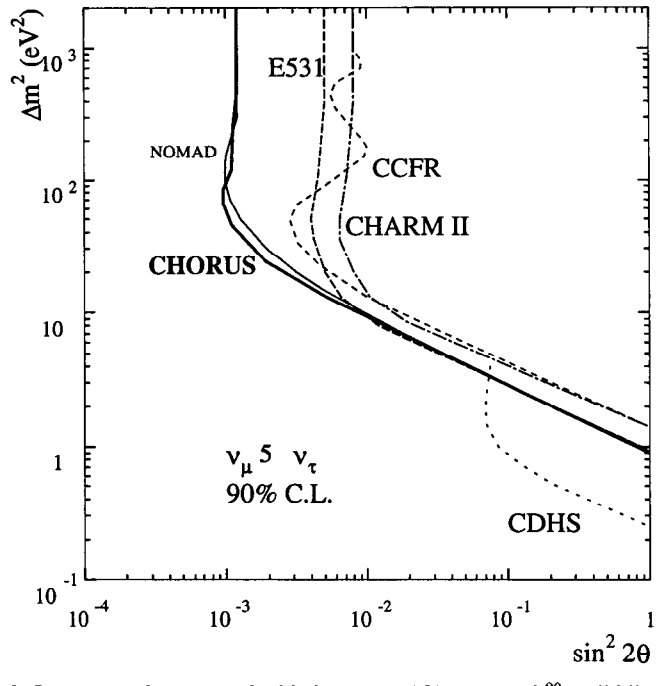
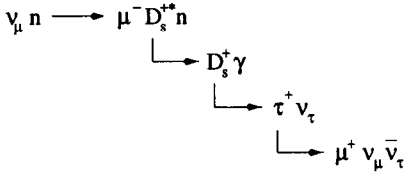


Figure 2: Present result compared with the recent NOMAD result[20] (solid line) and the previous limits (dotted lines).

4.2 The D_s^{+*} Observation

We have observed a neutrino-induced charged current interaction with two subsequent decays within 215 μm. A complete analysis of this event is possible because of the exceptional tracking capabilities of the CHORUS emulsion detector. Topological and kinematical reconstruction for the complete event are consistent with the decay chain



To find double kink events from the reaction $\nu_\mu N \to \mu^- D_s^+ (^* \to D_s^+ \gamma)(X^0)$, $D_s^+ \to \tau^+ \nu_\tau, \tau^+ \to \mu^+ \nu_\mu \bar{\nu}_\tau$, events with two opposite sign muons in the electronic detector are selected. To be sensitive to $D_s^{+*} \to D_s \gamma$ decays, we only allow electromagnetic energy in the calorimeter and veto hadronic activity. After finding the primary vertex, the event signature is manually examined for decay topologies with a double kink. One event recorded in the 1994 exposure fulfills all requirements for this event selection (see Fig. 3).

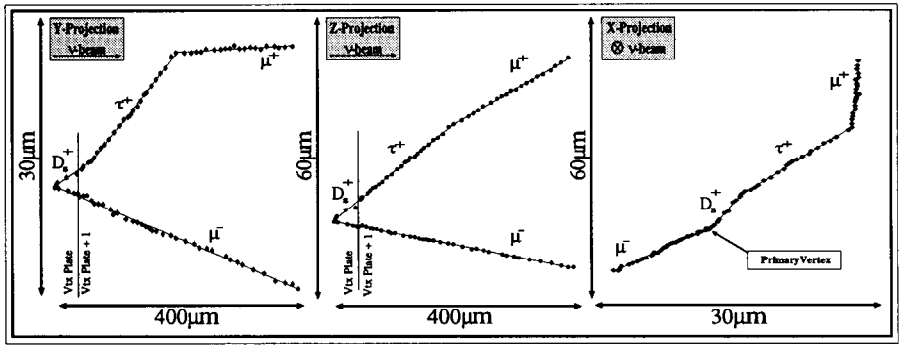


Figure 3: The double kink event in the emulsion.

In view of the short flight lengths with the measured momenta and decay angles, we assume the neutrino-induced production of a charmed particle (D or D_s meson) at the primary vertex. The probabilities for the different scenarios to take place and for detection in the CHORUS detector have been studied. The decay chain $D_s \to \tau \to \mu$ is clearly favored. Furthermore, the measured decay angles and momenta of the candidate event are consistent with this hypothesis.

Therefore, we may conclude that the observed double kink originates from the signature $D_s^+ \to \tau^+\nu_\tau$, $\tau^+ \to \mu^+\nu_\mu\bar{\nu}_\tau$.

At the primary vertex, no nuclear break up is observed. Low $Q^2 = (0.77 \pm 0.15)$ (GeV/c)^2 and the signal of a recoil neutron ($t = (0.89 \pm 0.39)$ (GeV/c)^2) point to a diffractive D_s^* production.

5 Conclusion

The emulsion scanning methods, previously described in Refs. 10 and 11, have been applied to a fraction of the 1994, 1995, and 1996 data. No τ^- decay candidate has been found, leading to a more stringent 90% C.L. upper limit on the $\nu_\mu \to \nu_\tau$ oscillation probability ($P_{\mu\tau} \leq 6.0 \cdot 10^{-4}$). During the second phase of the analysis (with better efficiencies, larger statistics, and faster automatic emulsion scanning), we plan to reach the design sensitivity ($P_{\mu\tau} \leq 1.0 \cdot 10^{-4}$).[1] The first direct observation of a neutrino-induced charged current interaction with a D_s^{+*} production detected at the vertex proves the exceptional capability of CHORUS in the detection of rare processes involving the decay of short-lived particles.

References

[1] CHORUS Collaboration, "A new search for $\nu_\mu \to \nu_\tau$," CERN/PPE/93-131.

[2] NOMAD Proposal, CERN-SPSC/91-21 (1991), SPSC P261 (1991).

[3] G. Aquistapace *et al.*, "The west area neutrino facility for CHORUS and NOMAD experiments: '94 to '97 operations," CERN/ECP/95-14.

[4] N. Ushida *et al.* (E531), Phys. Rev. Lett. **57**, 2897 (1986).

[5] S. Aoki *et al.*, "Fully automated emulsion analysis system, Nucl. Instrum. Methods B **51**, 466 (1990).

[6] S. Aoki *et al.*, "Scintillating fiber tracker with opto-electronics readout for the CHORUS neutrino experiment," Nucl. Instrum. Methods A **344**, 254 (1994).

[7] F. Bergsma *et al.*, "The hexagonal toroidal air-core magnet of the CHORUS detector," Nucl. Instrum. Methods A **357**, 243 (1995).

[8] S. Buontempo *et al.*, "Response to electrons and pions of the calorimeter for the CHORUS experiment," CERN-PPE/96-188, Nucl. Instrum. Methods A **378**, 222 (1996).

[9] E. Eskut *et al.*, CHORUS Collaboration, Nucl. Instr. Method A **401**, 7 (1997).

[10] E. Eskut *et al.*, CHORUS Collaboration, Phys. Lett. B **424**, 202 (1998).

[11] E. Eskut *et al.*, CHORUS Collaboration, Phys. Lett. B **434**, 205 (1998).

[12] N. Ushida *et al.*, E531 Collaboration, Phys. Lett. B **206**, 375 (1988).

[13] K. Kodama *et al.*, Fermilab P803 Proposal (1993), Appendix D.

[14] F. Baldassarre *et al.*, Nuovo Cimento XXI **3**, 459 (1961).

[15] A. Fassò, A. Ferrari, J. Ranft, and P. R. Sala, "New developments in FLUKA modelling of hadronic and EM interactions," in *Proceedings of the Third Workshop on Simulating Accelerator Radiation Environment,* SARE-3, KEK-Tsukuba, May 7–9, 1997, edited by H. Hirayama, KEK Report Proceedings 97-5, (1997) p. 32 .

[16] A. Ferrari, T. Rancati, and P. R. Sala, "FLUKA applications in high energy problems: from LHC to ICARUS and atmospheric showers," in *Proceedings of the Third Workshop on Simulating Accelerator Radiation Environment,*" SARE-3, KEK-Tsukuba, May 7–9 1997, edited by H. Hirayama, KEK Report Proceedings 97-5, (1997) p. 165 .

[17] B. Van de Vyver, Nucl. Instrum. Methods A **385**, 91 (1997).

[18] Particle Data Group, Phys. Rev. D **54**, 1 (1996).

[19] R. D. Cousins and V. L. Highland, Nucl. Instrum. Methods A **320**, 331 (1992).

[20] A. Marchionni for the NOMAD Collaboration, these Proceedings.

RESULTS FROM NOMAD

Alberto Marchionni
INFN, Sezione di Firenze
Largo E. Fermi 2, I-50125 Firenze, Italy

Representing the NOMAD Collaboration

ABSTRACT

The NOMAD experiment at CERN is conducting a search for $\nu_\mu \to \nu_\tau$ oscillations, for neutrino squared mass differences larger than about 1 eV^2. The ν_τ is searched for via its charged current interactions, where the τ lepton is identified through its decay products by means of kinematical criteria. Present results show no evidence for oscillations and allow us to set the 90% confidence level limit of $\sin^2 2\theta_{\mu\tau} < 1.2 \times 10^{-3}$ at large Δm^2.

© 1998 by Alberto Marchionni.

1 Introduction

In the Standard Model of electroweak interactions,[1] the existence of three different generations in the quark sector with corresponding generations in the leptonic sector is a solid experimental fact, but whose deep reason is still unexplained. The presence of mixing between the different generations in the quark sector has been known since a long time and it is presently described by the Cabibbo-Kobayashi-Maskawa matrix (see Ref. 2 for a review). In the leptonic sector, the conservation of a lepton family number has been assumed. Searches for processes of the type $\mu \rightarrow e\gamma$ have provided only upper limits so far.[2]

Neutrino oscillations,[3] first proposed by B. Pontecorvo in the late '50s, might be a more sensitive way to search for lepton family number violation. If neutrinos are massive, it is likely that the mass eigenstates are different from the weak interaction eigenstates, which could then be written as a linear superposition of the mass eigenstates. In a simplified two-generation approach, if $\mid \nu_1\rangle$ and $\mid \nu_2\rangle$ are the mass eigenstates with masses m_1 and m_2, respectively, and $\mid \nu_\alpha\rangle$ and $\mid \nu_\beta\rangle$ are the flavor eigenstates, then:

$$\begin{pmatrix} \nu_\alpha \\ \nu_\beta \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix},$$

where θ is the "mixing angle." The probability P that a neutrino of energy E, originally in a state $\mid \nu_\alpha\rangle$, is in a state $\mid \nu_\beta\rangle$ after a distance L from the point where it was generated, is given by:

$$P(\nu_\alpha \rightarrow \nu_\beta) = \sin^2(2\theta)\sin^2(1.27\Delta m^2 L/E),$$

where $\Delta m^2 = \mid m_1^2 - m_2^2 \mid \mathrm{eV}^2$ and L/E is expressed in km/GeV.

In order that the oscillation probability P be different from zero, it is necessary that $\theta \neq 0$ and $\Delta m^2 \neq 0$. The existence of the phenomenon of neutrino oscillations is thus directly connected to the question of whether neutrinos are massive and indeed it provides a way to measure the Δm^2 parameter.

Recent experimental results[4] from atmospheric and solar neutrino experiments, and from the LSND experiment give some indications for the existence of neutrino oscillations. The interpretation of these results in a complete three-generation approach is still far from clear, in particular for what concerns the Δm^2's involved in the process.

The NOMAD experiment[5,6] at CERN, together with its companion experiment CHORUS,[7] is investigating the region $\Delta m^2 > 1$ eV^2. The original motivation of NOMAD is the search for a cosmologically relevant neutrino mass.[8] Big Bang relic neutrinos, if there is at least one mass eigenstate with $m > 1$ eV, would play an important role in the evolution of the universe.

The NOMAD experiment, which has been carried out on the CERN wide-band neutrino beam, is investigating both $\nu_\mu \rightarrow \nu_\tau$ and $\nu_\mu \rightarrow \nu_e$ oscillation channels. It is an appearance experiment where the identification of ν_τ charged current (CC) interactions relies purely on kinematical criteria.

A description of the experimental method and a discussion of present results will be given for the $\nu_\mu \rightarrow \nu_\tau$ oscillation channel, which has been the primary motivation of the experiment. The search for $\nu_\mu \rightarrow \nu_e$ oscillations from the available data sample is not completed yet. The result depends critically on the knowledge of the systematic error on the flux and spectrum of the electron neutrino component in the beam. An effort to precisely determine the neutrino beam composition is currently underway.

2 The Experimental Apparatus

The NOMAD detector is installed on the neutrino beam of the West Area Neutrino Facility (WANF) at CERN. After an overview of the neutrino beam, a description of the detector is given.

2.1 The Neutrino Beam

The WANF wide-band neutrino beam[9] (see Fig. 1) is produced from the decays of secondary π and K mesons originated in the interactions of 450 GeV/c protons from the SPS on a Be target. In order to obtain a beam composed mainly of neutrinos, with a small contamination of antineutrinos, positively charged particles from the target

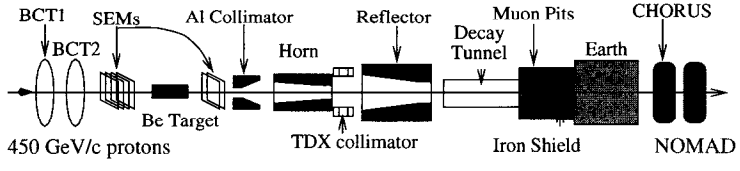


Fig. 1. Schematic layout of the West Area Neutrino Facility at CERN (not drawn to scale).

are focused by a pair of coaxial magnetic lenses (the horn and the reflector), while negative particles are defocused. The mesons are allowed to decay in a 290 m long vacuum tunnel, which is followed by a shield of iron and earth for the absorption of the undecayed hadrons.

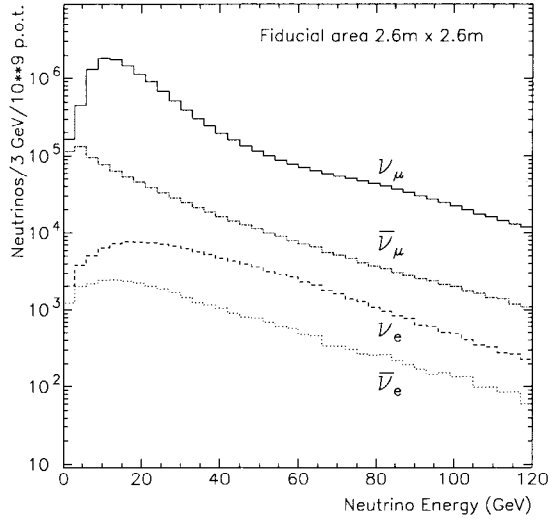


Fig. 2. Energy spectra of the different components of the neutrino beam at the NOMAD position.

The NOMAD detector is installed at 835 m from the Be target, at a mean distance of 620 m from the hadron decay position. The resulting neutrino spectra from π and K decays are shown in Fig. 2 at the position of the NOMAD detector, integrated over an area of 2.6×2.6 m^2. Table 1 summarizes mean energies and relative abundances of the different components of the neutrino beam and of the resulting interactions in the detector.

The beam is mainly composed of muon neutrinos, with an electron neutrino contamination of the order of 1%. Interactions of the "wrong sign" components $\overline{\nu}_\mu$ and $\overline{\nu}_e$ are present at the level of about 2% and 0.2% of the ν_μ interactions, respectively. The presence of prompt ν_τ in the beam, coming from charmed meson D_s decays, has been estimated[10] to be negligible. They amount to about 5×10^{-6} of the ν_μ interactions, thus

Table 1. Summary of mean energies and relative abundances of the different components of the neutrino beam and of the resulting interactions.

	Flux		Interactions	
Neutrino	$\langle E_\nu \rangle$ (GeV)	Rel. Abund.	$\langle E_\nu \rangle$ (GeV)	Rel. Abund.
ν_μ	23.6	1.0	44.1	1.0
$\overline{\nu}_\mu$	18.4	0.055	40.6	0.021
ν_e	37.5	0.009	56.6	0.015
$\overline{\nu}_e$	30.3	0.0025	50.7	0.0016
ν_τ	~ 35			$\simeq 5 \times 10^{-6}$

resulting in much less than one event detectable in NOMAD.

The average value of the L/E (traveled distance over energy) distribution of ν_μ neutrinos interacting in NOMAD is about 1.5×10^{-2} km/GeV. Thus the maximum Δm^2 sensitivity is at about 90 eV^2, making the experiment sensitive to the cosmologically relevant mass range.

2.2 The Detector

The NOMAD detector[6] has been designed to detect ν_τ CC interactions through different τ decay channels by means of kinematical criteria which require a detailed reconstruction of the event. Thus a detector is needed which is capable of a precise measurement of charged particles, a good identification of electrons and muons, an accurate measurement of electromagnetic showers and of neutral hadron detection.

The NOMAD detector (see Fig. 3) is located in a large dipole magnet which provides a field of 0.4 T over a volume of $7.5 \times 3.5 \times 3.5$ m^3. The magnetic field direction is horizontal and orthogonal to the beam line.

A set of 44 drift chambers constitute at the same time the target material for neutrino interactions and the tracking device of the experiment. Materials of low atomic number have been employed for the construction of the chambers in order to minimize multiple scattering of charged particles and photon conversions. Chamber panels made of Aramid fibers in a honeycomb structure sandwiched by two Kevlar-epoxy resin skins provide the necessary mechanical rigidity and flatness over a large surface area of 3×3 m^2. The total fiducial mass of the target amounts to 2.7 tons over a volume of

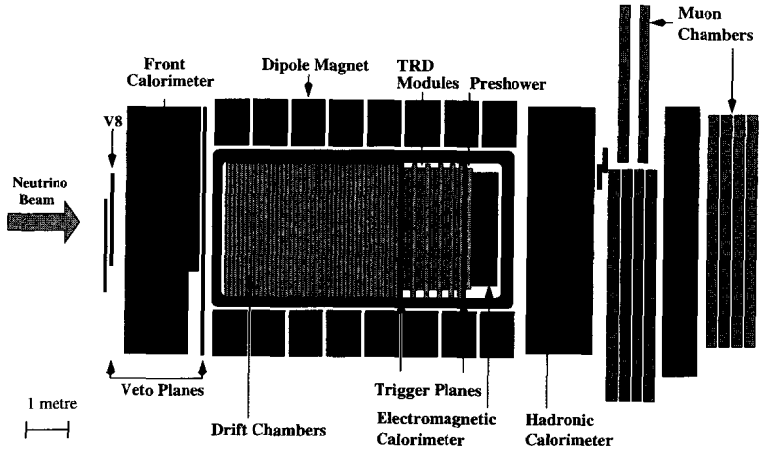


Fig. 3. Side view of the NOMAD detector.

$2.6 \times 2.6 \times 4$ m^3, with an average density of 0.1 g/cm^3. Each chamber contributes only 2% of a radiation length.

Each drift chamber consists of three wire planes with wires at $0°$, $+5°$, and $-5°$ with respect to the magnetic field direction. The spatial resolution for tracks at normal incidence on the chambers is about 150 μm. The resulting momentum resolution, which is a function of both momentum p and length L of the tracks, can be parameterized as:

$$\frac{\sigma_p}{p} \approx \frac{0.05}{\sqrt{L}} \oplus \frac{0.008p}{\sqrt{L^5}}.$$

For momenta less than 10 GeV/c the resolution is $\simeq 3.5\%$, where the dominant contribution comes from multiple scattering.

The drift chambers are followed by a transition radiation detector for electron identification. A set of nine modules, each consisting of a radiator of polypropylene foils followed by a plane of proportional straw tubes filled with a Xe–CH_4 mixture, gives a pion rejection factor greater than 10^3 for a 90% electron efficiency in a momentum range from 1 to 50 GeV/c. Five additional drift chambers are interleaved with the transition radiation modules to provide a precise track extrapolation from the drift chamber target to the following detectors, a preshower and an electromagnetic calorimeter.

The preshower, made of two planes of proportional tubes preceded by a lead plane of 1.6 radiation lengths, is followed by an electromagnetic calorimeter consisting of a matrix of 875 lead-glass Cherenkov counters. The energy resolution for electrons has been measured to be:

$$\frac{\sigma_E}{E} \approx \frac{3.2\%}{\sqrt{E(GeV)}} \oplus 1\%.$$

A hadron calorimeter, consisting of an iron-scintillator sampling calorimeter positioned behind the magnet, is intended to detect neutral hadrons and helps in distinguishing between muons and charged hadrons.

A set of ten drift chambers, each with an active area of 3.75×5.55 m^2 with two planes of drift tubes in the horizontal and two in the vertical directions, are arranged in two muon stations. The first, placed behind the hadron calorimeter, is followed by an 80 cm thick iron absorber and then by the second muon station. The muon momentum thresholds to reach the muon chambers (with 50% probability) are measured to be 2.3 GeV/c for the first muon station and 3.7 GeV/c for the second. The geometrical acceptance for muons to hit either one of the two muon stations is about 98%.

A ν_e CC candidate event detected in NOMAD is shown in Fig. 4. The track pointing to the large cluster in the electromagnetic calorimeter has been identified as an electron. An e^+e^- pair from the conversion of a bremsstrahlung photon emitted by the electron is also present.

3 Search for $\nu_\mu \rightarrow \nu_\tau$ Oscillations

The NOMAD experiment is conducting an appearance search for $\nu_\mu \rightarrow \nu_\tau$ oscillations, which is based on the identification of the τ lepton produced in ν_τ CC interactions. Differently from the CHORUS experiment,[7] where the τ identification is based on the decay topology directly observed in nuclear emulsions, in NOMAD the discrimination of the τ signal against the background of ν_e (ν_μ) CC and ν neutral current (NC) interactions relies on kinematical criteria.

The τ decay products are generally well isolated with respect to the hadronic jet. This can be expressed quantitatively by introducing the variable Q_T, defined as the component of the momentum of visible τ decay products, perpendicular to the total visible momentum vector of the event. Neutral current background, where a particle belonging to the hadronic jet is erroneously taken as coming from a τ decay, is characterized by small values of Q_T, while τ decays have larger Q_T values (see Fig. 5).

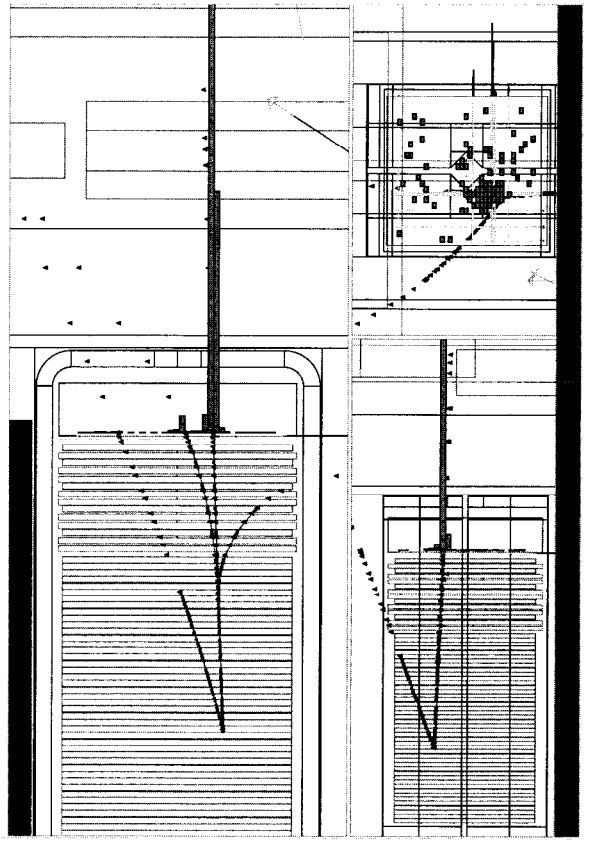


Fig. 4. A ν_e charged current event reconstructed in NOMAD. The longitudinal view orthogonal to the magnetic field is shown on the left. On the right side, the other longitudinal view is represented at the bottom and the transverse view at the top.

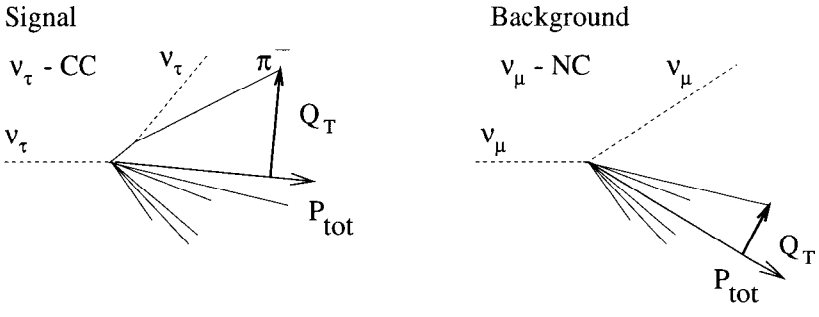


Fig. 5. Visualization of the Q_T variable for the ν_τ-CC signal (on the left), where the produced τ decays to $\tau^- \rightarrow \pi^-\nu_\tau$ and for the ν_μ-NC background (on the right).

The main feature of τ decays is that at least one neutrino carrying away some momentum is present. It is not possible to apply a detailed momentum balance along the neutrino beam direction on an event by event basis since the incoming neutrino momentum is not known. However in the plane transverse to the neutrino beam, as shown in Fig. 6, large missing transverse momentum $\vec{p}_T$ is expected in τ decays from the escaping neutrino(s).[11] In the case of $\nu_{e,\mu}$ CC background, a smaller missing transverse momentum is present due to measurement errors, losses of neutral hadronic particles, and nuclear effects (Fermi motion and nuclear reinteractions). The amount of missing transverse momentum, the angle ϕ_{mh} between $\vec{p}_T$ and the transverse momentum of the hadronic jet $\vec{p}_T^{\,h}$, and the angle ϕ_{lh} between the transverse momentum of the lepton $\vec{p}_T^{\,l}$ and $\vec{p}_T^{\,h}$, completely characterize the kinematics in the transverse plane. The angle ϕ_{mh} is generally larger in ν_τ CC interaction than in $\nu_{\mu,e}$ CC background, where it is almost uniformly distributed. On the other hand, the angle ϕ_{lh} is normally larger for $\nu_{\mu,e}$ background than for the ν_τ signal. Another variable which is useful to characterize the transverse plane kinematics is the transverse mass $M_T = \sqrt{(\mid \vec{p}_T \mid + \mid \vec{p}_T^{\,l} \mid)^2 - (\vec{p}_T^{\,h})^2}$, which is generally large for the neutral current background while it cannot exceed the τ mass in the case of τ decays.

These and many other variables can be used to separate the signal from the background. Each variable X_i provides only some degree of separation. The analysis can be optimized by building probability density functions L for the signal and the back-

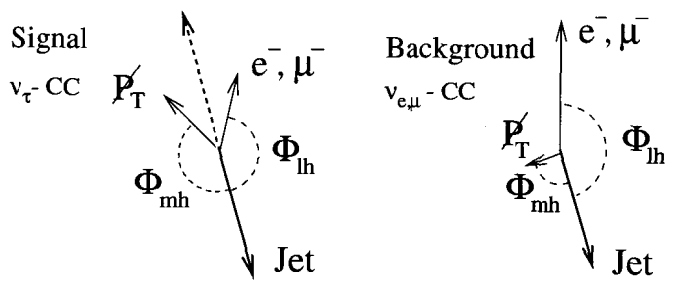


Fig. 6. Kinematical configuration in the transverse plane for ν_τ-CC (on the left), where the produced τ (dashed arrow) decays to electron or muon, and for $\nu_{e,\mu}$-CC interactions (on the right).

ground, approximated by the product of the probability density functions $P(X_i)$ of the set $X_{i=1,N}$ of variables used:

$$L = \prod_{i=1}^{N} P(X_i),$$

where $P(X_i)$ can be one or n-dimensional probability density functions. The likelihood ratio L_R defined as

$$L_R = \frac{L_{Signal}}{L_{Background}}$$

is then used to discriminate signal against background.

In order to avoid biases, the NOMAD collaboration has chosen to conduct a "blind analysis." First of all a comfortably large τ signal region is defined. The analysis procedure is adjusted on the Monte Carlo simulations and by examining the data events outside the signal region. Once reliable estimates of the background and of the signal efficiency have been reached, the analysis cuts are frozen and the events inside the signal region are looked at.

Detectable decays are the electron channel $\tau^- \to e^-\nu_\tau\overline{\nu}_e$, the one-prong hadronic channel $\tau^- \to h^-(+n\pi^0)\nu_\tau$, and the three-prong hadronic channel $\tau^- \to \pi^-\,\pi^-\,\pi^+\nu_\tau$, which have branching ratios of 17.8%, 49.5%, and 10.0%, respectively, thus covering a large part of the possible τ decays. The $\tau \to e$ channel is particularly favored because of the low contamination of ν_e in the beam and its energy spectrum shifted to higher energies with respect to the ν_μ component. The $\tau \to \mu$ channel is not included in the oscillation search because, as will be shown in the following, it is not possible to make a reliable estimate of the background in this case.

The searches in the $\tau^- \to e^-\nu_\tau\overline{\nu}_e$ and $\tau^- \to h^-(+n\pi^0)\nu_\tau$ decay channels will be described here in some detail. These analyses are based on the data collected from 1995 to 1997, as summarized in Table 2.

Table 2. Data sample collected in the years 1995–1997.

Year	ν_μ CC
'95	180,000
'96	380,000
'97	380,000
Total	940,000

3.1 The $\tau^- \to e^-\nu_\tau\overline{\nu}_e$ Channel

The search in the $\tau \to e$ channel begins with the selection of events with only one electron and no muons coming from the primary vertex. The electron, identified as such by the transition radiation detector, the preshower, and the electromagnetic calorimeter, is required to have the first hit of its track at a distance of at most 15 cm from the primary vertex position. A further rejection of electrons coming from photon conversions is achieved by requiring that the invariant mass between the e^- candidate and any positive track from the primary vertex be larger than 50 MeV.

The energy of the electron is carefully reconstructed by taking into account the energy released in the electromagnetic calorimeter and the conversions of bremsstrahlung photons in the drift chamber volume. Electrons with energies larger than 1.5 GeV are accepted as candidates.

Two different likelihood ratios, $L(\tau/CC)$ and $L(\tau/NC)$ (see Fig. 7), have been constructed to reject ν_e CC and ν NC backgrounds, respectively, along the criteria discussed in the previous section.

The likelihood ratio $L(\tau/NC)$ is mainly an estimate of the isolation of the electron candidate from the hadronic jet. It has been built with multi-dimensional probability density functions which make use of the following variables: Q_T and M_T, already defined; the minimum angle between the electron and any track of the hadronic system;

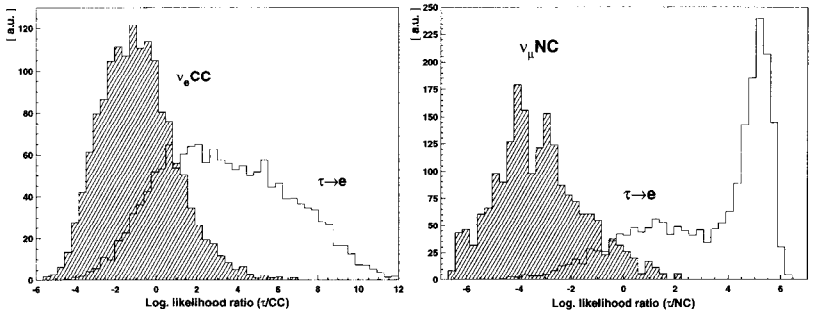


Fig. 7. The logarithm of the likelihood ratios $L(\tau/CC)$ (on the left) and $L(\tau/NC)$ (on the right) for the signal (solid line) and for the background (hatched area).

the angle $\theta_{\nu T}$ between the incident ν direction and the total visible momentum in the event; the angle between the incident ν direction and the total hadron momentum; the electron energy.

The likelihood ratio $L(\tau/CC)$, mainly related to the momentum imbalance in the transverse plane, has been built as a product of two three-dimensional probability density functions, symbolically denoted as:

$$L(\tau/CC) = [p_T^l, p_T^h, \phi_{lh}][E_T, \theta_{\nu T}, Q_{lep}],$$

where, in addition to variables already defined, E_T is the total visible energy and Q_{lep} is the component of the electron momentum perpendicular to the hadronic jet momentum.

The discrimination capability of the $\tau \rightarrow e$ signal against the ν_e CC background, fully expressed by the likelihood ratio $L(\tau/CC)$, can be understood from the inspection of the two-dimensional distributions of the variables which enter in the definition of $L(\tau/CC)$, as shown in Fig. 8.

The $\tau \rightarrow e$ events are both momentum imbalanced and characterized by an electron isolated from the hadronic jet. In a plot of $L(\tau/CC)$ versus $L(\tau/NC)$ (see Fig. 9) the $\tau \rightarrow e$ signal largely lies in the upper right corner, with the background somewhat spread along the diagonal. The thick lines in the upper right part of the plots define the $\tau \rightarrow e$ signal region.

A correct estimate of the background in the signal region cannot rely completely on the Monte Carlo simulation, which critically depends on the correct treatment of the hadronic system. For example, phenomena like nuclear reinteractions, which are

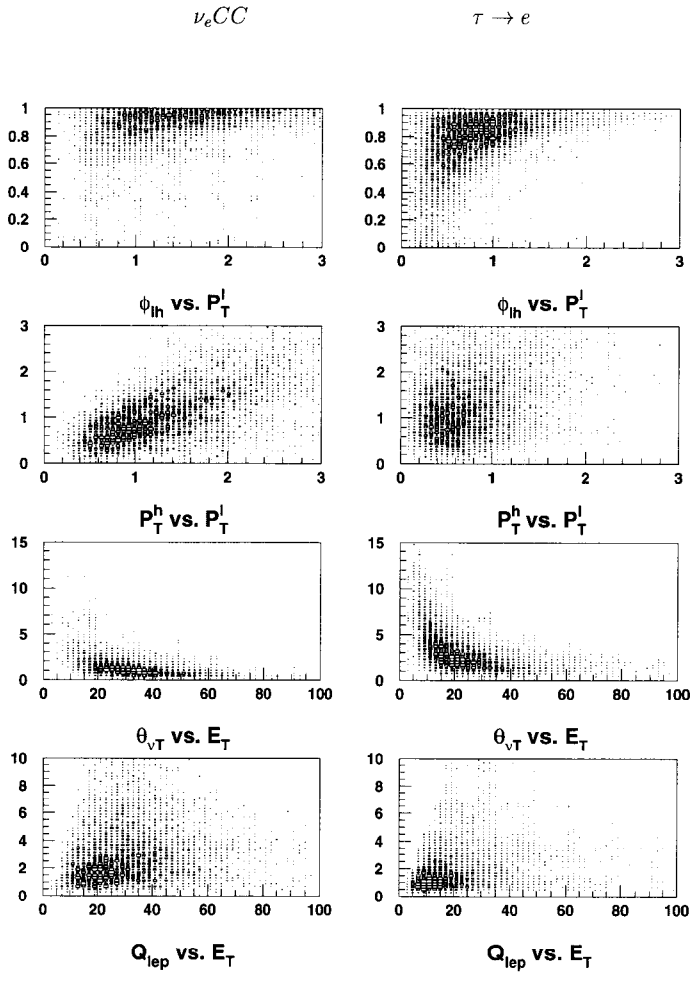


Fig. 8. Two-dimensional plots between the variables which enter in the definition of the likelihood ratio $L(\tau/CC)$, for the case of ν_eCC background (on the left) and for the $\tau \rightarrow e$ signal (on the right).

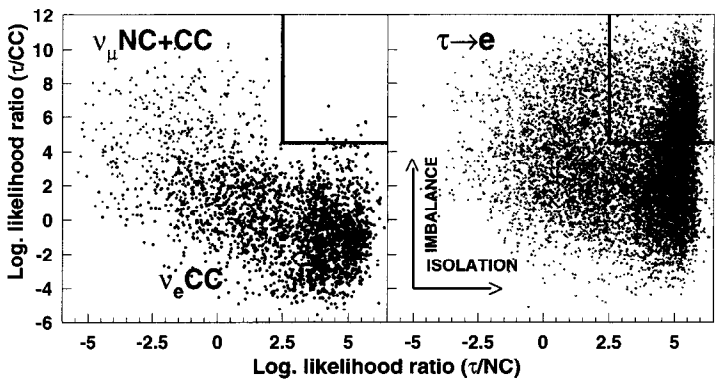


Fig. 9. Plot of $L(\tau/CC)$ versus $L(\tau/NC)$ for the background (on the left) and for the $\tau \rightarrow e$ signal (on the right).

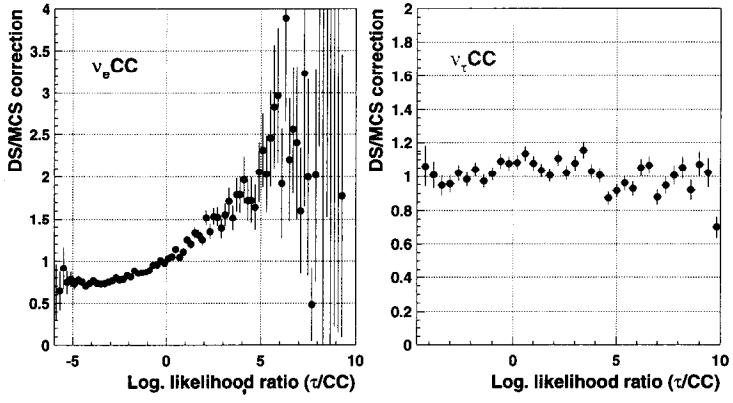


Fig. 10. DS/MCS corrections as a function of the logarithm of $L(\tau/CC)$ for the ν_eCC background (on the left) and for the ν_τCC signal (on the right).

hard to simulate, might affect the kinematics in the transverse plane. For this reason a "data simulator" procedure has been set up, where the data themselves have been used to correct the Monte Carlo estimates of signal efficiencies and background survival fractions. Starting from a measured ν_μ CC event, the identified muon is removed and replaced by another lepton, that is a τ, an electron, or a neutrino, thus getting a "data simulator" (DS) ν_τ CC, a ν_e CC, or a ν NC event, respectively. A similar procedure is applied to Monte Carlo ν_μ CC events in order to produce "Monte Carlo simulator" (MCS) events. DS and MCS events are processed through the same analysis chain as data and Monte Carlo events. Signal efficiencies and background survival fractions (here generically indicated by ϵ) are then obtained through the relation:

$$\epsilon = \epsilon_{MC} \times \frac{\epsilon_{DS}}{\epsilon_{MCS}}$$

where ϵ_{MC} is the Monte Carlo estimate without any "data simulator" correction. The corrections $\epsilon_{DS}/\epsilon_{MCS}$ in the case of ν_e CC background and ν_τ CC signal are shown in Fig. 10 as a function of $L(\tau/CC)$. While the correction is around unity for the ν_τ CC signal, in the case of ν_e CC background, it becomes quantitatively important in the signal region. Since it is not possible to compute data simulator corrections for $\tau \rightarrow \mu$, this channel is not included in the $\nu_\mu \rightarrow \nu_\tau$ oscillation search.

In order to optimize the analysis, a "signal box", where a possible τ signal may be visible, is carefully chosen by studying the sensitivity. This is "defined as the average upper limit one would get from an ensemble of experiments with the expected background and no true signal."[12] The computed sensitivity is shown in Fig. 11 as a function of the lower cut on the likelihood ratio $L(\tau/CC)$. The cut defining the "signal box" is set at the minimum of the sensitivity curve. The rapid increase of the sensitivity observed on the right side of Fig. 11 is due to a loss of signal efficiency in a region where the background is already negligible. Three different likelihood ratio cuts are shown in Fig. 11, the first to indicate the position of the full "signal box" and the other two to show its subdivisions into three different bins. Data events and background estimates are then compared in each of the bins in order to estimate the oscillation probability.

From a similar study on the sensitivity as a function of $L(\tau/NC)$, the full "signal box" is then defined by $L(\tau/NC) > 2.5$ and $L(\tau/CC) > 4.5$. Background estimates and signal efficiency in the "signal box" are shown in Table 3 after "data simulator" corrections. The number N_τ of expected signal events for unit $\nu_\mu \rightarrow \nu_\tau$ oscillation

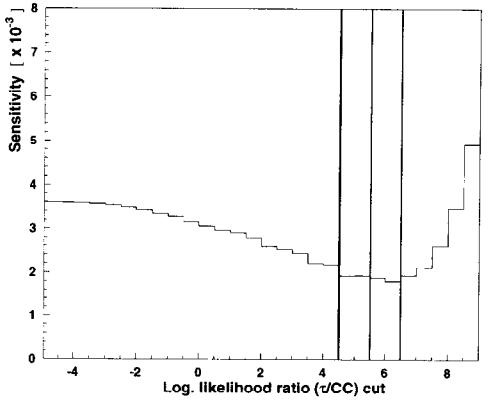


Fig. 11. Sensitivity as a function of the cut on the likelihood ratio $L(\tau/CC)$. The thick vertical line on the left indicates the position of the "signal box" and the other two lines its subdivision into three bins.

Table 3. Background estimates, signal efficiency and N_τ in the "signal box."

Background and efficiencies DS corrected			
ν NC + ν_μ CC backg.	ν_e CC backg.	$\epsilon_\tau(\%)$	N_τ
$0^{+0.4}$	6.6 ± 0.8	4.1 ± 0.3	2904

probability is also shown. N_τ is given by

$$N_\tau = N_\mu \times (\sigma_\tau/\sigma_\mu) \times Br \times \epsilon_\tau,$$

where σ_τ/σ_μ is the suppression factor of the ν_τ cross section (calculated to be 0.48), Br is the decay branching ratio of the $\tau \rightarrow e$ channel, and ϵ_τ is the corresponding τ selection efficiency.

The agreement between data and Monte Carlo predictions after "data simulator" corrections is checked on the electron data outside the "signal box" and on the positron sample (performing a τ^+ search, where no signal is expected in NOMAD). The results are shown in Fig. 12 as a function of the cut on the likelihood ratio $L(\tau/CC)$.

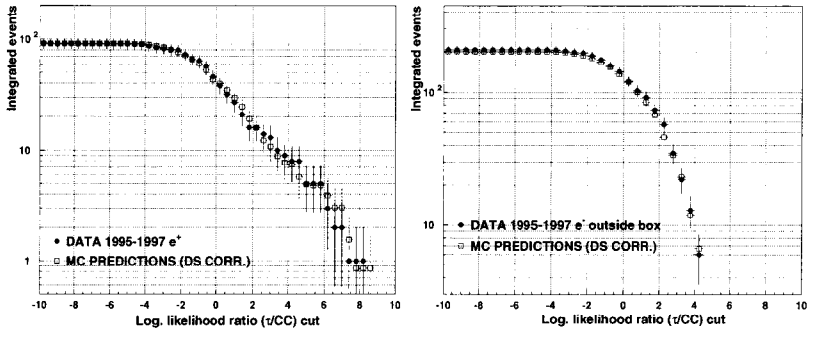


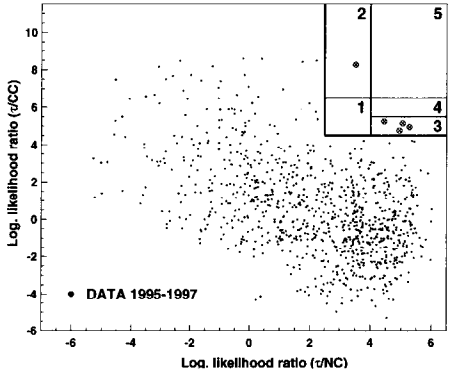
Fig. 12. Number of events surviving in the data (black dots) and corresponding Monte Carlo predictions (white squares) as a function of the cut on the likelihood ratio $L(\tau/CC)$, for e^+ data (on the left) and for e^- data outside the "signal box" (on the right).

Given the quite satisfactory agreement between data and Monte Carlo in Fig. 12, it is allowed to open the "signal box." The results are shown in Fig. 13. Five events are found in the box, with an expected background of 6.6 ± 0.8 events, showing no evidence for oscillations. The table on the right side of Fig. 13 exhibits a good agreement between data events and expected background in each of the bins into which the "signal box" has been subdivided. A final comparison of the likelihood ratio $L(\tau/CC)$ for the estimated background and the data is shown in Fig. 14, where the data follows the background distribution quite satisfactorily.

3.2 The $\tau^- \rightarrow h^-(+n\pi^0)\nu_\tau$ Channel

This decay channel requires the selection of events with no identified muon or primary electron. In addition, events with negative primary tracks with large transverse momentum ($p_T > 0.8$ GeV/c) pointing outside the acceptance of the lepton detectors (muon chambers and electromagnetic calorimeter) are rejected in order to veto charged current events where the lepton cannot be identified. The τ decay candidate is selected as the negative hadron with the highest or second highest p_T in the event.

The rejection of neutral and charged current background events is first accomplished by preliminary cuts on the transverse mass ($M_T < 4$ GeV) and the transverse momentum of the hadronic jet ($p_T^h > 1.3$ GeV/c). As shown in Fig. 15, each of the variables



Bin	N_τ	Exp. Backg.	Data
1	240	1.52 ± 0.38	0
2	273	0.23 ± 0.16	1
3	646	3.10 ± 0.57	4
4	543	1.31 ± 0.40	0
5	1202	0.46 ± 0.27	0
Tot.	2904	6.6 ± 0.8	5

Fig. 13. Data distribution in the $L(\tau/NC)$–$L(\tau/CC)$ plane. In the upper right part of the plot the "signal box" with its subdivisions is shown. Data points inside the "signal box" are indicated with large dots. The table on the right side shows data events, expected background, and N_τ in each bin of the "signal box" in the case of $\tau^- \to e^- \nu_\tau \overline{\nu}_e$.

