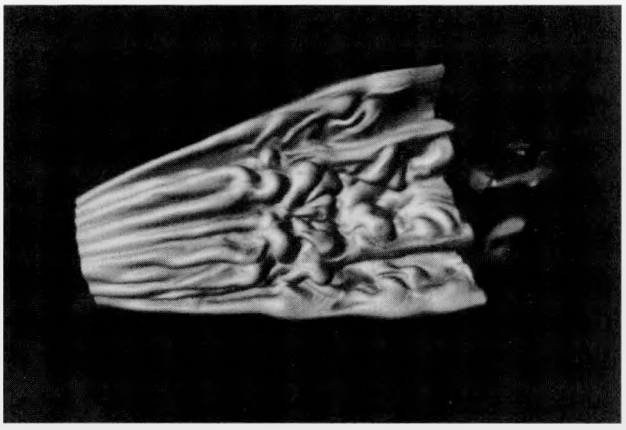
Thrust Area Report FY 89

ENGINEERING RESEARCH AND DEVELOPMENT

UCRL-53868-89

Lawrence Livermore National Laboratory

ABOUT THE COVER



The cover figure was adapted from Brian Cabral's article, "Interactive Volumetric Visualization." The figure depicts a jet of water flowing through a pipe at supersonic speed and abruptly slowing down as it reaches an ambient medium (air). Modeling such a flow normally takes hundreds of hours on a Cray, but a new finite-element mesh developed by the individuals listed below has reduced the time required by an order of magnitude (to approximately 80 hr). This means that we can model the Laboratory's complicated hydrodynamic problems faster (and cheaper).

As shown in the figure, the interface between the jet flow and the air highlights the structure of instabilities transverse to the main direction of flow (left to right). The image was achieved by rendering the results of a numerical computation of the jet using 3D adaptive-mesh refinement. A black and white medium was chosen to highlight the visual contrast between the unstable areas of the flow. A single layer was rendered in the area of greatest change (gradient). Shading was incorporated to enhance the visual complexity.

This image was the result of a cooperative project in which a number of engineers contributed to the numerical simulation software:

J. Bell	(LLNL)
M. Welcome	(LLNL)
M. Berger	(NYU)
J. Saltzman	(LANL)
T. Ferretta	(LLNL)

In addition, B. Cabral of LLNL developed the rendering software.

Acknowledgments

Scientific Editor
Edward Lafranchi

Lead Technical Editor
Mike Genin

Technical Editors
Paul Adye
June Canada
Wally Clements
Karinne Gordon
Herb Highstone
Marie Kotowski
Dallas Sceales

Compositor
Janice Brosius

Graphic Designer
Lionel Fernandez Jr.

Production Coordinators
Lionel Fernandez Jr.
Mark Lewis

Production Artists
David Hebenstreit
Stephen Krause
Mark Lewis

Production Manager
Betita Gamble

Work performed under the auspices of the U. S. Department of Energy by Lawrence Livermore National Laboratory under Contract W–7405–Eng–48.