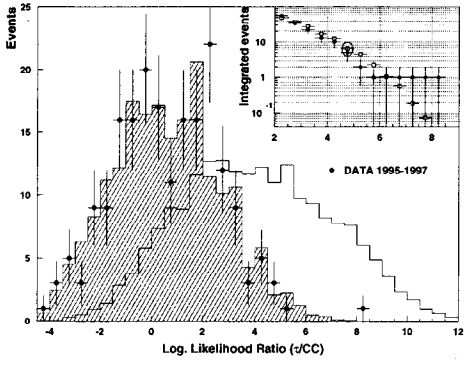


Fig. 14. Logarithm of the likelihood ratio $L(\tau/CC)$ for the data (black dots), estimated background (hatched area), and $\tau \to e$ signal (solid line). The inset shows the number of events surviving in the data (black dots) and the corresponding Monte Carlo predictions (white squares) as a function of the cut on $L(\tau/CC)$.

Q_T, the transverse mass M_T, and $\rho_m = \mid \vec{p}_T \mid /(\mid \vec{p}_T^{\,\pi} \mid + \mid \vec{p}_T^{\,h} \mid + \mid \vec{p}_T \mid)$ can achieve some isolation of the τ signal from the neutral and charged current background events. By making use of these quantities, a likelihood ratio is built (see Fig. 15) according to

$$L = ([Q_T, M_T, \rho_m], \frac{p^\pi}{E_{vis}}, p_T^h).$$

This is the product of the three-dimensional probability density function of the variables Q_T, M_T, and ρ_m with the one-dimensional probability density functions of the variables p_T^h and of the ratio $\frac{p^\pi}{E_{vis}}$ between the hadron candidate momentum and the visible energy.

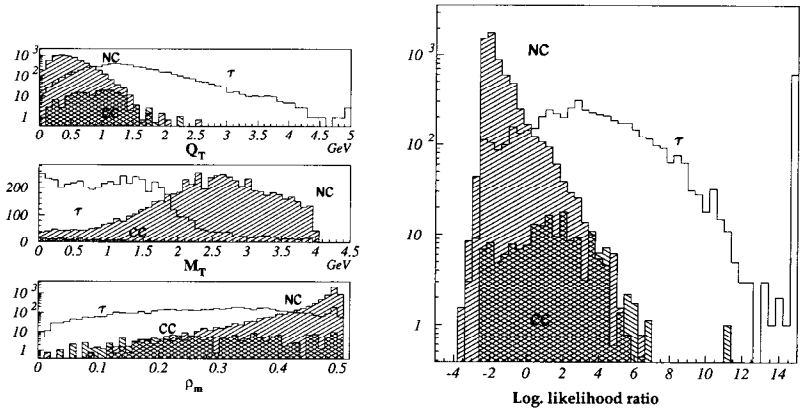


Fig. 15. Distributions of the variables Q_T, M_T, and ρ_m (on the left side), and of the logarithm of the likelihood ratio (on the right side) are shown for the τ signal, CC, and NC background.

The "signal box," defined by the logarithm of the likelihood ratio being larger than seven, is further subdivided into three bins, as shown in Fig. 16. Here the efficiency for $\tau^- \to h^-(+n\pi^0)\nu_\tau$, the background, and the sensitivity are shown as a function of the cut on the likelihood ratio.

The number of data events in the "signal box" is consistent with the expected background, as summarized in Table 4.

4 Present Results

Results from the analysis of the 1995 data have already been published,[13] obtaining a limit on the oscillation probability of $P(\nu_\mu \to \nu_\tau) < 2.1 \times 10^{-3}$ at 90% confidence level.

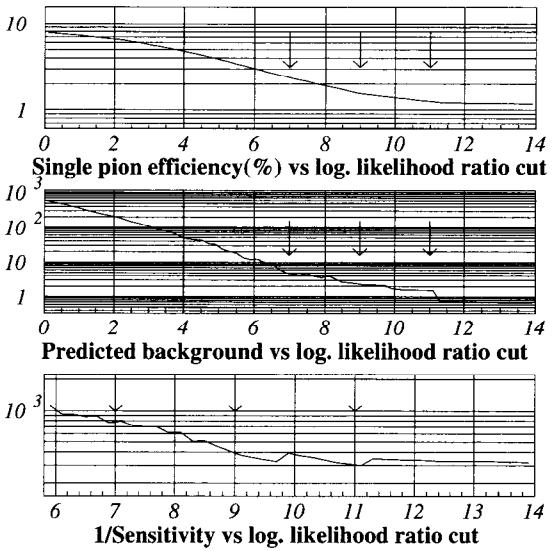


Fig. 16. The single π efficiency, the expected background, and the sensitivity are shown as a function of the likelihood ratio cut. The vertical arrows indicate the cuts defining the "signal box" and its subdivisions.

The present status of the analysis for the different τ decay channels is summarized in Table 5.

For each τ decay channel, two different analyses have been carried out. One, indicated as "DIS," starts from a sample of events with a hadronic jet momentum larger than 1.5 GeV/c, consisting mainly of deep inelastic events. The analyses discussed in the previous section are of this kind. For data events with the total hadron momentum smaller than 1.5 GeV/c, a specific analysis, here indicated as "LM," has been performed. This takes into account the low multiplicity of the hadronic system in this sample of events, which is enriched in quasi-elastic and resonance interactions. The "LM" analysis has not yet been extended to the '96–'97 data.

The expected total background is in good agreement with the number of events observed in the data. Following the statistical procedure based on the unified Feldman-Cousins approach,[12] we can set a limit on the oscillation probability of $P(\nu_\mu \to \nu_\tau) < 0.6 \times 10^{-3}$ at 90% confidence level. The approximate sensitivity of the experiment for the number of expected background events is 0.8×10^{-3}.

From the L/E (traveled distance over energy) distribution of ν_μ neutrinos interacting in NOMAD, an exclusion region in the $\Delta m^2 - \sin^2 2\theta$ plane can be derived as shown in Fig. 17. This gives an upper limit on $\sin^2 2\theta_{\mu\tau}$ at large Δm^2 of 1.2×10^{-3} at 90% confidence level. The CHORUS experiment[7] has obtained a comparable limit. Given the 1% contamination of electron neutrinos in the beam, it is possible to set a limit on the $\nu_e \to \nu_\tau$ oscillation probability. The preliminary upper limit on $\sin^2 2\theta_{e\tau}$ at large Δm^2 is 8×10^{-2} at 90% confidence level.

Table 4. Data events, expected background, and N_τ are shown in each bin of the "signal box" in the case of the $\tau^- \to h^-(+n\pi^0)\nu_\tau$ decay channel.

Likelihood	N_τ	Exp. Backg.	Data
7-9	664	2.3 ± 0.8	3
9-11	234	$1.1^{+0.8}_{-0.6}$	2
> 11	1133	$1.1^{+0.7}_{-0.5}$	0
> 7	2031	4.5 ± 1.2	5

Table 5. Summary of the analyses in the different τ decay channels for deep inelastic and low multiplicity interactions. The data sample used, the expected background, the number of events observed in the data, and N_τ are given for each decay channel.

Channel	Data	Exp. Backg.	N_{obs}	N_τ
$\tau^- \to e^-\nu_\tau\bar{\nu}_e$ DIS	'95–'97	6.6 ± 0.8	5	2904
$\tau^- \to e^-\nu_\tau\bar{\nu}_e$ LM	'95	$0.5^{+0.6}_{-0.2}$	0	218
$\tau^- \to h^-(+n\pi^0)\nu_\tau$ DIS	'95–'97	4.5 ± 1.2	5	2031
$\tau^- \to h^-(+n\pi^0)\nu_\tau$ LM	'95	$0.1^{+0.3}_{-0.1}$	1	198
$\tau^- \to \rho^-\nu_\tau$ DIS	'95–'97	$5^{+1.2}_{-0.9}$	5	1900
$\tau^- \to \pi^-\pi^-\pi^+\nu_\tau$ DIS	'95–'96	7.0 ± 2.7	5	1011
$\tau^- \to \pi^-\pi^-\pi^+(+n\pi^0)\nu_\tau$ LM	'95	$0.4^{+0.6}_{-0.4}$	0	108
Total		24.1 ± 3.3	21	8370

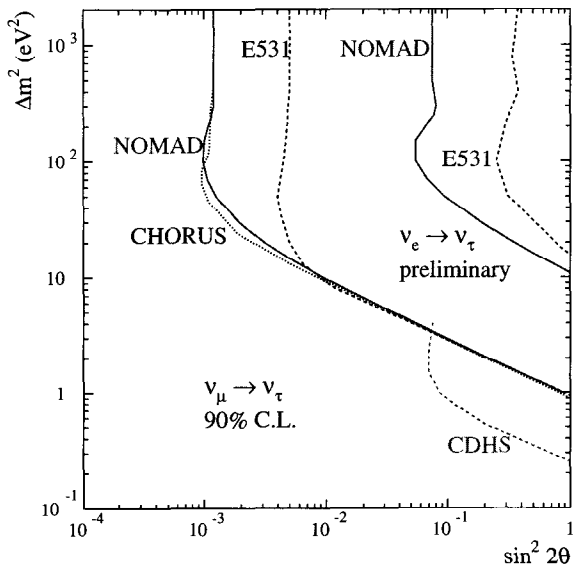


Fig. 17. NOMAD 90% C. L. exclusion plots for $\nu_\mu \rightarrow \nu_\tau$ and $\nu_e \rightarrow \nu_\tau$ oscillations (solid lines). The results of the CHORUS experiment[7] (dotted line) and of the previous E531 experiment[14,2] (dashed lines) using the technique of nuclear emulsions are shown.

5 Conclusions

NOMAD has set an upper limit on the $\nu_\mu \rightarrow \nu_\tau$ oscillation probability of 0.6×10^{-3} at 90% confidence level. The result has been achieved from the analysis of about 9.4×10^5 charged current interactions collected in the years from 1995 to 1997. The presence of a τ signal has been searched for in a large fraction of the possible τ decay channels. The quality of the event reconstruction which is currently being improved, the additional data collected in 1998, and the extension of all the analyses to the full available data sample will provide further sensitivity in the near future.

References

[1] E. D. Commins and P. H. Bucksbaum, *Weak Interactions of Leptons and Quarks* (Cambridge University Press, Cambridge, 1983); L. B. Okun, *Leptons and Quarks* (North-Holland, Amsterdam, 1982).

[2] Particle Data Group, C. Caso *et al.*, Eur. Phys. J. C **3**, 1 (1998).

[3] S. M. Bilenky and B. Pontecorvo, Phys. Rev. **4**, 225 (1978) and references therein.

[4] J. M. Conrad, to be published in *Proceedings of the* 29^{th} *International Conference on High-Energy Physics (ICHEP 98),* (Vancouver, Canada, July 1998); see also hep-ex/9811009; M. Nakahata, "Results from SuperKamiokande," these proceedings; G. Mills, "Results from LSND," these proceedings.

[5] NOMAD Collaboration, P. Astier *et al.*, CERN-SPSLC/91-21 (1991); Addendum 1, CERN-SPSLC/91-48 (1991); Addendum 2, CERN-SPSLC/91-53 (1991).

[6] NOMAD Collaboration, J. Altegoer *et al.*, Nucl. Instrum. Methods A **404**, 96 (1998).

[7] CHORUS Collaboration, E. Eskut *et al.*, Phys. Lett. B **424**, 202 (1998); CHORUS Collaboration, E. Eskut *et al.*, Phys. Lett. B **434**, 205 (1998); G. Catanesi, "Results from CHORUS," these proceedings.

[8] H. Harari, Phys. Lett. B **216**, 413 (1989).

[9] G. Acquistapace *et al.*, CERN-ECP/95-14 (July 1995).

[10] B. Van de Vyver, Nucl. Instrum. Methods A **385**, 91 (1997); M. C. Gonzalez-Garcia and J. J. Gomez-Cadenas, Phys. Rev. D **55**, 1297 (1997).

[11] C. H. Albright and R. E. Shrock, Phys. Lett. B **84**, 123 (1979); C. H. Albright, R. E. Shrock, and J. Smith, Phys. Rev. D **20**, 2177 (1979).

[12] G. J. Feldman and R. D. Cousins, Phys. Rev. D **57**, 3873 (1998).

[13] NOMAD Collaboration, J. Altegoer *et al.*, Phys. Lett. B **431**, 219 (1998).

[14] E531 Collaboration, N. Ushida *et al.*, Phys. Rev. Lett. **57**, 2897 (1986).

RESULTS ON NEUTRINOS FROM LSND

Geoffrey B. Mills

Physics Division
Los Alamos National Laboratory
Los Alamos, NM 87545
USA
e-mail: mills@lanl.gov

Representing the LSND Collaboration[1]

ABSTRACT

The LSND Collaboration has conducted a search for neutrino oscillations from a stopped muon neutrino source at the LANSCE facility at the Los Alamos National Laboratory. The collaboration finds evidence for an excess of events consistent with $\bar{\nu}_\mu \rightarrow \bar{\nu}_e$ oscillations. The collaboration also sees an excess of events consistent with $\nu_\mu \rightarrow \nu_e$ oscillations. An excess of $82.8 \pm 23.7 \pm 12.0$ events is observed with e^+ energy between 20 and 60 MeV for the $\bar{\nu}_\mu \rightarrow \bar{\nu}_e$ search. For the $\nu_\mu \rightarrow \nu_e$ search, an excess of $18.1 \pm 6.6 \pm 4.0$ events is observed with e^- energy between 60 and 200 MeV.

© 1998 by Geoffrey B. Mills.

1 Overview

The masses of the neutrinos are currently unknown quantities. On the other hand, the masses of the three generations of quarks and charged leptons are reasonably well measured. There remains the measurement of neutrino masses to complete the fermion generation mass picture. There is no reason to suppose that the masses of neutrinos are zero. We know from experiment that neutrino masses are small relative to their charged lepton partners, but the scale of this smallness is not known. Measurement of this scale would constitute a large step forward in our understanding of the fermion generations.

In order to establish that neutrinos have nonzero mass, one is required either to measure the mass directly, e.g., via conservation of energy, or to show that neutrinos oscillate with a mass-like phase advance. The phase advance in the latter case must be proportional to the difference in proper times of the neutrino eigenstates. Neutrino oscillations not only require massive neutrinos, they also require the neutrino mass eigenstates to be mixed. Neutrinos can mix in several possible ways, with noninteracting chiral partner states (sterile neutrinos), with different generation flavor states, or with a combination of these. The sterile oscillations are observed only in disappearance experiments while flavor oscillations can be observed in both disappearance and in appearance experiments.

There are three well-known experimental hints of neutrino mass, all having to do with the observation of neutrino oscillations. The deficits of solar ν_e neutrinos and atmospheric ν_μ neutrinos are well-documented disappearance results. In these results, the phase dependence seems to be observed; however, one cannot yet distinguish between sterile and flavor oscillations.

The third hint of neutrino mass is the observed anomalous appearance of $\bar{\nu}_e$ and ν_e by the LSND Collaboration.[2,3] The phase dependence in this case has not yet been verified; however, oscillations remain the favored explanation of the anomaly. That result is the subject of this manuscript.

2 Experimental Technique

The technique employed by LSND to observe neutrino oscillations is an appearance experiment. A beam of neutrinos is prepared at a beam dump, and the flavor composition of the beam is observed some distance away from the production source. An excess of any type of neutrino above calculated rates indicates a change in the flavor composition and indicates that oscillations may have occurred.

The LANSCE linear accelerator produces a beam of 200 MHz RF micro structure, on top of a 120 Hz macro-pulse structure. The macro-pulses are roughly 700 μs long, that gives a duty ratio of five to eight percent for normal running. The data between macro pulses serves to measure random cosmic ray backgrounds.

For LSND, the 1 mA beam of protons from the linear accelerator impacts upon a beam dump to produce charged pions. These pions either decay in flight (DIF), decay at rest (DAR), or (in the case of negative pions) are absorbed on nuclei in the dump. The neutrinos produced in those decays escape, and the resulting muons either decay, or (in the case of negative muons) undergo conversion on a nucleus in the beam dump. This sequence of decays produces a beam of neutrinos in the LSND detector. The composition of the beam dump has been varied over the course of the LSND running, from a low-Z water target to a high-Z tungsten target with no significant effect on the rate of anomalous events. Careful simulations of the beam dump area are used to calculate expected neutrino fluxes.[4]

The neutrino composition in such a beam is ideal for $\bar{\nu}_e$ appearance experiments. There are eight times as many positive pions produced as negative pions. Only five percent of those negative pions decay, and of those negative muons that result, only twelve percent decay. The ratio of negative muon decays to positive muon decays is thus less than 7.5×10^{-4}. The appearance of excess $\bar{\nu}_e$ is observable above this level.

At energies above the muon decay endpoint, the pion decay-in-flight neutrinos dominate the flux. The appearance of excess ν_e is then observable since normal decay channels are suppressed by the low (1.2×10^{-4}) branching ratio and by the long lifetime of the daughter muon relative to the flight time of the muon (5.0×10^{-4}).

The fluxes are calculated by using code[4] that simulates the beam stop geometry, materials, and particle production. The code was calibrated by using a test beam setup which was sensitive to pion production from 800 MeV protons. The systematic error on the DAR flux is found to be seven percent, and the systematic error on the DIF flux is found to be fifteen percent.

3 Experimental Apparatus and Neutrino Detection

The LSND detector[5] consists of a cylindrical tank filled with 167 tons of active mineral oil target fluid. The fluid is composed of hydrocarbon chains of roughly 25 carbon units

with a CH_2 stoichiometry. A small amount (0.031 g/l) of butyl-PBD is added to the oil in order to enhance the production of scintillation light in the fluid. This low scintillator concentration allows the detection of both Čerenkov light and scintillation light, and yields a relatively long attenuation length of more than 20 m for wavelengths greater than 400 nm (Ref. 6). A typical 45 MeV electron created in the detector produces a total of around 1500 photoelectrons, of which around 280 photoelectrons are in the Čerenkov cone.

This fluid is viewed by 1220 eight-inch photomultiplier tubes (PMT's) mounted on the inner surface of the tank. The tubes achieve a 25 percent coverage with a quantum efficiency of 25 percent. The electronic readout of those tubes yields both photon arrival time and charge information for each tube.

Neutrinos interact with the carbon nuclei, hydrogenic protons, and electrons inside the tank volume. The daughter products of these reactions produce light inside the tank. The timing and charge information is then used to reconstruct the position, time, and direction of the event. Calibration of the system is achieved by using a large sample of electrons and positrons from the decay of stopped cosmic-ray muons.

Relativistic particles produce both Čerenkov light and isotropic scintillation light. The former is emitted in a characteristic Čerenkov cone at an angle of 47 degrees. More than half of the Čerenkov light is isotropically scattered by the fluid due to the presence of the scintillation fluor in the liquid. The Čerenkov light is produced almost immediately, while the scintillation light has an exponential decay time of roughly 30 ns. Nonrelativistic particles produce only scintillation light and can be separated from relativistic particles by analyzing the time distribution of the detected light and by the presence of a Čerenkov cone in the angular distribution of the light.

The detector tank is surrounded by a veto shield[7] which is used to suppress cosmic-ray muon events at the trigger level. Trigger activities inhibit the trigger for 15.0 μs subsequent to the activity. This reduces the trigger rate to less than 50 Hz at the expense of a 19% loss in detector live time.

The detector trigger runs on a 100 ns clock and the number of PMT's are summed every 100 ns. Threshold levels of 18, 21, 75, and 150 hit PMT's for tank hits, and four or six hit PMT's for veto hits are recorded in a time stamped trigger word at every clock edge. A trigger processor then reads these words and decides whether or not an interesting event has occurred. If an interesting event is found, the trigger processor signals readout processors to send the event to the online reconstruction computer, which reconstructs the event and records it on tape. The trigger processor also watches for 1000 μs subsequent to the event and sends any events which might be 2.2 MeV gamma candidates in that interval. All activities which occur within 50 μs prior to the trigger event are also recorded. The trigger operates almost independently of whether the beam is on or off.

4 Neutrino Cross-Section Measurements

There are several neutrino cross sections which have been measured by LSND. These events may also be used to cross check the calculated neutrino fluxes used in the oscillation searches. They arise from the interactions of the dominant neutrino components in the beam, ν_e from μ^+ DAR, and, ν_μ from pion DIF.

The μ^+ DAR neutrinos have energies below the muon rest mass, so only the ν_e can interact via a charged current reaction on the neutrons inside a carbon nucleus, $^{12}C(\nu_e,\ e^-)\ ^{12}N$. There are two categories for this reaction. They are distinguished by the whether the resulting ^{12}N nucleus is in its ground state or not. The ground state ^{12}N nucleus will β decay back to the original ^{12}C configuration with a lifetime of 15.9 ms [$^{12}C(\nu_e,\ e^-)\ ^{12}N$]. This signature is observable in the detector as an event correlated in position and time with the neutrino event. The presence of the correlated β-decay event serves to reduce random backgrounds to this process. In addition, the nuclear matrix element for the neutrino cross section is directly related to the β decay matrix element, which is well measured. Thus, the number of observed ground state events is an accurate measure of the DAR flux.

In the case where the ν_e interaction leaves the ^{12}N nucleus in an excited state, $^{12}C(\nu_e,\ e^-)\ ^{12}N^*$, it is well-known that the excited nitrogen nucleus will decay via proton emission without returning to its ground state. Therefore, no subsequent correlated β decay will be observed.

The energies of the pion DIF neutrinos extend above the threshold for muon production on carbon of 123.1 MeV, so that the $^{12}C(\nu_\mu, \mu^-)\ ^{12}N$ cross sections are measurable. These reactions are also coincidence measurements in the sense that the subsequent decay of the muon and/or β decay is correlated to the neutrino interaction in position and time. They have low accidental backgrounds. The matrix element for the $^{12}C(\nu_\mu, \mu^-)\ ^{12}N_{GS}$ reaction is well determined by the $^{12}N_{GS}$ lifetime, however, the statistics are poor. The calculations of the $^{12}C(\nu_\mu, \mu^-)\ ^{12}N^*$ cross sections are less

certain. CRPA and shell model predictions range widely, and for the moment, the measured rate is below the predicted values, although only slightly below the shell model predictions.

The ν_e, $\bar{\nu}_\mu$, and ν_μ can both interact with the electrons in the tank fluid, and in the case of the ν_e, via both the s-channel and t-channel paths. These events are peaked forward in the angle with respect to the neutrino direction, and they can be distinguished from nuclear scattering processes.

The LSND measurements[8,9] of these cross sections are shown in Table 1. The ground state reaction measurements, $^{12}C(\nu_e, e^-)\ ^{12}N_{GS}$ and $^{12}C(\nu_\mu, \mu^-)\ ^{12}N_{GS}$, agree well with theoretical predictions and lend confidence to the flux calculations. The $^{12}C(\nu_e, e^-)\ ^{12}N^*$ and νe^- reactions agree well with theory.

The $^{12}C(\nu_\mu, \mu^-)\ ^{12}N$ inclusive cross section is complicated by the nuclear dynamics of the $^{12}N^*$ system, since the momentum transfer is larger than in the case of the $^{12}C(\nu_e, e^-)\ ^{12}N$. The CRPA model calculations[10] have been unable to reproduce the LSND measured values. The situation has been nearly resolved because of two new revelations. Improvements in our flux calculations have slightly raised the LSND measured value. Recent shell model calculations[12] yield a smaller number than the CRPA model. Correlations in the closed $p_{3/2}$ shell nucleons seem to have a larger effect than previously assumed, so that the theory is more compatible with the LSND results.

Table 1. Summary of LSND cross-section measurements.

Process	Measured	Prediction	Neutrino source
$^{12}C(\nu_e, e^-)\ ^{12}N_{GS}$	$(9.1 \pm 0.4 \pm 0.9) \times 10^{-42}$	9.3×10^{-42}	μ^+ DAR
$^{12}C(\nu_e, e^-)\ ^{12}N^*$	$(5.7 \pm 0.6 \pm 0.6) \times 10^{-42}$	6.3×10^{-42}	μ^+ DAR
νe^-	$(4.2 \pm 0.6 \pm 0.4) \times 10^{-42}$	3.97×10^{-42}	μ^+ DAR
$^{12}C(\nu_\mu, \mu^-)\ ^{12}N_{GS}$	$(6.6 \pm 1.1 \pm 1.0) \times 10^{-41}$	6.4×10^{-41}	π^+ DIF
$^{12}C(\nu_\mu, \mu^-)\ ^{12}N$	$(12.3 \pm 0.3 \pm 1.8) \times 10^{-40}$	$(13 - 20) \times 10^{-40}$	π^+ DIF

5 Neutrino Oscillation Searches

LSND has performed searches for signs of neutrino oscillations in two separate channels. The decay at rest flux is ideal for studying the $\bar{\nu}_\mu \rightarrow \bar{\nu}_e$ transition and the decay in flight flux is suited to study the $\nu_\mu \rightarrow \nu_e$ transition. In either case, the search is for events with an electron-like signature (positrons behave identically to electrons in LSND).

5.1 The $\bar{\nu}_\mu \rightarrow \bar{\nu}_e$ Search

The $\bar{\nu}_\mu \rightarrow \bar{\nu}_e$ search relies on the absence of $\bar{\nu}_e$ in the initial neutrino beam. The analysis is discussed elsewhere,[2] and the salient features will be discussed here. The data presented here are from the years 1993-1997, where the event selection has been modified slightly to account for differences in year-to-year calibrations. Background estimates are based on the original analysis, where the number of $^{12}C(\nu_e, e^-)\ ^{12}N_{GS}$ events is used to estimate the decay at rest flux.

The reaction that is searched for is the $\bar{\nu}_e p \rightarrow e^+ n$ inverse beta decay reaction. The matrix element for this cross section is known indirectly from the neutron decay time constant. The final state consists of a positron and a neutron. The neutron contributes little to the primary event; however, it rapidly thermalizes with a time constant of 186 μs. The neutron is then captured 35 percent of the time on carbon and 65 percent of the time on the hydrogen in the mineral oil. The result of the capture on the hydrogen is a 2.2 MeV gamma ray which is observed as an event subsequent to the neutrino interaction, correlated in space and in time.

Events are considered in the energy range from 20 MeV to 60 MeV and inside the tank fiducial volume (35 cm from the PMT surface inside the tank). The event selection reduces backgrounds from cosmic ray muons, neutrons, and other sources by requiring that the amount of prompt light, and reconstruction of the vertex and direction of the event, be consistent with a positron event. A large sample of Michel decays from stopped cosmic ray muons assures that the efficiency is well determined for this process.

Correlated events over the 18 hit trigger threshold, within the 1000 μs gamma window subsequent to the trigger event, are searched to see if they are consistent with a neutron capture. The correlated events will have the hit distribution characteristic of a 2.2 MeV energy deposition, a time correlation with a 186 μs time constant, and a spatial correlation with an RMS of 74 cm. A likelihood is formed for each future activity from the number of hits, primary and secondary event time difference, and distance between primary and secondary vertices. These distributions are shown in Fig. 1 for correlated and uncorrelated gamma samples from cosmic ray data.

Future activities have a likelihood to show that they were a correlated gamma ray

Table 2. LSND decay at rest oscillation search.

Selection	Beam on	Beam off	ν backgrounds	Excess signal
$36 < E_e < 60, R > 30$	29	5.2 ± 0.6	3.0 ± 0.6	20.8 ± 5.4
$20 < E_e < 60, R > 30$	61	15.6 ± 1.0	11.5 ± 0.6	33.9 ± 8.0
Fitted oscillation probability			$0.31 \pm 0.09 \pm 0.05\%$	

from neutron capture. They also have a likelihood that they were a random background gamma from another source. The ratio of those likelihoods is called R, the correlated-to-uncorrelated gamma likelihood ratio. This ratio is used to discriminate between correlated and uncorrelated gammas in the positron event sample. The data are analyzed in two sets, $R \geq 0$ (no cut on R) and $R > 30$. The requirement of the presence of a correlated neutron capture, $R > 30$, greatly reduces the backgrounds from a number of sources, both beam related and cosmic ray.

The number of events observed after the event selection is shown in Table 2 for two different sets of selection, a strict selection with $36 < E_e < 60$ MeV and $R > 30$, and a loose selection with $20 < E_e < 60$ MeV and $R > 30$. The energy, angle, and vertex (x, y) distributions for the events satisfying the strict selection is shown in Figs. 3, 4, and 5. The oscillation probability, defined as the observed number of events divided by the maximum possible number of events in the high Δm^2 limit, is measured by fitting the distribution in R for events in the $20 < E_e < 60$ MeV range. The results of this fit are shown in Fig. 2.

The data satisfying the loose selection criteria are binned into positron energy, R, angle with respect to the incident neutrino direction, and distance from the neutrino source. The Poisson likelihood for these binned data are calculated in a model that includes neutrino induced backgrounds, cosmic ray backgrounds, and oscillation events. The oscillation events are taken to be dependent on two parameters, Δm^2 and $\sin^2 2\theta$, in the usual form

$$P_{osc} = \sin^2 2\theta \, \sin^2(1.27\Delta m^2 L/E_\nu). \quad (1)$$

The log-likelihood function is then calculated as a function of Δm^2 and $\sin^2 2\theta$. Regions bounded by contours of 2.3 and 4.5 (90 and 99 percent) units from the peak log-likelihood value, indicating favored regions in the (Δm^2, $\sin^2 2\theta$) parameter space,

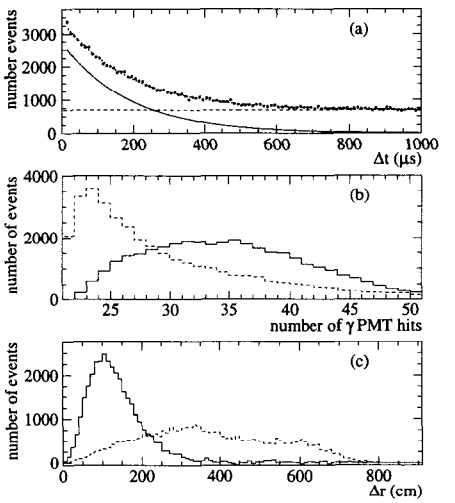


Fig. 1. Associated γ distributions for (a) time, (b) PMT hits, and (c) distance distributions for associated (solid) and accidental (dashed) γ's.

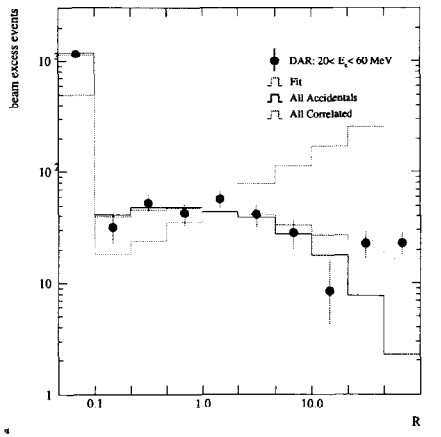


Fig. 2. The R distribution, beam-on minus beam-off excess, for events that have energies in the range $20 < E_e < 60$ MeV. Shown in the figure are the best fit to the data, together with the expectations for uncorrelated and correlated γ's.

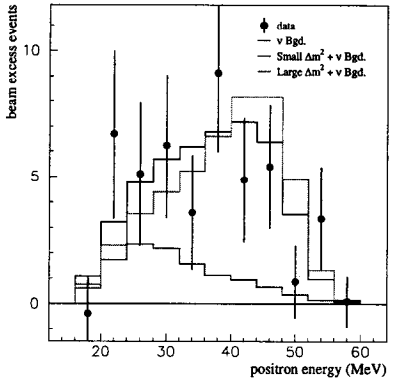


Fig. 3. The energy distribution for events with $R > 30$. Shown in the figure are the beam-excess data, estimated neutrino background, expected distribution for neutrino oscillations at large Δm^2 plus estimated neutrino background, and expected distribution for neutrino oscillations at small Δm^2 plus estimated neutrino background.

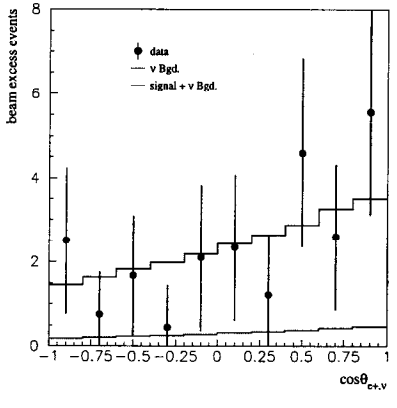


Fig. 4. The $\cos\theta_\nu$ distribution for events with $R > 30$ and $36 < E < 60$ MeV, where $\cos\theta_\nu$ is the angle between the neutrino direction and the reconstructed positron direction.

are calculated. Figure 6 shows regions that are a logical OR of a set of regions made by varying systematic errors.

5.2 The $\nu_\mu \rightarrow \nu_e$ Search

The $\nu_\mu \rightarrow \nu_e$ search uses the ν_μ flux from pion decay in flight and searches for $^{12}C(\nu_e, e^-)\ ^{12}N$ events at a level higher than expected from normal processes. The analysis is described in detail in prior publications[3] and is outlined here. The $\nu_\mu \rightarrow \nu_e$ search relies on the suppression of $\bar{\nu}_e$ in the initial neutrino beam. The data presented here are from the years 1993-1995.

The pion DIF flux is calculated in the same program that yields the DAR flux, where good agreement is achieved with the number of observed $^{12}C(\nu_e, e^-)\ ^{12}N$ events. It is checked by the measured number of $^{12}C(\nu_\mu, \mu^-)\ ^{12}N_{GS}$ events observed. That method is statistically limited with an error of around 15 percent, but the method is more reliable than the inclusive $^{12}C(\nu_\mu, \mu^-)\ ^{12}N$ events, where the cross section is not well-known.

The $^{12}C(\nu_e, e^-)\ ^{12}N$ cross section is calculated by using the CRPA model.[11] This model accurately predicts the $^{12}C(\nu_e, e^-)\ ^{12}N$ process at lower energies as discussed above, but for higher energy ranges, is known to differ from the shell model calculation.[12] This discrepancy is a source of uncertainty in the oscillation background estimates. In the present analysis, the larger values of theoretically possible cross sections are used, which give an overestimate of the backgrounds, although the final result is not too sensitive to the assumption.

Events consistent with the $^{12}C(\nu_e, e^-)\ ^{12}N$ process are searched for in the energy range from 60 MeV to 200 MeV and inside the tank fiducial volume (35 cm from the PMT surface inside the tank). The event selection reduces backgrounds from cosmic ray muons, neutrons, and other sources by loose requirements that the amount of prompt light, and reconstruction of the vertex and direction of the event, be consistent with the $^{12}C(\nu_e, e^-)\ ^{12}N$ event. A more complex likelihood method is then employed to reconstruct the event. The likelihood method both gives a likelihood for the electron hypothesis and gives the amount of Čerenkov light observed in the Čerenkov cone. Both of these are then used to further select electron events. Significant cosmic ray background persists in the event sample even after this selection and the fiducial volume of the selection is reduced further by requiring that events near the wall of the tank are not pointing into the tank.

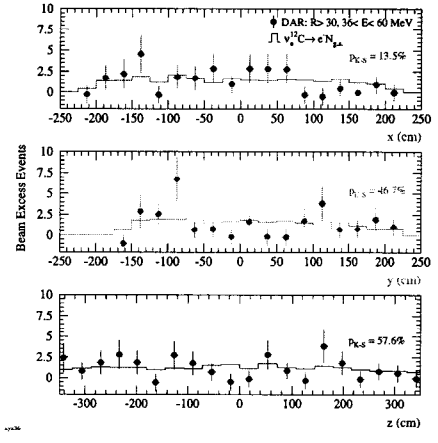


Fig. 5. The spatial distributions for events with $R > 30$ and $36 < E < 60$ MeV.

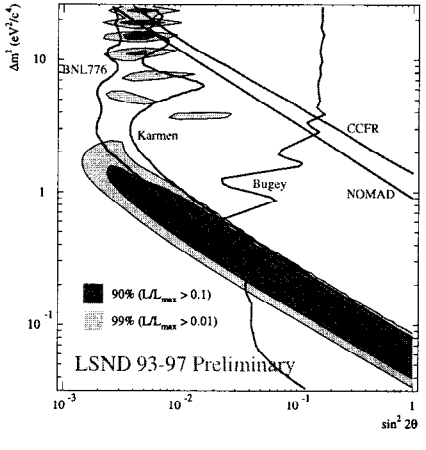


Fig. 6. Plot of the LSND Δm^2 *vs* $\sin^2 2\theta$ favored regions. They correspond to 90% and 99% likelihood regions after the inclusion of the effects of systematic errors. Also shown are 90% C.L. limits from KARMEN1 at ISIS, E776 at BNL, the Bugey reactor experiment, the CCFR experiment at FNAL, and the NOMAD experiment at CERN.

The results of two independent sets of software analyses are shown in Table 3. The number of observed events, after including systematic and statistical errors, is higher than that predicted by conventional processes. Figures 7 and 8 show the energy distribution for events that satisfy either analysis along with the predicted signal and background levels.

The chance for a fluctuation to have produced this effect is less than two percent. This, in combination with the DAR result, strongly suggests an unexplained source of electron- or positron-like events. A confidence region calculation was performed to check the consistency of the $\nu_\mu \rightarrow \nu_e$ result with the $\bar{\nu}_\mu \rightarrow \bar{\nu}_e$ result. The regions for the two analyses, shown in Fig. 9, overlap, which shows that the two results can be explained by a single neutrino oscillation hypothesis.

6 Conclusion

The LSND experiment has observed anomalous signals in the $\bar{\nu}_e$ and ν_e channels, which are consistent with the hypothesis of neutrino flavor oscillations. In order to demonstrate that the signal is due to neutrino flavor oscillations, a follow-on experiment is necessary which is capable of observing the quantum-mechanical phase dependence of the oscillation on L/E. Such an experiment has been proposed and approved at the Fermi National Laboratory, the Booster Neutrino Experiment[13] (BooNE). This experiment will use the rapid cycling 8 GeV proton booster synchrotron at Fermilab to generate a high-flux neutrino beam from focussed pions that decay in flight. The experiment will have a sensitivity an order of magnitude greater than LSND and will be able to measure the Δm^2 parameter of the oscillations if they are occurring.

Table 3. LSND decay-in-flight oscillation search.

Selection	Beam on(off)	Cosmic events	ν events	Excess signal	Efficiency	Oscillation probability
Analysis A	23(114)	8.0 ± 0.7	4.5 ± 0.9	10.5 ± 4.9	0.084	$(2.9 \pm 1.4) \times 10^{-3}$
Analysis B	25(92)	6.4 ± 8.5	8.5 ± 1.7	10.1 ± 5.3	0.138	$(1.7 \pm 0.9) \times 10^{-3}$

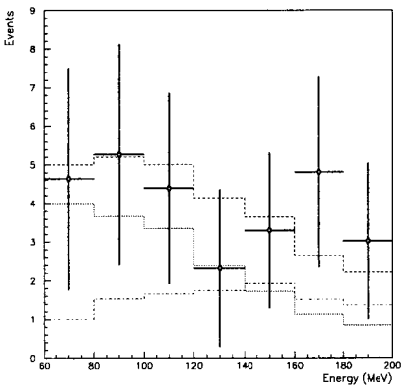


Fig. 7. The energy distribution for the final beam-excess decay-in-flight events. The expectation for backgrounds (dot-dashed histogram), the oscillation signal for most likely values of Δm^2 (dotted histogram), and the sum of the two (dashed histogram) are shown.

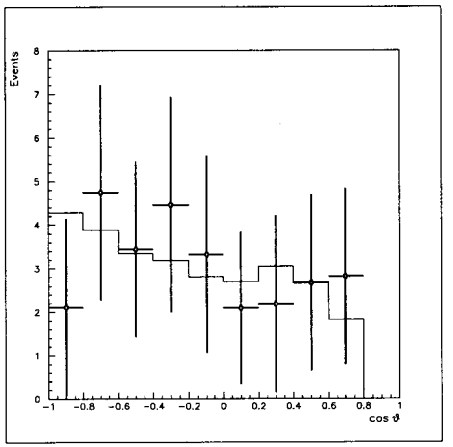


Fig. 8. The angular distribution for the final beam-excess decay-in-flight events. The expectation for neutrino background and an oscillation signal at most likely values of Δm^2 (solid histogram) is shown.

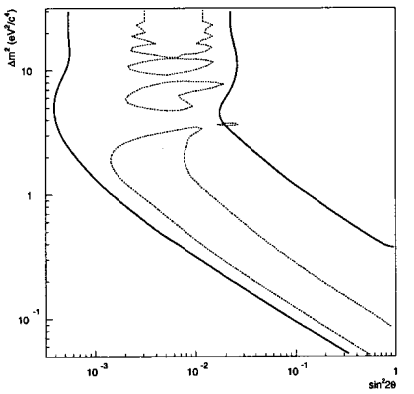


Fig. 9. The 95% confidence region for $\nu_\mu \rightarrow \nu_e$ oscillations (solid curve) along with the favored regions for $\bar{\nu}_\mu \rightarrow \bar{\nu}_e$ oscillations (dotted curve).

References

[1] The LSND Collaboration presently consists of the following people and institutions: E. D. Church, I. Stancu, W. Strossman, G. J. VanDalen (University of California, Riverside); W. Vernon (University of California, San Diego); D. O. Caldwell, S. Yellin (University of California, Santa Barbara); D. Smith (Embry-Riddle Aeronautical University); R. L. Burman, J. B. Donahue, G. T. Garvey, W. C. Louis, G. B. Mills, V. Sandberg, R. Tayloe, D. H. White (Los Alamos National Laboratory); R. Imlay, H. J. Kim, A. Malik, W. Metcalf, M. K. Sung, N. Wadia (Louisiana State University); K. Johnston (Louisiana Tech University); A. Fazely, R. M. Gunasingha (Southern University); L. B. Auerbach, R. Majkic (Temple University).

[2] C. Athanassopoulos *et al.*, Phys. Rev. C **54**, 2685 (1996); C. Athanassopoulos *et al.*, Phys. Rev. Lett. **77**, 3082 (1996).

[3] C. Athanassopoulos *et al.*, Phys. Rev. Lett. **81**, 1774 (1998); C. Athanassopoulos *et al.*, Phys. Rev. C **58**, 2489 (1998).

[4] R. L. Burman, M. E. Potter, and E. S. Smith, Nucl. Instrum. Methods A **291**, 621 (1990); R. L. Burman, A. C. Dodd, and P. Plischke, Nucl. Instrum. Methods in Phys. Rev. A **368**, 416 (1996).

[5] C. Athanassopoulos *et al.*, Nucl. Instrum. Methods A **388**, 149 (1997).

[6] R.A. Reeder *et al.*, Nucl. Instrum. Methods A **334**, 353 (1993).

[7] J.J. Napolitano *et al.*, Nucl. Instrum. Methods A **274**, 152 (1989).

[8] C. Athanassopoulos *et al.*, Phys. Review C **55**, 2078 (1997).

[9] C. Athanassopoulos *et al.*, Phys. Review C **56**, 2806 (1997).

[10] E. Kolbe, K. Langanke, F. K. Thielemann, and P. Vogel, Phys. Rev. C **52**, 3437 (1995).

[11] E. Kolbe, K. Langanke, and S. Krewald, Phys. Rev. C **49**, 1122 (1994).

[12] A. C. Hayes, *Elsevier Preprint*, 1998 (private communication).

[13] E. Church *et al.*, "A proposal for an experiment to measure $\nu_\mu \rightarrow \nu_e$ oscillations and ν_μ disappearance at the Fermilab Booster: BooNE," LA-UR-98-352, Fermilab experiment 898 (1997).

RESULTS FROM SUPER-KAMIOKANDE

Masayuki NAKAHATA
Kamioka Observatory, Institute for Cosmic Ray Research
University of Tokyo, Higashi-Mozumi, Kamioka-cho, Yoshiki-gun
Gifu, Japan, 506-1205

For the Super-Kamiokande Collaboration

ABSTRACT

The results from Super-Kamiokande on atmospheric neutrinos and solar neutrinos are presented. The contained event analysis of atmospheric neutrinos shows a small value of μ/e ratio and a large asymmetry in the zenith angle distribution. A distortion of zenith angle distribution is observed in the upward-through-going muons resulting from atmospheric neutrino interactions in the rock surrounding the Super–Kamiokande detector. The ratio of upward-stopping muons to upward-through-going muons is significantly smaller than the expectation. All those results on atmospheric neutrinos give evidence for ν_μ oscillations. The obtained neutrino oscillation parameters are Δm^2=$10^{-3} - 10^{-2}\mathrm{eV}^2$ and $\sin^2 2\theta > 0.8$.

Super-Kamiokande observed about 7000 solar neutrino events over 504 days of live time. The observed solar neutrino flux is 47% of the expectation from the standard solar model. No significant day/night difference is observed and a solar model independent limit on neutrino oscillation parameters is obtained. A comparison of the observed recoil electron spectrum with the expectation exhibits a poor agreement at the 4.8% confidence level.

© 1998 Masayuki Nakahata.