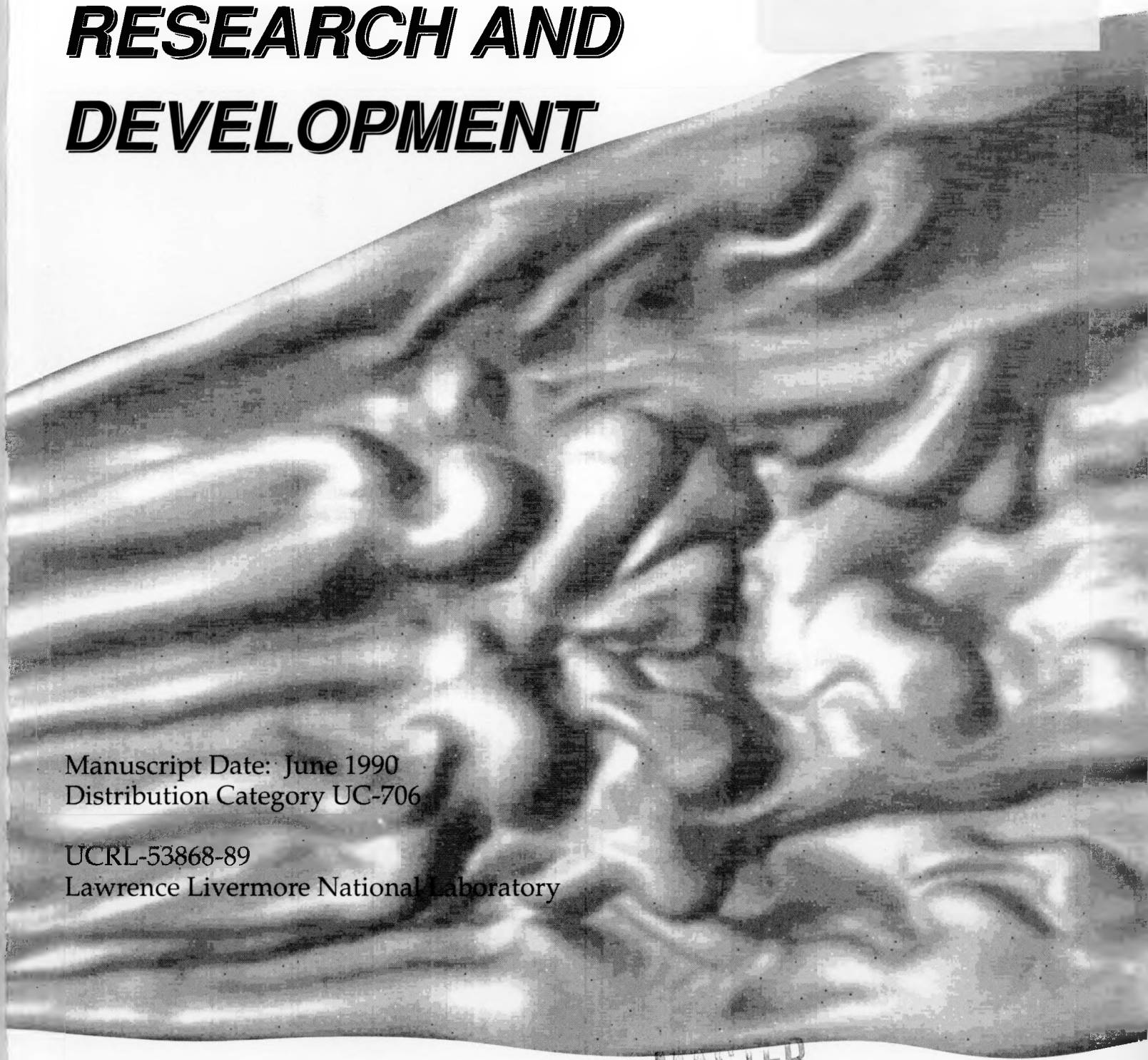
Printed in the United States of America. Available from the National Technical Information Service, U. S. Department of Commerce, 5285 Port Royal Road, Springfield, VA 22161. Price: printed copy A13, microfiche A01.

Thrust Area Report FY 89

ENGINEERING RESEARCH AND DEVELOPMENT

UCRL--53868-89

DE90 014728



Manuscript Date: June 1990
Distribution Category UC-706

UCRL-53868-89
Lawrence Livermore National Laboratory

Contents

Introduction v

1 Computational Electronics and Electromagnetics

Transient Wave Reconstruction from Finite Arrays ..... *James V. Candy* 1-1

Electronic Circuit Analysis by Computer Simulation ..... *Waldo G. Magnuson, Jr.* 1-6

Miniature Vacuum Tubes for Harsh Environments ..... *William J. Orvis* 1-9

Modeling the Performance of Photoconductors ..... *W. Travis White III* 1-13

Accelerator Cavity Modeling with TSAR ..... *Robert R. McLeod* 1-18

Localized Transmission of Wave Energy ..... *Richard W. Ziolkowski* 1-21

Modeling Pulse Driven Antenna Systems ..... *Marvin J. Barth* 1-25

A 3D Modified Finite Volume Technique for Maxwell's Equations ..... *Niel K. Madsen* 1-29