1 Introduction

Atmospheric neutrinos and solar neutrinos are extremely good sources for investigating neutrino oscillations. A probability of neutrino oscillations is given by the following formula:

$$P(\nu_\alpha \to \nu_\beta) = \sin^2 2\theta \sin^2(1.27\frac{L}{E_\nu}\Delta m^2), \quad (1)$$

where θ is the mixing angle of neutrinos, L is the distance from a neutrino source to a detector in units of kilometers, E_ν is the energy of neutrinos in units of GeV, and Δm^2 is the difference of mass squares in units of eV^2. As seen in the equation, one needs longer distances for investigating smaller Δm^2. The neutrino oscillation experiments with a baseline of sun-to-earth and earth-size, namely solar neutrino and atmospheric neutrino experiments, are best suited for searching for small Δm^2.

The atmospheric anomaly was first presented by the Kamiokande[1,2] and IMB[3] experiments. The anomaly suggested neutrino oscillations, but the statistics of those experiments were not enough to give convincing evidence for neutrino oscillations. Super-Kamiokande is almost ten times larger than those experiments and high statistics atmospheric neutrino data have given strong evidence for neutrino oscillations.

The solar neutrino experiments observed significantly lower flux than the expectations from the standard solar models, known as "the solar neutrino problem." One of the possible solutions of the solar neutrino problem is neutrino oscillations. In order to get evidence for neutrino oscillations in solar neutrinos, solar model independent evidence for oscillations should be searched for. Super-Kamiokande tests solar neutrino oscillations using short time variations, such as day/night difference of the flux, and the shape of the energy spectrum.

In this report, results from Super-Kamiokande on atmospheric neutrinos and solar neutrinos are described.

2 Super-Kamiokande Detector

Super-Kamiokande (SK) is a water Cherenkov detector located in the Kamioka Mine in Japan. A schematic view of the detector is shown in Fig. 1.

The water tank consists of inner and outer detectors. The inner detector is composed of 11,146 20-inch diameter photomultiplier tubes (PMTs) placed facing inward on the surface of a 33.8 m diameter by 36.2 m high cylinder. The tubes are spaced on a 0.707 m grid, and enclose 32,000 metric tons of water. The total photocathode surface area of

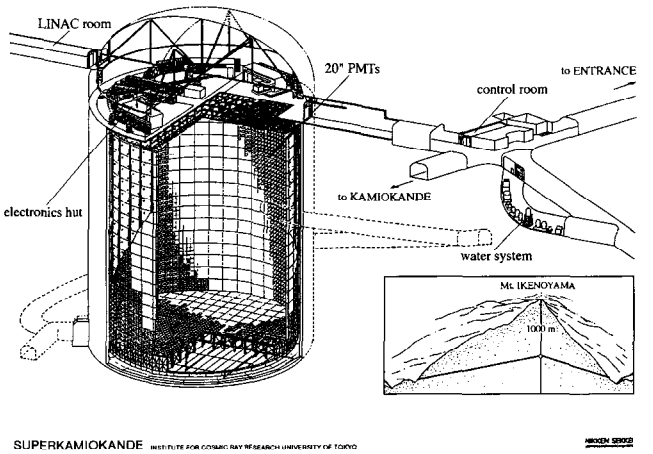


Fig. 1. Schematic view of the Super-Kamiokande detector.

all PMTs is 40% of the surface of the inner detector. A 4π solid-angle outer detector surrounds the inner detector. The outer detector is also a water Cherenkov counter with 1,885 8-inch PMTs facing outwards. The light collection of these smaller PMTs is enhanced by the addition of 60 cm square wavelength shifter plates around each PMT. The thickness of the outer detector is 2.6–2.75 meters. The outer detector is designed to reduce external gamma-rays from surrounding rocks for the solar neutrino analysis and to tag cosmic-ray muons. The fiducial volume of the detector is defined to be more than 2 m inwards from the inner detector phototube wall for the solar neutrino analysis and contained event analysis of atmospheric neutrinos, resulting in a 22-kton fiducial water volume.

A trigger of the detector is issued when more than 29 PMTs give signals within 200 nanoseconds, which corresponds to 5.7 MeV equivalent electron energy. The trigger efficiency for the event above the current analysis threshold of solar neutrinos (6.5 MeV) is better than 99%. The trigger threshold is low enough to observe atmospheric neutrino events.

3 Atmospheric Neutrinos

Atmospheric neutrinos are produced by the interaction of primary cosmic rays in the atmosphere. The hadronic showers resulting from the cosmic ray interaction produces electron and muon neutrinos through $\pi^+ \to \mu^+ + \nu_\mu$ followed by $\mu^+ \to e^+ + \bar{\nu}_\mu + \nu_e$ (and their charge conjugates). The ratio ($\equiv \nu_\mu/\nu_e$) of the flux of $\nu_\mu + \bar{\nu}_\mu$ to the flux of $\nu_e + \bar{\nu}_e$ is expected to be about two from the decay chain of $\pi^\pm$ (it is true for the energy below $\sim$1 GeV). As the energy increases, the ν_μ/ν_e ratio increases, because the fraction of the muons which reach earth's surface increases as their energy. The ν_μ/ν_e ratio is determined by the well-understood decay chain of mesons and the uncertainty of the ratio is less than 5% (Refs. 4 and 5), whereas the absolute flux of each neutrino has an uncertainty of $\sim$25%. Hence, the ν_μ/ν_e ratio is a robust value to test neutrino oscillations. Another way to test neutrino oscillations is the zenith angle of neutrinos. The zenith angle distribution is expected to be up/down symmetric for neutrinos above 1 GeV, where the effects due to the Earth's magnetic field on primary cosmic rays are small. The symmetry is determined by the simple geometry of neutrino production.

The measurements of the ν_μ/ν_e ratio are reported as $R \equiv (\mu/e)_{DATA}/(\mu/e)_{MC}$, where μ and e are the number of muon-like (μ-like) and electron-like (e-like) events observed in the detector for both data and Monte Carlo simulation. This ratio largely cancels experimental and theoretical uncertainties. $R = 1$ is expected if the physics in the Monte Carlo simulation accurately models the data. Measurements of significantly smaller values of R have been reported by Kamiokande,[1,2] IMB,[3] and recently by Super–Kamiokande.[6,7] Although measurements of R by early iron-calorimeter experiments, Fréjus[8] and NUSEX,[9] were consistent with one, the Soudan-2 iron-calorimeter experiment[10] has reported observation of a small value of R.

Super–Kamiokande detects atmospheric neutrinos as (1) contained events, (2) upward-through-going muons, and (3) upward-stopping muons. The contained events are neutrino interactions occuring in the fiducial volume of the inner detector. High energy atmospheric muon neutrinos interact in the rock surrounding Super–Kamiokande produce muons by charged current interactions. They are observed as through-going or stopping muons in the detector. Although the large number of cosmic ray downward going muons limit the observation only to upward-going muons, they cover the angular range from nearly horizontal to vertical directions. The following sub-sections describe results of each analysis.

3.1 Results of Contained Events

The contained events are separated into fully-contained (FC) events and partially-contained (PC) events. The FC events deposit all of their Cherenkov radiation in the inner detector, while PC events have exiting particles which deposit some Cherenkov radiation in the outer detector. For the contained event analysis, the neutrino interaction vertex is required to be reconstructed within the 22.5 kiloton fiducial volume, defined to be > 2 m from the PMT wall. Super–Kamiokande has observed 4353 FC events and 301 PC events in a 33.0 kiloton-year exposure. The FC contained events are divided into sub-GeV ($E_{vis} < 1.33$ GeV) and multi-GeV ($E_{vis} > 1.33$ GeV) samples. The number of Cherenkov rings is counted by an automatic ring counting program and a particle identification program is applied for single ring events. Electrons produce diffused Cherenkov ring patterns whereas muons make clear Cherenkov rings. The diffuseness is evaluated by a likelihood (L) and Fig. 2 shows the L distribution for single ring sub-GeV events. Positive and negative L are defined to be μ-like and e-like, respectively. The figure shows that the μ-like and e-like events are clearly separated and the shape of the L distribution is well reproduced by the Monte Carlo simulation. The particle identification is checked by using muon decay electrons and stopping cosmic ray muons. The misidentification probability is estimated to be 0.6% for sub-GeV events and 2% for multi-GeV events. The particle identification programs have also been tested using beams of electrons and muons incident on a water Cherenkov detector at KEK.[14]

Table 1 summarizes the number of observed events and expectations from the Monte Carlo simulation based on the flux calculation by Honda *et al.*[5] Details of the data analyses of sub-GeV and multi-GeV samples are described in Refs. 6 and 7. The R value for the sub-GeV sample is obtained to be

$$R(sub-GeV) = 0.63 \pm 0.03\,(stat.) \pm 0.05\,(sys.).$$

The PC events are classified as μ-like events, because a Monte Carlo simulation shows that 98% of PC events come from charged current interaction of ν_μ. In the Monte Carlo simulation, 88% (96%) of the sub-GeV e-like (μ-like) events were ν_e (ν_μ) charged-current interactions and 84% (99%) of the multi-GeV e-like (μ-like) events were ν_e (ν_μ) charged-current interactions. The R value for the multi-GeV FC sample and PC sample is obtained to be

$$R(multi-GeV; FC+PC) = 0.65 \pm 0.05\,(stat.) \pm 0.08\,(sys.).$$

The sources of the systematic error of R are uncertainty of the predicted ratio of ν_μ flux

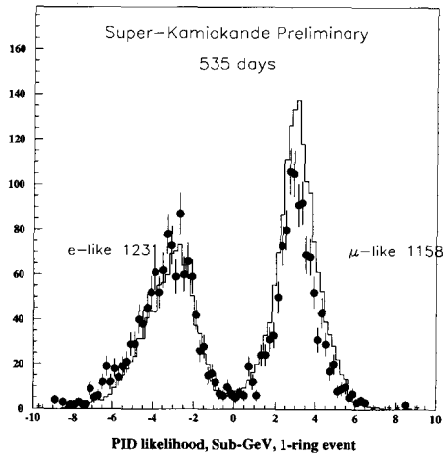


Fig. 2. The particle identification likelihood distribution for single ring sub-GeV events.

	Data	Monte Carlo
sub-GeV		
single-ring	2389	2622.6
e-like	1231	1049.1
μ-like	1158	1573.6
multi-ring	911	980.7
total	3300	3603.3
multi-GeV		
single-ring	520	531.7
e-like	290	236.0
μ-like	230	295.7
multi-ring	533	560.1
total	1053	1091.8
partially-contained	301	371.6

Table 1. Summary of the sub-GeV, multi-GeV, and PC event samples compared with the Monte Carlo prediction based on the neutrino flux calculation by Honda *et al.*[5]

to ν_e flux ($< 5\%$), particle identification efficiency described above, charged-current and neutral current cross sections (3~4%) and ring number counting (3% for sub-GeV and 6% for multi-GeV), etc. The probability that the observed μ/e ratios could be due to statistical fluctuation is less than 0.001% for sub-GeV R and less than 1% for multi-GeV R.

The zenith angle distributions of FC and PC events are shown in Fig. 3 together with the Monte Carlo simulations. The shape of the observed zenith angle distribution

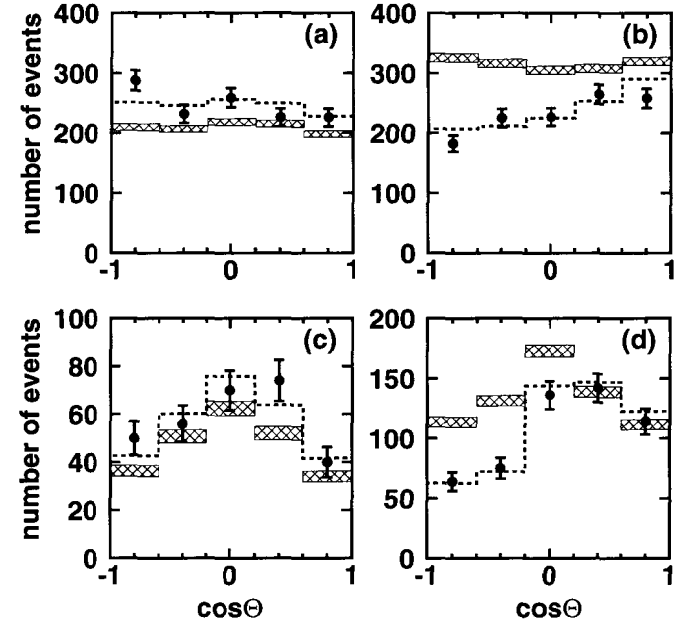


Fig. 3. Zenith angle distribution of atmospheric neutrino events: (a) sub-GeV e-like, (b) sub-GeV μ-like, (c) multi-GeV e-like, and (d) multi-GeV μ-like + PC. The shaded histograms show the expectations from the Monte Carlo simulation and the dotted histograms show the best fit assuming neutrino oscillations (see text).

of e-like events agrees well with the simulation, but the large distortion of the zenith angle distribution is observed in the μ-like events. The shape comparison of the zenith angle distribution between the observed data and Monte Carlo simulation gives a χ^2

of 2.8 with 4 d.o.f. (degrees of freedom) for e-like multi-GeV events, corresponding to 59% of confidence level (CL), and 30 for μ-like multi-GeV events, corresponding to 4.9×10^{-4}% CL. The large χ^2 value in μ-like events comes from the large deficit of upward-going events. Up/down asymmetry is defined as $A = (U-D)/(U+D)$ where U is the number of upward-going events ($-1 < \cos\Theta < -0.2$) and D is the number of downward-going events ($0.2 < \cos\Theta < 1$). The asymmetry is expected to be near zero independent of the flux model for $E_\nu > 1$ GeV as described above. The momentum dependence of the asymmetry is shown in Fig. 4. The asymmetry for e-like events is near

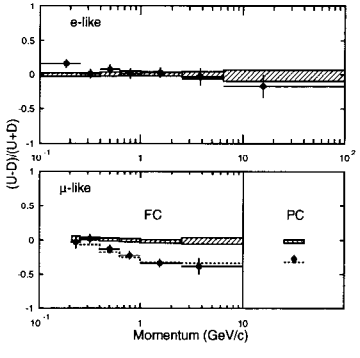


Fig. 4. Momentum dependence of the zenith angle asymmetry of the data (data points) and the Monte Carlo simulation (shaded histograms). The dotted histogram shows the best fit with neutrino oscillations.

zero and consistent with the expectation, while the μ-like events at higher momentum largely deviate from the expectation. The angular correlation between charged leptons and incident neutrinos is estimated by a Monte Carlo simulation and it is 15–20° for multi-GeV events and ~60° for sub-GeV events. Events with $E < 0.4$ GeV essentially do not remember the direction of the incident neutrinos. The main reason why the asymmetry of the μ-like event approaches to zero at lower energy is due to this effect.

We have examined the hypotheses of two-flavor neutrino oscillations of $\nu_\mu \leftrightarrow \nu_e$ and $\nu_\mu \leftrightarrow \nu_\tau$ using a χ^2 comparison of momentum and angular distributions of the data and the Monte Carlo. The contained events are divided into five zenith angle bins and seven momentum bins both for e-like events and μ-like plus PC events (70 bins in total). The χ^2 is defined as

$$\chi^2 = \sum_{cos\Theta,p}^{70} \left(\frac{N_{Data} - N_{MC}(\sin^2 2\theta, \Delta m^2, \epsilon_j)}{\sigma} \right)^2 + \sum_j \left(\frac{\epsilon_j}{\sigma_j} \right)^2 ,$$

where N_{Data} is the measured number of events in each bin. $N_{MC}(\sin^2 2\theta, \Delta m^2, \epsilon_j)$ is the expected number of Monte Carlo events, which is a function of $\sin^2 2\theta$, Δm^2; and ϵ_j, σ_j are parameters which are related to the systematic uncertainties. σ accounts for both data statistics and the weighted Monte Carlo statistics. The parameters for the systematic errors (and their uncertainties) are: E_ν spectral index (0.05), sub-GeV R (8%), multi-GeV R (12%), relative normalization of PC to FC (8%), L/E_ν (15%), up/down ratio for sub-GeV (2.4%), and multi-GeV (2.7%). The absolute normalization of the flux is a free parameter in the analysis. See reference [11] for more details. The best χ^2 is obtained to be 65 for $\nu_\mu \leftrightarrow \nu_\tau$ and 88 for $\nu_\mu \leftrightarrow \nu_e$ oscillations with 67 degrees of freedom. Hence, SK data favors $\nu_\mu \leftrightarrow \nu_\tau$ oscillations more than $\nu_\mu \leftrightarrow \nu_e$ as far as two neutrino oscillations are assumed. The best χ^2 is obtained at $\Delta m^2 = 2.2 \times 10^{-3}$ eV^2 and $\sin^2 2\theta = 1.0$ for $\nu_\mu \leftrightarrow \nu_\tau$ oscillations. The best fit zenith angle distribution is shown by the dotted histogram in Fig. 3.

Neutrino oscillations are a function of L/E_ν as shown in Eq. (1), where L is the neutrino travel path length and E_ν is the primary energy of neutrinos. Super–Kamiokande cannot measure E_ν on an event by event basis, but it can be estimated statistically. The relation between the lepton energy and the neutrino energy was obtained by a Monte Carlo simulation and the L/E_ν distribution is obtained as Fig. 5. The energy of the events is limited to E >0.4 GeV, because events below this energy do not have good resolution in L. The dotted histogram in the figure shows the expectation from the $\nu_\mu \leftrightarrow \nu_\tau$ oscillations with the best fit oscillation parameters. The L/E_ν distributions of e-like and μ-like events agree with the expectation from neutrino oscillations.

An allowed region of neutrino oscillation parameters is obtained as the thick-black curve in Fig. 6 for $\nu_\mu \leftrightarrow \nu_\tau$. The region $\Delta m^2 = 5 \times 10^{-4} - 6 \times 10^{-3}$ eV^2 and $\sin^2 2\theta > 0.83$ is allowed with 90% CL.

The anomalies of the small μ/e double ratio and the distortion of the zenith angle distribution could be due to $\nu_\mu \leftrightarrow \nu_s$ (sterile neutrinos) oscillations. We test $\nu_\mu \leftrightarrow \nu_\tau$ versus $\nu_\mu \leftrightarrow \nu_s$ oscillations using the ratio of the number of π^0 events to e-like events (π^0/e ratio). If the anomalies are due to $\nu_\mu \leftrightarrow \nu_\tau$ oscillations, the π^0 event rate duc

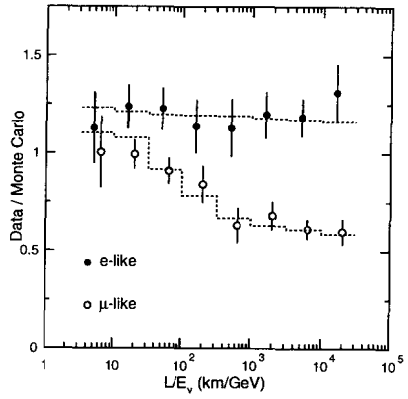


Fig. 5. L/E_ν distributions for e-like and μ-like events. The points show the ratio of observed data to MC expectations without neutrino oscillations. The dotted histogram shows the expected shape for $\nu_\mu \leftrightarrow \nu_\tau$ at $\Delta m^2 = 2.2 \times 10^{-3}\mathrm{eV}^2$ and $\sin^2 2\theta = 1$.

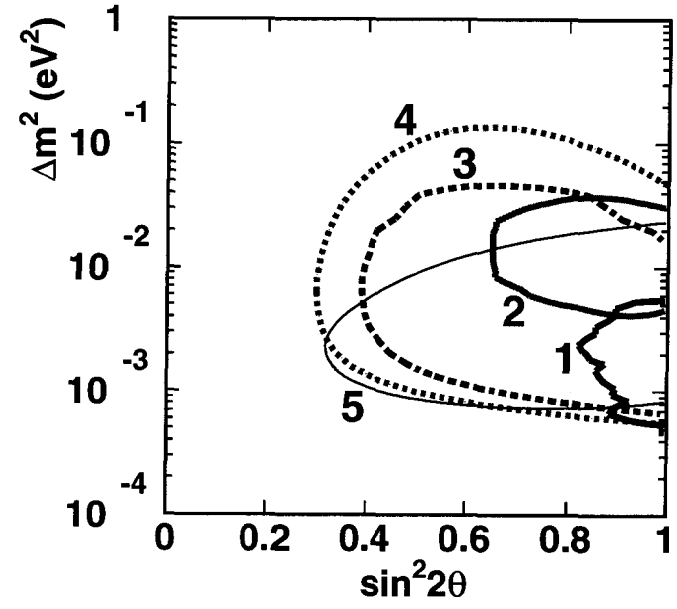


Fig. 6. Allowed regions of neutrino oscillation parameters obtained by the analyses of contained events [thick-black (1)], upward-through-going muons [thick-dotted (3)] and upward *stop/through* muons ratio [thin-solid (5)]. Also shown are results from Kamiokande contained event analysis [thick-gray (2)] and upward-through-going muons [thick-gray-dotted (4)].

to neutral current interactions does not change and the π^0/e ratio is the same as the expectation without neutrino oscillations. On the other hand, if it is due to $\nu_\mu \leftrightarrow \nu_s$ oscillations, the π^0/e ratio should be smaller than the expectation, because ν_s does not cause a neutral current interaction. The π^0/e ratio is expected to be $\sim$83% of the no-mixing expectation, for $\nu_\mu \leftrightarrow \nu_s$ with the oscillation parameters obtained above. Figure 7 shows invariant mass distribution for two e-like ring events. A clear signature of π^0 events are seen around the π^0 mass. The events in the mass range of 90–190 MeV/c^2 are selected as π^0 candidates.

The comparison of the π^0/e ratio of data and Monte Carlo gives

$$\frac{(\pi_0/e)_{data}}{(\pi_0/e)_{MC}} = \frac{210/1231}{192.5/1049.1} = 0.93 \pm 0.07\,(stat.) \pm 0.19\,(sys.).$$

The π^0/e double ratio is consistent with $\nu_\mu \leftrightarrow \nu_\tau$. But, because of the current large

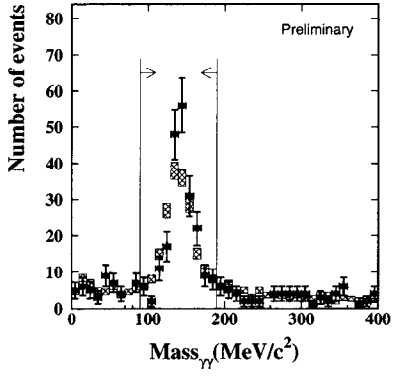


Fig. 7. Invariant mass distribution of two e-like ring events. The shaded histogram shows the Monte Carlo simulation and the vertical spread of the histogram shows the statistical error of the Monte Carlo simulation. The cut region for π^0 events is also shown.

systematic error, we cannot differentiate $\nu_\mu \leftrightarrow \nu_\tau$ and $\nu_\mu \leftrightarrow \nu_s$ at present. The main source of the systematic error is the uncertainty of the cross section of the neutral current π^0 production. Fortunately, we will start the K2K long baseline experiment next year. We will construct a one kilo-ton water Cherenkov (1kt) detector at KEK. The 1kt detector is able to measure the cross section quite precisely. Hence, in the near future we can discuss the π^0/e double ratio with a much smaller systematic error.

3.2 Upward-Going Muons

Atmospheric muon neutrinos which interact in the rock surrounding the SK tank produce muons by charged current interactions and the resulting muons pass through the SK detector. Although downward-going muons are dominated by cosmic-rays, upward-going muons are pure samples of atmospheric neutrinos. We analyze two types of upward-going muons. One is upward-through-going muons, which have signals of muon entrance and exit in the outer detector and pass through the inner detector. The path length of the through-going muons is required to be longer than seven meters in this analysis. The other type is upward-stopping muons, which is the same as upward-through-going muons but they have only entrance signals in the outer detector. Figure 8 shows calculated energy distributions of primary atmospheric neutrinos which cause upward-through-going or upward-stopping muons. The averaged neutrino energy is $\sim$100 GeV for upward-through-going muons and $\sim$10 GeV for upward-stopping muons. We expect a large reduction of flux for near vertical upward-through-going muons and small reduction for near horizontal direction with the oscillation parameters given by the contained event analysis. On the other hand, the flux of upward-stopping muons must be reduced for the entire zenith angle range by neutrino oscillations.

Super-Kamiokande observed 617 upward-through-going muons for 534 days of live time. The observed flux of atmospheric neutrinos is 1.75$\pm$0.07(stat.)$\pm$0.09(sys.) ($\times 10^{-13}/cm^2/sec/str$). On the other hand, the expected flux is 1.88 $\pm$ 0.4 ($\times 10^{-13}$ $/cm^2/sec/str$) using the flux calculation by Honda *et al.*[5] and 2.02 $\pm$ 0.4 ($\times 10^{-13}$ $/cm^2/sec/str$) using Bartol flux.[4] Neutrino interactions are simulated by using the structure function of GRV94DIS[12] and the energy loss of muons is based on the calculation by Lohmann *et al.*[13] Because of the $\sim$20% error in expected flux calculation, we use only the shape of the zenith angle distribution for the neutrino oscillation analysis for upward-through-going muons. The zenith angle distribution of the flux is shown in Fig. 9 together with the expectation using the Honda flux. The observed flux in the near horizontal direction agrees with the expectation, but most of the data points near the vertical direction are lower than the expectation. The shape comparison between the data and the expectation gives a χ^2 of 18.3 with 9 d.o.f., corresponding to 3.2% CL. By assuming the neutrino oscillations, the best fit of the shape of the zenith angle

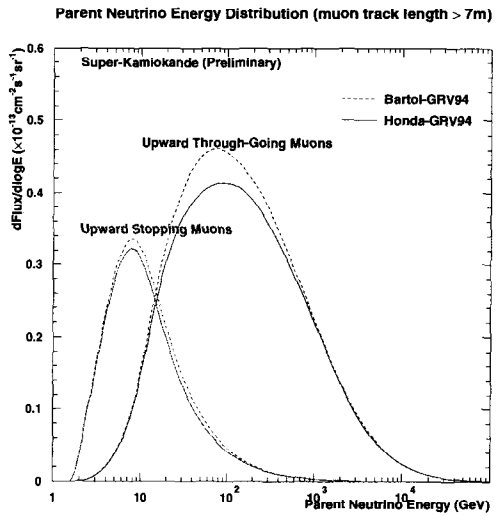


Fig. 8. Calculated primary neutrino energy distributions for upward-through-going and upward-stopping muons. The solid and dotted curves are based on the flux calculations by Bartol[4] and Honda *et al.*,[5] respectively.

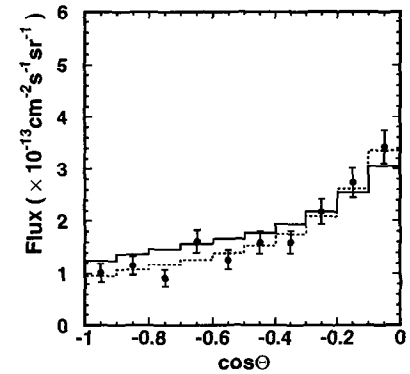


Fig. 9. Zenith angle distribution of the flux of upward-through going muons. Solid histograms show Monte Carlo expectations without neutrino oscillations. The dotted histogram shows the best fit with neutrino oscillations.

distribution is obtained at $\Delta m^2 = 2.5 \times 10^{-3}$ eV^2 and $\sin^2 2\theta = 1.0$. The χ^2 for the fit is 7.3 with 9 d.o.f. (61% CL). The absolute normalization of the flux is a free parameter in the fit. The best fit zenith angle distribution is shown by a dotted histogram in Fig. 9. An allowed region of the neutrino oscillation parameters is shown by the thick-dotted curve in Fig. 6 and the region overlaps with the one obtained by contained event analysis quite well.

The neutrino oscillation parameters obtained by the contained events analysis suggest that neutrinos with $E_\nu \lesssim 40$ GeV should be largely oscillated in traveling the distance of earth's diameter, whereas neutrinos with $E_\nu \gg 40$ GeV do not oscillate so much. Hence, the oscillation effect is larger for upward-stopping muons than the upward-through-going muons as seen in Fig. 8. Super–Kamiokande has observed 137 upward-stopping muons during 516 days' live time. We use the ratio of the numbers of stopping muons to through-going muons ($stop/through$ ratio) for the test of neutrino oscillations, because it cancels the absolute normalization of the flux. The $stop/through$ ratio is compared between data and a Monte Carlo simulation and:

$$\frac{(stop/through)_{data}}{(stop/through)_{MC}} = \frac{0.22 \pm 0.023 \pm 0.014}{0.39 \pm 0.05} = 0.56 \pm 0.10.$$

The value is significantly smaller than unity. Figure 10 shows the zenith angle distribu-

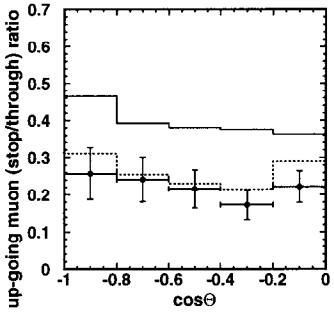


Fig. 10. Zenith angle distribution of the *stop/through* ratio of the upward going muons. The solid histogram shows the expected ratio without neutrino oscillations. The dashed histogram shows the expectation for $\nu_\mu \leftrightarrow \nu_\tau$ oscillations with $\sin^2 2\theta = 1.0$ and $\Delta m^2 = 3 \times 10^{-3}$ eV^2. The expected *stop/through* ratio has $\pm 14\%$ uncertainty.

tion of the *stop/through* ratio together with the predictions with and without neutrino oscillations. The prediction with neutrino oscillations agrees well with the data within the systematic uncertainty ($\pm 14\%$). The allowed region of neutrino oscillation parameters obtained by the *stop/through* ratio is shown by the thin-solid curve in Fig. 6. The allowed region overlaps with the one obtained by the contained event analysis.

3.3 Summary

The small μ/e ratio and the distortion of the zenith angle distribution in contained events give evidence for ν_μ oscillations. The analyses of the zenith angle dependence of upward-through-going muons and the *stop/through* ratio confirm the results obtained by the contained event analysis. The obtained neutrino oscillation parameters are Δm^2 of $10^{-3} - 10^{-2}$eV^2 and $\sin^2 2\theta > 0.8$.

4 Solar Neutrinos

Solar neutrinos had been observed by the Homestake experiment,[15] Kamiokande,[16] GALLEX,[17] and SAGE.[18] All these experiments observed significantly smaller solar neutrino flux than the expectations from standard solar models (SSMs),[19–21] known as "the solar neutrino problem." It is suggested that explaining the results of all four experiments in the framework of SSM has a difficulty, even if the input parameters of the SSM are changed. One possible way to explain all these results is neutrino oscillations.[22] Analysis taking into account the MSW mechanism of neutrino oscillations suggests neutrino oscillation parameters around $\Delta m^2 \sim 6 \times 10^{-6}$ eV^2 and $\sin^2 2\theta \sim 5 \times 10^{-3}$ (small mixing angle solution), $\Delta m^2 \sim 10^{-5} - 10^{-4}$ eV^2 and $\sin^2 2\theta \sim 0.6$ (large mixing angle solution), and $\Delta m^2 \sim 10^{-11}$ eV^2 and $\sin^2 2\theta \sim 1$ (just-so solution).[23,24] In order to test neutrino oscillations without relying on solar models, quantities which are independent from SSM expectations should be used. Short time variations of the solar neutrino flux, such as the day/night difference of the flux, could be observed in case of a large mixing angle solution and a part of a small mixing angle solution. And a distortion of the neutrino energy spectrum could be observed in case of small mixing angle and just-so solutions. Super-Kamiokande is a real-time detector and measures ^{8}B solar neutrinos by neutrino-electron scattering. Super-Kamiokande is able to measure precise neutrino flux variation and spectrum of solar neutrinos.

4.1 Calibration

The calibration of the detector is quite important for the solar neutrino measurement. Especially, absolute energy calibration is essential for the spectrum measurement. In order to perform precise calibration of absolute energy scale, energy resolution, and angular resolution, an electron linear accelerator (LINAC) was constructed near the SK tank.[25] An electron beam from the LINAC is transported into the SK tank. The LINAC injects monoenergetic electrons with a tunable energy ranging from 5 to 16 MeV. The absolute energy of the beam is measured by a germanium detector, which was in turn calibrated by gamma-ray sources and internal-conversion electrons from a ^{207}Bi source. The uncertainty of the beam energy is less than 20 keV over the energy range covered by the LINAC. LINAC data were taken at eight representative positions in the Super–Kamiokande inner volume with seven different momenta ranging from 4.89 to 16.09 MeV. Figure 11 shows energy distributions of the LINAC data taken at a

typical position overlaid with corresponding Monte Carlo (MC) simulations.

The LINAC data and MC distributions agree with each other quite well. The peak

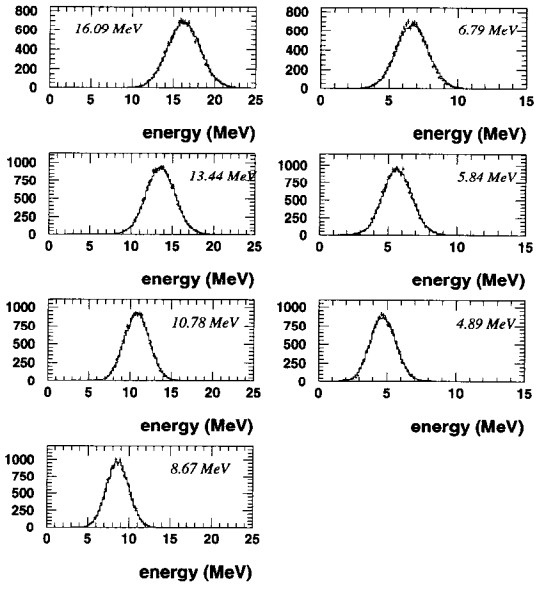


Fig. 11. LINAC (crosses) and MC (histogram) energy distributions for LINAC calibration data.

position of the distributions indicates the absolute energy scale. Figure 12 shows the absolute energy scale difference between data and MC simulation. The detector position dependence of the energy scale is estimated to be 0.5% in rms using the LINAC calibration. The uncertainty of the absolute energy in the LINAC calibration system itself is 0.55% at 6 MeV and 0.3% at 10 MeV. The LINAC calibration data was taken only in the downward-going direction at the moment. The energy scale for all directions are checked using beta-rays from spallation products induced by cosmic-ray muons and the directional dependence of the absolute energy scale is measured to be less than 0.5%. Adding all possible sources of uncertainties, the absolute energy scale is calibrated with

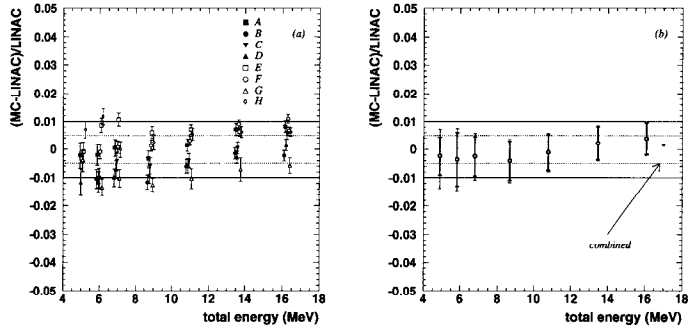


Fig. 12. Comparison of absolute energy scales for LINAC and MC. (a) the difference of the absolute energy scale for each LINAC position; (b) position average. The last point in (b) represents the total average over all positions and beam energies.

an accuracy of 0.8% at 10 MeV.

The energy resolution is also calibrated using the LINAC. The observed energy resolution is, for example, 1.58 MeV for 10.78 MeV electrons, and the measured energy resolution agrees with the MC simulation better than 2%. Angular resolution is also calibrated by the LINAC. Figure 13 shows the opening angle between the reconstructed electron direction and the direction of beam injection for MC and LINAC at a typical position. The measured angular resolution is 2–3% smaller than the one obtained by the current MC simulation. The difference is corrected in extracting solar neutrino flux from the observed data.

The stability of the detector is monitored by using decay electrons from stopping cosmic-ray muons and beta-decay nuclei from spallation products. The time variation of the water transparency is monitored by using the decay electrons and the obtained water transparency is used to calculate energy for each event. Applying the water transparency correction, the stability of the energy scale is better than $\pm 0.5\%$ for over two years. The absolute energy calibration is cross-checked by using the decays of ^{16}N produced by stopping muon capture on oxygen. The difference in energy scales between that obtained by ^{16}N decay beta spectrum and the MC tuned by LINAC is $0.2\ ^{+0.6}_{-0.8}\%$.

In summary, the energy response and angular response of the SK detector are cali-

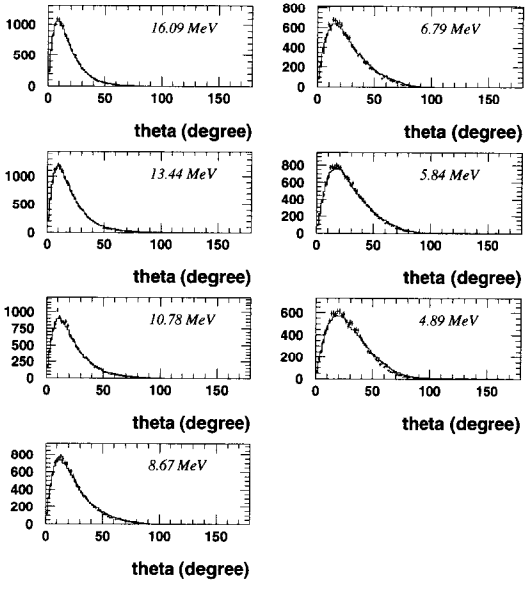


Fig. 13. Angular distributions for a typical LINAC calibration data. The histogram is MC, points LINAC.

brated quite precisely and they are enough for testing flux time variations and the energy spectrum of solar neutrinos at the present statistical level.

4.2 Data Analysis

SK has obtained 504 live days' data between 31 May 1996 and 25 March 1998. The obtained data were reduced by applying the following cuts[26]: (1) less than 20 hits in the outer detector (i.e., select contained events), (2) less than 1000 photoelectrons, corresponds to 110~120 MeV, and (3) time to the previous event is more than 20 μs (to reject decay electrons from stopping cosmic-ray muons). The event rate after the reduction is ~57 events/day/kton above 6.5 MeV. The efficiency of the reduction for solar neutrino signals is 94.2%. The main source of the background after the reduction is muon–induced spallation products. They are rejected by using the time from the preceding muon, distance between the vertex of low energy event and the muon track, and pulse height of the muon. The spallation cut reduces the event rate to ~12 events/day/kton. The dead time induced by this cut is 20.0%, which gives maximum significance, defined to be solar neutrino signal/$\sqrt{\text{background}}$. The data was finally reduced by removing the gamma-ray background coming from the wall of the detector by requiring more than 4.5 m in the length of the backward extrapolated vector from the event vertex to the inner detector wall. This cut gives a final event rate of ~7.6 events/day/kton with an estimated loss of 7.8% of solar neutrino signal.

Figure 14 shows the angular distribution of the final events with respect to the direction of the sun. A clear enhancement of solar neutrino signal is seen in the direction of the sun. The solar neutrino signal amounts to $6820^{+150}_{-130}(stat.)$ events for the 504 days' data. Using the expected energy shape of ^{8}B solar neutrinos, the flux of ^{8}B solar neutrinos is obtained to be $[2.44\pm0.05(stat.)^{+0.09}_{-0.07}(syst.)] \times 10^6/\text{cm}^2/\text{s}$. Major sources of the systematic error in the flux are (1) uncertainty in absolute energy scale and resolution ($^{+2.2}_{-2.1}$% error in flux), (2) uncertainty in angular resolution (+2.2%), (3) uncertainty in the ^{8}B spectral shape ($^{+2.2}_{-2.1}$%). Comparing the obtained flux with the SSM predictions, the ratio of flux between the SK data and SSM prediction (data/SSM) is $0.474^{+0.010}_{-0.009}(stat.)^{+0.017}_{-0.014}(syst.)$ using the latest SSM by Bahcall *et al.*[19] The obtained ratio is significantly smaller than unity. The time variation of solar neutrino flux is shown in Fig. 15. The curve in the figure shows expected flux variation due to the eccentricity of the earth. The absolute flux of the expectation is reduced to 47%. Comparing the observed time variation of the flux with the expectation, χ^2 is 8.78 with 14 degrees of

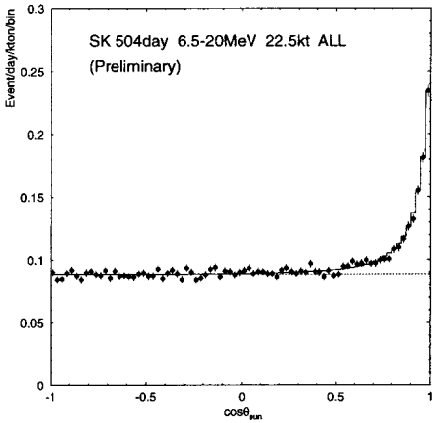


Fig. 14. Angular distribution of the final events with respect to the direction of the sun.

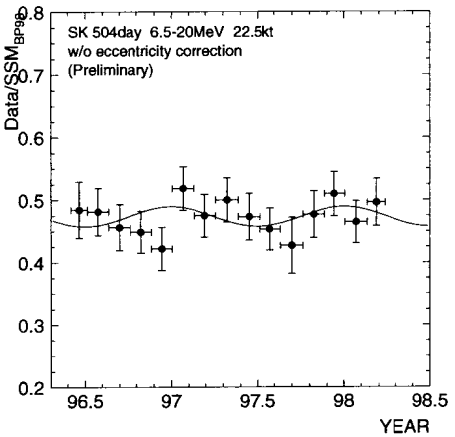


Fig. 15. Time variation of the solar neutrino flux. The curve in the figure shows expected flux variation due to the eccentricity of the earth.

freedom (d.o.f.), which corresponds to 84.4% CL. Hence, no significant time variation of the flux over the expected variation is observed.

4.3 Day/Night Analysis

The obtained data was divided into day and night samples, where day and night are defined to be whether the sun is above or below horizon. The observed solar neutrino flux is $[2.37\pm0.07(stat.)^{+0.08}_{-0.07}(syst.)] \times 10^6/\mathrm{cm}^2/\mathrm{s}$ for day-time and $[2.48^{+0.07}_{-0.06}(stat.)^{+0.09}_{-0.07}(syst.)] \times 10^6/\mathrm{cm}^2/\mathrm{s}$ for night-time. The difference of the flux between day-time(D) and night-time(N) is given as

$$\frac{D-N}{D+N} = -0.023 \pm 0.020 \pm 0.014.$$

Thus, there is no significant day/night difference. The night-time data was subdivided into five bins according to the zenith angle of the sun (θ_z) at the time when the neutrino event is observed, N1 ($0 < \cos(\theta_z) \leq 0.2$), N2 ($0.2 < \cos(\theta_z) \leq 0.4$), ..., N5 ($0.8 < \cos(\theta_z) \leq 1.0$). Neutrinos pass through the mantle for N1 to N4 and through the core of the earth in the case of N5. The flux of each data set is shown in Fig. 16. The χ^2 for the night 5 bins assuming a flat distribution is 5.87 with 4 d.o.f. (20.9 % CL). Therefore, no significant variation of the flux during the night-time is observed.

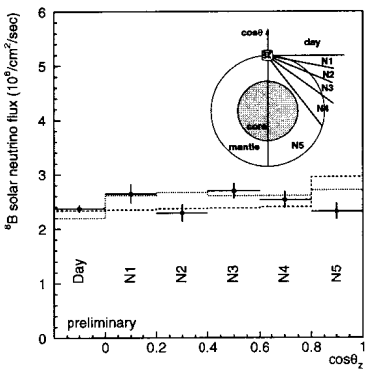


Fig. 16. Flux of solar neutrinos for day-time and night-time samples. The histograms are expectations from neutrino oscillations (see text).

Using the day/night comparison, a solar model independent constraint on neutrino oscillation parameters is obtained. The dotted and dashed histograms in Fig. 16 show expected flux variations for the examples of the large mixing angle solution ($\sin^2 2\theta$ = 0.56, $\Delta m^2 = 1.2\times10^{-5}$ eV^2) and the small mixing angle solution ($\sin^2 2\theta = 0.01$, Δm^2 = 6.3×10^{-6} eV^2) assuming $\nu_e \rightarrow \nu_\mu(\nu_\tau)$ oscillations. Large variation of the flux during night-time is expected for the large mixing angle solution and increase of flux in N5 is expected for the small mixing angle solution. A flux independent constraint on neutrino oscillation parameters is given by using the following χ^2:

$$\chi^2 = \sum_{i=D,N1,\cdots,N5} \left\{ \frac{\phi_i - \alpha \times \phi_{SSM} \times R_i(\sin^2 2\theta, \Delta m^2)}{\sigma_i} \right\}^2,$$

where ϕ_i is the observed flux, ϕ_{SSM} is the flux expected from the SSM, $R_i(\sin^2 2\theta, \Delta m^2)$ is the ratio of the event rate with and without neutrino oscillations, σ_i is the error of observed flux (statistical errors and systematic errors are added quadratically), and α is a normalization factor of the solar neutrino flux. α is a free parameter in the χ^2. The region where $\chi^2 > 15.09$ is excluded with 99% CL as shown in Fig. 17. As a reference, the shaded regions show allowed regions of oscillation parameters which were obtained by comparing the observed flux by Homestake, GALLEX, SAGE, and SK with the corresponding expectations from SSM.[19] A bottom part of the large mixing angle solution and a small part of the small mixing angle solution are excluded by the model independent analysis using day/night flux variation.