Calculations of Optical and Transport Properties of Solid-State Superlattice and Quantum-Well Structure Materials ..... *Jick H. Yee* 1-33

Transient Induced Surface Breakdown ..... *Jick H. Yee* 1-37

AMOS Development and Applications ..... *John F. DeFord* 1-47

2 Computational Mechanics

Using the NIKE2D Code to Model Severely Strained Ductile Metals ..... *Bruce Engelmann* 2-1

Code Acquisition and Development Project of the Thermo-Fluid Mechanics Group ..... *Werner Stein* 2-7

New Nonlinear Solution Algorithm for the TOPAZ2D Heat Transfer Code ..... *Arthur B. Shapiro* 2-10

Elastoplastic Shell Analysis in DYNA3D ..... *Robert G. Whirley* 2-12

3 Diagnostics and Microelectronics

Integrated Optics ..... *Glen M. McWright* 3-1

Microfabrication Development ..... *Dino R. Ciarlo* 3-4

Improvements in Molecular Beam Epitaxy Technology ..... *Raymond P. Mariella, Jr.* 3-8

Quantum Well Lasers ..... *Raymond P. Mariella, Jr.* 3-10

Wavelength Deflection Technique for Recording Microwave Signals ..... *Charles F. McConaghy* 3-12

Calibration of Optical Transient Detectors ..... *Philip M. Armatis* **3-18**

Ion Assisted Etching for Microelectronics and Microoptics ..... *Dale E. Miller* **3-21**

Comparison of Electro-Optic Diagnostic Systems ..... *Karla G. Hagans* **3-26**

GaAs CCDs for High-Speed Transient Recording ..... *Gregory A. Cooper* **3-30**

4 Fabrication Technology

Ductile Grinding of Glass ..... *Kenneth L. Blaedel* **4-1**

Molecular Dynamics Modeling Applied to Indentation and Metal Cutting Problems ..... *James F. Belak* **4-8**

Cathodic-Arc Process for Thin-Film Coatings ..... *David M. Sanders* **4-17**

Diamond Turning of Optical Single Crystals ..... *Chol K. Syn* **4-21**

Evaluation of Electroformed Copper for Shaped-Charge Applications ..... *Jack W. Dini* **4-29**

5 Materials Engineering and Science

Creep Model for Fiber Reinforced Epoxy-Resin Composites ..... *William W. Feng* **5-1**

Experimental Solutions to Residual Stresses in Carbon/Epoxy Composites ..... *Jay Brentjes* **5-5**

Failure Modeling in Finite Element Analysis ..... *Daniel J. Nikkel, Jr.* **5-10**

Characterizing the Failure of Composite Structures ..... *Scott E. Groves* **5-14**

Modeling and Experimental Measurement of Residual Stress ..... *Elane C. Flower* **5-23**

6 Microwave and Pulsed Power

A Long-Pulse, 50-kV Electron-Gun Modulator ..... *Ron Kihara* **6-1**

High-Power Backward-Wave Oscillator ..... *J. Frank Camacho* **6-4**

The Development of Mega-ampere, Kilocoulomb Commercial Ignitrons ..... *Ron Kihara* **6-9**

Magnetic Switch Modeling ..... *Jan-Mark Zentler* **6-14**

High-Voltage, Subnanosecond Photoconductive Switching ..... *Michael D. Pocha* **6-21**

7 Nondestructive Evaluation

X-Ray Optics for Microimaging Scanning Microscopy and Other Applications ..... *Richard W. Ryon* **7-1**

New Techniques for Ultrasonic Detection of Surface and Subsurface Damage in Laser Materials ..... *D. Kent Lewis* **7-3**

Contents

Non-Destructive Optical Characterization of Subsurface Damages in Laser Materials ..... *Chol K. Syn* **7-7**

Computerized Tomography ..... *Harry E. Martz* **7-15**

8 Remote Sensing, Imaging, and Signal Engineering

Robust Reconstruction Methods for Acoustic and Seismic Tomography ..... *James G. Berryman* **8-1**