4.4 Energy Spectrum Analysis

The obtained final data sample is sub-divided into 16 energy bins, every 0.5 MeV from 6.5 to 14.0 MeV and one bin combining energies from 14.0 to 20.0 MeV. The number of solar neutrino events in each energy bin is extracted from the data by fitting each angular distribution to the sun with a background angular distribution and the expected angular distribution of solar neutrino signals. The background angular distribution is obtained by mapping the remaining background to the direction of the sun. The uncertainty of the flux due to the background angular distribution is estimated to be 0.1%. The number of solar neutrino events thus obtained is shown in Fig. 18. The histogram in the figure shows the expected spectrum from the SSM, which was calculated using the total ^{8}B solar neutrino flux from Ref. 19 (BP98) (5.15×10^6/cm^2/s), the calculation of ^{8}B neutrino spectral shape from Ref. 27, and the electron spectrum of ν-e scattering from Ref. 28, in which radiative corrections are taken into account. The response of the

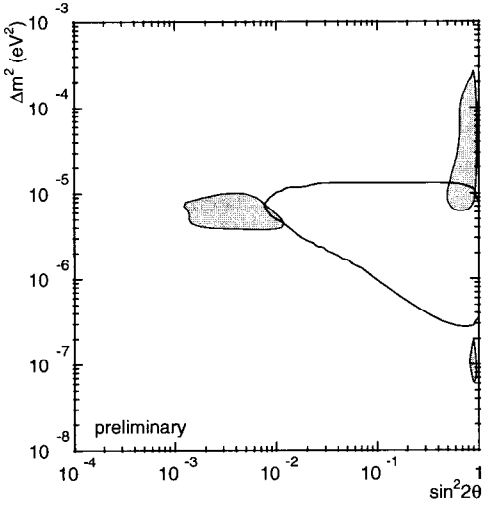


Fig. 17. Excluded region of neutrino oscillation parameters using the model independent day/night analysis with 99% CL for $\nu_e \rightarrow \nu_\mu(\nu_\tau)$ oscillations (white region). The shaded regions show allowed regions of oscillation parameters obtained by using the measured flux of solar neutrino experiments (see text).

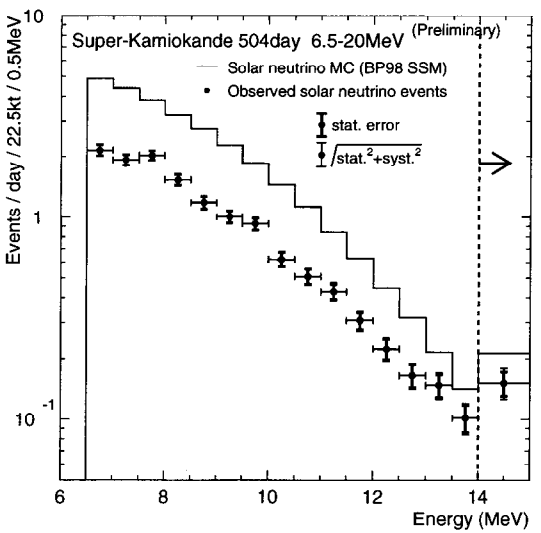


Fig. 18. Energy spectrum of recoil electrons. The inner error shows statistical error and outer error shows systematic error. The histogram shows the expected energy spectrum (see text).

detector, such as the smearing of the electron energy spectrum by the detector energy resolution, is also taken into account by a full MC simulation. In order to compare the shape of the observed energy spectrum with the expected spectrum, the ratio of the spectra are taken as shown in Fig. 19. Systematic errors in the energy shape comparison

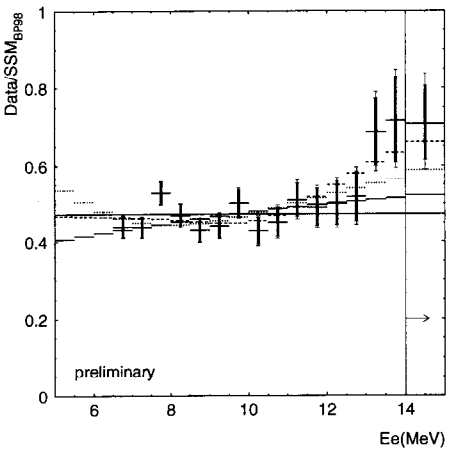


Fig. 19. Ratio of observed energy spectrum to expected spectrum. Inner and outer error bars show statistical and systematic errors, respectively. The dashed histogram shows the best fit assuming neutrino oscillations obtained with $\sin^2 2\theta = 0.93$, $\Delta m^2 = 4.2\times10^{-10}$ eV^2. The solid and dotted histograms show a typical small mixing angle solution ($\sin^2 2\theta = 5\times10^{-3}$, $\Delta m^2 = 7.1\times10^{-6}$ eV^2) and a just-so solution ($\sin^2 2\theta = 0.83$, $\Delta m^2 = 7.1\times10^{-11}$ eV^2), respectively. $\nu_e \rightarrow \nu_\mu(\nu_\tau)$ oscillations are assumed here.

are classified into energy-bin-correlated (called "correlated" from now on) errors and energy-bin-uncorrelated ("uncorrelated") errors. The sources of correlated errors are uncertainties in the absolute energy scale and energy resolution, and uncertainty in the expected energy shape of neutrinos from ^{8}B decays. For example, the correlated errors for the 13.0 to 13.5 MeV energy bin is $^{+6.9}_{6.2}$% from the uncertainties of energy scale and resolution, and $^{+4.7}_{-3.8}$% from the uncertainty of the expected energy shape. The sources of uncorrelated errors are uncertainties in trigger efficiency, data reduction efficiency,

angular resolution, etc. The sum of uncorrelated errors is estimated to be $^{+2.1}_{-1.7}$% for the 6.5 to 7.0 MeV energy bin and ± 1.7% for other energy bins. The sum of correlated and uncorrelated systematic errors is shown as outer error bars in Fig. 19. The statistical significance of the possible distortion of the energy spectrum is discussed using the following χ^2:

$$\chi^2 = \sum_{i=1}^{16} \left\{ \frac{(\frac{data}{SSM})_i - \alpha/(1+\delta_{i,corr} \times \beta)}{\sigma_i} \right\}^2 + \beta^2,$$

where $\delta_{i,corr}$ is the 1σ error of the correlated error described above, σ_i is an 1σ error for each energy bin defined as a sum of the statistical error and uncorrelated errors added quadratically, and α is a free parameter which indicates the relative flux of ^{8}B solar neutrinos. β is a free parameter to constrain the magnitude of the correlated error. The minimum χ^2 is obtained to be 25.1 with 15 d.o.f., which corresponds to 4.8% CL.

The possible contribution of high energy solar neutrinos from $^3He + p \rightarrow {}^4\mathrm{He} + e^+ + \nu_e$ (hep) is estimated using the best estimate flux in BP98, $2.1 \times 10^3/\mathrm{cm}^2/\mathrm{s}$. The expected number of hep events during the 504 days' period is 2.1 for energy above 13.0 MeV, corresponding to 1.0% of the observed solar neutrino events. The error of the expected hep neutrino flux is not precisely estimated in BP98, but unless the expectation is drastically wrong, the contribution of the hep neutrino is negligible at the present statistical level.

Preliminary neutrino oscillation analysis is performed using the spectral shape together with the day/night difference. χ^2 for the oscillation analysis is defined to be

$$\chi^2 = \sum_{D,N} \sum_{i=1}^{i=16} \left\{ \frac{(\frac{data}{SSM})_i - \alpha \times R_i(sin^2 2\theta, \Delta m^2)/(1+\delta_{i,corr} \times \beta)}{\sigma_i} \right\}^2 + \beta^2,$$

where $R_i(sin^2 2\theta, \Delta m^2)$ is the ratio of electron energy spectrum with and without neutrino oscillations. $\nu_e \rightarrow \nu_\mu(\nu_\tau)$ oscillations are assumed here. The minimum χ^2 is obtained to be 38.7 with 31 d.o.f. (16.1 % CL) at $\sin^2 2\theta = 0.93$, $\Delta m^2 = 4.2 \times 10^{-10}$ eV^2. The best fit energy spectrum is shown by the dotted histogram in Fig. 19. The χ^2 calculated by combining day and night spectra is 13.2 with 15 d.o.f. (58% CL). Typical small mixing angle and just-so solutions are also shown in Fig. 19, and χ^2 (day/night combined) for them are 23.0 (8.5% CL) and 18.9 (22% CL), respectively. Allowed regions of neutrino oscillation parameters are obtained as Fig. 20 with 99% CL. The current analysis on the solar neutrino spectrum is not able to give a definite statement on neutrino oscillations, because the current statistical significance is not high enough

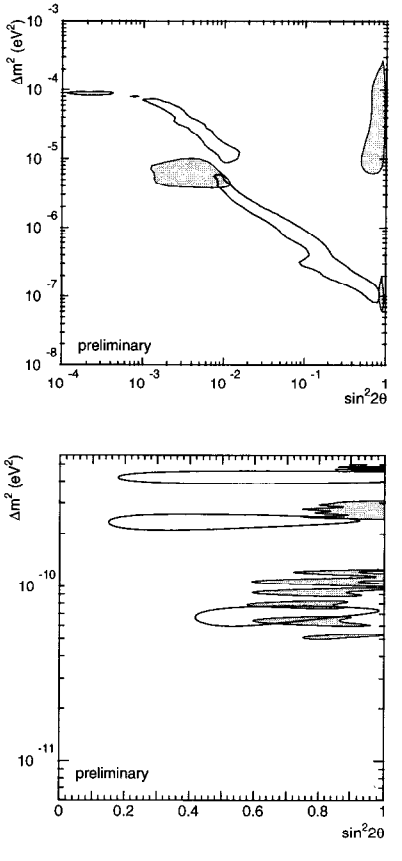


Fig. 20. White regions show allowed regions of neutrino oscillation parameters obtained by using the spectral shape analysis with 99% CL. The shaded regions show the allowed region obtained by the flux comparison between solar neutrino experiments and SSM predictions. $\nu_e \rightarrow \nu_\mu(\nu_\tau)$ oscillations are assumed.

to discuss possible solutions. But, in the future, increasing statistics and lowering the analysis energy threshold should enable us to pinpoint neutrino oscillations using the shape of the energy spectrum. In fact, the spectrum below 6.5 MeV should differentiate possible parameters of neutrino oscillations as seen in Fig. 19.

5 Conclusions

Results of atmospheric neutrinos observed by contained events, upward-through-going muons, and upward-stopping muons are presented. The small μ/e ratio is observed in the contained events and large zenith angle asymmetry is observed in μ-like events. Those anomalies can be explained consistently by neutrino oscillations. The zenith angle distortion of upward-through-going muons and the small *stop/through* ratio of upward-going muons confirm the results of the contained event analyses. The allowed region of neutrino oscillation parameters obtained by the three independent analyses overlap quite well and the obtained neutrino oscillation parameters are Δm^2 of $10^{-3} - 10^{-2}\mathrm{eV}^2$ and $\sin^2 2\theta > 0.8$.

Preliminary results from 504 days' solar neutrino data of Super-Kamiokande is presented. The observed flux of ^{8}B solar neutrinos is $[2.44\pm0.05(\mathrm{stat.})^{+0.09}_{-0.07}(\mathrm{syst.})] \times 10^6/\mathrm{cm}^2/\mathrm{s}$, which corresponds to 47% of the value expected from the standard solar model.[19] Analysis of the day/night difference and the flux variation during the night-time does not show significant flux variations. This result gives a model independent excluded region on neutrino oscillation parameters as shown in Fig. 17. The spectral shape comparison between the observation and the expectation gives a χ^2 of 25.1 (corresponding to 4.8%) with 15 d.o.f. above 6.5 MeV. A future increase of statistics and lowering the analysis energy threshold should provide fruitful information on neutrino oscillations.

References

[1] K. S. Hirata *et al.*, Phys. Lett. B **205**, 416 (1988); K. S. Hirata *et al.*, Phys. Lett. B **280**, 146 (1992).

[2] Y. Fukuda *et al.*, Phys. Lett. B **335**, 237 (1994).

[3] D. Casper *et al.*, Phys. Rev. Lett. **66**, 2561 (1991); R. Becker-Szendy *et al.*, Phys. Rev. D **46**, 3720 (1992).

[4] G. Barr *et al.*, Phys. Rev. D **39**, 3532 (1989); V. Agrawal *et al.*, Phys. Rev. D53, 1313 (1996); T. K. Gaisser and T. Stanev, Proc. 24th Int. Cosmic Ray Conf. (Rome) Vol. 1, 694 (1995).

[5] M. Honda *et al.*, Phys. Lett. B **248**, 193 (1990); M. Honda *et al.*, Phys. Rev. D **52**, 4985 (1995).

[6] Y. Fukuda, *et al.*, Phys. Lett. B **433**, 9 (1998).

[7] Super-Kamiokande Collaboration, Y. Fukuda *et al.*, Phys. Lett. B **436**, 33 (1998).

[8] K. Daum *et al.*, Z. Phys. C **66**, 417 (1995).

[9] M. Aglietta *et al.*, Europhys. Lett. **8**, 611 (1989).

[10] W. W. M. Allison *et al.*, Phys. Lett. B **391**, 491 (1997); T. Kafka, Proceedings of 5th Int. Workshop on Topics in Astroparticle and Underground Physics, Gran Sasso, Italy, Sept. 1997.

[11] Y. Fukuda *et al.*, Phys. Rev. Lett. **81** (1998) 1562.

[12] M. Glück, E. Reya, and A. Vogt, Z. Phys. C **67**, 433 (1995).

[13] W. Lohmann, R. Kopp, and R. Voss, CERN Yellow Report No. 85-03.

[14] S. Kasuga *et al.*, Phys. Lett. B **374**, 238 (1996).

[15] R. Davis, Prog. Part. Nucl. Phys. **32**, 13 (1994); B. T. Cleveland *et al.*, Nucl. Phys. B (Proc. Suppl.) **38**, 47 (1995).

[16] K. S. Hirata *et al.*, Phys. Rev. Lett. **65**, 1297 (1990); K. S. Hirata *et al.*, Phys. Rev. D **44**, 2241 (1991); D **45**, 2170E (1992); Y. Fukuda *et al.*, Phys. Rev. Lett. **77**, 1683 (1996).

[17] P. Anselmann *et al.*, Phys. Lett. B **327**, 377 (1994): B*342*, 440 (1995).

[18] J. N. Abdurashitov *et al.*, Phys. Lett. B **328**, 234 (1994).

[19] J. N. Bahcall, S. Basu, and M. Pinsonneault, Phys. Lett. B **433**, 1 (1988).

[20] J. N. Bahcall and M. Pinsonneault, Rev. Mod. Phys. **67**, 781 (1995).

[21] S. Turck-Chieze and I. Lopes, Ap. J. **408**, 347 (1993).

[22] S. P. Mikheyev and A. Y. Smirnov, Sov. Jour. Nucl. Phys. **42**, 913 (1985); L. Wolfenstein, Phys. Rev. D **17**, 2369 (1978).

[23] N. Hata and P. Langacker, Phys. Rev. D **56**, 6107 (1997).

[24] P. Langacker, talk presented in NEUTRINO '98, Proceedings of the XVIII International Conference on Neutrino Physics and Astrophysics, Takayama, Japan, 4-9 June 1998, edited by Y. Suzuki and Y. Totsuka, Nucl. Phys. B (Proc. Suppl.) **77**, 241 (1999).

[25] M. Nakahata *et al.*, Nucl. Instrum. Meth. A **421**, 113 (1999).

[26] Y. Fukuda *et al.*, Phys. Rev. Lett. **81**, 1158 (1998).

[27] J. N. Bahcall *et al.*, Phys. Rev. C **54**, 411 (1996).

[28] J. N. Bahcall *et al.*, Phys. Rev. D **51**, 6146 (1995).

PRELIMINARY RESULTS ON CP-VIOLATING PHENOMENA IN THE KTEV EXPERIMENT AT FERMILAB

Richard Kessler
Enrico Fermi Institute
University of Chicago
Chicago, Illinois 60637

ABSTRACT

These proceedings will discuss some preliminary results from the KTeV experiment at Fermilab, with an emphasis on CP violation in the kaon system. Direct CP-violating topics include the measurement of $\mathrm{Re}(\epsilon'/\epsilon)$ and searches for the decays $K_L \to \pi^0\nu\bar{\nu}$ and $K_L \to \pi^0 e^+ e^-$. A new indirect CP-violating observable in the decay $K_L \to \pi^+\pi^- e^+ e^-$ is also discussed.

© 1998 by Richard Kessler.

1 Introduction

Since the early 1960s, there have been three observations of CP-violating phenomena in particle physics: $K_L \to 2\pi$, the rate asymmetry between $\Gamma(K_L \to \pi^+ e^- \nu)$ and $\Gamma(K_L \to \pi^- e^+ \nu)$ (and similarly for $K_L \to \pi\mu\nu$), and the interference between $K_{L,S} \to \pi^+\pi^-\gamma$. All of these CP-violating effects are in the kaon system and can be described by a tiny asymmetry in the $K^0 \leftrightarrow \overline{K^0}$ mixing, which is usually referred to as "indirect" CP violation (CPV). The interesting question that KTeV is designed to study is whether CPV can be seen "directly" in decay amplitudes. With at least three families of quarks and leptons, the Standard Model (SM) allows for direct CPV which can be conveniently parameterized by the CKM parameter η.

The KTeV experiment is searching for three effects that are expected to exhibit direct CPV and provide information on the value of the CP-violating parameter η. Most of our data taking was dedicated to collecting $K_{L,S} \to \pi^+\pi^-$ and $K_{L,S} \to \pi^0\pi^0$ decays to measure $\mathrm{Re}(\epsilon'/\epsilon)$. We have also searched for the decays $K_L \to \pi^0\nu\bar{\nu}$ and $K_L \to \pi^0 e^+ e^-$; the former is expected to be "purely" direct CP violating, i.e., the rate is proportional to $|\eta|^2$, while the latter is expected to have both CP-conserving (CPC) and CP-violating contributions. New data on the $K_L \to \pi^0\gamma\gamma$ decay are presented, which helps unravel the CPC contribution to $K_L \to \pi^0 e^+ e^-$. In addition to our direct CP-violating searches, we have also observed the decay $K_L \to \pi^+\pi^- e^+ e^-$ which is expected to have an indirect CP-violating asymmetry in the angle between the $\pi^+\pi^-$ and e^+e^- decay planes.

2 The KTeV Detector

The kaon beams are produced by an 800 GeV primary proton beam that strikes a BeO target ($3 \times 3 \times 305$ mm^3). Roughly half of the 3×10^{12} protons per spill interact in the target while the other half are absorbed in a beam dump. A large sweeping magnet downstream from the beam dump effectively sweeps muons from both the primary target and the dump so that the muon flux hitting the detector is less than 100 kHz. Two identical sets of collimators produce two secondary neutral beams at ± 0.8 mrad in the horizontal view. A set of beryllium and lead absorbers, 20 meters from the target, reduced the fraction of neutrons and photons in the beam.

The two kaon beams that enter the fiducial volume are illustrated in Fig. 1, which shows the detector configuration used to measure $\mathrm{Re}(\epsilon'/\epsilon)$. The vacuum decay region is from about 120 to 159 meters from the primary target. Following the vacuum region, helium bags are used to fill the volumes between detector elements. The K_L beam is realized by placing the detector sufficiently far from the primary target so that the K_S component has decayed away; the K_S beam is made by sending one of the K_L beams into a fully active 1.7 meter long regenerator made of a plastic scintillator.

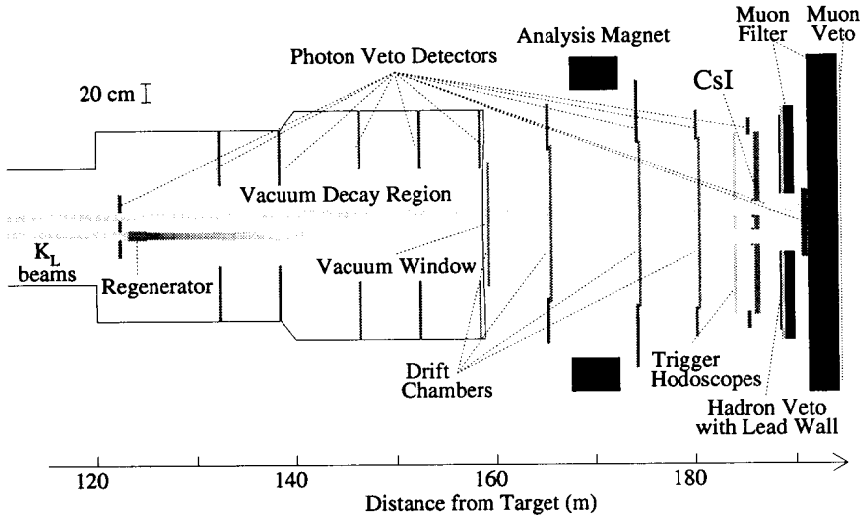


Fig. 1. The KTeV detector configuration used for the measurement of $\mathrm{Re}(\epsilon'/\epsilon)$.

The momenta of charged tracks are measured with four drift chambers, two on each side of an analysis magnet (0.4 GeV/c kick in X-view). A CsI electromagnetic calorimeter is used to measure the position and energy of photons and electrons. The inner 1.2×1.2 m^2 region consists of 2232 small crystals, each $0.025 \times 0.025 \times 0.500$ m^3; the outer region consists of 868 large crystals, each $0.050 \times 0.050 \times 0.500$ m^3. A nearly hermetic photon veto system is used to veto events in which one or more particles escape the detector. A set of trigger hodoscopes just upstream of the CsI calorimeter is used to trigger charged particles. Both the CsI calorimeter and the trigger hodoscopes

have two 0.15×0.15 m^2 holes to allow the high intensity neutral beams to pass through.

Rare kaon decay searches used the detector configuration shown in Fig. 1 with the following differences: (i) the regenerator and most upstream photon vetoes were removed to provide two identical K_L beams and a longer decay region, (ii) a set of nine Transition Radiation Detectors (TRD) were installed downstream from the last drift chamber to provide additional separation between pions and electrons, (iii) the K_L flux in the decay volume was several times higher, and (iv) the magnet kick was reduced from 0.4 GeV/c to 0.2 GeV/c in order to increase the acceptance for many-body decays.

3 Direct CP-Violating Searches

3.1 Re(ϵ'/ϵ)

We do not yet have a result on the measurement of Re(ϵ'/ϵ) so I will just make a few comments about our analysis and the critical systematic issues. We are aiming to extract a first result using about 25% of our total data set. The calibration and alignment has been completed for the CsI, drift chambers, and photon vetos. The statistics of this data set is shown in Table 1 and leads to a statistical uncertainty of 3×10^{-4} on Re(ϵ'/ϵ).

ϵ'/ϵ mode	Events	Bkg./signal
Vac ($K_L \to$) $\pi^0\pi^0$	1.0×10^6	0.8%
Reg ($K_S \to$) $\pi^0\pi^0$	1.6×10^6	1.5%
Vac ($K_L \to$) $\pi^+\pi^-$	1.9×10^6	0.1%
Reg ($K_S \to$) $\pi^+\pi^-$	3.9×10^6	0.1%

Table 1. $K \to \pi\pi$ statistics under analysis for the first result on Re(ϵ'/ϵ).

The most serious systematic issue is related to understanding the $K_{L,S} \to \pi^+\pi^-$ acceptance for tracks that hit the beam region of the drift chambers. The size of this region is $\sim 0.1 \times 0.1$ m^2 in each of the two beams, and about one-fourth of the $K_{L,S} \to \pi^+\pi^-$ events have at least one track that goes through this region of a drift chamber. A combination of high intensity and damage to the wires in the beam regions resulted in higher inefficiencies, and late hits in which the primary drift electrons did not fire the discriminator. These two effects cause tracks to be lost in the reconstruction and it adds tails to the distributions that we cut on the analysis. This results in a slight deficit of events in the beam region when compared to the Monte Carlo simulation, and is illustrated in Fig. 2; the beam region shows a percent level loss in the data that is not predicted by the Monte Carlo. This loss is not seen for events that lie outside the beam regions.

We are currently attempting to improve our simulation using detailed maps of the various chamber pathologies. With no understanding of the chamber performance we would have to assign a systematic uncertainty which is comparable to the statistical uncertainty; therefore, even a modest improvement in the simulation may help reduce the uncertainty for our first result.

The critical systematic issues for $K_{L,S} \to \pi^0\pi^0$ are related to the CsI calibration at low energies and the background. The CsI calibration below 4 GeV is not as well determined because the 0.4 GeV/c magnet kick swept away low-energy electrons from $K_L \to \pi e\nu$ decays that were used to calibrate the CsI response. About 25% of our $K_{L,S} \to \pi^0\pi^0$ events have at least one photon with an energy between 2 and 4 GeV. We are using $K_L \to \pi^+\pi^-\pi^0$ and $K_L \to 3\pi^0$ events to improve our understanding of low energy photons in the CsI.

The other neutral mode issue is related to the $K_{L,S} \to \pi^0\pi^0$ background, which is at the 1% level, and therefore, must be understood to within several percent of itself. To achieve this we have done detailed simulations of regenerator scatters, $K_L \to 3\pi^0$ decays with two undetected photons, collimator scatters, and hadronic $2\pi^0$ production off the edge of the regenerator. The largest background to $K_{L,S} \to \pi^0\pi^0$ is from kaons that scatter in the regenerator, but whose center of energy still lies within the coherent (i.e., 0^0 scatter) beam profile. These small scatters cannot be identified in $2\pi^0$ decays, but they are easily detected in the charged $K^0 \to \pi^+\pi^-$ decays. The acceptance-corrected scatter distribution from the charged decays was parameterized as a function of its transverse momentum and decay time, and the result was fed into a Monte Carlo to simulate $2\pi^0$ decays from scattered kaons. This Monte Carlo background prediction for $2\pi^0$ decays is shown in Fig. 3 for the Y-view center of energy. The shape of the background distribution is based on the charged decays; the background scale comes from normalizing the data to the simulation for events in a region that has only background and no coherent signal (i.e., the side-bands in Fig. 3).

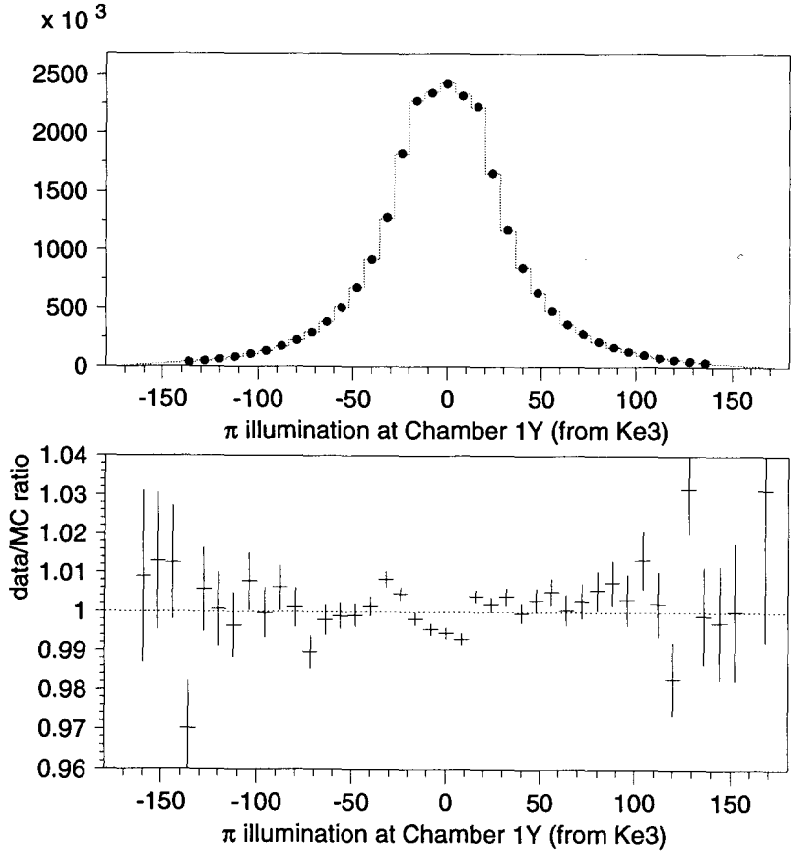


Fig. 2. The Y-view illumination at the first drift chamber is shown for pions from $K^0 \to \pi e\nu$. A cut requiring the beam region in the X-view has been applied. The dots are data, the histogram is Monte Carlo *without* detailed chamber simulations, and the ratio is shown in the lower plot. Each bin corresponds to half of a drift chamber cell (3.175 mm). The beam region occupies the central ± 16 half-cells.

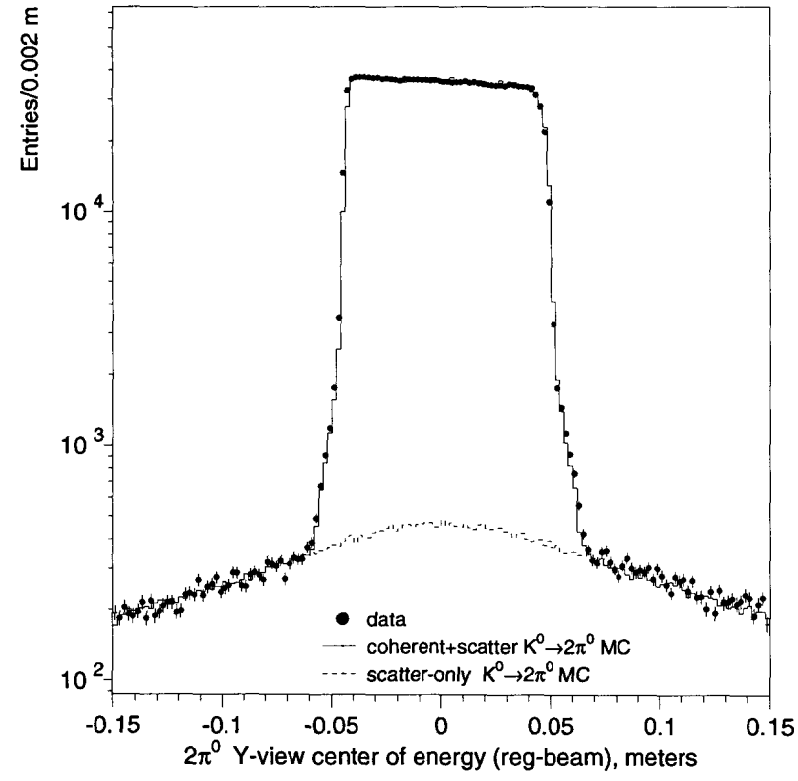


Fig. 3. *Preliminary* four-photon center of energy (meters) in the Y-view is shown for $2\pi^0$ decays in the regenerator beam. All other offline analysis cuts have been applied, including a cut on the X-view center of energy. The data are shown by the solid dots. The dashed histogram is from a background Monte Carlo simulation of kaon scatters in the regenerator. The histogram overlay is the Monte Carlo background (dashed) plus the coherent signal.

3.2 $K_L \to \pi^0\nu\bar{\nu}$

For the decay $K_L \to \pi^0\nu\bar{\nu}$, we have made two searches using different techniques. The first method used the decay $\pi^0 \to \gamma\gamma$ (Ref. 1), and the other method used the Dalitz decay $\pi^0 \to e^+e^-\gamma$ (Ref. 2). The advantage of the first method is that the branching ratio into two photons is 80 times larger and the acceptance is several times larger than for the Dalitz decay; the advantage of the second method is that the charged vertex information helps to reduce backgrounds. The $\pi^0 \to \gamma\gamma$ method resulted in one candidate which was consistent with the expected background, and gave a 90% confidence upper limit on the branching fraction of 1.6×10^{-6}. The $\pi^0 \to e^+e^-\gamma$ method resulted in no candidates and a 90% confidence upper limit of 5.9×10^{-7}. It should be noted that the $\pi^0 \to e^+e^-\gamma$ method required 50 times more data collection time and 50 times more K_L flux than that used for the $\pi^0 \to \gamma\gamma$ method, yet it only improved the upper limit by a factor of three. This suggests that the path to reaching the SM prediction of $B_{SM}(K_L \to \pi^0\nu\bar{\nu}) \sim 3 \times 10^{-11}$ is via the $\pi^0 \to \gamma\gamma$ mode. The $\pi^0 \to \gamma\gamma$ technique was limited in our experiment by background due to hadronic $2\pi^0$ production off the vacuum window, in which two photons escaped detection and the other two photons reconstructed as a π^0 inside the decay volume.

I would like to finish this section with a few comments about future prospects for $K_L \to \pi^0\nu\bar{\nu}$ searches using the $\pi^0 \to \gamma\gamma$ decay. A dedicated $K_L \to \pi^0\nu\bar{\nu}$ experiment will have to remove as much material as possible between the decay volume and the detector to reduce the hadronic $2\pi^0$ background from neutron interactions. Another challenge for the $\pi^0 \to \gamma\gamma$ method will be to sufficiently suppress backgrounds from the dominant neutral decay modes $K_L \to 2\pi^0$ and $K_L \to 3\pi^0$, which will require efficient photon detection down to tens of MeV. The final and perhaps most challenging task is to make detectors that operate in a high rate environment with good signal efficiency. In particular, the photon veto system must be very efficient in detecting low-energy photons from background sources, yet at the same time, it should not reject too much signal due to accidental losses.

The difficulty in the photon veto system is well illustrated by our own analysis with two very different kaon fluxes. The $K_L \to \pi^0\nu\bar{\nu}$ search using $\pi^0 \to \gamma\gamma$ used just one kaon beam with reduced size (0.04×0.04 m^2 at the CsI) to improve the transverse momentum measurement of the two detected photons. A beryllium absorber was placed far upstream to reduce the neutron/kaon ratio, and further reduced the kaon flux by a factor of three. The resulting kaon flux for this measurement was $\sim$ 0.3 MHz. The $K_L \to \pi^0\nu\bar{\nu}$ search using $\pi^0 \to e^+e^-\gamma$ used two large kaon beams (0.1×0.1 m^2 at the CsI) with no beryllium absorber and a slightly higher proton beam intensity, resulting in a 12 MHz kaon flux. The photon veto signal loss with a 0.3 MHz kaon flux was about 20%; had we used the same photon veto cuts in the $\pi^0 \to e^+e^-\gamma$ method at high intensity, the signal loss would have been 96%! Fortunately, the charged vertex constraint was very effective at reducing the background, which allowed the photon veto cuts to be relaxed in order to increase the signal acceptance. Most of the potential 96% signal loss at 12 MHz flux is from the photon veto in the beam hole behind the CsI (see Fig. 1). The need to increase the kaon flux by yet another large factor makes this problem even more severe. A faster detector signal from either Cherenkov media (if there is enough light for low-energy photons) or from signal shaping will be necessary for a $K_L \to \pi^0\nu\bar{\nu}$ search to reach the SM sensitivity. Reducing the electronic noise will also help improve the signal efficiency of the photon veto system.

3.3 $K_L \to \pi^0 e^+e^-$

Using our full data set from the rare decay configuration (see end of Sec. 2) our single event sensitivity for the decay $K_L \to \pi^0 e^+e^-$ is $\sim 10^{-10}$, which is a factor of 10 to 100 away from reaching the SM sensitivity. The background from the decay $K_L \to e^+e^-\gamma\gamma$ is also at the 10^{-10} level, which means that $K_L \to \pi^0 e^+e^-$ will be dominated by background unless it is enhanced by nonstandard model effects. The background from $K_L \to \pi e\nu$ with a misidentified pion and two accidental photons is at the 10^{-12} level; this small background is due to the excellent pion rejection of both the CsI electromagnetic calorimeter and the TRD system.

4 Indirect CP Violation in $K_L \to \pi^+\pi^-e^+e^-$

The decay $K_L \to \pi^+\pi^-\gamma$ has both a CPC direct emission M1 amplitude and a CP-violating inner brem (IB) amplitude. The interference between these two amplitudes results in a CP-violating polarization that can be measured if the photon converts to an e^+e^- pair. The decay $K_L \to \pi^+\pi^-e^+e^-$ internally converts the photon, and thus, provides another opportunity to observe indirect CPV.

Indirect CPV in $K_L \to \pi^+\pi^-e^+e^-$ is expected to cause an asymmetry in the angle between the $\pi^+\pi^-$ and e^+e^- decay planes.[3] Defining this angle to be ϕ, CP invariance requires that the number of events at ϕ is the same as the number of events at $-\phi$. The

asymmetry in the ϕ distribution is defined to be

$$A_\phi \equiv \frac{N(0 < \phi < \pi/2) - N(\pi/2 < \phi < \pi)}{N(0 < \phi < \pi/2) + N(\pi/2 < \phi < \pi)}, \quad (1)$$

where $N(\pi/2 < \phi < \pi) = N(-\pi/2 < \phi < 0)$ due to the symmetry $dN/d\phi(\phi) = dN/d\phi(\phi + \pi)$.

The predicted asymmetry and decay amplitudes are shown in Fig. 4 as a function of the e^+e^- energy (i.e., the virtual photon energy) in the kaon center-of-mass frame. The IB amplitude dominates at small virtual photon energies, while the M1 amplitude dominates at higher virtual photon energies. The asymmetry results from the interference between these two amplitudes, and is maximal when the two amplitudes have approximately the same magnitude. The asymmetry integrated over all phase space is predicted to be 14% so the effect should be large. The predicted ϕ distribution[3] is

$$\frac{dN}{d\phi} = \Gamma_1 \cos^2(\phi) + \Gamma_2 \sin^2(\phi) + \Gamma_3 \sin(2\phi), \quad (2)$$

and is shown in Fig. 5 without the inner brem term because our experiment has very little sensitivity for IB events, which tend to have e^+e^- tracks that are close together. The pure M1 term (dashed) in Fig. 5 is symmetric about $\pi/2$, and thus, has no net asymmetry. The interference term is clearly asymmetric about $\pi/2$, and thus, has a nonzero value of A_ϕ. The sum of the two terms is shown by the thick solid curve, and qualitatively indicates the CP-violating signal that we are searching for experimentally.

Now that the expected CP-violating behavior for $K_L \to \pi^+\pi^-e^+e^-$ has been outlined, we can look for this effect in our data. For this search, we used the configuration for rare decays as explained at the end of Sec. 2.

The first step in the analysis is to cleanly identify the $K_L \to \pi^+\pi^-e^+e^-$ decay mode, which has previously not been observed. Based on a small sample of 36 events, a branching fraction of $[3.2 \pm 0.6(stat.) \pm 0.4(syst.)] \times 10^{-7}$ has already been reported by our collaboration.[4] A preliminary $\pi^+\pi^-e^+e^-$ invariant mass spectrum is shown in Fig. 6(a) for our full data set after all cuts, and contains 1766 $K_L \to \pi^+\pi^-e^+e^-$ events on a background of 45 events. The small background is due mainly to $K_L \to \pi^+\pi^-\pi^0_D$, where π^0_D indicates the Dalitz decay $\pi^0 \to e^+e^-\gamma$.

The ϕ distribution without acceptance corrections is shown in Fig. 6(b) for the full data sample, and is compared with two Monte Carlo samples. The histogram is from a Monte Carlo using the SM matrix element that predicts an angular asymmetry. The dashed curve is from a Monte Carlo using the standard matrix element except that CPV

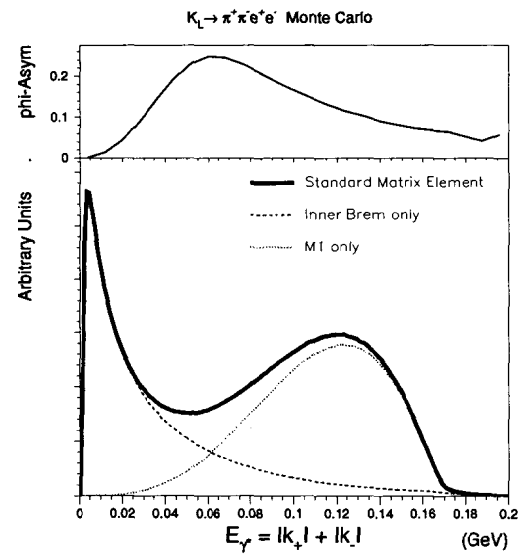


Fig. 4. The predicted angular asymmetry is shown as a function of the virtual photon energy in the kaon rest frame (top). The bottom plot shows the contributions to the matrix element from inner brem and direct emission amplitudes.

has been turned off by setting the relative phase between the M1 and IB amplitudes to be 90^0. It is clear that our data favor the model with CPV. At the time of this conference a result was not available, but since then a preliminary result has been announced[5] and is $A_\phi = 13.5 \pm 2.5(stat.) \pm 3.0(syst.)\%$.

5 $K_L \to \pi^0\gamma\gamma$

The decay $K_L \to \pi^0\gamma\gamma$ is of interest because it can be used to estimate the CPC contribution to $K_L \to \pi^0 e^+e^-$. The CPC two-photon intermediate state $K_L \to \pi^0\gamma\gamma \to K_L \to \pi^0 e^+e^-$ is very small if only the S-wave $\gamma\gamma$ state is considered using p^4 calculations in Chiral Perturbation Theory (ChPT). This lead to the idea in the late 1980s that $K_L \to \pi^0 e^+e^-$ might be dominated by direct CPV if the SM is correct.[6,7] However, the first measurements of $B(K_L \to \pi^0\gamma\gamma)$ (Refs. 8 and 9) were about three times larger than the prediction for this mode, which cast doubt on the validity of the CPC predictions for $K_L \to \pi^0 e^+e^-$.

Shortly after the first measurements of $B(K_L \to \pi^0\gamma\gamma)$, ChPT calculations including p^6 corrections and vector meson exchange predicted a branching fraction consistent with experiment. However, these new calculations also predicted a much larger CPC contribution to $K_L \to \pi^0 e^+e^-$, such that the CPC and CPV contributions were comparable. The increase in the CPC amplitude arises from the spin 2 (D-wave) component of the vector meson and p^6 corrections, and it is not helicity suppressed as was for the case of the p^4 (spin 0) predictions. Another new prediction is the presence of a low-side tail in the $m_{\gamma\gamma}$ distribution, where the $\gamma\gamma$ subscript refers to the two photons not associated with the π^0. The previous experiments did not have enough sensitivity to see this tail. Our new measurement in KTeV has sufficient statistics to observe the low-side $m_{\gamma\gamma}$ tail and to compare with ChPT. The data presented here were collected simultaneously with the $K_{L,S} \to \pi^0\pi^0$ events used in the measurement of $\mathrm{Re}(\epsilon'/\epsilon)$.

The detection of $K_L \to \pi^0\gamma\gamma$ is difficult because there are not enough kinematic constraints to make a traditional invariant mass plot that clearly identifies the signal. We therefore rely heavily on a Monte Carlo simulation to predict the background levels. The background from $K_{L,S} \to \pi^0\pi^0$ is removed by discarding events that have an $m_{\gamma\gamma}$ value within a 0.06 GeV window around the π^0 mass. The dominant background is from $K_L \to 3\pi^0$, in which two photons "fuse" with other photons and appear as a four-photon event. The strategy for reducing this background is to use the transverse cluster shapes at the CsI calorimeter. Using simulated photon showers as a template,

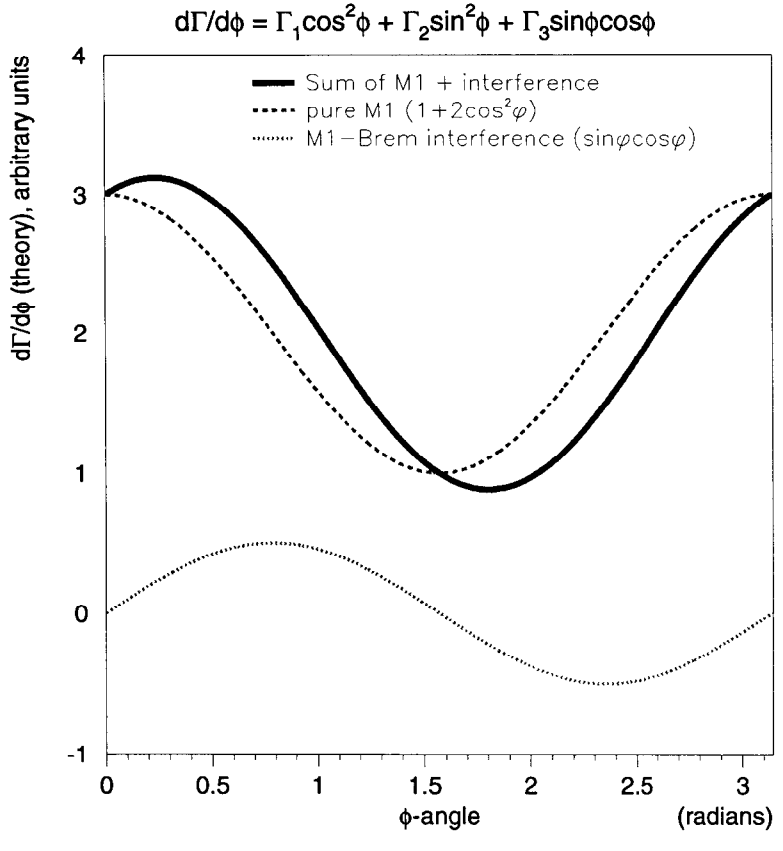


Fig. 5. The predicted ϕ distribution is shown for M1 direct emission (dashed), M1-brem interference (dots), and the sum of M1 and interference (solid).

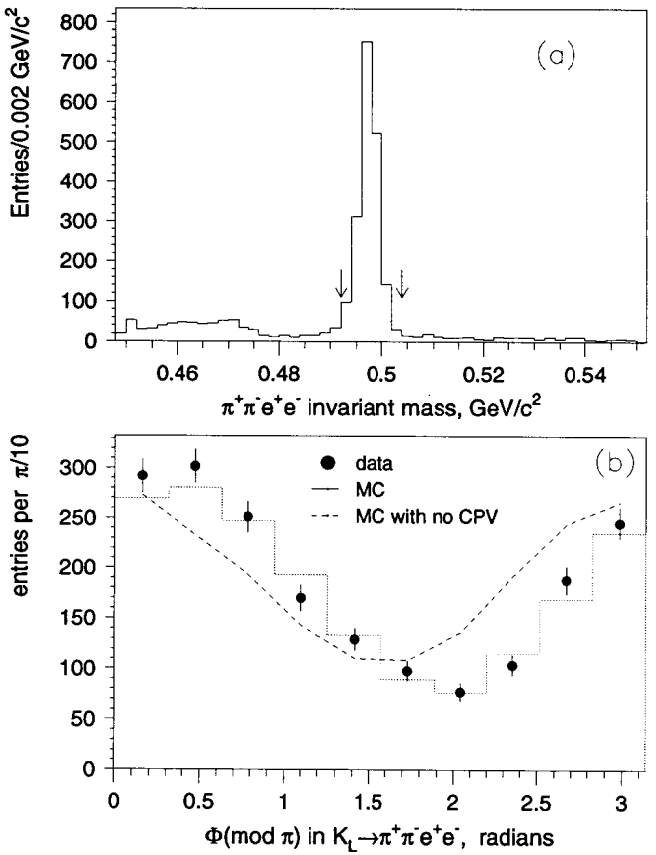


Fig. 6. Our preliminary $\pi^+\pi^-e^+e^-$ invariant mass spectrum is shown in (a) after all the analysis cuts. The arrows show the mass window (0.492 to 0.504 GeV/c^2) used to select $K_L \to \pi^+\pi^-e^+e^-$ events for the angular distribution shown in (b). Events at ϕ and $\phi + \pi$ are plotted in the same bin. The dots are for data, the histogram is the Monte Carlo using the SM matrix element, and the dashed curve is a Monte Carlo with no CPV in the matrix element.

we have computed a "shape" χ^2 for each cluster using the central 3×3 crystal energies. The maximum shape χ^2 of the four clusters is shown in Fig. 7 for each $K_L \to \pi^0\gamma\gamma$ candidate. A peak is clearly seen at small values of the shape χ^2, and we interpret this peak as the $K_L \to \pi^0\gamma\gamma$ signal. The histogram in Fig. 7 shows the Monte Carlo prediction of the $K_L \to 3\pi^0$ background, which is absolutely normalized to the number of $K_L \to 2\pi^0$ decays. After applying a cut of shape $\chi^2 < 2.0$, a preliminary $m_{\gamma\gamma}$ spectrum is shown in Fig. 8. There are 796 signal events on a background of 111 ± 14. This is a factor of 12 larger than the statistics used in the previous best measurement. Normalizing this sample to $K_L \to 2\pi^0$ gives a preliminary branching fraction of

$$B(K_L \to \pi^0\gamma\gamma) = 1.76 \pm 0.06(stat.) \pm 0.09(syst.) \times 10^{-6}, \quad (3)$$

where about half of the systematic uncertainty is related to the determination of the $K_L \to 3\pi^0$ background, and the remaining uncertainty is from the $K_L \to 2\pi^0$ branching ratio (2.1%) and the offline analysis cuts (3.3%).

The number of events with $m_{\gamma\gamma} < 0.240$ GeV/c^2 is $83 \pm 14 \pm 9$, and it is the first observation of the low-mass events for this decay mode. This corresponds to about 14% of the total decay rate after correcting for the unobserved events near the π^0 mass. The shape χ^2 distribution for these low-mass events is shown in the the inset of Fig. 7, and clearly shows a signal peak. The next step is to use the data to extract the vector meson coupling, $a_V = -0.8 \pm 0.1$. This coupling is then used to calculate the CPC part of $K_L \to \pi^0e^+e^-$ and results in

$$B_{CPC}(K_L \to \pi^0e^+e^-) = 1 - 2 \times 10^{-12}, \quad (4)$$

which is a few times smaller than the expected rate from direct CPV.

There is also an indirect CPV contribution from the small CP-even part of the K_L, which can be determined from a measurement of $B(K_S \to \pi^0e^+e^-)$. The current 90% upper limit is $B(K_S \to \pi^0e^+e^-) < 1.1 \times 10^{-6}$ from NA31 at CERN.[10] During the collection of $K^0 \to \pi\pi$ events for the Re(ϵ'/ϵ) measurement we saved events from the regenerator beam that had the topology of a $K_S \to \pi^0e^+e^-$ decay. Preliminary analysis indicates that we have a sensitivity of $\sim 10^{-7}$; if there is no background, this translates into a sensitivity of a *few* $\times 10^{-10}$ on the indirect CPV contribution to $K_L \to \pi^0e^+e^-$.

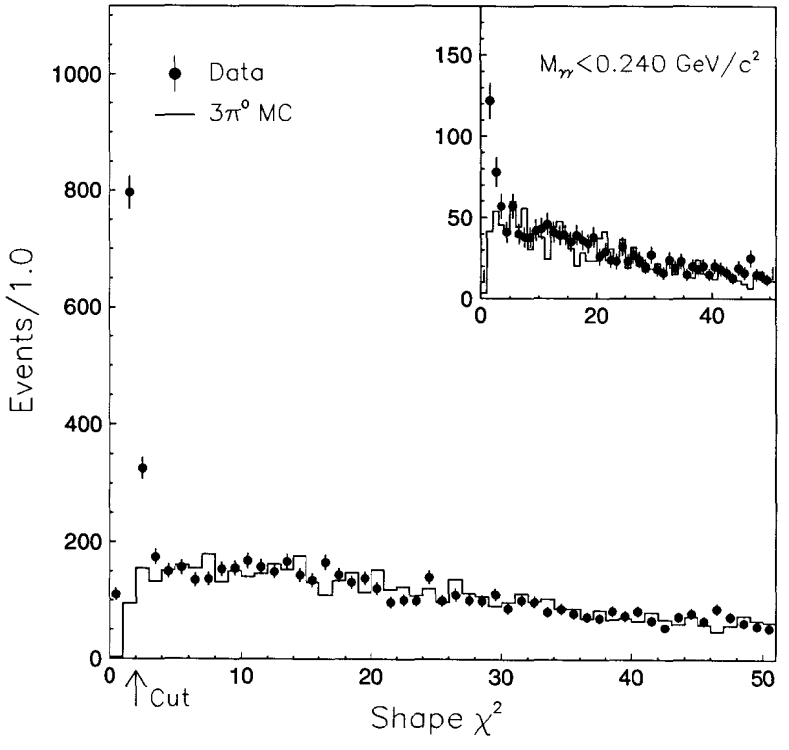


Fig. 7. The maximum transverse shape χ^2 of the four photons for each $K_L \to \pi^0\gamma\gamma$ candidate after all offline software cuts except the χ^2 cut (*preliminary*). The inset shows the same distribution for events with $m_{\gamma\gamma} < 0.240$ GeV/c^2. The data are shown by dots; the histogram is from a Monte Carlo simulation of $K_L \to 3\pi^0$ events.

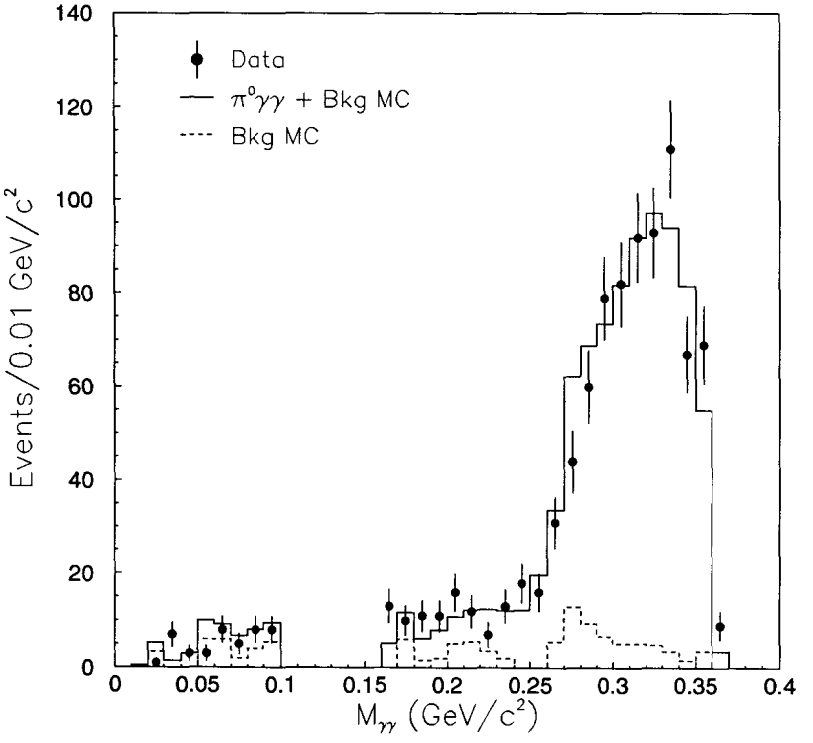


Fig. 8. *Preliminary* $m_{\gamma\gamma}$ mass spectrum for $K_L \to \pi^0\gamma\gamma$ candidates. These two photons are not associated with the π^0.

6 Conclusions

To summarize, the impact and status of KTeV on CP-violating phenomenon in the kaon system is shown in Table 2. A gauge of progress is shown in the column labeled "additional sensitivity needed to reach the SM."

Physics		*Preliminary* KTeV result	Additional sensitivity to reach SM
$\mathrm{Re}(\epsilon'/\epsilon)$	D_{CPV}	not yet	< 10 ?
$K_L \to \pi^0\nu\bar{\nu}$ with $\pi^0 \to \gamma\gamma$	D_{CPV}	$< 1.6 \times 10^{-6}$	10^5
$K_L \to \pi^0\nu\bar{\nu}$ with $\pi^0 \to e^+e^-\gamma$	D_{CPV}	$< 5.9 \times 10^{-7}$	10^5
$K_L \to \pi^0 e^+e^-$	$D_{CPV} + I_{CPV}$	not yet	10^2
$K_L \to \pi^0\gamma\gamma \to \pi^0 e^+e^-$	CPC	$1 - 2 \times 10^{-12}$	—
CP-violating asymmetry in $K_L \to \pi^+\pi^-e^+e^-$	I_{CPV}	"evidence"	none

Table 2. Summary of CPV searches in the KTeV experiment. D_{CPV} and I_{CPV} refer to Direct and Indirect CPV.

For the $\mathrm{Re}(\epsilon'/\epsilon)$ measurement, we expect to soon release a result based on 25% of our data. The statistical precision will be 3×10^{-4}. The systematic uncertainty will depend on our ability to understand the beam region of the drift chambers, and on the CsI calibration for photons with energies below 4 GeV.

On the $K_L \to \pi^0\nu\bar{\nu}$ "golden" mode, we are a dismal five orders of magnitude away from reaching the SM sensitivity using the $\pi^0 \to \gamma\gamma$ technique, and seven orders of magnitude away from the ultimate goal of observing ~ 100 events in order to measure the CKM parameter η with several percent accuracy. Two orders of magnitude in sensitivity can be gained by running for 100 days instead of just one day, and another order of magnitude can be gained using currently available proton intensities. How to obtain the remaining four orders of magnitude is still not clear. The physics potential in this mode is fantastic but it is not evident that current technology will work. New ideas will most likely be a necessary ingredient for observing this mode at the SM level.

Although our sensitivity to $K_L \to \pi^0 e^+e^-$ is still a factor of 100 away from the SM, we have made progress on determining its CPC contribution. Our measurement of the $K_L \to \pi^0\gamma\gamma$ branching fraction and of its $m_{\gamma\gamma}$ spectrum can be combined with ChPT calculations to predict the CPC contribution to $K_L \to \pi^0 e^+e^-$. We find that $B_{CPC}(K_L \to \pi^0 e^+e^-) = 1 - 2 \times 10^{-12}$.

Finally, we have shown new evidence of indirect CPV in the angular distribution of the $K_L \to \pi^+\pi^-e^+e^-$ decay planes.[5] This is the fourth and largest observed manifestation of CPV since the initial discovery of $K_L \to \pi^+\pi^-$.

References

[1] T. Nakaya, presented at the *Flavor Changing Neutral Current Conference*, Santa Monica, CA, 1997; KTeV Collaboration, hep-ex/9806007 (1998).

[2] K. Hanagaki, "Search for the Decay $K_L \to \pi^0\nu\bar{\nu}$," Ph. D. thesis, Osaka University, 1998.

[3] L. M. Sehgal and M. Wanninger, "CP Violation in the Decay $K_L \to \pi^+\pi^-e^+e^-$," Phys. Rev. D **46**, 1035 (1992).

[4] J. Adams *et al.*, "Measurement of the Branching Fraction of the Decay $K_L \to \pi^+\pi^-e^+e^-$," Phys. Rev. Lett. **80**, 4123 (1998).

[5] A quantitative result on the $K_L \to \pi^+\pi^-e^+e^-$ angular asymmetry has recently been announced by our collaboration at the *HQ98 Conference* at Fermilab (October 1998).

[6] G. Ecker, A. Pich, and E. De Rafael, Phys. Lett. **189**, 363 (1987).

[7] G. Ecker, A. Pich, and E. De Rafael, Nucl. Phys. B **303**, 665 (1988).

[8] G. Barr *et al.*, Phys. Lett. B **284**, 440 (1992).

[9] V. Papadimitriou *et al.*, Phys. Rev. D **44**, 573 (1991).

[10] G. Barr *et al.*, Phys. Lett. B **304**, 381 (1993).

EVIDENCE AND PROSPECTS FOR THE RARE DECAY $K^+ \to \pi^+\nu\bar{\nu}$ AT BNL E787

George Redlinger*
TRIUMF, 4004 Wesbrook Mall
Vancouver, B.C., Canada V6T 2A3

Representing the E787 Collaboration

ABSTRACT

In 1997, E787 at Brookhaven National Laboratory (BNL) reported evidence for the rare decay $K^+ \to \pi^+\nu\bar{\nu}$ based on the observation of one event in the 1995 data set. The current status of this measurement is described.

*Supported by the Natural Sciences and Engineering Research Council of Canada.

© 1998 by George Redlinger.

1 Introduction

In the Standard Model (SM), quark mixing in weak decays and CP violation are related.[1] Detailed exploration of this relationship between two longstanding mysteries of particle physics is one of the major themes of experimental work in the field over the next decade(s). The rare decay $K^+ \to \pi^+\nu\bar{\nu}$ has drawn interest in this effort due to the theoretically clean relationship between the branching ratio and the poorly measured quark mixing parameter $|V_{td}|$. The intrinsic theoretical uncertainty in the branching ratio (arising mainly from QCD corrections to the charm contribution to the process) is estimated[2] to be about 7% for a given set of SM input parameters m_t, $|V_{cb}|$, and $|V_{td}|$. Stated differently, once the branching ratio for $K^+ \to \pi^+\nu\bar{\nu}$ is known, $|V_{td}|$ can be determined to $\sim 5\%$ (given perfect knowledge of m_t and $|V_{cb}|$). In the SM, the branching ratio is expected to be[3] $\mathrm{B}(K^+ \to \pi^+\nu\bar{\nu}) = (0.6 - 1.5) \times 10^{-10}$, using current data on m_t, m_c, V_{cb}, $|V_{ub}/V_{cb}|$, ϵ_K, and $\mathrm{B_d} - \overline{\mathrm{B_d}}$ mixing.

The E787 Collaboration at BNL presented evidence for $K^+ \to \pi^+\nu\bar{\nu}$ based on the observation of one clean event from data collected in the 1995 run of the Alternating Gradient Synchrotron (AGS).[4] The expected level of background in the signal region was 0.08 ± 0.03 events. However, the event also satisfied the most demanding criteria designed in advance for candidate evaluation; this put the event in a region where the expected background was 0.008 ± 0.005 events. If the event is due to $K^+ \to \pi^+\nu\bar{\nu}$, this would imply a branching ratio $B(K^+ \to \pi^+\nu\bar{\nu}) = 4.2^{+9.7}_{-3.5} \times 10^{-10}$, which is consistent with the SM range, although the central value is higher by a factor of about four.

Amongst the vast literature on non-SM physics with the signature $K^+ \to \pi^+$ "nothing," recent attention has focused on implications for electroweak symmetry breaking. R-parity conserving supersymmetry (SUSY) has been considered in many papers[5]; conclusions vary depending on assumptions, but a branching ratio two to three times above the SM level seems to be possible without going to exotic models. In R-parity violating SUSY,[6] $K^+ \to \pi^+\nu\bar{\nu}$ yields the best constraints on couplings relating the first two generations. Effects from top-color models[7] can also be significant.

The two-body decay $K^+ \to \pi^+X^0$, where X^0 is a long-lived noninteracting object, is indistinguishable experimentally from $K^+ \to \pi^+\nu\bar{\nu}$, except for the two-body kinematics. An interesting possibility is where X^0 is a Nambu-Goldstone boson from breaking some global symmetry, a well-known example of which is the familon[8] from the breaking of global family symmetry. The branching ratio is expected to be $B(K^+ \to \pi^+f) = 2.7 \times 10^{13}\ \mathrm{GeV}^2/F^2$, where F is the scale at which the symmetry is broken. Cosmological considerations lead to an upper bound[9] $F < 10^{12}$ GeV, implying $B(K^+ \to \pi^+f) \gtrsim 10^{-11}$. A new class of axion models has also recently been considered.[10]

In the late '60s, it was the absence of the process $K^+ \to \pi^+\nu\bar{\nu}$ (among others) at the then-expected rate which led to the "new" physics (GIM mechanism, charm, etc.) of the time and the eventual establishment of the SM. It would be poetic indeed if now a higher rate than expected were to be confirmed, leading the way beyond the SM.* The remainder of this paper is organized as follows. We first remind the reader of the $K^+ \to \pi^+\nu\bar{\nu}$ detection strategy employed in E787, and of the results from the 1995 data. We then discuss the 1996 and 1997 runs, followed by a brief report on the status of the ongoing 1998 run and future plans.

2 Detection Strategy

2.1 The E787 Detector

The signature for $K^+ \to \pi^+\nu\bar{\nu}$ is a K^+ decay to a π^+ of momentum $P < 227$ MeV/c and no other observable product. Definitive observation of this signal requires suppression of all backgrounds to well below the sensitivity for the signal. Furthermore, reliable estimates of the residual background levels are needed.

Major sources of background include the two-body decays $K^+ \to \mu^+\nu_\mu$ ($K_{\mu 2}$) with a 64% branching ratio and $P = 236$ MeV/c, and $K^+ \to \pi^+\pi^0$ ($K_{\pi 2}$) with a 21% branching ratio and $P = 205$ MeV/c. The charged particle spectrum of the major K^+ decay modes is shown in Fig. 1 together with the spectrum from $K^+ \to \pi^+\nu\bar{\nu}$. The $K^+ \to \pi^+\nu\bar{\nu}$ signal can be observed in the region away from these two kinematic peaks. The search described here concentrates on the region between the $K_{\pi 2}$ and $K_{\mu 2}$ peaks.

The only other important background sources are scattering of pions in the beam (either from kaon decay/interaction or pions from the K^+ production target) and K^+ charge exchange (CEX) reactions resulting in decays $K_L^0 \to \pi^+l^-\overline{\nu}$, where $l = e$ or μ.

The detector has been described in detail elsewhere[11]; here, we sketch the main ideas. To achieve the large kinematic suppression of the monochromatic peaks $K_{\pi 2}$ and $K_{\mu 2}$, we work in the kaon rest frame, slowing down a 790 MeV/c kaon beam in a BeO

*See the amusing exchange between B. Winstein and G. Kane in the Sept. 1998 issue of *Physics Today* on a fictitious look back from the year 2011 to the history of particle physics from the late 1990s onward.

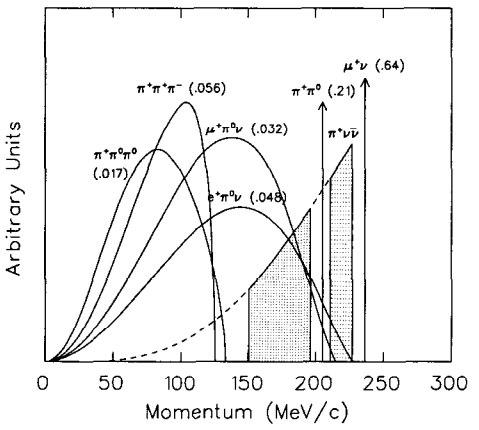


Fig. 1. Charged-particle momentum spectrum in the K^+ rest frame for the major K^+ decay modes and for $K^+ \to \pi^+\nu\bar{\nu}$.

degrader and stopping it in a finely segmented, fully active, scintillating-fiber target. This drives the geometry of the detector, which is cylindrical with endcaps covering the polar regions, much like a colliding-beam detector. Since vetoing of extra energy in the event is a key to the detection strategy, the detector is almost fully active.

The fiducial volume is defined by the Range Stack (RS) scintillators in the barrel region, covering about 50% of the solid angle. The pions from K^+ decay are stopped in the RS scintillator. This allows redundant measurements of track kinematics, namely kinetic energy and range. It also allows for a powerful technique for π/μ separation, which requires the positive identification of the $\pi^+ \to \mu^+ \to e^+$ decay sequence. This is achieved with flash-ADC-based 500-MHz transient digitizers (TD)[12] which digitize the phototube outputs. In addition, dE/dx separation of π from μ can be utilized. Further kinematic rejection is achieved with a momentum measurement in a 1 T solenoidal field, using a low-mass drift chamber[13] with a resolution $\Delta P/P \sim 0.9\%$ at $P = 205$ MeV/c. Efficient rejection of modes with photons is achieved with nearly 4π photon veto coverage, typically about 15 radiation lengths thick, employing a variety of technologies: lead-scintillator sandwich in the barrel region (BV), undoped CsI in the endcaps (EC),[14] lead-glass, lead-scintillator sandwich, and "Shashlik"-type detectors in the beam region. Signals are digitized with 500-MHz CCD transient digitizers[15] based on Ga-As technology. Photon veto time windows are typically a few ns, with energy thresholds ranging from 0.2 to about 3 MeV. For the incoming beam, K/π separation is achieved with a Cherenkov counter with a lucite radiator. Finely segmented tracking and good timing of the beam is achieved with MWPC's (1.27 mm wire spacing), scintillator hodoscopes and a scintillating fiber kaon stopping target (5 mm square fiber), the latter two also digitized by CCD's transient digitizers.

At typical rates, we took about 5 MHz of incoming kaons. These were analyzed by a two-level trigger. The first level provided a rejection factor of about 800 by requiring a kaon stopping in the target, a decay 1.5 ns later, no photons in the BV, EC, or RS, a range longer than $K^+ \to \pi^+\pi^+\pi^-$ and shorter than $K_{\mu 2}$. The second level required a $\pi^+ \to \mu^+$ decay in the counter, where the charged particle stopped and no energy was deposited in the counter radially outward from the stopping counter; this provided a rejection factor of about 20.

The DAQ[16] is Fastbus-based, with front-end readout into SLAC Scanner Processors (SSP). The data is transferred via the cable segment to VME processors and then to a Silicon Graphics (SGI) machine (model number improves with year). The data transfer capability is currently > 25 Mbytes/sec. The deadtime, dominated by the readout of the front-end modules into the SSP's, is currently about 17% per MHz of stopped kaons (down from 28% in 1995).

2.2 Offline Analysis

To elude rejection, $K_{\mu 2}$ and $K_{\pi 2}$ events have to be reconstructed incorrectly in range, energy, and momentum. In addition, any event with a muon has to have its track misidentified as a pion. The most effective weapon here is the TD analysis, requiring observation of the $\pi^+ \to \mu^+ \to e^+$ decay sequence; this provides a muon rejection factor of about 10^5. Events with photons, such as $K_{\pi 2}$ decays, are efficiently eliminated by the photon veto; the rejection factor for events with π^0's is around 10^6. A scattered beam pion can survive the analysis only by misidentification as a K^+ and if the track is mismeasured as delayed, or if the track is missed entirely by the beam counters after a valid K^+ stopped in the target. CEX background events can survive only if the K_L^0 is produced at low enough energy to remain in the target for at least 2 ns, if there is no visible gap between the beam track and the observed π^+ track, and if the additional charged lepton goes unobserved.

The data are analyzed with the goal of reducing the total expected background to significantly less than one event in the final sample. The same data set is used for background studies and for the signal search so that any hardware failures, or time-dependent effects, are naturally accounted for in the background estimates. The offline analysis is performed "blind," in the sense that the signal region is always hidden (by inverting one or more cuts) while cuts are developed and background levels estimated. Without this procedure, it would be difficult to avoid the systematic error, whereby a signal event might be eliminated with minimal acceptance loss by a slight tweaking of one of the more than 50 cuts in the analysis.

To develop the cuts, we take advantage of redundant, independent constraints available on each source of background to establish two independent sets of cuts. One set of cuts is relaxed or inverted to enhance the background (by up to three orders of magnitude) so that the other group of cuts can be evaluated to determine its power for rejection.

For example, $K_{\pi 2}$ is studied by measuring the rejection of the kinematic cuts on a sample of events failing the photon veto, while the photon veto rejection is measured on a sample of events which are kinematically consistent with $K_{\pi 2}$. This is illustrated in Fig. 2. Figure 2(a) shows the range versus energy for events tagged by the presence of photon activity in the detector after removing $K_{\mu 2}$ and beam backgrounds. Figure 2(b) shows the momentum distribution for events kinematically consistent with $K_{\pi 2}$, again after removing $K_{\mu 2}$ and beam backgrounds. The three histograms show the momentum spectra at different levels of photon vetoing: raw spectrum, after the trigger and first pass of the offline analysis, and finally after the final photon veto cuts. To determine the background in the final signal region, the number of events in the signal region in Fig. 2(a) (i.e., two events) is corrected for the tagging efficiency and then divided by the photon veto rejection from Fig. 2(b).

$K_{\mu 2}$ (including $K^+ \to \mu^+ \nu_\mu \gamma$) is studied by separately measuring the rejections of the TD particle identification cuts and the kinematic cuts. The background from beam pion scattering is evaluated by separately measuring the rejections of the beam counter and timing cuts. For the CEX background, events with K^+ charge exchange in the stopping target were collected with a special trigger, triggering on the two pion decay of the K_S^0; these events are then used as input to Monte Carlo studies, replacing the K_S^0 with a Monte Carlo K_L^0.

Small correlations in the separate groups of cuts are investigated for each background source and corrected for, if they exist. An example of the absence of correlation

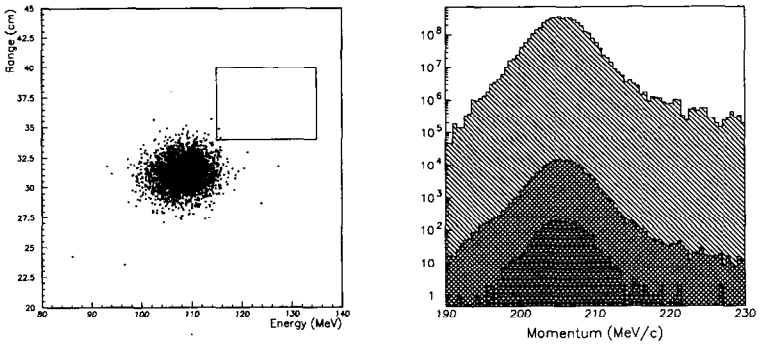


Fig. 2. $K_{\pi 2}$ analysis with the 1995 data. (a) shows the range vs energy for events tagged by the presence of photon activity in the detector after removing $K_{\mu 2}$ and beam backgrounds. (b) shows the momentum distribution for events kinematically consistent with $K_{\pi 2}$ after the removal of $K_{\mu 2}$ and beam backgrounds. The three histograms show the momentum spectra at different levels of photon vetoing: raw spectrum, after the trigger and first pass of the offline analysis, and finally, after the final photon veto cuts.

in the $K_{\pi 2}$ study can be seen in Fig. 2 by noting that in the right plot [Fig. 2(a)], the $K_{\pi 2}$ lineshape does not change as tighter photon veto cuts are applied.

Before looking in the final signal region, likelihood functions for each background type are constructed, using the method described above. For example, the $K_{\pi 2}$ kinematic likelihood function is measured with a sample of events failing the photon veto, while the $K_{\pi 2}$ photon veto likelihood is measured with a sample of events kinematically consistent with $K_{\pi 2}$. These likelihood functions are used to predict the shape of the background distributions outside the final signal region. The prediction is not guaranteed to be satisfied because the final measurement involves the simultaneous application of all likelihood functions; the success of the predictions, therefore, tests the independence of the individual likelihood functions. Finally, the background likelihood functions are used to assess the background likelihood of any candidate events.

3 1995 Results

The results from the 1995 analysis have been published.[4] The key points are summarized here.

The background levels in the 1995 data were estimated with the above procedure to be $b_{K_{\mu 2}} = 0.02 \pm 0.02$, $b_{K_{\pi 2}} = 0.03 \pm 0.02$, $b_{beam} = 0.02 \pm 0.01$, and $b_{CEX} = 0.01 \pm 0.01$ for a total background of $b = 0.08 \pm 0.03$ events.

As a cross-check of the background estimates and of the shapes of the background distributions near the signal region, the background likelihood functions were used to estimate the number of events expected to appear when the cuts were relaxed in predetermined ways, so as to allow orders of magnitude higher levels of all background types. At approximately the $20 \times b$ level, we observed two events where 1.6 ± 0.6 were expected. At the level $150 \times b$, we found 15 events where 12 ± 5 were expected. The events which were admitted by the relaxed cuts were consistent with being due to the known background sources.

Figure 3(a) shows the range versus energy for the events surviving all other analysis cuts. Only events with measured momentum in the accepted region $211 < P < 230$ MeV/c are plotted. The rectangular box indicates the signal region specified as range $34 < R < 40$ cm of scintillator and energy $115 < E < 135$ MeV. One event was observed in the signal region. The residual events below the signal region clustered at $E = 108$ MeV were due to $K_{\pi 2}$ decays where both photons had been missed. The number of these events was consistent with estimates of the photon detection inefficiency.

The event satisfied the tightest background likelihood criteria designed in advance for candidate evaluation. This put the event in a region where $b' = 0.008 \pm 0.005$ events would be expected from known background sources while 55% of the final acceptance for $K^+ \to \pi^+ \nu\bar{\nu}$ would be retained.

The acceptance for $K^+ \to \pi^+ \nu\bar{\nu}$ was $A = 0.0016 \pm 0.0001(stat.) \pm 0.0002(sys.)$, determined mainly from calibration data taken simultaneously with the physics data. Monte Carlo was used to obtain the acceptance factors only for solid angle, $K^+ \to \pi^+ \nu\bar{\nu}$ phase space, and losses from π^+ nuclear interactions and decays in flight. Figure 3(b) shows the simulated R versus E distribution for $K^+ \to \pi^+ \nu\bar{\nu}$ with the final analysis cuts applied. The acceptance was cross-checked by a parallel measurement of the $K_{\pi 2}$ branching ratio, comparing it to the world-average value. The total exposure of the 1995 run was $N_{K^+} = 1.49 \times 10^{12}$ kaons entering the stopping target. The resulting branching ratio is $B(K^+ \to \pi^+ \nu\bar{\nu}) = 4.2^{+9.7}_{-3.5} \times 10^{-10}$, assuming that the one event observed is from $K^+ \to \pi^+ \nu\bar{\nu}$.

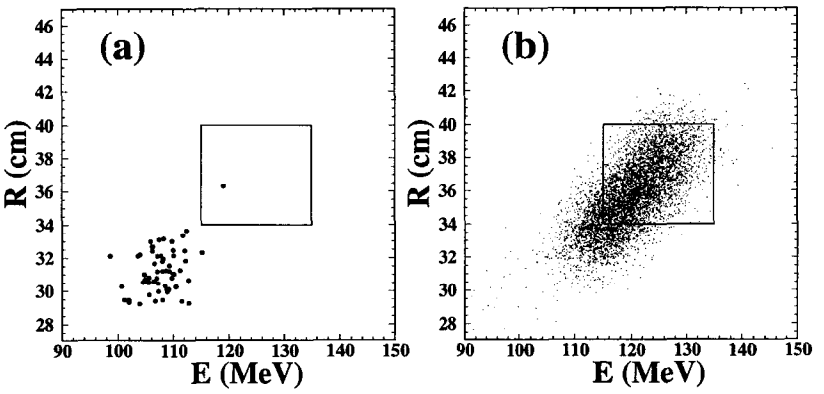


Fig. 3. (a) Range R vs energy E distribution for the $K^+ \to \pi^+ \nu\bar{\nu}$ data set (1995) with the final cuts applied. The box enclosing the signal region contains a single candidate event. (b) The Monte Carlo simulation of $K^+ \to \pi^+ \nu\bar{\nu}$ with the same cuts applied.

The candidate event was kinematically inconsistent with a massless X^0. The χ^2 C.L. for consistency with the hypothesis of a massless X^0 was 0.8%. The acceptance was $A_{K^+ \to \pi^+ X^0} = 0.0052 \pm 0.0003(stat.) \pm 0.0007(sys.)$. Based on no events observed in the region $221 < P < 230$ MeV/c, we set a 90% C.L. upper limit $B(K^+ \to \pi^+ X^0) < 3.0 \times 10^{-10}$ for a massless X^0.

4 Post-1995 Developments

4.1 Online Improvements

To increase the sensitivity per hour, the basic strategy in the post-1995 running of E787 has been to run with a lower K^+ momentum and remove BeO degrader material. This decreases the probability that the incoming kaon is lost in the BeO degrader (from multiple Coulomb scattering and interactions), reducing the accidental rate in the detector and increasing the fraction of kaons reaching the stopping target. Simultaneously, the proton flux on the kaon production target was increased to maintain the overall kaon flux. In addition, improvements were made to the trigger, reducing the deadtime frac-

tion from 28% to 17% per MHz of stopped kaons. The efficiency of the second-level trigger ($\pi^+ \to \mu^+$) was increased by a factor of 1.27 over 1995. A further gain of a factor of 1.05 was achieved by extending the time range recorded by the TD's to look for the electron from $\mu^+ \to e^+$ decay.

In the ongoing 1998 run, the AGS duty factor was improved by a factor of 1.1 by going from a 1.6 sec spill every 3.6 sec to a 2.2 sec spill every 4.5 sec. The number of protons per spill was increased to maintain the kaon flux. In addition, a better operating point was found by careful optimization in the space of K momentum, spill length and protons on target. The result of these factors is summarized in Fig. 4, which shows the increase in sensitivity/hour from year to year. The sensitivity/hour for the ongoing run is about a factor of two better than the 1995 running conditions. The sensitivity/hour is limited by proton intensity, currently about 30×10^{12} protons per spill (30 Tp) onto our kaon production target. Intensities which would increase the sensitivity/hour by a factor of three times better than the 1995 value have been demonstrated to be within the current capabilities of the AGS.

The total exposure in the 1996 and 1997 runs amounted to about 1.5 times the 1995 exposure. The increase in sensitivity/hour was offset by a decrease in running time: 17 weeks in 1996 followed by nine weeks in 1997, compared to 22 weeks in 1995. The current run, however, is expected to last 27 weeks and to result in an exposure five times that of the 1995 run. This would imply a total exposure in 1995-1998 which is 7.5 times that of the 1995 run alone, or a single-event sensitivity of 0.7×10^{-10}.

4.2 Offline Analysis

The goal of the offline analysis is to increase the rejection to maintain the signal-to-noise ratio as the overall sensitivity grows, while maintaining (or possibly even increasing) the acceptance at the same time.

Better range resolution in both the Gaussian core and in the tails was achieved with tracking improvements in the range stack and target. The range tail for $K_{\pi 2}$ events is shown in Fig. 5; a significant reduction in the background level is apparent.

For the $K_{\mu 2}$ background, a more sophisticated dE/dx analysis was developed for the range stack, reducing the number of $K_{\mu 2}$ events downshifting in range and energy due to nuclear interactions. Tracking improvements in the drift chamber and better tracking quality cuts reduced the momentum tail. On the TD side, better understanding of the different sources of TD background in the $K_{\mu 2}$ peak compared to $K^+ \to \mu^+ \nu_\mu \gamma$ and

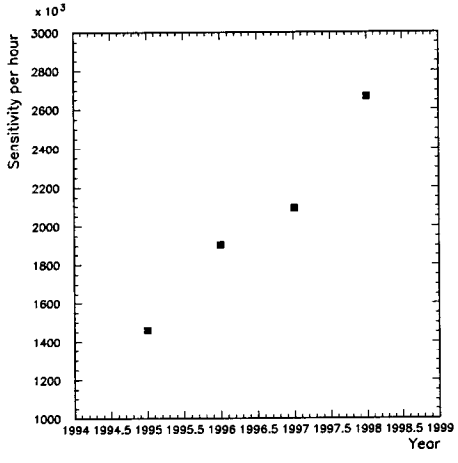


Fig. 4. Sensitivity per hour as a function of year.

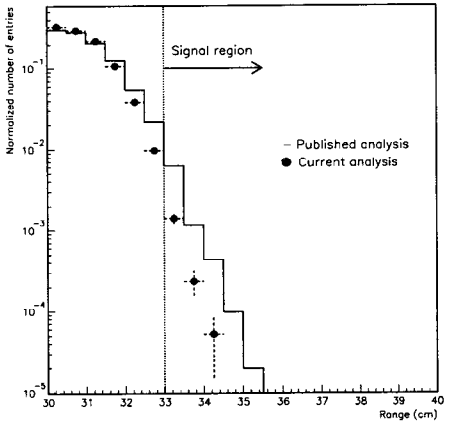


Fig. 5. $K_{\pi 2}$ range tail in the current analysis, compared to the published analysis.

$K^+ \to \mu^+\pi^0\nu_\mu$, and a better algorithm for finding the electron from $\mu^+ \to e^+$ decay resulted in a 17% gain in acceptance for the same level of rejection as the published analysis. The beam background was lowered by improved cuts against kaons decaying in flight in the stopping target, improved use of the CCD information in the target, and better cuts against two kaons entering the detector closely spaced in time. The CEX background was lowered by the above-mentioned improvements in the target tracking and by a likelihood analysis utilizing information in the target and beam hodoscopes.

In total, a preliminary assessment indicates that the overall background levels can be reduced by a factor of three with no loss in acceptance. This would imply a background level, stated as a branching ratio, of 1.1×10^{-11}.

5 Conclusion

The E787 Collaboration has published evidence for the decay $K^+ \to \pi^+\nu\bar{\nu}$ based on one clean event seen in the 1995 data sample. Analysis is proceeding on the 1996-1997 data set, which has an exposure 1.5 times larger than the 1995 data set alone.

In the ongoing 1998 run, extrapolating the current running conditions, the exposure is expected to reach five times that of 1995, for a total sensitivity (1995-1998) reaching 0.7×10^{-10}. Preliminary indications are that background levels can be held down to the level of 1.1×10^{-11} with no loss in acceptance compared to the 1995 run.

A new experiment, based largely on E787 with only modest upgrades and further improvement in the duty factor (and proton intensity), aims to reach a sensitivity of $(0.07 - 0.13) \times 10^{-12}$ with two years of AGS running in the RHIC era.[17] This proposal (E949) was recently approved by the BNL PAC.

References

[1] M. Kobayashi and T. Maskawa, Prog. Theor. Phys. **49**, 652 (1973).

[2] A. J. Buras and R. Fleischer, in *Heavy Flavors II*, edited by A. J. Buras and M. Lindner (World Scientific, 1997). Also hep-ph/9704376.

[3] G. Buchalla *et al.*, Rev. Mod. Phys. **68**, 1125 (1996).

[4] S. Adler *et al.*, Phys. Rev. Lett. **79**, 2204 (1997).

[5] Among many examples, Y. Nir and M. Worah, Phys. Lett. B **423**, 319 (1998); A. J. Buras, A. Romanino, and L. Silvestrini, Nucl. Phys. B **520**, 3 (1998); G. Colangelo and G. Isidori, J. High Energy Phys. **9809**, 9 (1998).

[6] K. Agashe and M. Graesser, Phys. Rev. D **54**, 4445 (1996).

[7] G. Buchalla *et al.*, Phys. Rev. D **53**, 5185 (1996).

[8] F. Wilczek, Phys. Rev. Lett. **49**, 1529 (1982).

[9] J. Preskill, M. B. Wise and F. Wilczek, Phys. Lett. B **120**, 127 (1983); L. F. Abbott and P. Sikivie, *ibid*, 133 (1983); M. Dine and W. Fischler, *ibid*, 137 (1983).

[10] M. Hindmarsh and P. Moulatsiotis, hep-ph/9807363.

[11] M. S. Atiya *et al.*, Nucl. Instrum. Methods A **321**, 129 (1992) describes the original E787 detector. References 12-16 describe most of the upgrades for the current detector.

[12] M. S. Atiya *et al.*, Nucl. Instrum. Methods A **279**, 180 (1989).

[13] E. W. Blackmore *et al.*, Nucl. Instrum. Methods A **404**, 295 (1998).

[14] I.-H. Chiang *et al.*, IEEE Trans. Nucl. Sci. **NS-42**, 394 (1995); T. K. Komatsubara *et al.*, Nucl. Instrum. Methods A **404**, 315 (1998); M. Kobayashi *et al.*, Nucl. Instrum. Methods A **337**, 355 (1994).

[15] D. A. Bryman *et al.*, Nucl. Instrum. Methods A **396**, 394 (1997).

[16] M. Burke *et al.*, IEEE Trans. Nucl. Sci. **NS-41**, 131 (1994); C. Witzig and S. Adler, Real-Time Comput. Appl. **123** (1993); S. Adler, Int. Conf. Electr. Part. Phys. **133** (1997); C. Zein *et al.*, Real-Time Comput. Appl. **103** (1993).

[17] See "http://www.phy.bnl.gov/e949/e949.html".

T VIOLATION IN $K^+ \to \pi^0 \mu^+ \nu_\mu$ AT KEK

Jun Imazato
Institute of Particle and Nuclear Studies
High Energy Accelerator Research Organization
Tsukuba-shi, Ibaraki-ken, 305-0801 Japan

Representing the E246 Collaboration[1]

ABSTRACT

The transverse muon polarization in $K^+ \to \pi^0 \mu^+ \nu_\mu$ is a measure of T violation and a good probe of CP violation from physics beyond the Standard Model. The KEK E246 experiment is searching for the transverse polarization using a new technique with stopped kaons in conjunction with a Superconducting Toroidal Spectrometer. With such a high performance detector, a measurement with low backgrounds and small systematic errors can be realized. Data taking began in 1996 and analysis has now been completed for the data obtained until the end of 1997. There is no sign of T violation found within the experimental errors. The sources of systematic errors have been carefully studied with the conclusion that the total systematic error is far below the current statistical error.

© 1998 by Jun Imazato.

1 Transverse Muon Polarization in $K^+ \to \pi^0 \mu^+ \nu_\mu$

$K^+ \to \pi^0 \mu^+ \nu_\mu$ decay ($K_{\mu 3}$) provides a unique possiblity to test the violation of time reversal invariance (T violation) in kaon decay through the transverse muon polarization (P_T), which is the polarization component normal to the decay plane. P_T is expressed in terms of a triple vector correlation of the muon spin $\vec{\sigma}_\mu$ and the momentum vectors $\vec{p}_\pi$, $\vec{p}_\mu$, of the pion and muon as

$$P_T = \frac{\vec{\sigma}_\mu \cdot (\vec{p}_\pi \times \vec{p}_\mu)}{|\vec{p}_\pi \times \vec{p}_\mu|}, \tag{1}$$

which is T-odd; thus nonvanishing P_T would be a clear sign of T violation.[2] The spurious effect from final state interactions is known to be as small as 10^{-6} in this charged kaon decay case[3] because there is only one charged particle in the final state. Under the assumption of CPT invariance, the search for T violation can address the origin of CP violation. Since the contribution from the Standard Model is only from higher orders which are negligibly small $\sim O(10^{-6})$, P_T can be a sensitive probe of additional Non-Standard Model sources of CP violation. Various theoretical models, such as the multi-Higgs model and the leptoquark model, can produce a sizable effect in P_T at the level of 10^{-3}, which is experimentally detectable, without conflicting with the present constraints from other experimental channels.[4,5] More recently, a class of supersymmetric models,[6,7] such as the R-parity violation model, have also been examined and shown to have a potentially large contribution to P_T. From these theoretical points of view, a P_T search with a sensitivity of 10^{-3} is of great interest.