Radar Ocean Imaging ..... *J. Patrick Fitch* **8-6**

Atmospheric Compensated Imaging ..... *Taylor W. Lawrence* **8-9**

New Algorithms for Deconvolution and Identification ..... *Dennis M. Goodman* **8-12**

Computational Recovery of Holographic Images ..... *Thomas J. Yorkey* **8-16**

9 New Initiatives

SPRINT 2 as a Computing Engine for Electromagnetic Simulations ..... *W. Travis White III* **9-1**

Scientific Database Project ..... *Everett W. Wheelock* **9-9**

An Expert System for Tuning Particle Beam Accelerators ..... *Darrel L. Lager* **9-13**

Interactive Volumetric Visualization ..... *Brian K. Cabral* **9-22**

Scientific Visualization Development ..... *Stephen G. Azevedo* **9-25**

Two-Dimensional Visualization System ..... *Brian K. Cabral* **9-31**

X-Ray Imaging System ..... *H. Joseph Weaver* **9-33**

Indexes

Author **Index-1**

Subject **Index-3**

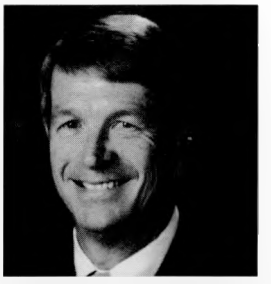
Introduction

The mission of the Engineering Research Program is to develop the people and technology needed to support current and future Laboratory programs. To accomplish this mission Engineering Research has two important goals - to identify key technology areas important to the Laboratory and to conduct high quality research and development. To help focus our efforts in achieving these goals we identify key technologies called *thrust areas* and select technical leaders for each area. The thrust areas are an integrated Engineering activity and, rather than being primarily discipline based, they are staffed by researchers from EE, ME and other Laboratory organizations as appropriate.

The thrust area leaders are accountable to me for the quality and progress of their research activities but have significant latitude to manage the resources allocated to their thrust area. They are expected to establish strong links to Laboratory program leaders, to use outside and inside experts to review the quality and direction of the research, to use university contacts to supplement and complement their efforts, and to be certain that we are not duplicating the work of others. The thrust area leader is also responsible for carrying out the development work that follows from the Research Program so that the results of the research can be applied as early as possible to the needs of the programs.

This annual report, organized by thrust area, describes research and development activities conducted within the Engineering Research Program for fiscal year 1989. Our intent is to provide timely summaries of objectives, theories, methods, and results. Readers desiring more information are encouraged to contact the individual authors.

Roger W. Werne
Associate Director for Engineering



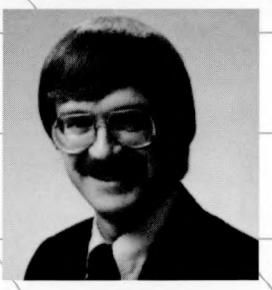
ENGINEERING RESEARCH AND DEVELOPMENT

Thrust Area Report FY 89

Computational Electronics and Electromagnetics

The Computational Electronics and Electromagnetics (CEE) thrust area focuses Engineering research and development activities in areas of electronics and electromagnetics of importance to current and future Laboratory programs. The goal of our thrust area is to develop theoretical and numerical models of electronic and electromagnetic phenomena of interest to LLNL programs. Based on these models, we then write computer codes that predict the behavior of solid-state devices and plasmas exposed to those fields. We also have the responsibility of upgrading these codes so that Lab researchers will find them easier to use. In FY 89 we worked on a number of projects in three general areas: modeling electromagnetic and acoustic phenomena, modeling semiconductor materials and devices, and modeling the interactions of EM waves and plasmas.

Richard W. Ziolkowski
Thrust Area Leader



Transient Wave Reconstruction from Finite Arrays

James V. Candy and Richard W. Ziolkowski
Engineering Research Division
Electronics Engineering

We are investigating the design of sampling arrays for the reception (transmission) of transient waves propagating in an electromagnetic or acoustic medium. The transmission and reception of these waves create unique problems because they are essentially transient wavefronts in *both* time and space. Thus, new methods must be developed for analysis and processing. We are developing ways to design receiving arrays and to process and reconstruct these waves from noisy measurement data.