The $K_{\mu 3}$ process is described in the V-A scheme in terms of a decay amplitude with two form factors $f^+(q^2)$ and $f^-(q^2)$ as

$$M \propto \frac{G_F}{2} \sin\theta_C [f^+(q^2)(\tilde{p}_K^\lambda + \tilde{p}_\pi^\lambda) + f^-(q^2)(\tilde{p}_K^\lambda - \tilde{p}_\pi^\lambda)] \cdot [\bar{u}_\mu \gamma_\lambda (1-\gamma_5) u_\nu], \tag{2}$$

with four momenta $\tilde{p}_K$ and $\tilde{p}_\pi$ of the kaon and pion, respectively. A complex phase between $f^+(q^2)$ and $f^-(q^2)$ induces T violation. Using the definition, $\xi \equiv f^-/f^+$, P_T can be written as

$$P_T = Im\xi \cdot \frac{m_\mu}{m_K} \cdot \frac{|\vec{p_\mu}|}{E_\mu + |\vec{p_\mu}| \vec{n_\mu} \cdot \vec{n_\nu} - m_\mu^2/m_K}. \tag{3}$$

Thus, $Im\xi$ is the measure of T violation in $K_{\mu 3}$ decay independent of decay kinematics. In the three-Higgs doublet model with natural flavor conservation[4,5,8] for example, the decay can proceed via charged Higgs exchange, contributing mainly to f^-, and a sizable effect on P_T appears due to the interference with the normal W exchange term. $Im\xi$ can be expressed as

$$Im\xi = Im(\alpha_1 \beta_1^*) \cdot \frac{v_2^2}{v_3^2} \cdot \frac{m_{K^+}^2}{m_{H^+}^2}, \tag{4}$$

where α_i and β_i are the coupling coefficients of the interaction in the mass eigenstate of the Higgs boson with down-type quarks and up-type quarks, respectively, and v_2/v_3 is the ratio of the Higgs field vacuum expectation values. P_T thus constrains the two parameters, $Im(\alpha_1\beta_1^*)$ and v_2/v_3. The former parameter is solely constrained by non-leptonic processes, such as the neutron electric dipole moment[9] and $b \to s\gamma$.[10] However, $Im\xi$ can constrain it more stringently for the large v_2/v_3 region, which may be relevant in a scenario[4] where the v_i's are proportional to the fermion masses to which they couple, i.e., with $v_2/v_3 \propto m_t/m_\tau \sim 100$. The semileptonic $B \to \tau \nu_\tau X$ decay can provide similar constraints to the parameters as does $Im\xi$ (Ref. 10).

A search for T violation in $K_{\mu 3}$ decay has been attempted before in both K^+ and K_L^0 decays. The most recent data for K^+ come from a BNL-AGS experiment more than 15 years ago,[11] giving a limit of P_T(lab)= $(-0.31 \pm 0.53) \times 10^{-2}$ and $Im\xi = -0.016 \pm 0.025$. This K^+ experiment used an in-flight decay method detecting a photon from the pion decaying in the direction of the beam. A similar K_L^0 decay experiment, also done at BNL-AGS, gave a similar limit, P_T(lab)= $(0.21 \pm 0.48) \times 10^{-2}$ and $Im\xi = 0.009 \pm 0.030$ (Ref. 12). However, in the case of the K_L^0 decay, one suffers from large final state interaction effects of nearly 10^{-2}. Therefore, K^+ decay is uniquely suited for a high-precision experiment at the present time.

2 KEK E246 Experiment

2.1 Stopped Beam Method and Setup

In a P_T search experiment, it is necessary to detect a tiny P_T component in the presence of the large T conserving in-plane polarization components, P_L and P_N, which are parallel and perpendicular to the muon momentum vector, respectively. In order to overcome this difficulty, the KEK E246 experiment adopted a stopped K^+ method in conjunction with a Superconducting Toroidal Spectrometer. Using these two basic strategies we are aiming not only for an improvement of the statistical accuracy compared to the previous limit, but also for a measurement with a substantial reduction of the many systematic errors to the 10^{-3} level for ΔP_T. The advantages of using a

stopped kaon beam are manifold, but the most important is the near total coverage of $K_{\mu 3}$ decay kinematics by the detector. This enables us to carry out a measurement of the decay phase-space region with an opposite contribution to P_T. Namely, we can perform a double-ratio measurement with enhanced sensitivity to P_T, while suppressing systematic errors drastically. Moreover, the decay phase-space includes a region with vanishing P_T, providing us with the opportunity for a null-effect check. Other merits of this method, compared with in-flight decay experiments, are the small influence of a finite beam emittance, low backgrounds, smaller counting rates in each detector element, and a relatively short length of the detector.

Figure 1 shows the cross section and end view of the setup. The Superconducting Toroidal Spectrometer has a large acceptance as well as a high momentum resolution for the analysis of charged particles from the stopped kaon decay in the central region of the magnet. This allows a high statistics measurement with efficient background rejection. Furthermore, the magnet has precise 30-degree rotational symmetry, which is essential in an asymmetry experiment in order to reduce a number of systematic effects. A positive kaon beam with a momentum of 660 MeV/c was slowed down through an Al+BeO degrader and stopped in an active target made of a bundle of scintillating fibers at the center of the magnet. The $K_{\mu 3}$ decay was identified by detecting both the π^0 and the μ^+. The pion was detected as two-γ or one-γ by a calorimeter which comprises 768 CsI(Tl) crystals surrounding the target. The muon was accepted by one of the 12 magnet gaps, momentum-analyzed in a 0.9 T field, and then stopped in a muon stopper located at the exit of the magnet gap, where the polarization measurement takes place. At the stopper, there was a field applied to hold the transverse component of the polarization, while precessing the in-plane components.

2.2 Double-Ratio Experiment

The P_T of the stopped muon manifests itself as either a clockwise (cw) or counter-clockwise (ccw) asymmetry in the polarimeter, depending on the emission angle of the pion relative to the beam axis. It is measured by means of the decay positron cw/ccw asymmetry as

$$\frac{\sum_{i=1}^{12} N_i(cw)}{\sum_{i=1}^{12} N_i(ccw)} \cong 1 + 2\alpha\langle\cos\theta_T\rangle P_T. \quad (5)$$

Here, α is the analyzing power including four factors: (i) the intrinsic asymmetry of 1/3, (ii) the attenuation due to finite counter solid angle, (iii) the attenuation factor of P_T due to precession in the imperfect parallel field, and (iv) the effects of positron

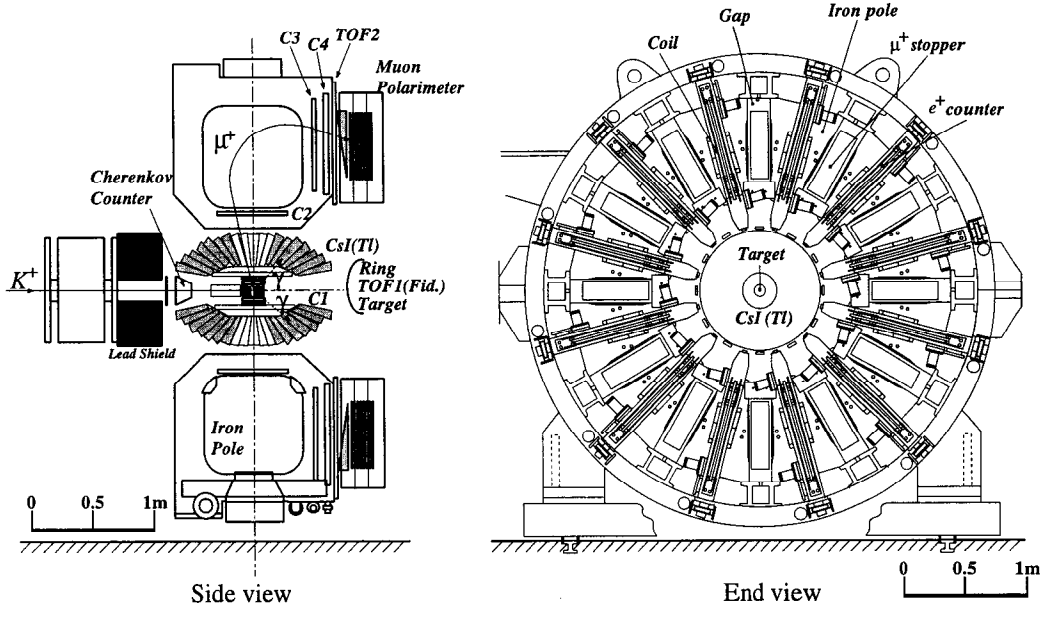


Fig. 1. Cross-section view and end view of the E246 detector.

scattering and absorption. The factor $\langle \cos\theta_T \rangle$ is the kinematic attenuation of P_T due to the distribution of decay planes in the finite acceptance of the detector. The magnetic field at the stopper was carefully trimmed to assure a symmetric distribution of the field across the magnet median plane, so that there should be no effect from P_L and P_N in total. By summing over the 12 gaps, several systematics, such as the effect of an asymmetric kaon stopping distribution in the target and unequal detection efficiency of positron counting, can be cancelled. In the actual measurement, we take the double ratio of Eq. (5) for the cases of π^0 going into the forward and the backward directions

$$\frac{[\sum_{i=1}^{12} N_i(cw)/\sum_{i=1}^{12} N_i(ccw)]_{fwd}}{[\sum_{i=1}^{12} N_i(cw)/\sum_{i=1}^{12} N_i(ccw)]_{bwd}} \cong 1 + 4\alpha\langle\cos\theta_T\rangle P_T, \qquad (6)$$

which provides further cancellation of systematic errors, and is only possible in a stopped beam experiment. The sensitivity to $Im\xi$ for a given P_T is large at the high muon energy boundary on the Dalitz plane, but the decay intensity is large for the high pion energy region with the result that the figure-of-merit ($\sqrt{N}P_T/Im\xi$) distribution is concentrated in the central part of the Dalitz plane. The acceptance of the detector can cover this high figure-of-merit region by selecting the appropriate spectrometer magnetic field.

3 Detector Performance and Analysis

The main part of the E246 detector comprises μ^+ tracking systems, a π^0 detector, and muon polarimeters. It was designed to detect all the decay particles with high resolution and large acceptance. It was also designed to reduce the influence of accidental background arising from pions in the beam which are several times more numerous than the kaons.

3.1 μ^+ Tracking and Identification

The charged particles from the kaon decay are tracked by four sets of chambers C1, C2, C3, and C4, a scintillating fiber target, and a ring counter system surrounding the target. C1 is a cylindrical drift chamber with a helical cathode strip readout of the z (beam axis) coordinate together with anode signals for azimuthal information. The other chambers, located at the entrance and exit of the magnet gaps, are planar MWPC's with cathode readout for both x and y directions. The ring system[13] consists of 32 ring scintillator counters with a 118 mm inner diameter, 5 mm thickness, and 6 mm width, covering the 20 cm fiducial region of the target. This provides us with additional z-coordinate information under high rate conditions. Readout is by means of a wavelength-shifter fiber glued into a groove cut in each ring. A high light yield of 46 photoelectrons (pe) for a minimum ionizing particle was achieved. The active stopping target, consisting of 256 5×5 mm^2 scintillating fibers with direct individual PMT readout, serves both as one of the tracking elements as well as a means of decay vertex determination. The kaon stopping is identified as a signal with large energy deposit, while decay muons leave dE of minimum ionization. Figure 2 illustrates the target and ring assembly. The main parameters of the tracking system are summarized in Table 1.

Table 1. Main parameters of the target and ring systems.

Target		Rings	
Fiber size	5 mm × 5 mm × 2 m	Size of a ring	$5^T \times 6^W \times 119$ mm^{ID}
Number of fibers	256	Number of rings	32
Diameter	106 mm	Total length	20 cm
Fiducial length	20 cm	Readout	single turn WLS fiber
Information	x and y coordinate	Information	z coordinate

Figure 3 shows the momentum spectrum of K-decay charged particles under a realistic trigger condition requiring a photon detection in the CsI(Tl) detector and a decay positron detection in the polarimeter. The resolution is dominated by multiple scattering in air and in the chambers, but is sufficient for the purpose of the $K_{\mu 3}$ selection. The predominant background peak from $K_{\pi 2}$ pions stopping in the muon stopper can be readily rejected by setting a momentum threshold at 190 MeV/c. The $K_{\mu 3}$ events are distributed in the bump below 190 MeV/c together with K_{e3} positrons, which can be rejected by using TOF information between a counter at the exit of the magnet gap and the target fiducial counter. Otherwise, these K_{e3} positrons, when scattered by the copper degrader and the stopper, contribute to the background in the muon-decay positron time spectrum as a prompt timing peak. The performance of the TOF separation is shown in Fig. 4.

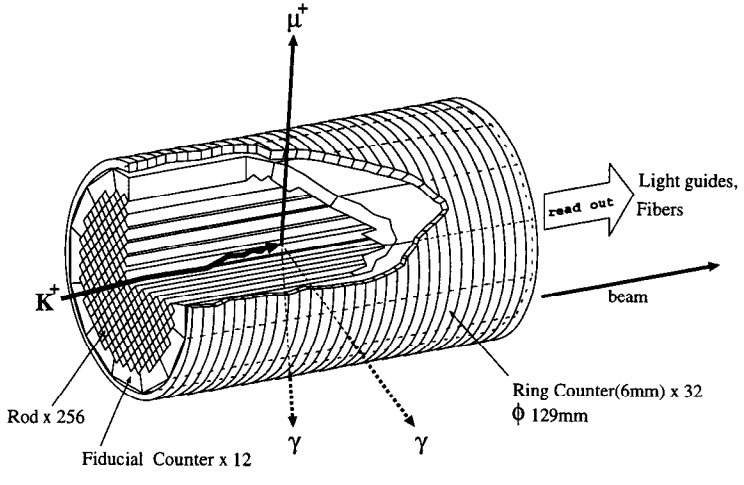


Fig. 2. Schematic view of the active target and ring counter system.

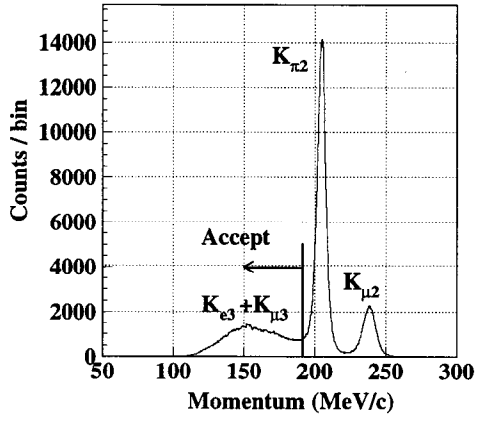


Fig. 3. Momentum spectrum of charged particles under the trigger condition of photon detection in the CsI(Tl) and positron detection in the positron counters. The $K_{\mu 2}$ peak is due to an accidental photon in the CsI(Tl).

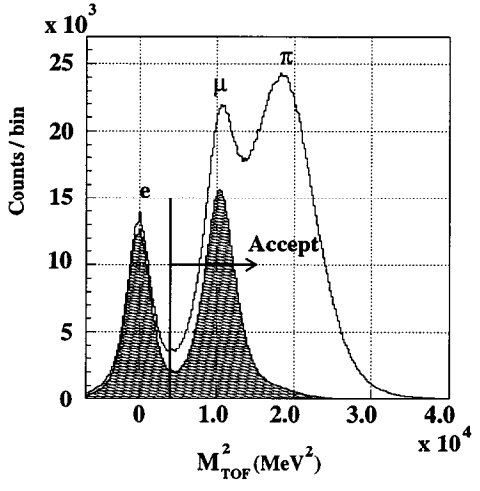


Fig. 4. TOF mass spectrum of charged particles before momentum selection (solid line) and after momentum selection (shaded spectrum).

3.2 CsI(Tl) π^0 Detector

The π^0 detector[14,15] is a barrel array of 768 CsI(Tl) crystals surrounding the target with a solid angle of about 3/4 of 4π. The barrel is segmented with $\Delta\theta = \Delta\phi = 7.5°$, both for polar and azimuthal directions, except for one layer nearest to the beam axis where $\Delta\phi = 15°$. There are 12 rectangular holes for charged particles to enter the magnet gaps and two holes for the beam entrance and exit (Fig. 5). The main parameters of each crystal module and the whole detector are summarized in Table 2. Because of space problems in the central region of the magnet and the existence of a fringing field, PIN photodiode readout was employed. Very high performance of these CsI modules was achieved with an average light yield as high as 11,000 pe/MeV with an equivalent noise level (EVL) of 65 keV.[15] This low-noise level allowed us to lower the photon detection threshold to 5 MeV. The signal from each diode is read by a peak-sensing ADC after pulse shaping and at the same time by a pulse shape analyzer by means of a transient digitizer.[16] A timing resolution of 13 ns (FWHM) was obtained in the relevant energy region by means of a constant fraction discriminator trigger.[17] This excellent timing performance was very powerful in identifying photon clusters and rejecting backgrounds. In spite of the large lateral shower leakage, a relatively narrow invariant mass spectrum is observed. Figure 6 shows the spectra for π^0 in the forward or backward directions in which photons are likely to hit a nonhole region, and in the normal direction in which photons hit mostly muon holes, respectively.

The clean π^0 events consisted of two photons in the CsI. In order to recover some of the π^0 acceptance lost due to the holes in the CsI array, we also accepted events with only one–photon cluster, satisfying a high-energy threshold (70 MeV), for which most of the π^0 kinematic information was preserved.

3.3 Muon Polarimeter

The structure of the muon polarimeter is shown schematically in Fig. 7. Muons are bent in the magnet gap, slowed down through a wedge-shaped copper degrader, and stopped in a muon stopper made of plates of pure Al. In order to minimize the effect of scattering and absorption for the decay positrons, the plates are spaced to give an effective density of 0.4 of bulk Al. The material has been studied in a muon spin rotation measurement and it was confirmed that there is no significant depolarization during the stopping process nor any relaxation due to solid state effects. At the stopper, a magnetic field was applied using pure-iron shim plates to conduct a fraction of the

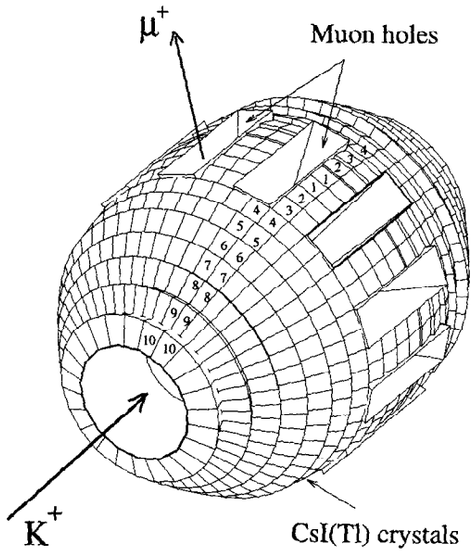


Fig. 5. CsI(Tl) π^0 detector. There are 12 rectangular holes for the charged particles to enter the magnet gaps.

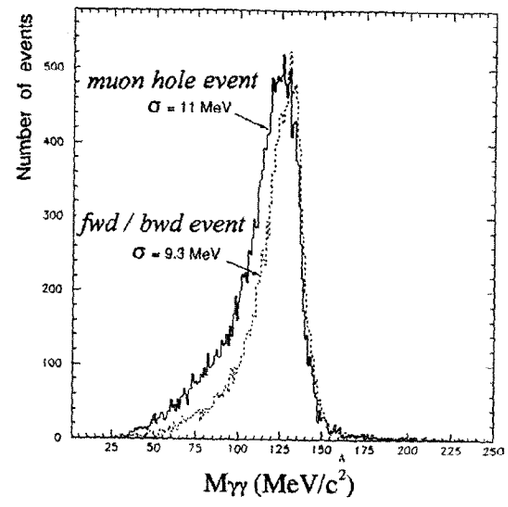


Fig. 6. Invariant mass spectrum for two-γ; the dotted line is for the events with a pion in the forward or backward directions, and the solid line is for the events with a pion in the direction perpendicular to the beam.

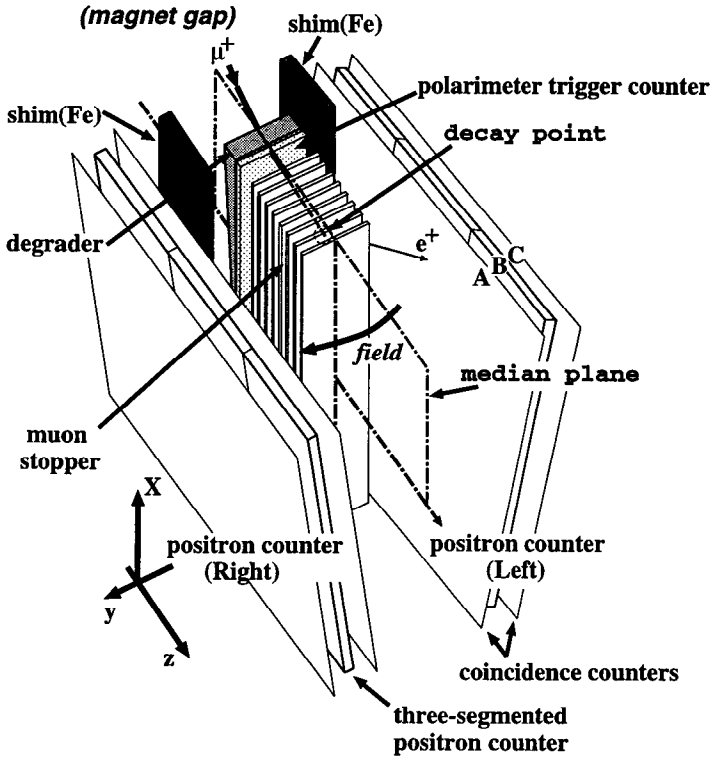


Fig. 7. Schematic view of the polarimeter of one gap.

Table 2. Main parameters of the CsI(Tl) π^0 detector.

Structure		Performance	
Number of crystals	768	Readout	PIN photodiode
Segmentation	$\Delta\theta = \Delta\phi = 7.5°$	Light yield	11000 pe/MeV
Length of crystal	25 cm	ENL	65 keV
Typical size	$3 \times 3 - 6 \times 6$ cm	Time resolution	6.5 ns (σ)
Inner radius	20 cm	Energy resolution	3.0% at 200 MeV
Outer radius	50 cm	Position resolution	1.0 cm
Total weight	1.7 ton	Energy threshold	5 MeV

superconducting magnet fringing field. The shim plates were designed so as to make a strong field on the stopper and also to accomplish symmetric distribution of the field relative to the magnet median plane. Thus, the average field direction is parallel to the average direction of P_T, and the spurious effects arising from P_L and P_N should be cancelled in total, although the field flux is curving between the two shim plates. Field mapping was performed with a specially designed high-precision device with a three-dimensional Hall probe,[18] and the excellent symmetry of the field was confirmed.

Decay positrons were measured by counter systems located between each pair of adjacent muon stoppers. Each counter system consists of three layers of plastic scintillators, the middle one being segmented into three parts. The alignment of the polarimeter elements was done with particular care. The rear and outer sides of the stoppers were surrounded by veto counters to suppress background events from muons stopping outside the stoppers and from accidental room background. A large fraction of the background associated with the beam could be suppressed by a veto counter system surrounding the central region of the polarimeter. Figure 8 shows the time spectrum of positrons in good events recorded by a multistop TDC before and after the background suppression. A very low level of constant background of 13% for the signal region up to 8 μsec was achieved. There is a small structure seen at the beginning of the decay period due to the spin precession of the in-plane component. The $K_{\mu 3}$ events were extracted by subtracting the constant background from the signal region of the time spectrum.

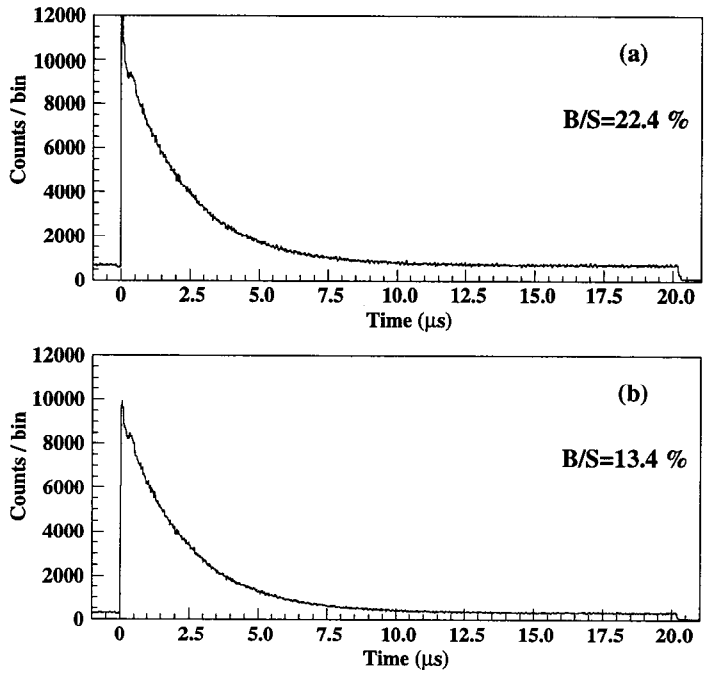


Fig. 8. Positron time spectra: (a) before applying veto counters, and (b) after applying veto counters. B/S was calculated for the signal region from 0 to 8μsec.

Table 3. Main parameters of the muon polarimeter.

Stopper		Counters	
Material	pure Al	e^+ detection	three fold of plastics
Size of a plate	$0.6^T \times 55^H \times 16^W$cm^3	Size	$84^H \times 30^W$cm^2
Number of plates	8	Segmentation	three middle layers
Spacing	0.8 cm	Solid angle /4π	0.07
Effective ρ	0.43 $\rho_{\rm Al}$	Muon gate	20 μsec

3.4 Analysis

Good $K_{\mu 3}$ events were selected by setting cuts on the (a) momentum spectrum, (b) invariant mass spectrum and photon total energy (in the case of two-γ), (c) single photon energy (in the case of one-γ), (d) mass spectrum from TOF, and (e) kaon decay time spectrum. A kinematic cut on the opening angle between the muon and the pion was very useful for rejecting any $K_{\pi 2}$ pions remaining from the momentum cut and their decay-in-flight muons. Finally, the missing-mass cut was tuned to reject events from kaon decay-in-flight. Events were analyzed separately for the two-γ and one-γ cases. The regions of analysis, π^0–forward and π^0–backward, were determined to be $|\cos\theta_{\pi^0}|$, $|\cos\theta_\gamma| > 0.34$ both for two-γ and one-γ events by maximizing the figure-of-merit function $\langle\cos\theta_T \times P_T/Im\xi\rangle \times \sqrt{N}$; here, θ_{π^0} and θ_γ are the polar angles of the π^0 or one-γ, respectively. The asymmetry, A_T, was extracted from the ratio of cw and ccw positron counting, and P_T was calculated as

$$P_T = \frac{A_T}{\alpha\langle\cos\theta_T\rangle}. \tag{7}$$

The analyzing power α was deduced to be 0.198 by measuring the asymmetry from the in-plane polarization P_N by selecting events with the π^0 going into the Cs(Tl) barrel perpendicular to the beam axis. The kinematic attenuation factor $\langle\cos\theta_T\rangle$ was calculated in a Monte Carlo simulation taking into account the beam accidental background, which could be estimated from $K_{\mu 2}$ events in coincidence with accidental photons. P_T was then converted to $Im\xi$ using an average sensitivity factor $\langle P_T/Im\xi\rangle$ also estimated by a Monte Carlo calculation,

$$Im\xi = \frac{P_T}{\langle P_T/Im\xi\rangle}. \tag{8}$$

4 Current Status of E246

4.1 Statistical Limit

After the detector construction from 1992 to 1995, we started data acquisition in early 1996. Using the low momentum K^+ beam from the K5 beam channel at the KEK proton synchrotron with intensity of 1.1×10^5/sec, we obtained a $K_{\mu 3}$ event rate of 1.0 /sec. By the summer of 1998, we had accumulated a total of 3.6 million events in the π^0–forward and π^0–backward regions. We are performing two independent analyses in the group. This should make the result more reliable because a consistency test can be done, and it also gives us a means of estimating the systematic errors associated with the analysis. The major differences in the two analyses lie in the charged particle tracking algorithm and in the photon clustering algorithm in the CsI(Tl) array. The near equivalence of the two analyses was checked by comparing good final events in both analyses and their correlations. However, some fraction of the good events belong only to one analysis, and the difference of the central values is attributed to the statistical fluctuations of this part. Both analyses have finished processing all the 1996 and 1997 data.[19–21] Both preliminary results are consistent with null-T violation within the statistical error

$$P_T = (-2.55 \pm 5.68) \times 10^{-3},\ Im\xi = (-0.77 \pm 1.86) \times 10^{-2} \qquad (9)$$

for one analysis, and

$$P_T = (-0.25 \pm 5.72) \times 10^{-3},\ Im\xi = (-0.01 \pm 1.80) \times 10^{-2} \qquad (10)$$

for the other. We are currently engaged in a discussion on how to best combine these two results. When the 1998 data has been analyzed, a statistical limit of $\Delta Im\xi = 0.011$ is expected.

4.2 Backgrounds

Although the main part of the $K_{\pi 2}$ background was rejected in the momentum spectrum as well as by an opening-angle cut, the decay-in-flight of pions into muon remains as an admixed background estimated to be a few percent of the good events. However, the average contribution to P_T for these events should be zero, and moreover, the forward/backward cancellation is expected to work, with no significant influence on the result. The contamination of other backgrounds, $K^+ \to \pi^+\pi^0\pi^0$ and $K^+ \to \pi^0 e^+ \nu$, was suppressed below 0.5% and 0.1%, respectively. These backgrounds should not produce any spurious asymmetry. The contamination of kaon decay-in-flight events was also suppressed below the 3–4% level by a cut on the missing mass and it was checked that there was no bias induced by its admixture. The amount of accidental background in the CsI(Tl) detector which is due to scattered pions in the beam was estimated to be 7–9% in both analyses, and its effect was taken into account as a degradation of the kinematic attenuation factor $\langle \cos\theta_T \rangle$ arising from the uncertainty in the decay plane determination.

4.3 Systematic Errors

The systematic errors were evaluated using real experimental data as well as Monte Carlo simulation. The most serious source is the asymmetry of the muon spin rotation field distribution. However, this effect could be estimated to be $\Delta P_T \leq 6 \times 10^{-4}$ from the measured tilt of the field distribution in each gap and from the uncertainty of the angular determination in the field mapping. The effects of a misalignment of the detector elements could be tested by examining the distributions of the decay plane orientation relative to the polarimeter coordinate frame. The average values of the plane tilt around the radial axis x and the beam axis z induce mixing of P_L and P_N, respectively, resulting in a fake effect. However, these effects were shown to be $\Delta P_T \approx 3 - 4 \times 10^{-4}$ each. The effective cancellation factor in the forward/backward ratio was evaluated from an artificial asymmetry distribution introduced by setting asymmetric cuts on the real data and then calculating the cancellation. The relatively large asymmetry of the kaon stopping distribution in the target due to the beam momentum dispersion was shown to have no first order influence on the result thanks to the cancellation mechanism when summing over the 12 gaps. The total systematic error was, thus, estimated to be $\Delta P_T = 1.0 \times 10^{-3}$, which is more than five times smaller than the current statistical error. The overall systematics could be further checked in (1) the null asymmetry with the pion integrated over all directions, (2) the gap dependence of the asymmetry, and (3) the polar angle dependence of the asymmetry; no systematic bias nor any strange magnet gap dependence was found.

5 Summary

A transverse muon polarization in $K^+ \to \pi^0 \mu^+ \nu_\mu$ would be a clear sign of T violation and is a sensitive probe for new physics beyond the Standard Model. The three Higgs-doublet model, the leptoquark model, and some supersymmetric models can give a sizable effect without conflicting with other experimental constraints. The KEK E246 experiment, using a stopped kaon beam, is aiming for an eventual statistical sensitivity of ΔP_T at the 0.1% level. All elements of the detector system are working well and the level of systematic errors has been shown to be small. The current statistical accuracy of the analyzed data is $\Delta P_T = 6 \times 10^{-3}$, or $\Delta Im\xi = 0.018$. Our results are consistent with null transverse polarization at the present time. Our experiment is scheduled to run further for 1600 hours to improve the statistics.

References

[1] KEK-PS E246 Collaboration: M. Abe,[a] M. Aoki,[b] I. Arai,[c] Y. Asano,[a] T. Baker,[d] M. Blecher,[e] M. D. Chapman,[b] D. Dementyev,[f] P. Depommier,[g] M. Grigorjev,[f] P. Gumplinger,[h] M. Hasinoff,[i] R. Henderson,[h] K. Horie,[c] W. S. Hou,[j] H. C. Huang,[j] Y. Igarashi,[c] T. Ikeda,[c] A. Ivashkin,[f] J. Imazato,[b] J-M. Lee,[k] L. S. Lee,[l] G. Y. Lim,[l] J. H. Khang,[k] W. Keil,[h] M. Khabibullin,[f] A. Khotjantsev,[f] Y. Kudenko,[f] Y. Kuno,[b] J. A. Macdonald,[h] D. R. Marlow,[m] C. R. Mindas,[m] O. Mineev,[f] C. Rangacharyulu,[d] S. K. Sahu,[j] S. Sekikawa,[c] H. M. Shimizu,[b] S. Shimizu,[n] K. Shibata,[c] Y.-M. Shin,[d] Y. H. Shin,[k] K. S. Sim,[l] A. Suzuki,[c] T. Tashiro,[c] A. Watanabe,[c] D. H. Wright,[h] T. Yokoi[o]

Address: [a] Institute of Applied Physics, University of Tsukuba, Ibaraki 305-0006, Japan.

[b] IPNS, High-Energy Accelerator Research Organization (KEK), Ibaraki 305-0801, Japan.

[c] Institute of Physics, University of Tsukuba, Ibaraki 305-0006, Japan.

[d] Department of Physics, University of Saskatchewan, Saskatoon, Canada S7N 0W0.

[e] Department of Physics, Virginia Polytechnic Institute and State University, VA 24061-0435, USA.

[f] Institute for Nuclear Research, Russian Academy of Sciences, Moscow 117312, Russia.

[g] Laboratoire de Physique Nucleaire, University de Montreal, Montreal, Canada H3C 3J7.

[h] TRIUMF, Vancouver, British Columbia, Canada V6T 2A3.

[i] Department of Physics and Astronomy, University of British Columbia, Vancouver, Canada V6T 1Z1.

[j] Department of Physics, National Taiwan University, Taipei, Taiwan.

[k] Department of Physics, Yonsei University, Seoul 120-749, Korea.

[l] Department of Physics, Korea University, Seoul 136-701, Korea.

[m] Department of Physics, Princeton University, NJ 08544, USA.

[n] Department of Applied Physics, Tokyo Institute for Technology, Tokyo 152, Japan.

[o] Department of Physics, University of Tokyo, Tokyo 113, Japan.

[2] J. J. Sakurai, Phys. Rev. **109**, 980 (1958).

[3] A. R. Zhitnitskii, Yad. Fiz. Sov. J. Nucl. Phys. **31**, 1024 (1980); *ibid* **31**, 529 (1980).

[4] R. Garisto and G. Kane, Phys. Rev. D **44**, 2038 (1991).

[5] G. Bélanger and C. Q. Geng, Phys. Rev. D **44**, 2789 (1991).

[6] M. Fabbrichesi and F. Vassani, Phys. Rev. D **55**, 5334 (1997).

[7] G. H. Wu and J. N. Ng, Phys. Lett. B **392**, 93 (1997).

[8] S. Weinberg, Phys. Rev. Lett. **37**, 657 (1976).

[9] H. Y. Cheng, Phys. Rev. D **42**, 2329 (1990).

[10] Y. Grossman, Nucl. Phys. B **426**, 355 (1994).

[11] S. R. Blatt *et al.*, Phys. Rev. D **27**, 1056 (1983).

[12] W. M. Morse *et al.*, Phys. Rev. D **21**, 1750 (1980).

[13] D. V. Dementyev *et al.*, Nucl. Instrum. Methods A **379**, 499 (1996).

[14] A. P. Ivashkin *et al.*, Nucl. Instrum. Methods A **394**, 321 (1997).

[15] Y. Kudenko and J. Imazato, KEK-Report 92-15 (1992).

[16] R. L. Wixted *et al.*, Nucl. Instrum. Methods **386**, 483 (1997).

[17] Yu. G. Kudenko, O. V. Mineev, and J. Imazato, Nucl. Instrum. Methods A **411**, 437 (1998).

[18] T. Ikeda *et al.*, Nucl. Instrum. Methods A **401**, 243 (1997).

[19] T. Yokoi, Ph.D. thesis, University of Tokyo, 1998 (unpublished).

[20] C. Mindas, Ph.D. thesis, Princeton University, 1998 (unpublished).

[21] A. Ivashkin, Ph.D. thesis, Institute for Nuclear Research, Russian Academy of Sciences, 1998 (unpublished).

GAMMA-RAY BURSTS: CURRENT OBSERVATIONAL PICTURE

Jay P. Norris
Laboratory for High Energy Astrophysics
NASA/Goddard Space Flight Center
Greenbelt, MD 20706

ABSTRACT

We now know that long (duration > 2 s) gamma-ray burst (GRB) sources lie at distances comparable to the "radius" of the Universe and that they are momentarily the most luminous known objects: several GRBs have now been observed to have associated X-ray, optical, and radio counterparts following power-law decays that fit theoretical expectations for relativistic blastwaves. Three GRBs (and/or their host galaxies) have determined redshifts larger than z = 0.8, and virtually all GRBs with associated optical transients appear to lie in distant galaxies. Spectroscopic analysis of afterglows and modeling of the GRB brightness distribution indicate that GRB sources lie in star-forming regions. The population statistics of GRBs as well as their energetic, spectral, and temporal aspects—also consistent with the cosmological distance scale—are difficult but not impossible to explain via neutron star or neutron star-black hole mergers. The suggested association of smooth, single-pulse GRBs with supernovae in nearby redshift space can only involve an insignificant fraction of the total GRB population; moreover the association is contraindicated by the random sky distribution of single-pulse GRBs. Future observations of GRBs with rapid temporal variations at supra-GeV energies by GLAST, the Gamma-ray Large Area Space Telescope, may provide a unique cosmic experiment for constraining the scale of quantum gravity. Detection of a gravitationally lensed GRB would be evidence of a non-zero value for Λ, the cosmological constant.

© 1998 by Jay P. Norris.

1 Introduction: Gamma-Ray Bursts Become Highly Visible

Gamma-Ray Bursts (GRBs) are currently enjoying the best public relations since their discovery 30 years ago. Even Sky and Telescope (February 1998), which reaches a large popular scientific audience, featured a fantastic looking cover picture of the suggested progenitor object moments before the GRB explosion. The object which looks like a crackling bowling pin, frozen in time, represents the merger of a neutron-star (NS) binary. Of course the ends of the bowling pin would be rotating at a sizable fraction of the speed of light during the final moments before coalescence. The feature article,[1] appearing one year after the discovery of the first X-ray and optical counterparts to a GRB, concerned the gigantic energy output of GRBs and the theoretical burst of cosmic rays which would follow the electromagnetic (EM) radiation of the GRB in a matter of days. Since the sources of GRBs (at least those of the "long" duration variety) lie at cosmological distances—billions of light-years from our Galaxy—GRBs must be luminous in the extreme. The rate of detected GRBs in the Universe (~ 10^{11} galaxies) —a few per day—translates into ~ one GRB per galaxy per ~ 10^7 years. The total rate would be much higher if the radiation pattern were strongly beamed into a small solid angle: only a small fraction of GRBs would then be detected by Earth observers. Thus, every two rotations of the Sun about the Galactic Center (a half billion years), a GRB would be close enough (~ 1 kiloparsec) for the accompanying cosmic ray burst to annihilate all life on Earth. This is just one of several recent cosmological theories which might provide a sufficiently large energy release to power a GRB.[2] We shall also examine the recent putative connection between type Ic supernovae (SNe) and GRBs, and find that only a very small fraction of GRBs can be connected with SNe.

Observations of NS binaries have already provided interesting tests of physical theory. Table 1 lists several confirmed and suspected NS systems, along with their distances from Earth, rotational periods, and orbital decay times. Included is the famous Hulse-Taylor binary pulsar, which provided the first quantitative evidence for gravitational radiation—energy derived from orbital decay. The distance column is relevant for the GRB/cosmic-ray burst extinction theory, but the decay, or coalescence, timescale provides an astrophysical constraint: When the components of the binary each go supernova, a large velocity kick, ~ 10^3 km s^{-1}, can be imparted to the system. Assuming coalescence timescales of 0.1 – 10 Gyr (still comparable to the age of the Universe), we would look to find at least some of these systems ~ 100 kpc – 10 Mpc outside their parent galaxies. Note however, the systems listed in Table 1 are those we have detected; systems with much shorter decay timescales are much less likely to be discovered before coalescence.

The central questions of source progenitor population and their distribution in redshift, as well as the mechanisms for generation of γ-radiation and the presumed beaming into a narrow solid angle (probably completely distinct from processes governing the counterpart afterglows), all will be constrained—some in ways we do not yet apprehend—not only by evidence from afterglows, but also from the extrinsic characteristics of GRBs and their spectral/temporal behavior.

Table 1: Pairs of Neutron Stars in Our Galaxy

Name	Distance (kiloparsecs)	Period (hours)	Decay Time (10^9 years)
PSR B1534+12	0.5	10.1	2.73
PSR B1913+16 [1]	7.3	7.75	0.30
PSR B2127+11C [2]	10.6	8.05	0.22
PSR B2303+46 [3]	2.3	296	4000
PSR J1518+4904 [3]	0.7	207	2400

[1] Hulse-Taylor pulsar (first direct evidence for gravitational radiation)

[2] In Globular Cluster M13

[3] Suspected binary neutron stars – masses not firmly established

The picture of GRBs at gamma-ray energies as revealed by current instruments on the Compton Gamma Ray Observatory is reviewed in Section 2. We then proceed to examine some more definitively derived constraints imposed by the recent detections at longer wavelengths in Section 3. The apparent association of at least one GRB with an unusual SN is critiqued in Section 4. Section 5 describes two future measurements: a constraint on the energy scale of quantum gravity, using GRBs detected by GLAST; and a constraint on Λ via detection of gravitationally lensed GRBs.

2 GRB Extrinsic Characteristics and GRBs Themselves

2.1 The Distance Scale: Isotropy and the Brightness Distribution

Prior to the discovery of GRB counterparts at longer wavelengths, the extrinsic characteristics of GRBs were endlessly debated, within a milieu of few definitive clues, and mostly under the assumption that GRBs were galactic in origin. The most important extrinsic descriptor — the GRB distribution in space—was originally unconstrained, since the first small GRB detectors were not sensitive enough to sample the entire population. Progressively less oblate spatial distributions (our Galaxy's disk – very much oblate; the Oort Cloud—slightly flattened; our Galaxy's putative "extended halo" —possibly in a common envelope with our sister galaxy, Andromeda) were eliminated as BATSE on the Compton Gamma Ray Observatory detected more and more GRBs, which appeared at random positions on the sky. This isotropic celestial distribution, combined with the fact that BATSE does not detect enough dim bursts compared to the expectation for a homogeneous Euclidean space — indicating that the edge of the spatial distribution is now probed—strongly suggested that GRB sources lie at cosmological distances. All other extrinsic distributions corroborated this appearance, and even before the discovery of counterparts, predicted that GRB sources should lie at redshifts z ~ unity and higher. One distance scale that was never suggested by the data—the "Supergalactic Plane" out to a few times the Virgo cluster distance, ~ 100 Mpc—is now claimed by some investigators as the distance scale for a small subset of GRBs associated with SNe of type Ib/Ic. There is now some evidence against this SNe/GRB connection (see Section 4).

The power outputs from GRBs were then estimable, assuming their cosmological origin. If the power in gamma rays that we observe from the more luminous GRBs is being radiated isotropically, then their instantaneous luminosities, ~ 10^{53} ergs s^{-1} ~ 10^{19} solar luminosities ~ 10^{9} L_* galaxies, would be a few percent of the Universe's power output. If GRBs had turned out to be in the galactic disc, their power output would have been relatively puny, somewhat less than the peak luminosity of a type II supernova, ~ few × 10^{42} ergs s^{-1}. Total GRB radiated energies, integrated over their (unpredictable) durations of 0.01 – 1000 s, would range up to ~ few × 10^{54} ergs, again assuming the unlikely isotropic case.

The web page "www.batse.msfc.nasa.gov" shows the BATSE sky map for the accumulated sample of seven plus years. Occasionally, articles appear suggesting that some subset of GRBs is distributed anisotropically. Indeed, since this map shows individual detections and therefore is necessarily uncorrected for exposure, the equatorial plane is barely perceptible as a zone of slight avoidance. When corrected for exposure with respect to any arbitrary plane, the dipole and quadrupole moments are consistent with isotropy for many subsets of GRBs.[3]

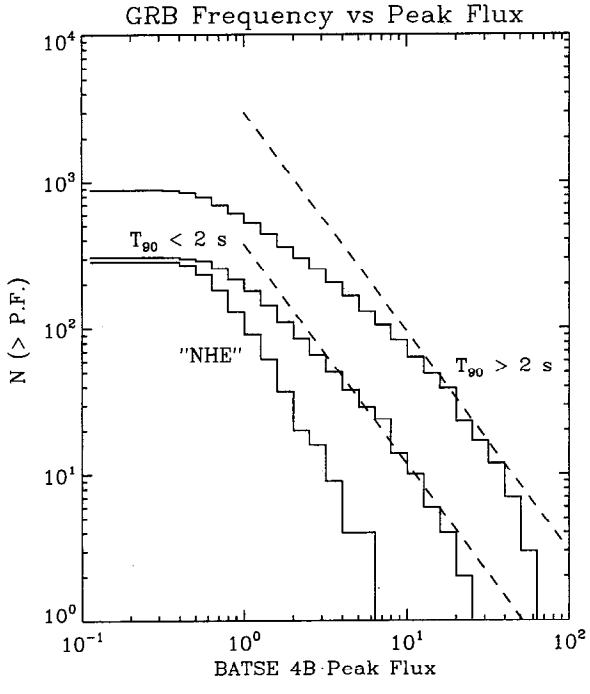


Fig. 1 – GRB brightness distributions for long and short bursts (duration break point at 2 s), and bursts with no high-energy emission ("NHE"), above 300 keV. Dashed lines are -3/2 power laws, which would be expected for uniform homogeneous distributions. Above a peak flux of ~ 1 photon cm^{-2} s^{-1} (50 – 300 keV), BATSE detection efficiency is nearly complete.