This research supports new energy transmission technology development. It is a part of the ongoing inquiry into the basic concepts underlying wave propagation, transmission, and reception. It will impact communications in an oceanic environment, radar communications, medical imaging, and numerous other types of data transmission. Potential applications at the Laboratory include microscopes and telescopes with extremely large depths of field, low-loss power transmission, secure communications, remote sensing, and directed energy weapons.

Introduction

Recently, interest has been kindled in the possibility of transmitting wave energy in space in nonstandard manners. This interest has led to the discovery of pulses with enhanced localization and energy fluence characteristics. We define a localized wave as a directed energy pulse that is transient in both time and space. We have already demonstrated that the localized transmission of energy is possible in many practical applications. Localized waves can be recovered approximately from an array; we have verified them experimentally.[1] Localized waves have a vast number of potential applications at the Laboratory, including microscopes and telescopes with extremely large depths of field, low-loss power transmission, secure communications, remote sensing, and directed energy weapons.

In this project, we are concerned with a specific class of the transient-wave solutions to the wave equation—the so-called modified power spectrum (MPS) pulses. Our goal is to design sampling arrays capable of reconstructing the MPS from noisy measurement data (see **Fig. 1**). This problem is complicated because the pulse is transient in both time and space, which leads to a broadband response in both the spatial and temporal frequency domains.

As shown in Ref. 2, fundamental Gaussian pulse fields can be used as basis functions to represent new transient solutions of the wave equation. In particular, the general acoustic directed-energy pulse train (ADEPT) solution

$$f(\rho, z, t) = \int_0^\infty \phi_k(\rho, z, t)\, F(k)\, dk = \frac{1}{4\pi i [z_0 + i(z - ct)]} \int_0^\infty dk\, F(k)\, e^{-ks(\rho, z, t)},$$

where

ρ = transverse coordinate
z = longitudinal coordinate along the direction of propagation
t = time,

and where

$$s(\rho, z, t) = \frac{\rho^2}{z_0 + i(z - ct)} - i(z + ct),$$

is an exact source-free solution of the wave equation. Clearly, different spectra $F(k)$ lead to different wave equation solutions, and hence, to different solutions of the scalar wave equation. One particularly interesting spectrum selection is the MPS:

$$F(k) = \begin{cases} 4\Pi i \beta \dfrac{(\beta k - b)^{\alpha - 1}\, e^{-\alpha(\beta k - b)}}{\Gamma(\alpha)} & k > \dfrac{b}{\beta} \\ 0 & 0 \le k > \dfrac{b}{\beta} \end{cases}.$$

It is so called because it is derived from the power spectrum $F(k) = k^{\alpha-1}e^{-ak}$ by a scaling and a truncation. This choice of spectrum leads to the MPS pulse*:

$$f(\rho, z, t) = \frac{1}{z_0 + i(z - ct)} \frac{1}{\left(\frac{s}{\beta} + a\right)^{\alpha}} e^{-bs/\beta}.$$

We performed a simulation using the ADEPT transient wave with parameters $a = 1.0$ m, $\alpha = 1.0$, $b = 600$ m^{-1}, $\beta = 300$, and $z_0 = 4.5 \times 10^{-4}$ m through a 2.5-m aperture. Analysis of the frequency-wave-number spectra reveals some interesting properties. The overall spectrum is bound by the rectangle [± 0.5 MHz, ± 20 m^{-1}] with a 3-dB bandwidth of [± 0.5 MHz, ± 2 m^{-1}], thus indicating that the MPS pulse is broadband both temporally and spatially.

Designing arrays to receive (transmit) a localized wave is difficult because the wave is transient. We can think of this problem as one of designing a spatial filter to pass certain frequencies and reject others; however, the frequency-wave number spectrum of the MPS pulse is quite complex and varies at each z. Our research addresses the design of arrays, or more properly—*sampling* arrays,** to distinguish them (in principle) from the typical designs of narrowband antenna arrays for planar/spherical wavefronts. Our main goal is to investigate designs that can be used for the reception (transmission) of transient waves.