Figure 1 shows the much discussed and modeled Log N(umber) – Log P(eak) F(lux) relation for GRBs, also known as the brightness or size distribution. The number of GRBs with PF greater than the abscissa value is plotted on the ordinate. Suppose that GRBs uniformly filled a Euclidean space; then from the simple consideration of three-space, their

brightness distribution would follow a -3/2 power law. The two dashed lines are -3/2 power laws. Shown separately are distributions for long ($T_{90} > 2$ s) and short bursts ($T_{90} < 2$ s); see section 2.4 for a discussion of GRB durations. T_{90} is the interval between the times to accumulate 5% and 95% of the total counts in a GRB. Clearly the long bursts depart more markedly on the dim end from the -3/2 power law, suggesting that we see through more nearly to the edge of their spatial distribution than we do for the short bursts. Said differently, short bursts would appear to be dimmer and not sampled as deeply in redshift space. This is significant, since we have yet to see an X-ray, optical, or radio counterpart for a short burst (the GRB instrument on the BeppoSAX satellite—which has signaled the alert for almost all GRB counterparts so far—does not trigger on short bursts). So short bursts could still be galactic. But, short and long bursts separately have isotropic sky distributions.

The third curve labeled "NHE" (No High Energy) is the brightness distribution for long bursts that have no significant emission above ~ 300 keV. Because of its steeper, nearly -3/2 slope, this subsample has been claimed to be in nearer redshift space, and less luminous because they are supposedly not beamed directly towards Earth.[4] This would constitute the main observational evidence for beaming. However, the apparent NHE class—that is, the approximate slope and normalization of the NHE curve in Figure 1—can be synthetically realized by equalizing the signal-to-noise levels (s/n) of bright bursts (practically all of which have HE emission) to successively dimmer steps in the log N – log PF, and redshifting their spectra (dimmer bursts, presumably at larger redshifts, have softer energy spectra). In this explanation, the steepness of the NHE curve then is due to the combined rapid onset of dimmer and "redder" bursts at lower PF —essentially a brightness bias.[5]

Careers have been spent modeling the GRB brightness distribution, including such extrinsic effects as cosmological time dilation and rate-density evolution with cosmic time; and effects like a GRB luminosity distribution, which may also depend on cosmic time; while also attempting to account for such instrumental effects as incomplete sampling on the dim end and dependence on spectral response. Clearly, we do not presently know enough about GRBs' intrinsic attributes and extrinsic distributions to hope for these efforts to be entirely successful, but we will know enough soon, after accumulation of a sufficient number of redshifts (and therefore luminosities) from counterparts that will calibrate the brightness distribution. The current wisdom in one camp is that the GRB rate-density follows that of star formation,[6,7] which apparently peaked near redshift epoch $z \sim 1.25 - 1.5$, and then declined to the present time. But other modelers[8] claim that a constant GRB rate-density is equally acceptable. Models including luminosity distributions are more realistic (but a broad luminosity range may not be needed) and are consistent with the peak fluxes and redshifts of GRBs 970508, 971214, and 980703 (Ref. 9). It may also be judicious to consider the log N – log PF for long and short bursts separately.[10]

For the cosmological distance scale, it is easy to understand qualitatively why the dim end of the GRB brightness distribution should depart from a -3/2 power law: assume a constant number of GRBs per year, per Mpc^3 everywhere in the Universe. Then two effects diminish the observed number of dim bursts on Earth. First, time dilation makes the "per year" longer for the observer, and second, gravitational lensing of distant volumes by the nearby matter distribution makes the "per Mpc^3" appear larger.

2.2 Cosmological Time Dilation and Spectral Redshift

Before the modern era of counterparts, two additional pieces of evidence supported the cosmology camp's position. (The incredible energetics required of nature was the primary obstacle for the cosmological distance scale, but that is manifestly no longer a problem for nature.) Cosmological time dilation (TD) comes into play not only in log N – log PF as part of the rate-density consideration, but also for the innards of bursts: pulse widths, intervals between pulses, and overall durations. Since GRBs are immensely varied in temporal appearance, there is ample opportunity to measure these timescales. However, measurement difficulties ensue since the dynamic ranges of GRB durations and pulse widths are large. Also, brightness bias and instrument trig gering effects enter the game (GRB monitors usually are half asleep, waiting in low temporal resolution mode for a GRB to occur; then the instrument switches to higher time resolution modes to better record the fast time structure).

We have measured the relative, average widths of dimmer bursts (presumably more distant) compared to those of the brightest bursts (presumably closer) by a variety of methods. All our TD measures include a preparatory step which approximately equalizes the s/n levels of all burst time profiles to that of the dimmest bursts in the sample.[11,12] One particularly "common-sense" method is to group bursts by PF, then within a group, bring the time profiles into temporal registration with the highest peak in each burst positioned at t_0. Estimating the peak time in an unbiased manner is not necessarily straightforward for the artificially dimmed bursts, but is facilitated by a wavelet "denoising" approach. One then measures, e.g., the FWHM of the average profile for each peak-flux grouping and compares these widths with that of the brightest group. This is the observed time-dilation factor (TDF), and would be equal to $(1 + z)/(1 + z_{brightest})$, where z is the average redshift for a given peak-flux group, except that some corrections need to be estimated and applied.

Figure 2 shows the redshift-corrected "peak-aligned" TD measure for ten peak-flux groups from the long burst ($T_{90} > 2$ s) sample only. There are ~ 92 bursts per group. PF is in the standard units for BATSE, photons cm^{-2} s^{-1} (50 – 300 keV). The bootstrap error sizes are mostly due to sample variations (varied temporal appearance, burst to burst), rather than counting statistics. When the brightest group is subdivided further, no significant differences in TDFs result. We have constructed many variants on this sort of TD measure and others (e.g., durations)—varying thresholds, energy ranges, denoising parameters, etc.—and the basic results are always significant and approximately similar: the observed, relative TDF between the brightest and dimmest BATSE bursts is ~ 2. Other investigators report similar results.[13,14,15]

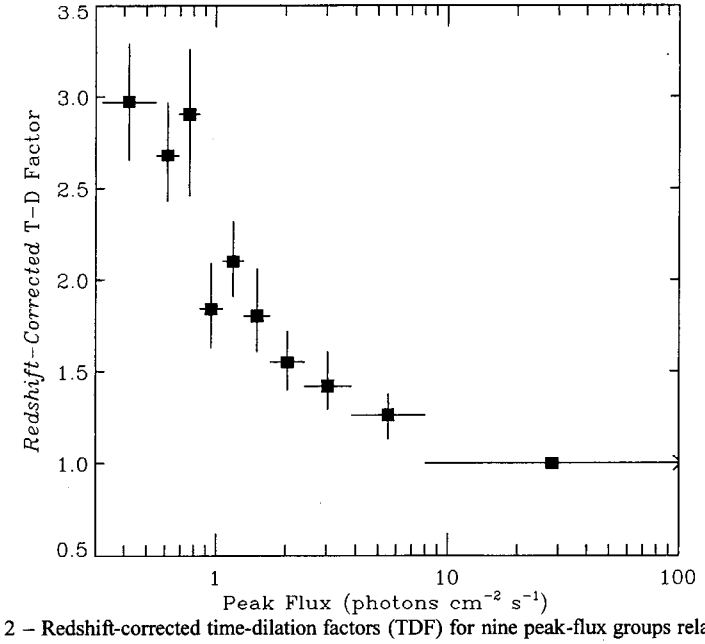


Fig. 2 – Redshift-corrected time-dilation factors (TDF) for nine peak-flux groups relative to the brightest group, for the peak-aligned average profile method.[11,15] There are ~ 92 bursts per group. TDF measures for dimmest three groups utilize only the decaying portion of average profile in attempt to estimate corrections for less than 100% detection efficiency of slowly rising, dim bursts.

At least two correction factors need to be applied to the observed TDFs to realize measures appropriate for comparison with other observables (e.g., log N – log PF). First, because burst temporal structures are intrinsically shorter at higher energy, those portions of the dimmer bursts' profiles that are redshifted into the BATSE detectors' energy band will be narrower than if we were observing the bursts in their comoving frame.[16] We estimate corrections for this complication by s/n equalizing, redshifting, and time-dilating—on a trial grid of TDFs—the bright burst sample to each dimmer sample in succession. Second, BATSE's on-board burst trigger algorithm is "causal": it samples the background for ~ 17 s, then asks if there is a significant (~ 5.5 σ) fluctuation in a 64-ms, 256-ms, or 1024-ms bin succeeding the 17-s integration, in any two of its eight identical detectors (see the web page "cossc.gsfc.nasa.gov/cossc/cossc.html" for descriptions of the Compton GRO instruments and data archive). Thus for dim bursts that rise slowly, the comparison is between a background level and not-so-different levels for the test bins; therefore, BATSE tends to miss slow, dim risers[17]—but not slow, dim decayers, unless they are also slow risers. For PF > 1. photon cm^{-2} s^{-1}, BATSE sampling is ~ 98% complete. For the three dimmer peak-flux groups in Figure 2 we estimated the TDFs by using only the decaying portion of the average time profiles.

The TDF measures plotted in Figure 2 are then an estimate of $(1 + z)/(1 + z_{bright})$, ostensibly with instrumental and redshift effects removed.[18] The dimmest bursts' average profiles are ~ factor of three wider than those of the brightest group. The TDF relation can be combined with the log N – log PF relation to constrain the distribution of GRBs within different model Universes.[8] Results indicate that redshifts of the dimmest bursts (PF ~ 0.1 photon cm^{-2} s^{-1}) detected by BATSE may be as high as $z \sim 5 - 8$. For comparison, the BATSE PFs are ≈ 1.2, 2.3, and 2.6 for the three bursts with associated redshifts $z = 0.835$, 3.412, and 0.966, respectively. The brightest bursts are factors of 10 – 50 brighter, and are predicted to lie at redshifts of a few tenths, if the constant GRB rate-density hypothesis is more nearly correct. If the GRB rate-density follows the star formation rate, then the sources would lie at redshifts of order unity, in which case a broad luminosity distribution may be obtained.

To appreciate the analogous study of burst spectra as a function of PF, we first compare the spectra of high-energy transients in general. From Figure 3, a theorist's $\nu \bullet f(\nu)$ spectral plot, we see that the most intense GRBs are, far and away, brighter and spectrally harder than the most intense X-ray bursters and black-hole binaries, whose spectra usually peter out below ~ 100 keV. The spectra of Soft Gamma Repeaters[19] peak at ~ 10 – 20 keV and disappear above ~ 200 keV. GRB spectra are routinely even harder than the Crab (pulsar +

nebula) power-law spectrum, and have been mapped to EGRET energies, ~ 20 GeV. The spectral peak in $\nu \bullet f(\nu)$ provides a break in the otherwise featureless spectra, with which to measure relative redshift of the sources. Redshift is time dilation of the EM radiation itself. Mallozzi and colleagues performed the analysis,[20] utilizing a sample of long and short bursts, down to a PF ~ 1 photon cm^{-2} s^{-1}, where BATSE detection is nearly complete. From their Figure 1b, a plot of E_{peak} vs. PF, we see that dimmer bursts have progressively lower average $\nu \bullet f(\nu)$ measures, consistent with time-dilation results.

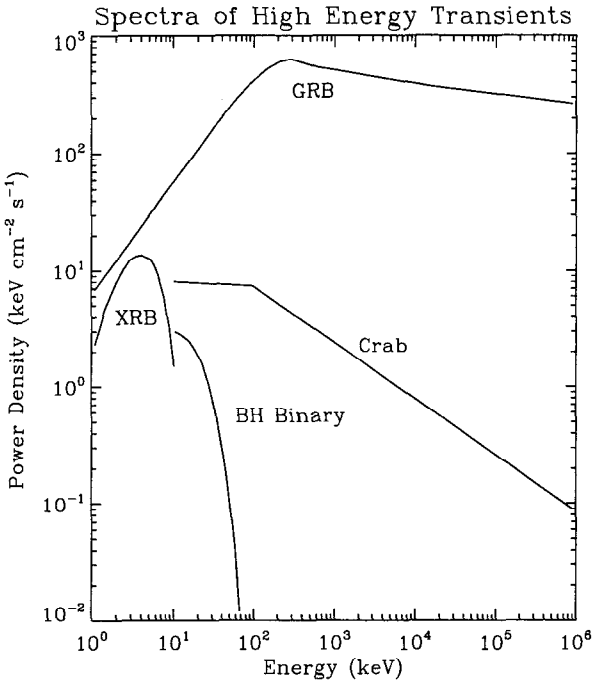


Fig. 3 – Canonical spectra ($\nu \bullet f[\nu]$) of high-energy transients: Gamma-Ray Bursts, X-Ray Binaries, Black Hole Binaries, and the Crab.

2.3 Summary, GRB Extrinsic Characteristics

This much was clear for long bursts, before the detection of counterparts: isotropy of the GRB celestial distribution, the deficit of dim bursts in the log N – log PF, time-dilation and $\nu \bullet f(\nu)$ trends with PF were mutually consistent with the cosmological distance scale hypothesis. None of these indicators individually, nor all combined, proved anything, given that subclasses of GRBs with different distributions could be invoked in models. There are still reasons to entertain possible subclasses.

For example, the distance scale for short bursts has not been established definitively. Short bursts' localizations by themselves are isotropically distributed, as are those of long bursts. But the log N – log PF relation for short bursts does not indicate a definitive deficit at low peak fluxes, since the leveling off on the dim end is more attributable to BATSE detection efficiency: some short bursts are so short that they do not fill the shortest trigger accumulation bin (64 ms). To my knowledge, no one has troubled to model their log N – log PF, probably because of the more severe detection completeness problem. Moreover, short bursts may not be sampled sufficiently deeply in redshift space to manifest curvature, thus modeling may not provide robust constraints given measurement uncertainties.

Some work on time dilation of short burst profiles has been reported.[13] Our work on short bursts does not reveal acceptably stable results as the sample enlarges, or as adjustable parameters are varied. The whole treatment is more difficult: the profiles are so short, containing little information; the definition of peak flux on short timescales is more arbitrary than for long bursts; and unbiased time-dilation measures are difficult to construct and calibrate. The optimum BATSE data type for short bursts, time-tagged two μs photon data, is of limited use due to its 32K maximum event buffer size, which effectively windows the burst record, such that the start and stop times depend on burst brightness. The $\nu \bullet f(\nu)$ vs. PF relation for short bursts shows a barely significant trend (R. Mallozzi, private comm.). Worst for short bursts, BeppoSAX has not detected any, and so there are no X-ray or optical counterparts to go chase, so far.

2.4 GRB Duration Distribution

The brightness distributions for short and long bursts might suggest that they are truly different groups. In fact, their spectral distributions are similar and overlapping, with short bursts tending to be slightly harder. Of course the defining difference is temporal: the logarithmic duration distribution is clearly bimodal, with maxima near 25 s and 300 – 500 ms

(Ref. 21). Distributions for two different duration measures are shown in Figure 4. The dashed histogram is the BATSE team's measurements of the original time profiles for th e "4B" catalog.[22] The solid histogram is our brightness-independent measure, which tends to yield shorter durations for the bright bursts, whose decaying tails are lost upon equalization of s/n levels (analysis as in Ref. 12, using a larger sample here). We measure durations at 64-ms resolution, whereas the BATSE team makes use of the time-tagged event data with two μs

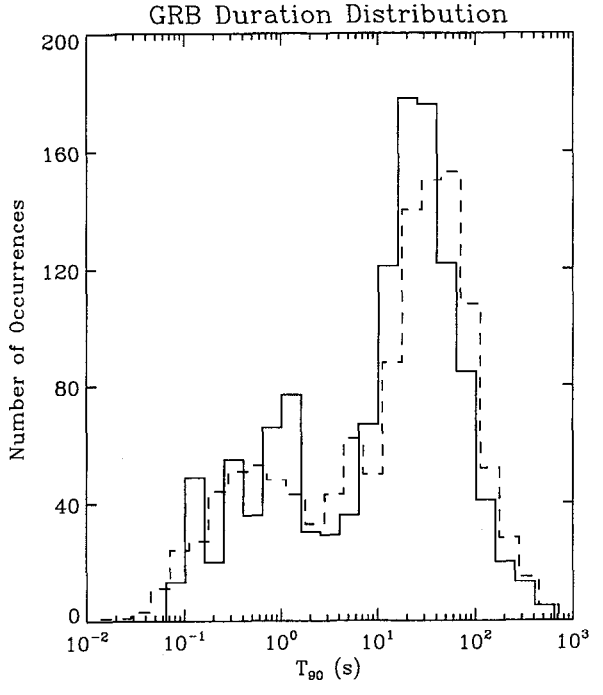


Fig. 4 – GRB duration distributions measured by two different methods (see text). The dashed histogram is derived from data in the BATSE 4B catalog. The solid histogram follows the noise equalization method of Ref. 12, but with a larger sample. T_{90} is the interval between the times to accumulate 5% and 95% of the total counts.

resolution; therefore, their characterization of the short end of the distribution is smoother and of higher fidelity. The valley near 2 – 3 s has been significant since the pre-BATSE era. A few physical explanations for the duration bimodality have been proposed. Recently, Yi and Blackman[23] proposed that the two classes correspond to accretion-induced collapse of white dwarfs to millisecond pulsars formed spinning above or below a gravitationally unstable limit, with the collapse followed by either gravitational radiation (short bursts) or electromagnetic dipole radiation (long bursts), respectively. The gravitational radiation from white dwarf collapse to pulsar may be rarely detectable for the very nearest short bursts.[23,24] NS binary coalescence may be detectable by LIGO out to a few times the Virgo cluster distance (tens of Mpc), but the event rate would be low, perhaps one event per decade.[25]

2.5 GRB Temporal Profiles

Since my interest concentrates on characterization of the widely varied temporal behavior of GRBs, I will avoid dwelling on this aspect. For the reader who would undertake serious temporal investigations, the most expeditious route to the BATSE data is through the GRO guest investigator web pages.[26] Here, discussion of two time profiles will suffice. Figure 5 illustrates a very short burst, GRB 930131, famous for its wavefront passing by the Earth during Superbowl XXXII. The triangles represent individual photons detected by the EGRET spark chamber; the histogram is the BATSE light curve, at energies > 25 keV. Because the spark chamber deadtime is ~ 100 ms per photon, the actual evolution at EGRET energies (> 30 MeV) is not measured during the intense ~ 200 ms spike-like pulse. Such fast, narrow GRBs might afford a natural experiment for probing the quantum gravity energy scale since photon arrival times may be dispersed as a function of energy (see section 5).

Visually, short bursts appear to have one or just a few emission episodes, or "pulse structures." Often a short GRB consists of a solitary pulse. However amongst short bursts can be found examples with many pulses and pulse structures, and these appear self-similar to the long bursts. The definition of pulses in short bursts is limited by statistics. Perhaps short bursts as a group would appear even more similar to long bursts if short bursts were observed with much larger detectors—then the pulse fluctuations would be as well-defined as for long bursts. Figure 2 of Ref. 27 illustrates GRB 940217, a complex burst from the long end of the duration distribution. This intense burst, also detected by EGRET, is famous for a 20 GeV photon which arrived 5000 s after the burst faded at lower energies, as detected by BATSE and Ulysses.

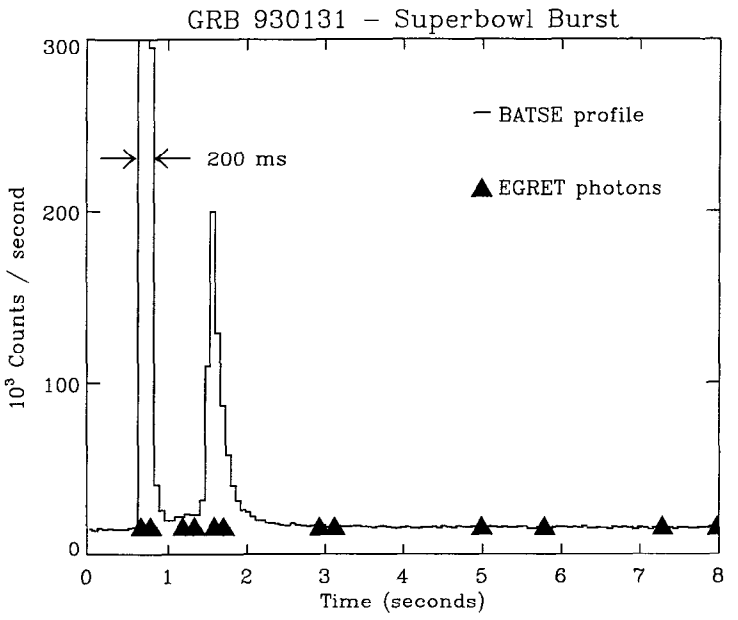


Fig. 5 – GRB 930131 (whose wavefront passed by the Earth during Superbowl XXXII) as detected by BATSE and EGRET. Vertical axis is truncated: maximum BATSE count rate was $> 7 \times 10^5$ s^{-1} , and limited by deadtime effects. EGRET spark chamber deadtime per event is ~ 100 ms. Thus neither instrument recorded actual fluxes during the initial intense spike.

Short and long bursts are similar in two other respects. Both tend to exhibit hard -to-soft spectral evolution, within a pulse and across the entire duration of an event: Pulses tend to be longer and their onsets later at lower energies, and bursts tend to be longer at lower energies. However, both groups exhibit the "asymmetry pulse paradigm": the more asymmetric a pulse, the wider it is, with lower energies peaking at later times. Conversely, in nearly symmetric pulses, the peaks are aligned across energy bands.[28,29]

Since the radiation mechanisms of GRBs may be completely separate from processes governing their counterpart afterglows, the temporal/spectral development at gamma-ray energies will remain central to eventual understanding. Recent temporal analysis work employing objective algorithms for the determination of the number of pulses in a burst, as well as pulse rise and decay times, has been done by Scargle[30], using a newly formulated "Bayesian Block" approach to estimate position s of change-points —where the mean changes significantly. Constraints that GRB time profiles impose on internal vs external shock models, via considerations of kinematics of the expanding emission front, have been expounded by Fenimore and co-workers.[31]

3. GRB Counterparts at Longer Wavelengths

3.1 Discoveries of GRB Counterparts and High Redshifts

On February 28, 1997, the gamma-ray burst community was surprised by the breakthrough discovery of an X-ray counterpart to GRB 970228, detected by the BeppoSAX GRB monitor and the narrower field X-ray instruments. The evidence from two observations, separated by three days, revealed the appearance and decay of a previously unknown X-ray source consistent with the position of the GRB.[32] Subsequent X-ray detections of GRBs, outnumbering optical detections by about two to one, have established a clear pattern of power-law temporal decays, predicted by Meszaros and Rees as the signature of cosmological fireballs/blastwaves in the aftermath of a GRB.[33]

The X-ray localization accuracy (~ 3 arc minutes) from BeppoSAX's wide field camera was sufficiently refined to allow deep pointed observations with large optical telescopes. The path to The Holy Grail of GRBs was then established by van Paradijs, Sahu, and Galama,[34,35,36] upon observing the X-ray error circle on Mar 1 and Mar 8 with the William Herschel (4.2 m) and Isaac Newton (2.5 m) telescopes, and finding a decaying optical source at visual magnitude ~ 21—extremely faint compared to previous claims of possible archival optical associations with GRBs. Two weeks later the field was observed by HST. The closeup revealed the still-bright GRB counterpart, compared to an extended nebulosity—the host galaxy. Comparison with a second HST observation made on Apr 2 led Caraveo and colleagues[37] to declare that the object had manifested a proper motion of 18 milliarcseconds, and was probably a compact star moving at ~ 250 km s^{-1} at a distance of ~ 100 pc; ~ 0.1 pixel centroiding was claimed, and possible to achieve.

HST observed the transient again in September and one year after the GRB.[38] By September the afterglow had faded considerably. No measurable proper motion had occurred in six months. (The positional analysis problem, as explained by D. Macchetto of the HST Institute, involved delicate subtraction of the underlying galaxy's light, which had a gradient in the direction of the GRB.) HST images revealed details of a foreground galaxy with the faint transient superposed, and other distant galaxies at redshifts of ~ unity, each of angular size ~ 1 arc second. There was the small probability of chance superposition, but with each subsequent optical counterpart to a GRB, evidence of a host has emerged—spatial coincidence, temporal decay pattern, and/or spectral reddening—indicating that the GRB source is embedded in an obscuring region, i.e., galaxian dust and gas. (Note, galaxies at redshifts z ~ 1 and higher tended to be smaller, closer together, and less organized. So whereas it would require less time for a supernova-kicked NS binary to escape from its smaller galaxy with weaker gravitational pull, the binary would soon find itself near another galaxy, at least in projection on the sky. So the question of transient/galaxy association was pertinent with this sample of one, and may still be for some subclass of GRB.)

During Mar 31–Apr 2, 1997, Tonry and colleagues obtained a 5500-s spectrum of the transient with the Keck II 10-m telescope; no obvious emission lines were identified, and no redshift was found.[39] After significant fading (Aug 13, 5.5 months after the GRB), Kulkarni and colleagues[40] performed an 1800-s integration with the Keck II to acquire the spectrum of transient and galaxy; still no strong emission lines were observed, and thus no redshift was obtained. The Holy Grail was missing.

Meanwhile, GRB 970508 was detected by BeppoSAX, and its behavior followed somewhat different patterns. The X-ray temporal decay revealed a significant bump in an otherwise power-law decay, at a delay of ~ 1 – 3 days. This relatively long lasting bump (compared to the GRB duration) contained ~ 25% of the fluence of the X-ray afterglow, and ~ 5% of the total energy released by the GRB. Piro and colleagues[41] suggested that afterglow and gamma rays arise from the same primary process, just on timescales differing by ~ 10^4. Optical flaring occurred nearly simultaneously with the X-ray bump. Fenimore and colleagues[31] used these kinds of temporal features, both during the gamma-ray events and the X-ray afterglows, to argue on a kinematic basis that external shocks could not be the driving radiative mechanism. Internal shocks for the GRB itself are favored on theoretical grounds.

The optical light curve of GRB 970508 showed the brightness remaining practically constant during the first day, and then rising to maximum light (R magnitude ~ 20) before beginning the power-law decay.[42] Near 100 days, hint of a departure from power-law decay appeared. This departure was observed to continue to day ~ 300, and is well-fitted by a power-law plus constant. The temporal fit and the photometric colors implied a small, faint host galaxy of the starburst variety, with ~ 3 kpc diameter.[43] In August 1998, the galaxy was imaged by HST, revealing a small (0.07 arc second) but extended, elliptical object with R magnitude ≈ 25.3 (Ref. 44). And of course, GRB 970508 was the Holy Grail: The light of the transient itself showed an absorption line system and oxygen emission,[45] both with redshift z = 0.835, roughly half the distance back to the start of spacetime.

GRB 970508 was also detected in the radio by Frail and colleagues[46] and revealed the signature of an (initially) adiabatically expanding fireball.[47] The radio signal was first detected at one week delay, at 4.9 and 8.5 GHz. Variations over the succeeding month, and their subsequent quenching, are indications of diffractive scintillation due to the initial small angular size (analogous to stars twinkling as their light passes through the atmosphere) and then expansion to an apparent size of ~ 10^{17} cm (~ 6×10^3 AU), with expansion velocity near the speed of light. The observed behavior at a several-week delay indicates evolution from highly relativistic to sub-relativistic, $\Gamma \sim 2$, as the fireball expands transversely, becoming less beamed and more spherical. Taking into account the probable nonspherical expansion, the implied fireball energy was ~ 2×10^{51} ergs (Refs. 46 and 47).

GRB 971214, the third with an optical counterpart, was imaged by HST on May 7, 1998 (Ref. 48). The transient had faded by then but a faint host galaxy, R magnitude ≈ 25.5, was found at the transient's position. The probability of chance superposition is estimated ~ 10^{-2} – 10^{-3}: ~ 4×10^5 galaxies per degree^2, brighter than 25.5 magnitudes, and the galaxy's apparent radius is ~ 0.35 arc seconds. The index for the afterglow power-law decay was $\alpha \sim 1.4 \pm 0.2$, slightly steeper than those for GRBs 970228 and 970508, with $\alpha \sim 1.1$ and 1.2, respectively.[49] GRB 971214's optical transient was fainter and redder than the previous two, suggesting a dense star-forming environment for the GRB progenitor.

Kulkarni and colleagues[48] determined the redshift of the apparent host galaxy after the GRB faded (a usable spectrum of the transient itself could not be obtained). The galaxy's light revealed a Lyman α emission line (consistent with drop in the blueward continuum, due to intervening hydrogen absorption), at a redshift identical to that of several galaxian absorption lines, species routinely found in high redshift, star-forming galaxies. The measured redshift is z = 3.148, ~ 90% of the distance back to the Big Bang, considerably higher than that for GRB 970508 at z = 0.835. The gamma-ray peak flux of GRB 971214 was ~ 2 × that of GRB 970508. Luminosity goes as $(1 + z)^3$, so GRB 971214 would be ~ 25 × more luminous than GRB 970508— the first observational evidence of a GRB intrinsic luminosity distribution. If the burst emission were isotropic, then the total gamma-ray fluence in the burst would be ~ 3×10^{53} ergs. The GRB lasted about 20 s, so the isotropic

average flux would be ~ 10^{52} ergs s^{-1}. The energy in the optical afterglow is estimated as ~ 2 × 10^{51} ergs, largely independent of beaming considerations, since the fireball expansion is presumed to be subrelativistic and nearly symmetric. And since the afterglow is best understood as an adiabatic phenomenon, a very small fraction of the energy available in the fireball is radiated. This is one energetics basis for the claim[50] that NS-NS or NS-BH mergers come up short, given the likely conversion efficiency to γ's (< few percent) and the available rest mass of the binary, ~ 5 × 10^{33} gm.

3.2 Summary: GRB Counterparts

Table 2 summarizes the nine optical counterparts to GRBs identified through July 1998. Some assumptions about conversions between energy bands and cross-instrument calibrations were necessary. Host galaxy magnitudes for the more recent GRBs are in the process of revision as the optical transients continue to decay. So not all entries are definitive. Variations in the γ/X peak flux ratio and X-ray spectra are taken as evidence of varying amounts of obscuration near the GRB source, indicative of star-forming regions.[51] In each case we have imaging, spectroscopic, and/or light-curve evidence of a host galaxy. The images are very faint extended/elongated smudges, but coincident in position with the optical counterpart. Given precise optical positions, radio counterparts are now routinely detected for a good fraction of GRBs. A third redshift (z = 0.966) for the optical transient associated with GRB 980703 has been obtained. The absorption lines in its spectra also indicate the presence of a star-forming galaxy.[52] Note that GRBs 970508 and 980703 have comparable redshifts, and their gamma-ray peak fluxes differ by a factor of only ~ 2, an indication that the GRB luminosity distribution may not be too broad. However, Fruchter[53] argues that the IR-optical colors for the GRB 980329 transient are best interpreted as absorption due to the Lyman α forest extending to a redshift z ~ 5, implying a luminosity distribution in peak flux greater than 100.

The combined astrometric and spectroscopic evidences indicate that those GRB sources with associated optical counterparts are inside star-forming galaxies. Bloom and colleagues[54] derive a median merger timescale of ~ 10^{8} years, and thus the sites of NS-NS mergers should be in close association with star formation regions, rather than outside host galaxies. Van den Heuvel maintains that many NS-NS binaries may merge on even shorter timescales. His calculation extrapolates from the observed properties of B-emission X-ray binaries, which comprise one neutron star and a massive companion, to arrive at the expected distribution of orbital periods after the second supernova. The result is that ~ 50% of NS - NS binaries may form with sufficiently short orbital periods such that they merge within 10^{7} years, leaving GRB sources well inside star-forming regions in host galaxies.[55] However, optical counterparts have been detected only for the two-thirds of GRBs in which rapid, longer wavelength observation strategies were possible; the other third may be those where the progenitor long ago escaped the host. We would hypothesize that optical emission requires a dense environment and so is not generated in those cases.

Table 2: Gamma-Ray Bursts with Optical Counterparts.

GRB	Peak Fluxes [7]					Host Galaxy	
	γ-ray [1]	X-ray [2]	γ/X Ratio	Optical [3]	Radio [4]	Brightness [5]	z [6]
970228	3.5	2.3	80	20.5	—	24.6	—
970508	1.2	3.0	25	19.8	1.2	25.8	0.835
971214	2.3	2.5	56	21.7	—	25.5	3.412
980326	1.3	4.7	17	21.0	—	25.3	—
980329	13.3	70.0	12	23.6	0.25	~ 29 ?	~ 5 ??
980425	1.1	2.6	26	13.7	49	14.3	0.0085
980519	4.7	2.9	100	20.4	0.1	24.7	—
980613	0.63	0.7	57	22.9	—	24.4	—
980703	2.6	4.0	40	20.1	1.0	23.0	0.966

[1] photon cm^{-2} s^{-1} (50–300 keV); conversion factor, photons to ergs, ≈ 6.15 × 10^{-7}.

[2] ×10^{-8} ergs cm^{-2} s^{-1} (2–10 keV).

[3] R band magnitude.

[4] milli Jansky, at 8.4 GHz (10 GHz for GRB 980425).

[5] R band magnitude, corrected for galactic extinction.

[6] Redshift.

[7] See these two web pages for many current references on GRB afterglows: GRB Coordinates Network notices, used to distribute GRB news rapidly, gcn.gsfc.nasa.gov/gcn/ ; and www.obs.aip.de/~jcg/grbgen.html.

4. SN 1998bw ≡ GRB 980425?

The clearly outstanding burst in Table 2 is GRB 980425 (optical peak flux ~ magnitude 14!!), for which the claimed optical counterpart—an unusual SN of type Ic —is suspect. Figure 6 illustrates the BATSE time profile of this relatively weak, monolithic burst in four energy bands (~ 25–50, 50–100, 100–300, > 300 keV) at 0.5-s resolution, with fits to the lower three energy bands using a single lognormal pulse model. It is probably not remarkable that the burst is not detected above 300 keV, since "no high-energy emission" has been shown to be attributable to brightness bias.[5] The tendency for high energy to lead low energy is obvious. The BeppoSAX Wide Field Camera localized this burst within an error circle ~ 15 arc minutes in diameter. Contained within the circle, and imaged by the Narrow Field Instruments, were two X-ray sources with error circles ~ 2 arc minutes in diameter. One of these sources faded away by the second NFI observation, and that's the last we know of it – was it the true counterpart? Also, an unusual supernova—1998bw—was found adjacent to but consistent with neither X-ray error region. Radio emission from SN 1998bw was also detected, ~ 10 × stronger than that for any other type Ic SN (but only a handful have radio detections), leading investigators to conjecture that the association is not by chance.[56]

To make the connection even more appealing, the extrapolated optical t_0 for the SN is plus or minus a couple days of the GRB. And, an available kinetic energy of few × 10^{52} ergs is suggested from modeling the optical light curve, assuming ejection of 0.7 $M_\odot$ of ^{56}Ni, about 10 × that usually produced in core-collapse-induced SNe[57]. Some SN aficionados are having a field day, going so far as to assert that all GRBs may be associated with unusual SNe, and that we are merely seeing ~ 10^{-3} of the total GRB population: The distant, very luminous GRBs that we detect are beamed at us in a narrow cone, and therefore we see a small fraction of the total, but from a much larger volume. GRB 980425 would be one of the more frequent ones in this large volume, but much closer and therefore visible, and we see a different kind of optical counterpart by virtue of its isotropic, lower intensity emission.[58] The redshift of SN1998bw is z = 0.0085, corresponding to a distance of 38 Mpc, about twice the distance to the Virgo cluster. The chance of finding this kind of SN inside the BeppoSAX WFC error circle is now very small, given the unusual energetics, and the close temporal coincidence. Too small to calculate confidently, or to ignore.

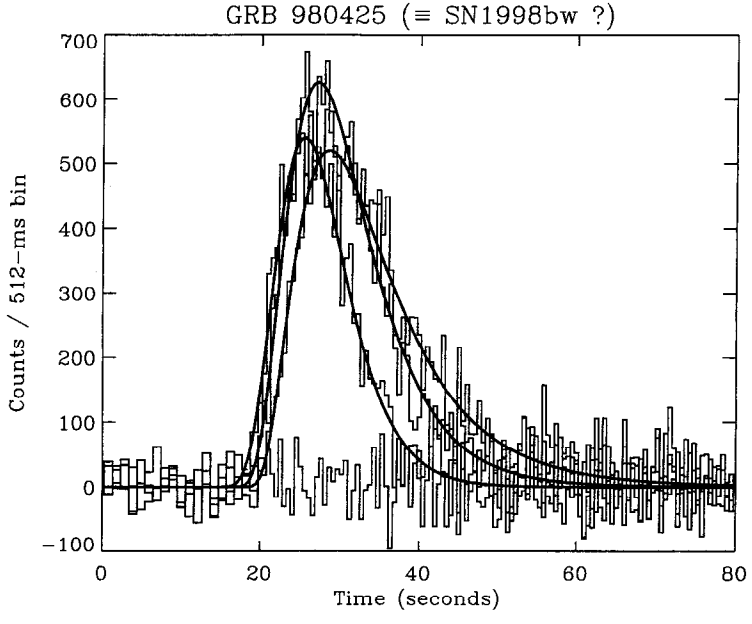


Fig. 6 – BATSE time profiles and lognormal fits for GRB 980425—a relatively weak, smooth, single-pulse burst. The fitted peak occurs progressively earlier at increasing energy (channels 1–3: 25–50, 50–100, 100–300 keV), an example of hard-to-soft spectral evolution in a pulse. The absence of emission above 300 keV is not remarkable, since "no high-energy emission" is likely attributable to brightness bias.[5]

Bloom and colleagues[59] summarize the attributes of SN1998bw and GRB 980425, and come to the conclusion that perhaps ~ 1% of BATSE GRBs could be SN type Ic. The required GRB attributes are a single, broad pulse exhibiting hard-to-soft spectral evolution with longer decay than rise; possibly no high-energy emission; and no X-ray afterglow. The upper limit for any set of GRBs being associated with nearby SNe must be ~ 10%, given isotropy. BATSE's detection limit for type Ic SN is estimated to be ~ 100 Mpc (Ref. 59). The nearby luminous matter distribution to ~ 100 Mpc is highly nonuniform, following a linear feature from the Pegasus to Virgo to Coma clusters—the Supergalactic Plane—the one distance scale never previously demonstrated as a possible home for GRBs.

We visually searched our collection of background-subtracted, noise-equalized (to the peak intensity of GRB 980425) time profiles of long bursts, liberally including in a first-cut sample anything remotely resembling a single-pulse event. In automated passes, we fitted a lognormal pulse model, excluding from further consideration those bursts with poor χ^2 or correlated residuals in channels 1, 2, or 3—considering such to be evidence of more than one pulse. The automated pass largely confirmed our expectations from the visual inspection. We also restricted consideration to those bursts with FWHM within a factor of three of GRB 980425's (4 – 35 s). Thirty-two bursts passed these cuts, and about five of these have "no high-energy emission." Most are dim bursts, and they have a clear resemblance to GRB 980425. Most but not all exhibited hard-to-soft spectral evolution.

For the two sets, with and without high-energy emission, we searched an exhaustive list of 900 SN (most are the more luminous SN type II or type Ia) from the BATSE era. The list is more complete by ~ factor of two during the last two to three years due to heightened interest in high z objects. We found one or two possible SN associations for the NHE + HE set, given the time constraint that the GRB must precede the first SN observation in any case. Correcting for SN detection incompleteness in previous years, we estimated that ~ 5 GRBs could be associated with SN type Ib/Ic.

We were able to find redshifts for two-thirds of the type Ib/Ic SNe, and they are indeed nearby: the redshifts are $z <\sim 0.0175$, i.e., up to roughly twice the distance of SN 1998bw ($z = 0.0085$). Both the nearby SNe and the single-pulse GRBs might be expected to show a preference for the Supergalactic Plane. In Figure 7 the SNe are plotted in the supergalactic coordinate system. The continuous line is the Galactic Plane. Note that 16 of these 21 SNe fall within 30° of the Supergalactic Plane and that they tend to avoid the region near the Galactic Plane (due to obscuration). As expected, the quadrupole moment, $\langle \sin^2 b - 1/3 \rangle$, for the SN distribution with respect to the Supergalactic Plane (b = supergalactic latitude) is significantly oblate at the 2-σ level, $-0.138^{+0.067}_{-0.043}$, where errors are estimated via a bootstrap approach. Figure 8 shows the 32 single-pulse GRBs in the same coordinate system. There appears to be no tendency for any subset of single-pulse GRBs to fall near the Supergalactic Plane. The quadrupole moment for the entire GRB single-pulse set is $-0.003^{+0.05}_{-0.05}$ – not significantly different from zero.

For economy of hypotheses, we suggest that some source population other than nearby SNe generally accounts for the smooth, single-pulse GRBs. We conclude that either the SN1998bw – GRB980425 association is coincidental, or events of this type comprise a very rare and distinct phenomenon.[60]

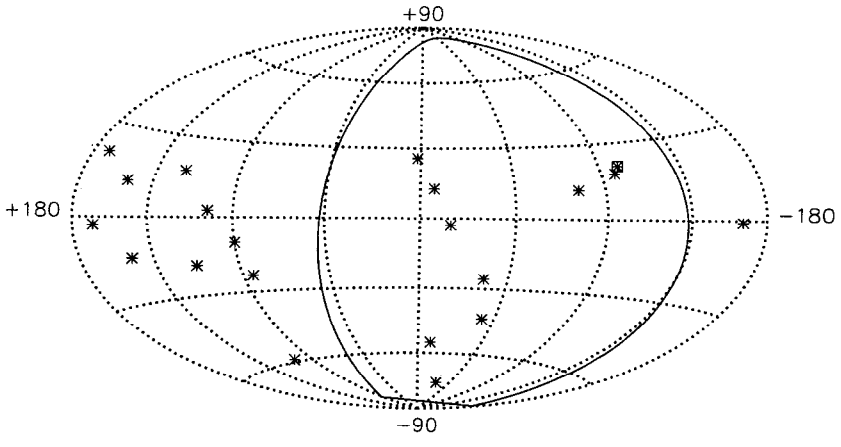


Fig. 7 – The 21 SNe of type Ib or Ic found in the IAUC supernovae catalog, plotted in the supergalactic coordinate system. The continuous line is the Galactic Plane. Sixteen SNe lie within 30° of the Supergalactic Plane. The symbol inside the square indicates SN1998bw. The quadrupole moment for the SN distribution with respect to the Supergalactic Plane is significantly oblate at the 2-σ level, $-0.138^{+0.067}_{-0.043}$.

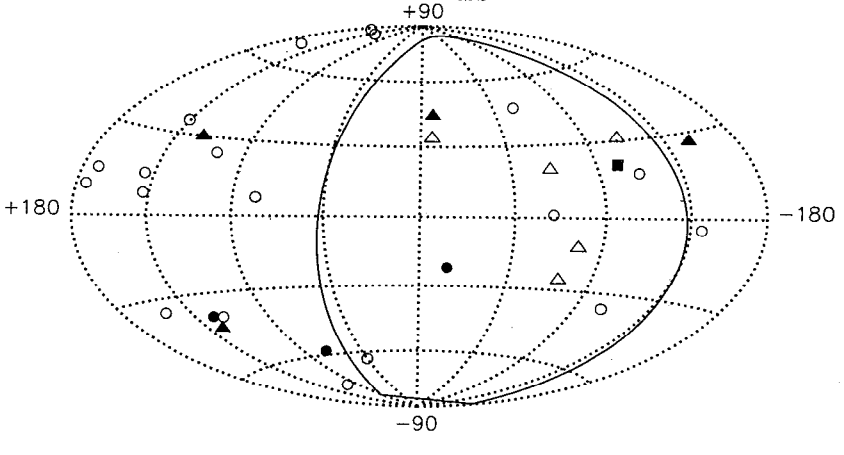


Fig. 8 – The 32 single-pulse GRBs in the supergalactic coordinate system. Circles indicate GRBs with peak > 720 counts s^{-1}; triangles indicate bursts below this threshold—too dim to determine reliably the number of pulses present. Filled (open) symbols are NHE (HE). No subset of single-pulse GRBs shows a tendency to cluster near the Supergalactic Plane.

5. *Future High-Energy GRB Instrument*

GLAST—Gamma-ray Large Area Space Telescope, the next generation high-energy gamma-ray telescope—comprises a pair-production tracker with a hodoscopic calorimeter and a segmented anti-coincidence shield. Using the most powerful particle accelerators in the Universe as cosmic laboratories, GLAST will explore the link between gravitation and quantum physics in the extreme environments of supermassive black holes, neutron stars, and gamma-ray bursts. On cosmological scales, GLAST will explore the era of star formation in the Universe, the physics of dark matter, and the creation and evolution of galaxies. Most important from my parochial point of view is that GLAST will detect GRBs to supra-GeV energies. Recall the Superbowl Burst plotted in Figure 5. Because GLAST will have insignificant deadtime and will be much more sensitive than EGRET, it offers the possibility of probing the quantum gravity scale, E_{QG}: Photon arrival times may be energy dependent, by virtue of dispersion in the vacuum of spacetime, hypothesized to be granular at the Planck scale.[61] Using the very fast, narrow rise features in some bursts, and taking advantage of the high-energy GRB radiation that travels gigaparsecs, we may be able to search for this distance-dependent and energy-dependent effect using GLAST: The brightest ~ 25 GRBs per year should be detectable at energies ~ 100 GeV, with dispersions of ~ 10 ms GeV^{-1} Gpc^{-1}. *Just as important for probing star and galaxy formation in the early Universe, the tracker's* angular resolution will be good enough to localize ~ 100 GRBs per year to better than ~ 10 arc minutes, as shown in Figure 9—sufficiently accurate to enable followups at all longer wavelengths.

Finally, detection of one or two cosmological lenses ("images" separated temporally by months to years) amongst the dimmest, untriggered BATSE bursts would be expected (G. Marani, private comm.) if the cosmological constant $\Lambda \sim 0.6$ and if the GRB rate-density follows that of star formation in the Universe.

Acknowledgments

I am very grateful to Jerry Bonnell and Gabriela Marani for many enlightening conversations and for much help in assembling GRB counterpart flux measurements.

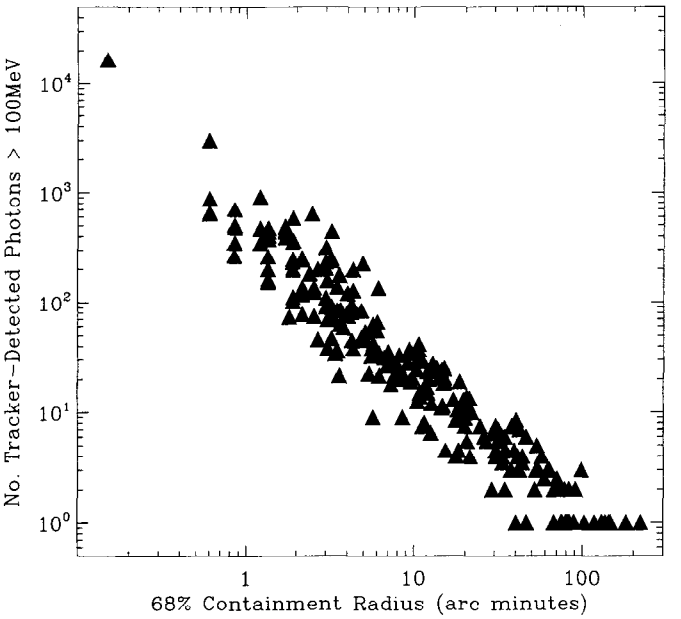


Fig. 9 – Simulated GRB localization accuracy for GLAST's silicon strip tracker: two-dimensional 68% containment angle vs number of detected photons above 100 MeV. The simulation algorithm takes into account measured GRB duration, brightness, and spectral shape distributions, assumes satellite zenith-pointing mode, and utilizes GLAST tracker reconstruction methods. Approximately 100 GRBs per year should be localized to better than *ten arc minutes.*

References

[1] P. J. T. Leonard and J. T. Bonnell, Sky and Telescope, p. 28 (February 1998).

[2] N. J. Shaviv and A. Dar, MNRAS 277, p. 287 (1995).

[3] M. F. Briggs et al., Astrophys. J. 459, p. 40 (1996).

[4] G. N. Pendleton et al., Astrophys. J. 464, p. 606 (1996).

[5] J. T. Bonnell and J. P. Norris, Astrophys. J. (1998) (submitted).

[6] P. Madau, MNRAS 282, p. 1388 (1996).

[7] R. A. M. J. Wijers et al., MNRAS 294, p. L13 (1996).

[8] E. E. Fenimore and J. S. Bloom, Astrophys. J. 453, p. 25 (1995).

[9] M. R. Krumholz et al., Astrophys. J. (1998) (astro-ph/9807117) (submitted).
[10] J. M. Kommers et al., Astrophys. J. (1998) (astro-ph/9809300) (submitted).
[11] J. P. Norris et al., Astrophys. J. 424, p. 540 (1994).
[12] J. T. Bonnell et al., Astrophys. J. 490, p. 79 (1997).
[13] H. Che et al., Astrophys. J. 483, p. L25 (1997).
[14] M. Deng and B. E. Schaefer, Astrophys. J. (1998) (astro-ph/9806010) (submitted).
[15] M. L. Litvak et al., in Gamma-Ray Bursts, AIP 428, p. 176 (1998).
[16] J. P. Norris, AIP 384, p. 13 (1996).
[17] J. M. Kommers et al., Astrophys. J. 491, p. 704 (1997).
[18] J. P. Norris et al., Astrophys. J. (1999) (to be submitted).
[19] C. Kouveliotou et al., Nature 393, p. 235 (1998).
[20] R. S. Mallozzi et al., Astrophys. J. 454, p. 597 (1995).
[21] C. Kouveliotou et al., Astrophys. J. 413, L101 (1993).
[22] C. A. Meegan et al., Current BATSE GRB Catalog: www.batse.msfc.nasa.gov/data/grb/catalog/.
[23] I. Yi and E. G. Blackman, Astrophys. J. 494, p. L163 (1998).
[24] D. Lai and S. L. Shapiro, Astrophys. J. 442, 259 (1995).
[25] C. S. Kochanek and T. Piran, Astrophys. J. 417, L17 (1993).
[26] Compton GRO Guest Investigator Resources (GRB temporal profiles): http://cossc.gsfc.nasa.gov/cossc/batse/4Bcatalog/4b_basic.html.
[27] K. Hurley et al., Nature 372, p. 652 (1994).
[28] J. P. Norris et al., Astrophys. J. 459, p. 393 (1996).
[29] J. P. Norris, Astrophys. Space Sci. 231, p. 95 (1995).
[30] J. D. Scargle, Astrophys. J. (1998) (astro-ph/9711233) (submitted).
[31] E. E. Fenimore et al., Astrophys. J. (1998) (astro-ph/9802200) (submitted).
[32] E. Costa et al., Nature 387, p. 783 (1997).
[33] P. Meszaros and M. J. Rees, Astrophys. J. 476, p. 232 (1997).
[34] J. van Paradijs et al., Nature 386, p. 686 (1997).
[35] K. C. Sahu et al., Nature 387, p. 476 (1997).
[36] T. Galama et al., Nature 387, p. 479 (1997).
[37] P. Caraveo et al., IAU Circ. 6629 (1997).
[38] A. S. Fruchter et al., Astrophys. J. (1998) (astro-ph/9807295) (submitted).
[39] J. L. Tonry et al., IAU Circ. 6620 (1997).
[40] S. R. Kulkarni et al., IAU Circ. 6732 (1997).
[41] L. Piro et al., A&A 331, L41 (1998).
[42] H. Pedersen et al., Astrophys. J. 496, p. 311 (1998).
[43] S. V. Zharikov, V. V. Sokolov, and Yu. V. Baryshev, A&A 337, p. 356 (1998).
[44] J. S. Bloom et al., Astrophys. J. (1998) (astro-ph/9807315) (submitted).
[45] M. R. Metzger et al., Nature 387, p. 878 (1997).
[46] D. A. Frail et al., Nature 389, p. 261 (1997).
[47] E. Waxman, S. R. Kulkarni, and D. A. Frail, Astrophys. J. 497, p. 288 (1997).
[48] S. R. Kulkarni et al., Nature 393, p. 35 (1998).
[49] J. P. Halpern et al., Nature 397, p. 41 (1998).
[50] A. N. Ramaprakesh et al., Nature 393, p. 35 (1998).
[51] A. Owens et al., Astronomy and Astrophys. (1998) (astro-ph/9809356) (submitted).
[52] S. G. Djorgovski et al., Astrophys. J. (1998) (astro-ph/9808188) (submitted).
[53] A. S. Fruchter, Astrophys. J. (1998) (astro-ph/9810224) (submitted).
[54] J. S. Bloom et al., Astrophys. J. (1998) (astro-ph/9805222) (submitted).
[55] E. P. J. van den Heuvel, in X-Ray Binaries and Recycled Pulsars (Kluwer Acad. Publishers, 1992), p. 233.
[56] S. R. Kulkarni et al., Nature (1998) (astro-ph/9807001) (in press).
[57] K. Iwamoto et al., Nature (1998) (astro-ph/9806382) (in press).
[58] L. Wang and J. G. Wheeler, Astrophys. J. (1998) (astro-ph/9806212) (submitted).
[59] J. S. Bloom et al., Astrophys. J. (1998) (astro-ph/9807050) (submitted).
[60] J. P. Norris, J. T. Bonnell, and K. Watanabe 1998, Astrophys. J. (1998) (astro-ph/9807322) (submitted).
[61] G. Amelino-Camelia et al., Nature 393, p. 763 (1998).

GAMMA-RAY BURSTS: THE FIREBALL SHOCK MODEL AND ITS IMPLICATIONS

P. Mészáros*
Department of Astronomy and Astrophysics
Pennsylvania State University, University Park, PA 16802

ABSTRACT

Major advances and discoveries in the field of gamma-ray bursts have been made at an increasingly fast pace over the last two years. The successful discovery of X-ray, optical and radio afterglows, which were predicted by theory, has made possible the identification of host galaxies at cosmological distances. The energy release inferred in these outbursts rivals that of supernovae, while its photon energy output may considerably exceed it. Current models envisage this to be the outcome of a cataclysmic stellar event leading to a relativistically expanding fireball, in which particles are accelerated at shocks and produce non-thermal radiation. The substantial agreement between observations and the theoretical predictions of the fireball shock model provide confirmation of the basic aspects of this scenario. The continued observations show a diversity of behavior, providing valuable constraints for more detailed models. Crucial questions now being addressed are the degree of beaming and its implications for the energetics, the time structure of the afterglow, its dependence on the central engine or progenitor system behavior, and the role of the environment on the evolution of the afterglow.

*Supported by NASA NAG5-2857.

© 1998 by P. Mészáros.

1 Introduction

With the first discovery[1] of the afterglow of a Gamma-Ray Burst (GRB) at X-ray and optical wavelengths in GRB 970228 in February 1997, a new and revolutionary phase has been entered in the investigation of these mysterious sources. The rapid successive detection of a number of other afterglows, extending in some cases to radio and microwave wavelengths as well, has made it possible to follow some of these sources over much longer time scales of many months, making the identification of counterparts and host galaxies possible (see the review by Norris in this volume). The study of afterglows has provided confirmation for much of the earlier work on the generic fireball shock model of GRB, in which the γ-ray emission arises at radii of 10^{13}–10^{16} cm (Refs. 2-7). In particular, this model led to the prediction of the quantitative nature of the signatures of afterglows,[8,9] in advance of the observational studies,[1,10] and as it turned out, in substantial agreement with these.[11–14] The first measurement of a redshift[15] in GRB 970508, provided confirmation of the hypothesis that these were at cosmological distances.

More recently, significant interest was aroused by the report[16] of an afterglow for the burst GRB 971214 at a redshift $z = 3.4$, whose fluence corresponds to a γ-ray energy of $10^{53.5}(\Omega_\gamma/4\pi)$ erg, where Ω_γ is the solid angle into which the gamma-rays are beamed. Such energies were discussed[17] in the context of compact mergers, such as neutron star-neutron star (NS-NS) or black hole-neutron star (BH-NS) mergers, which can power a relativistic fireball resulting in the observed radiation. In some of the detected afterglows, there is evidence for a relatively dense gaseous environments, as suggested, e.g., by evidence for dust[18] in GRB 970508, the absence of an optical afterglow and presence of strong soft X-ray absorption[19,20] in GRB 970828, the lack of an optical afterglow in the (radio-detected) afterglow[21] of GRB 980329, and spectral fits to low energy portion of the X-ray afterglow.[22] The latter observations may be suggestive of "hypernova" models,[23,24] involving the collapse of a massive star or its merger with a compact companion. While it is at present unclear which, if any, of these progenitors is responsible for GRB, or whether perhaps different progenitors represent different subclasses of GRB, there is general agreement that they all would be expected to lead to the generic fireball shock scenario mentioned above. Much of the current effort centers around trying to identify such progenitors more specifically, and trying to determine what effect, if any, they have on the observable fireball and afterglow characteristics.

2 The Fireball Shock Scenario

The isotropy of the angular distribution of GRB suggested already early on that GRB's were outside our own galaxy,[25] and the lack of structure associated with nearby galaxies indicated that their distance would be such that non-Euclidean and evolutionary effects may indeed be important. This was also indicated by studies of the counts of the number of bursts per unit fluence or peak flux (the brightness distribution),[25,26] and by evidence for time dilation in the bursts.[27] Fits to cosmological counts including curvature, luminosity function, and evolution effects indicated that the latter could be important.[28] In particular, the evolution may be related to the star formation rate as a function of cosmological age.[29–31] The cosmological distances of some bursts have been, since then,[15,16] directly measured. This, of course, poses significant constraints on the type of sources that may be able to produce this. The typical numbers required for the luminosity, total energy per bursts, and event rate are:

$$L \sim 10^{52}(\Omega/4\pi)\ \mathrm{erg\ s^{-1}}, \tag{1}$$

$$E \sim 10^{53}(\Omega/4\pi)\ \mathrm{erg}, \tag{2}$$

$$\mathcal{R} \sim 10^{-6}(\Omega/4\pi)^{-1}\ \mathrm{galaxy^{-1}yr^{-1}}, \tag{3}$$

where Ω is the solid angle into which the energy is channeled.

An early suggestion for a source of GRB at cosmological distances was the merger of NS-NS binaries (e.g., Refs. 32 and 33) whose estimated numbers and merger rates are in fact quite close to the observed GRB rate.[34,35] A related possibility are BH-NS binary mergers,[34,38,42] which have only slightly different event rates. An alternative suggestion[36] was a "failed Supernova Ib," resulting from the collapse of a fast rotating massive star which fails to produce a core collapse SN. Both a compact merger and a fast rotating stellar collapse were known from numerical simulations to lead to a fast rotating torus around a central, high-density object, which eventually would develop into a black hole. The binding energy liberated in both of these events is of order of 10^{54} ergs, a good fraction of which will be carried away in a short pulse of $\nu\nu$ and gravitational waves.

If a small fraction of this emerges as electromagnetic energy, some of which is deposited in a region with sufficiently low-baryon density, a relativistic fireball would form, whose radiation spectrum in the observer frame will peak in the MeV range.[32,33,37] One possible channel for converting some of this energy into electromagnetic form is given by the $\nu\bar{\nu} \rightarrow e^+e^-$ process.[39] A low-baryon load condition (as required to make the fireball highly relativistic) can occur naturally in a binary merger since a centrifugal barrier develops along the orbital symmetry axis, which is relatively free of baryons and provides an escape route for the $e^{\pm}, \gamma$ fireball.[40] This would also imply a collimation of the relativistic fireball, which would enhance the apparent flux (relative to what it would be if it were isotropic) by factors which conservatively might be of the order of 10–100. The same processes were estimated to be able to produce a collimated relativistic fireball in the failed SN Ib model.[36] Another suggestion for powering a fireball was that this might result from magnetic flaring activity[41] in the disrupted torus produced around the merged binary.

Irrespective of the details of the progenitor, the resulting fireball is expected to be initially highly optically thick. From causality considerations, the initial dimensions must be of order of $ct_{var} \lesssim 10^7$ cm, where t_{var} is the variability timescale, and the luminosities must be much higher than a solar Eddington limit. Since most of the spectral energy is observed above 0.5 MeV, the optical depth against $\gamma\gamma \rightarrow e^+e^-$ is large, and an $e^{\pm}$, γ fireball is expected. Due to the highly super-Eddington luminosity, this fireball must expand. Since in many bursts one observes a large fraction of the total energy at photon energies $\epsilon_\gamma \gtrsim 1$ GeV, somehow the flow must be able to avoid degrading these photons [$\gamma\gamma \rightarrow e^+e^-$ would lead, in a stationary or slowly expanding flow, to photons just below 0.511 MeV (Ref. 43)]. In order to avoid this, it seems inescapable that the flow must be expanding with a very high Lorentz factor, since in this case the relative angle at which the photons collide is less than Γ^{-1} and the threshold for the pair production is effectively diminished. Thus, photons with energy

$$\epsilon_{\gamma,\mathrm{MeV}} \lesssim 10^4 \epsilon_{t,\mathrm{MeV}}^{-1} \Gamma_2^2 \qquad (4)$$

are able to escape, where Γ_2 is the bulk Lorentz factor in units of 10^2, and ϵ_t is the energy of the target photons.[44] Thus, simply from observations and general physical considerations, a relativistically expanding fireball is expected. However, the observed γ-ray spectrum is generally a broken power law, i.e., highly nonthermal. The optically thick $e^{\pm}\gamma$ fireball cannot, by itself, produce such a spectrum (it would tend rather to produce a modified blackbody). In addition, the expansion would lead to a conversion of internal energy into kinetic energy of expansion, so even after the fireball becomes optically thin, it would be highly inefficient, most of the energy being in the kinetic energy of the associated protons, rather than in photons.

The most likely way to achieve a nonthermal spectrum in an energetically efficient manner is if the kinetic energy of the flow is reconverted into random energy via shocks, after the flow has become optically thin. This is a plausible scenario, in which two cases can be distinguished. In the first case (a) the expanding fireball runs into an external medium (the ISM, or a pre-ejected stellar wind.[2,3,6,7] The second possibility (b) is that (even before such external shocks occur) internal shocks develop in the relativistic wind itself, faster portions of the flow catching up with the slower portions.[4,5] This is a completely generic model, which is independent of the specific nature of the progenitor, as long as it delivers the appropriate amount of energy ($\gtrsim 10^{52}$ erg) in a small enough region ($\lesssim 10^7$ cm). This model has been successful in explaining the major observational properties of the gamma-ray emission, and is the main paradigm used for interpreting the GRB observations.

External shocks will occur in an impulsive outflow of total energy E_o in an external medium of average particle density n_o at a radius

$$r_{dec} \sim 10^{17} n_o^{1/3} E_{53}^{1/3} \eta_2^{-2/3} \text{ cm}, \qquad (5)$$

where the lab-frame energy of the swept-up external matter ($\Gamma^2 m_p c^2$ per proton) equals the initial energy $E_o = E_{53} \cdot 10^{53}$ ergs of the fireball, and $\eta = \Gamma = 10^2 \eta_2$ is the final bulk Lorentz factor of the ejecta. The typical observer-frame dynamic time of the shock (assuming the cooling time is shorter than this) is $t_{dec} \sim r_{dec}/c\Gamma^2 \sim$ seconds, for typical parameters, and $t_b \sim t_{dec}$ would be the burst duration (the impulsive assumption requires that the initial energy input occur in a time shorter than t_{dec}). Variability on timescales shorter than t_{dec} may occur on the cooling timescale or on the dynamic timescale for inhomogeneities in the external medium, but generally this is not ideal for reproducing highly variable profiles.[45] However, it can reproduce bursts with several peaks,[46] and may therefore, be applicable to the class of long, smooth bursts.

The same behavior $\Gamma \propto r$ with comoving temperature $\propto r^{-1}$, followed by saturation $\Gamma_{max} \sim \eta$ at the same radius $r/r_l \sim \eta$ occurs in a wind scenario, if one assumes that a lab-frame luminosity L_o and mass outflow $\dot{M}_o$ are injected at $r \sim r_l$ and continuously maintained over a time t_w; here $\eta = L_o/\dot{M}_o c^2$. In such a wind model, internal shocks will occur at a radius[4]

$$r_{dis} \sim c t_{var} \eta^2 \sim 3 \times 10^{14} t_{var} \eta_2^2 \text{ cm}, \quad (6)$$

where shells of different energies $\Delta\eta \sim \eta$ initially separated by ct_v (where $t_v \leq t_w$ is the timescale of typical variations in the energy at r_l) catch up with each other. In order for internal shocks to occur above the wind photosphere $r_{ph} \sim \dot{M}\sigma_T/(4\pi m_p c\Gamma^2) = 1.2 \times 10^{12} L_{51} \eta_2^{-3}$ cm, but also at radii greater than the saturation radius (so that most of the energy does not come out in the photospheric quasi-thermal radiation component) one needs to have $3 \times 10^1 L_{51}^{1/5} t_{var}^{-1/5} \lesssim \eta 10^2 L_{51}^{1/4} t_{var}^{-1/4}$. This type of models has the advantage[4] that it allows an arbitrarily complicated light curve; the shortest variation timescale $t_{var} \gtrsim 10^{-3}$ s is limited only by the dynamic timescale at r_l, where the energy input may be expected to vary chaotically. Such internal shocks have been shown explicitly to reproduce (and indeed even be required by) some of the more complicated light curves.[45,47]

3 Progenitors and Central Engines

Even before the measurement of a high redshift in a GRB afterglow, the difficulty in detecting the host galaxies of bright bursts (e.g., Ref. 48) motivated the exploration of ways of increasing the possible total energy budget of GRB. The first explicit model to do this[17] involved converting a large fraction of the binding energy of a black hole and torus system ($\sim 10^{54}$ ergs) into a fireball outflow, through MHD torques which power a Poynting jet outflow. Such a system would naturally arise from a NS-NS or a BH-NS merger (and it may also arise from a failed SN Ib). In the last year, a number of other possible energy sources have been considered as possible candidates for powering GRB.[49,50] A fact which is not widely realized[55] is that *all* plausible GRB progenitors suggested so far (e.g., NS-NS or NS-BH mergers, Helium core–black hole [He-BH] or white dwarf–black hole [WD-BH] mergers, and the wide category labeled as hypernova or collapsars including failed supernova Ib [SNe Ib], single or binary Wolf-Rayet [WR] collapse, etc.) are expected to lead to a BH plus debris torus system. An important point is that the overall energetics from these various progenitors do not differ by more than about one order of magnitude.

Two large reservoirs of energy are available in the generic merger or collapse scenario: the binding energy of the orbiting debris, and the spin energy of the black hole.[17] The first can provide up to 42% of the rest mass energy of the disk, for a maximally rotating black hole, while the second can provide up to 29% of the rest mass of the black hole itself. The $\nu\bar{\nu} \rightarrow e^+e^-$ process can tap the thermal energy of the torus produced by viscous dissipation. For this mechanism to be efficient, the neutrinos must escape before being advected into the hole; on the other hand, the efficiency of conversion into pairs (which scales with the square of the neutrino density) is low if the neutrino production is too gradual. Typical estimates suggest a fireball of $\lesssim 10^{51}$ erg (Ref. 50 and 56), except perhaps in the "collapsar" or failed SN Ib case where estimates[50] indicate up to $10^{52.3}$ ergs for optimum parameters. If the fireball is collimated into a solid angle Ω_j, then of course, the apparent "isotropized" energy would be larger by a factor $(4\pi/\Omega_j)$, but unless Ω_j is $\lesssim 10^{-2} - 10^{-3}$, this may fail to satisfy the apparent isotropized energy of $10^{53.5}$ ergs implied by a redshift $z = 3.4$ for GRB 971214. An alternative way to tap the torus energy is through dissipation of magnetic fields generated by the differential rotation in the torus.[17,38,41,57] Even before the BH forms, a NS-NS merging system might lead to the winding up of the fields and dissipation in the last stages before the merger.[9,40] The above mechanisms tap the energy available in the debris torus or disk. However, a hole formed from a coalescing compact binary is guaranteed to be rapidly spinning, and being more massive, could contain more energy than the torus; the energy extractable in principle through MHD coupling to the rotation of the hole by the Blandford and Znajek[58] effect could then be even larger than that contained in the orbiting debris.[17,23] Collectively, any such MHD outflows have been referred to as Poynting jets.

The various progenitors differ only slightly in the mass of the BH and that of the debris torus they produce, and they may differ more markedly in the amount of rotational energy contained in the BH. Strong magnetic fields, of order of 10^{15} G, are needed to carry away the rotational or gravitational energy in a timescale of tens of seconds.[60,59] If the magnetic fields do not thread the BH, then a Poynting outflow can at most carry the gravitational binding energy of the torus. For a maximally rotating and for a nonrotating BH, this is 0.42 and 0.06 of the torus rest mass, respectively. The torus, or disk mass, in a NS-NS merger is[61] $M_d \sim 0.1 M_\odot$,

and for a NS-BH, He-BH, WD-BH merger, or binary WR collapse, it may be estimated at[23,49] $M_d \sim 1M_\odot$. In the HeWD-BH merger and WR collapse the mass of the disk is uncertain due to lack of calculations on continued accretion from the envelope, so $1M_\odot$ is just a rough estimate. The largest energy reservoir is therefore *prima facie*, associated with NS-BH, HeWD-BH, or binary WR collapse, which have larger disks and fast rotation, the maximum energy being $\sim 8 \times 10^{53}\epsilon(M_d/M_\odot)$ ergs; for the failed SNe Ib (which is a slow rotator) it is $\sim 1.2 \times 10^{53}\epsilon(M_d/M_\odot)$ ergs, and for the (fast rotating) NS-NS merger it is $\sim 0.8 \times 10^{53}\epsilon(M_d/0.1M_\odot)$ ergs, where ϵ is the efficiency in converting gravitational into MHD jet energy. Conditions for the efficient escape of a high-Γ jet may, however, be less propitious if the "engine" is surrounded by an extensive envelope.

If the magnetic fields in the torus thread the BH, the rotational energy of the BH can be extracted via the Blandford and Znajek[58] (B-Z) mechanism.[17] The extractable energy is $\epsilon f(a)M_{bh}c^2$, where ϵ is the MHD efficiency factor and $a = Jc/GM^2$ is the rotation parameter, which equals one for a maximally rotating black hole. $f(a) = 1 - \sqrt{\frac{1}{2}[1+\sqrt{1-a^2}]}$ is small unless a is close to one, where it sharply rises to its maximum value $f(1) = 0.29$, so the main requirement is a rapidly rotating BH, $a \gtrsim 0.5$. For a maximally rotating BH, the extractable energy is therefore $0.29\epsilon M_{bh}c^2 \sim 5 \times 10^{53}\epsilon(M_{bh}/M_\odot)$ ergs. Rapid rotation is essentially guaranteed in a NS-NS merger, since the radius (especially for a soft equation of state) is close to that of a black hole and the final orbital spin period is close to the required maximal spin rotation period. Since the central BH will have a mass[61] of about $2.5M_\odot$, the NS-NS system can thus power a jet of up to $\sim 1.3\times10^{54}\epsilon(M_{bh}/2.5M_\odot)$ ergs. The scenarios less likely to produce a fast rotating BH are the NS-BH merger (where the rotation parameter could be limited to $a \leq M_{ns}/M_{bh}$, unless the BH is already fast rotating) and the failed SNe Ib (where the last material to fall in would have maximum angular momentum, but the material that was initially close to the hole would have less angular momentum). A maximal rotation rate may also be possible in a He-BH merger, depending on what fraction of the He core gets accreted along the rotation axis as opposed to along the equator,[49] and the same should apply to the binary fast-rotating WR scenario, which probably does not differ much in its final details from the He-BH merger. For a fast rotating BH of $3M_\odot$ threaded by the magnetic field, the maximal energy carried out by the jet is then $\sim 1.6 \times 10^{54}\epsilon(M_{bh}/3M_\odot)$ ergs.

Thus, in the accretion powered jet case, the total energetics between the various models differs at most by a factor of 20, whereas in the rotationally (B-Z) powered cases they differ by at most a factor of a few, depending on the rotation parameter. For instance, even allowing for low total efficiency (say 30%), a NS-NS merger whose jet is powered by the torus binding energy would only require a modest beaming of the γ-rays by a factor $(4\pi/\Omega_j) \sim 20$, or no beaming if the jet is powered by the B-Z mechanism, to produce the equivalent of an isotropic energy of $10^{53.5}$ ergs. The beaming requirements of BH-NS and some of the other progenitor scenarios are even less constraining.

There is also the apparent coincidence of GRB 980425 with the SN Ib/Ic 1998bw.[51] A simple but radical interpretation[52] is that all GRB may be associated with SNe Ib/Ic and differences arise only from different viewing angles relative to a very narrow jet. The difficulties with this are that it would require extreme collimations by factors of 10^{-3}–10^{-4}, and that the statistical association of any subgroup of GRB with SNe Ib/Ic (or any other class of objects, for that matter) is so far not significant.[53] If, however, the GRB 980425/1998bw association is real,[54] then we may be in the presence of a new subclass of GRB with lower energy $E_\gamma \sim 10^{48}(\Omega_j/4\pi)$ erg, which is only rarely observable even though its comoving volume density could be substantial. In this, more likely interpretation, the great majority of the observed GRB would have the energies $E_\gamma \sim 10^{54}(\Omega_j/4\pi)$ ergs as inferred from high redshift observations.

4 The Afterglows of Gamma-Ray Bursts

Just as one can interpret supernova remnants without fully understanding the initiating explosion, one can also understand the dynamics of the afterglows of gamma-ray bursts, despite the uncertainties recounted in the previous section. The simplest hypothesis is that the afterglow is due to a relativistic expanding blast wave. The complex time structure of some bursts suggests that the central trigger may continue for up to 100 seconds. However, at much later times all memory of the initial time structure would be lost: essentially all that matters is how much energy and momentum has been injected; the injection can be regarded as instantaneous in the context of the much longer afterglow. Detailed calculations and predictions from such a model[8] preceded the observations of the first afterglow detected, GRB970228.[1]

The simplest spherical afterglow model has been remarkably successful at explaining the gross features of the GRB 970228, GRB 970508, and other afterglows.[14] This has led to the temptation to take the assumed sphericity for granted. For instance, the lack of a break in the light curve of GRB 970508 prompted the inference[16] that all afterglows are essentially isotropic, leading to the very large (isotropic) energy estimate of $10^{53.5}$ ergs in GRB 971214. The multiwavelength data analysis has in fact advanced to the point where one can use observed light curves at different times and derive,[12,62] via parametric fitting, physical parameters of the burst and environment, such as the total energy E, the magnetic and electron-proton coupling parameters ϵ_B and ϵ_e, and the external density n_o. However, what these fits constrain is only the energy per unit solid angle $\mathcal{E} = (E/\Omega_j)$.

In the simplest departure from a spherical model the blast wave energy may be channeled into a solid angle Ω_j. Then one might expect[63] a faster decay of Γ after it drops below $\Omega_j^{-1/2}$. A simple calculation using the usual scaling laws leads indeed to a steepening of the flux power law in time. The lack of such an observed afterglow downturn in the optical has been interpreted as further supporting the sphericity of the entire fireball. There are several important caveats, however. The first one is that the above argument assumes a simple, impulsive energy input (lasting $\lesssim$ the observed γ-ray pulse duration), characterized by a single energy and bulk Lorentz factor value. Estimates for the time needed to reach the nonrelativistic regime, or $\Gamma < \Omega_j^{-1/2} \lesssim$ few, could then be under a month,[9] especially if an initial radiative regime with $\Gamma \propto r^{-3}$ prevails. It is unclear whether, even when electron radiative time scales are shorter than the expansion time, such a regime applies, as it would require strong electron-proton coupling.[55] Furthermore, even the simplest reasonable departures from a top-hat approximation (e.g., having more energy emitted with lower Lorentz factors at later times, which still do not exceed the gamma-ray pulse duration) would drastically extend the afterglow lifetime in the relativistic regime, by providing a late "energy refreshment" to the blast wave on time scales comparable to the afterglow time scale.[28] The transition to the $\Gamma < \Omega_j^{-1/2}$ regime occurring at $\Gamma \sim$ few could then occur as late as six months to more than a year after the outburst, depending on details of the brief energy input. Even in a simple top-hat model, more detailed calculations show that the transition to the nonrelativistic regime is very gradual ($\delta t/t \gtrsim 2$) in the light curve. Also, even though the flux from the head-on part of the remnant decreases faster, this is more than compensated by the increased emission measure from sweeping up external matter over a larger angle, and by the fact that the extra radiation, which arises at larger angles, arrives later and refills the steeper light curve. The sideways expansion thus actually can slow down the flux decay[65] rather than making for a faster decay.

The ratio L_γ/L_{opt} (or L_γ/L_x) can be quite different from burst to burst. The fit of Wijers and Galama[62] for GRB 970508 indicates an afterglow (X-ray energies or softer) energy per solid angle $\mathcal{E}_{52} = 3.7$, while at $z = 0.835$ with $h_{70} = 1$, the corresponding γ-ray $\mathcal{E}_{52\gamma} = 0.63$. On the other hand, for GRB 971214 at $z = 3.4$, the numbers are $\mathcal{E}_{52} = 0.68$ and $\mathcal{E}_{52\gamma} = 20$. The bursts themselves require ejecta with $\Gamma > 100$. The gamma-rays we receive come only from material whose motion is directed within one degree of our line of sight. They therefore provide no information about the ejecta in other directions: the outflow could be isotropic, or concentrated in a cone of angle, say, 20 degrees (provided that the line of sight lies inside the cone). At observer times of more than a week, the blast wave would be decelerated to a moderate Lorentz factor, irrespective of the initial value. The beaming and aberration effects are less extreme so we observe afterglow emission not just from material moving almost directly towards us, but from a wider range of angles.

The afterglow is thus a probe for the geometry of the ejecta—at late stages, if the outflow is beamed, we expect a spherically-symmetric assumption to be inadequate; the deviations from the predictions of such a model would then tell us about the ejection in directions away from our line of sight. It is quite possible, for instance, that there is relativistic outflow with lower Γ (heavier loading of baryons) in other directions; this slower matter could even carry most of the energy.[14,23] This hypothesis is, in fact, supported to some degree by the fits of Wijers and Galama[62] mentioned above.

One expects afterglows to show a significant amount of diversity. This is expected both because of a possible spread in the total energies (or energies per solid angle as seen by a given observer), and also from the fact that GRB may be going off in very different environments. The angular dependence of the outflow, and the radial dependence of the density of the external environment can have a marked effect on the time dependence of the observable afterglow quantities.[55] So do any changes of the bulk Lorentz factor and energy output during even a brief energy release episode.[64] The afterglow light curves are also affected by the degree of coupling between electrons and protons in the outflow.[55,66] Detailed

model fits[67] to the X-ray, optical and radio light curves of GRB 970228 and GRB 970508 show that the shock physics may be a function of the shock strength, and also indicate that dust absorption may be needed to simultaneously fit the X-ray and optical fluxes (the latter being affected more severely). The effects of beaming (outflow within a limited range of solid angles) can be significant,[65] but are coupled with other effects, and a careful analysis is needed to disentangle them. Finally, both the outflowing ejecta[68] and the external medium[69,70] are expected to provide discernible atomic edge and line features in the X-ray and optical spectrum of afterglows. These may be used as diagnostics for the outflow Lorentz factor, or as alternative measures of the GRB redshift.

The location of the afterglow relative to the host galaxy center can provide clues both for the nature of the progenitor and for the external density encountered by the fireball. A hypernova model would be expected to occur inside a galaxy, in fact inside a high density ($n_o > 10^3 - 10^5$ cm^{-3}). Some bursts are definitely inside the projected image of the host galaxy, and some also show evidence for a dense medium at least in front of the afterglow.[22] On the other hand, for a number of bursts there are strong constraints from the lack of a detectable, even faint, host galaxy.[71] In NS-NS mergers one would expect a BH plus debris torus system and roughly the same total energy as in a hypernova model, but the mean distance traveled from birth is of order several Kpc (Bloom, Sigurdsson, and Pols[73]), leading to a burst presumably in a less dense environment. The fits of Wijers and Galama[62] to the observational data on GRB 970508 and GRB 971214 in fact suggest external densities in the range of $n_o = 0.04$–0.4 cm^{-3}, which would be more typical of a tenuous interstellar medium. These could arise within the volume of the galaxy, but on average one would expect as many GRB inside as outside. This is based on an estimate mean NS-NS merger time of 10^8 years; other estimated merger times (e.g., 10^7 years[72]) would give a burst much closer to the birth site. BH-NS mergers would also occur in timescales $\lesssim 10^7$ years, and would be expected to give bursts well inside the host galaxy.[73]

5 Conclusions

The simple blast wave model seems able to accommodate the present data on afterglows remarkably well. However, the constraints on the angle-integrated γ-ray energy are not strong, and beaming effects remain uncertain. A relatively brief (1-100 s), probably modulated energy input appears likely, although in some progenitor scenarios there may be delayed effects. We need to be open minded about the possibility of there being more subclasses of classical GRB than just short ones and long ones. For instance, GRB with no high-energy pulses (NHE) appear to have a different (but still isotropic) spatial distribution than those with high-energy (HE) pulses.[74] Some caution is needed in interpreting this, since selection effects could lead to a bias against detecting HE emission in dim bursts.[10]

Much progress has been made in understanding how gamma rays can arise in fireballs produced by brief events depositing a large amount of energy in a small volume, and in deriving the generic properties of the long wavelength afterglows that follow from this. There still remain a number of mysteries, especially concerning the identity of their progenitors, the nature of the triggering mechanism, the transport of the energy, and the timescales involved. Nevertheless, even if we do not yet understand the intrinsic gamma-ray burst central engine, they may be the most powerful beacons for probing the high redshift ($z > 5$) universe. Even if their total energy is reduced by beaming to a "modest" $\sim 10^{52} - 10^{52.5}$ ergs in photons, they are the most extreme phenomena that we know about in high energy astrophysics. The modeling of the burst mechanism itself will continue to be a formidable challenge to theorists and to computational techniques. However, they do not appear insurmountable, and the prospects for significant progress in the near future are realistic.

Acknowledgments

I am grateful to several colleagues for collaborations in this subject, in particular Martin Rees, Ralph Wijers, and Alin Panaitescu, and to the organizers of the XXVI SLAC Summer Institute on Particle Physics for an excellent School.

References

[1] E. Costa *et al.*, Nature **387**, 783 (1997).
[2] M. J. Rees and P. Mészáros, Mon. Not. R. Astron. Soc. **258**, 41 (1992).
[3] P. Mészáros and M. J. Rees, Astrophys. J. **405**, 278 (1993).
[4] M. J. Rees and P. Mészáros, Astrophys. J. **430**, L93 (1994).
[5] B. Paczyński and G. Xu, Astrophys. J. **427**, 708 (1994).

[6] J. Katz, Astrophys. J. **422**, 248 (1994).

[7] R. Sari and T. Piran, Astrophys. J. **455**, L143 (1995).

[8] P. Mészáros and M. J. Rees, Astrophys. J. **476**, 232 (1997a).

[9] M. Vietri, Astrophys. J. **478**, L9 (1997a).

[10] J. Norris in *Proceedings of the 1998 SLAC Summer Institute*, edited by A. M. Breaux *et al.*, (SLAC, Stanford, CA, 1998).

[11] M. Tavani, Astrophys. J. **483**, L87 (1997).

[12] E. Waxman, Astrophys. J. Lett. (1997).

[13] D. Reichart, Astrophys. J., 1997 (to be published).

[14] R. A. M. J. Wijers, M. J. Rees, and P. Mészáros, Mon. Not. R. Astron. Soc. **288**, L51 (1997).

[15] M. Metzger *et al.*, Nature **387**, 878 (1997).

[16] S. Kulkarni *et al.*, Nature **393**, 35 (1998).

[17] P. Mészáros and M. J. Rees, Astrophys. J. **482**, L29 (1997b).

[18] D. Reichart, Astrophys. J. **495**, L99 (1998).

[19] P. Groot *et al.*, in *Gamma-Ray Bursts*, edited by C. Meegan *et al.*, (Am. Inst. Phys., New York), p. 557 (1997).

[20] T. Murakami *et al.*, in *Gamma-Ray Bursts*, edited by C. Meegan *et al.*, (Am. Inst. Phys., New York), p. 435 (1997).

[21] G. B. Taylor *et al.*, Nature **389**, 263 (1997).

[22] A. Owen et al., Astron. and Astrophys. (astro-ph/9809356), 1998 (to be published).

[23] B. Paczyński, J. Astrophys. **494**, L45 (1998).

[24] C. Fryer and S. Woosley (astro-ph/9804167), 1997 (submitted to Astrophys. J. Lett.).

[25] G. Fishman and C. Meegan, Ann. Rev. Astr. Ap. **33**, 415 (1995).

[26] J. M. Horack, R. S. Mallozzi, and T. M. Korshut, Astrophys. J. **466**, 21 (1996).

[27] J. P. Norris *et al.*, Astrophys. J. **439**, 542 (1995).

[28] D. E. Reichart and P. Mészáros, Astrophys. J. **483**, 597 (1997).

[29] T. Totani, Astrophys. J. **486**, L71 (1997).

[30] R. Wijers *et al.*, Mon. Not. R. Astron. Soc. **294**, L17 (1998).

[31] M. Krumholz, S. Thorsett, and F. Harrison, preprint (astro-ph/9807117), 1998.

[32] B. Paczyński, Astrophys. J. Lett. **308**, L43 (1986).

[33] J. Goodman, Astrophys. J. Lett. **308**, L47 (1986).

[34] R. Narayan, T. Piran, and A. Shemi, Astrophys. J. Lett. **379**, L17 (1991).

[35] E. S. Phinney, Astrophys. J. Lett. **380**, L17 (1991).

[36] S. Woosley, Astrophys. J. **405**, 273 (1993).

[37] A. Shemi and T. Piran, Astrophys. J. Lett. **365**, L55 (1990).

[38] B. Paczyński, Acta. Astron. **41**, 257 (1991).

[39] D. Eichler, M. Livio, T. Piran, and D. N. Schramm, Nature **340**, 126 (1989).

[40] P. Mészáros and M. J. Rees, Astrophys. J. **397**, 570 (1992).

[41] R. Narayan, B. Paczyński, and T. Piran, Astrophys. J. **395**, L83 (1992).

[42] P. Mészáros and M. J. Rees, Mon. Not. R. Astron. Soc. **257**, 29P (1992b).

[43] A. K. Harding and M. G. Baring, in *Gamma-ray Bursts*, edited by G. Fishman *et al.*, p. 520 (Am. Inst. Phys. 307, NY) (1994).

[44] P. Mészáros, in *17th Texas Symposium on Relativistic Astrophysics*, edited by H. Böhringer *et al.*, (Acad. Sci., New York, NY, 1995), p. 440.

[45] R. Sari and T. Piran, Astrophys. J. **485**, 270 (1998).

[46] A. Panaitescu and P. Mészáros, Astrophys. J. **492**, 683 (1998a).

[47] A. Panaitescu and P. Mészáros (astro-ph/9810258), 1998a (submitted to Astrophys. J.).

[48] B. E. Schaefer *et al.*, Astrophys. J. **489**, 693 (1997).

[49] C. Fryer and S. Woosley (astro-ph/9804167), 1998 (submitted to Astrophys. J. Lett.).

[50] R. Popham, S. Woosley, and C. Fryer (astro-ph/9807028), 1998 (submitted to Astrophys. J.).

[51] T. Galama *et al.*, Nature (astro-ph/9806175), 1998. (to be published).

[52] L. Wang and J. C. Wheeler (astro-ph/9806212), 1998 (submitted to Astrophys. J.).

[53] R. M. Kippen, *et al.* (astro-ph/9806364), 1998 (submitted to Astrophys. J.).

[54] S. Woosley, R. Eastman, and B. Schmidt (astro-ph/9806299), 1998 (submitted to Astrophys. J.).

[55] P. Mészáros, M. J. Rees, and R. Wijers, New Astron. (astro-ph/9808106), 1998 (to be published).

[56] M. Ruffert, H.-T. Janka, K. Takahashi, G. Schaefer, Astron. Astrophys. **319**, 122 (1997).

[57] J. I. Katz, Astrophys. J. **490**, 633 (1997).

[58] R. D. Blandford, and R. L. Znajek, Mon. Not. R. Astron. Soc. **179**, 433 (1977).

[59] C. Thompson, Mon. Not. R. Astron. Soc. **270**, 480 (1994).

[60] V. V. Usov, Mon. Not. R. Astron. Soc. **267**, 1035 (1994).

[61] M. Ruffert *et al.*, Astron. Astrophys. **319**, 122 (1997).

[62] R. A. M. J. Wijers and T. Galama (astro-ph/9805341), 1998 (submitted to Astrophys. J.).

[63] J. Rhoads, preprint (1997).

[64] M. J. Rees and P. Mészáros, Astrophys. J. Lett. **496**, L1 (1998).

[65] A. Panaitescu and P. Mészáros (astro-ph/9806016), 1998c (submitted to Astrophys. J.).

[66] A. Panaitescu and P. Mészáros, Astrophys. J. **501**, 772.

[67] A. Panaitescu, P. Mészáros, and M. J. Rees, Astrophys. J. **503**, 314 (1998b).

[68] P. Mészáros, and M. J. Rees, Astrophys. J. Lett. 502, L105 (1998a).

[69] R. Perna and A. Loeb, Astrophys. J. Lett. **503**, L135 (1998).

[70] P. Mészáros and M. J. Rees, Mon. Not. R. Astron. Soc. (astro-ph/9806183), 1998b (to be published).

[71] B. E. Schaefer, Astrophys. J. in press (astro-ph/9810424), 1998.

[72] E. van den Heuvel, in *X-ray Binaries and Recycled Pulsars* edited by E. P. J. van den Heuvel and S. A. Rappaport (Kluwer Ac. Pub., 1992), pp 233-256.

[73] J. Bloom, S. Sigurdsson, and O. Pols, Mon. Not. R. Astron. Soc. (astro-ph/9805222), 1998.

[74] G. Pendleton *et al.*, Astrophys. J. **464**, 606 (1996).

Author Index

G. Apollinari....367
N. Arkain-Hamed....299
G. Berdardi....301
R. Briere....343
G. Catanesi....447
L.R. Cominsky....55
A. Deckel....163
R. Drucker....249
S. Finn....165
G. Goldhaber....289
J. Imazato....513
D. Jackson....327
U. Katz....393
R. Kessler....495
J. Kotcher....227
W. Lorenson....437
A. Marchionni....455
W. Marciano....279
M. Martinez....229
P. Mészáros....539
G. Mills....467
M. Nakahata....477
J. Norris....524
M. Peskin....549
J. Polchinski....209
G. Redlinger....505
S. Rosier-Lees....409
P. Saulson....113
L. Sorrel....423
F. Taylor....259
T. Tyson....89
R. Wagoner....1
C. Will....15