The constraints on sampling array designs for transient waves are based on the property that these designs minimize spatial aliasing and enable reconstruction of the wave at the receiver. In this report, we limit our discussion to designs specifically for the transient MPS pulse. The following properties of the MPS pulse will affect our design:

- Transient wave (non-planar wavefront)

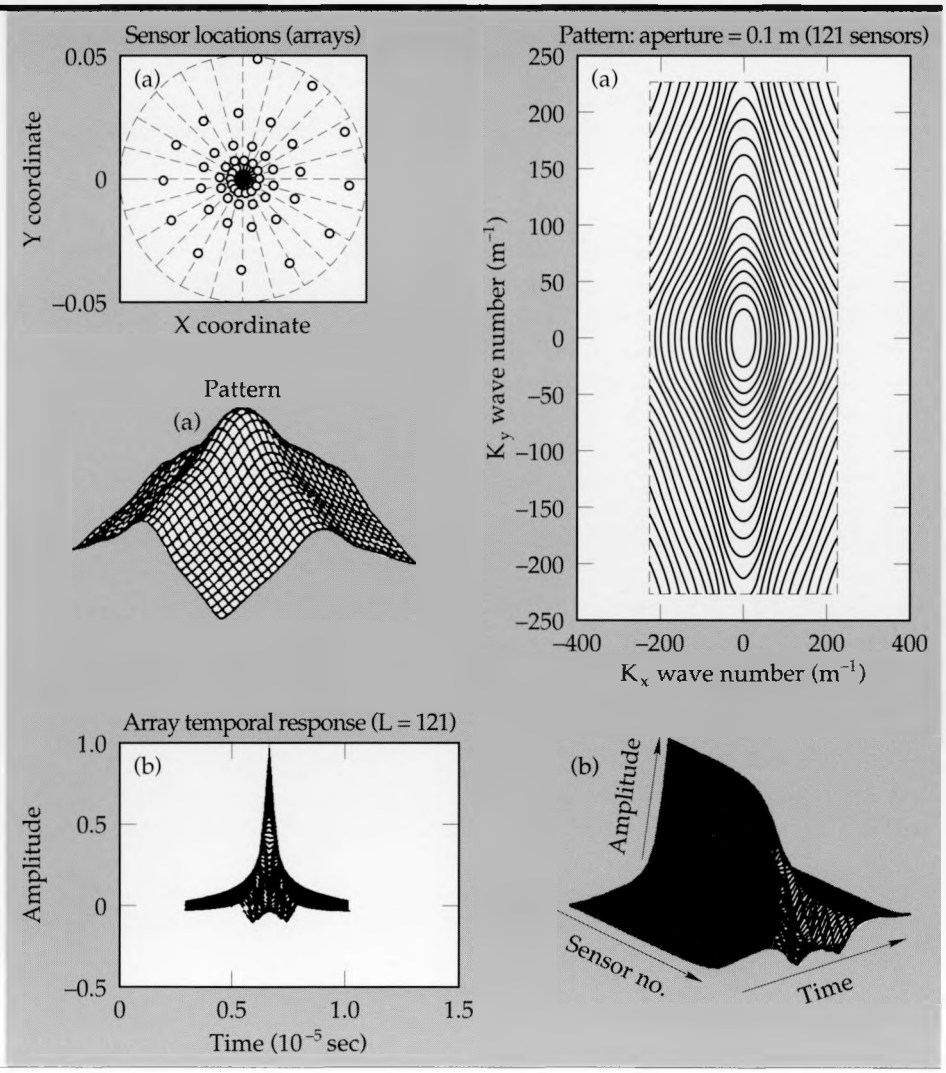


Figure 1. MPS pulse array design for a log-spiral array; (a) array and pattern, (b) sampled MPS pulse.

*For the rest of this report the MPS pulse, $f(\rho,z,t)$, will refer to the real part, and the direction of propagation will be taken along the positive z-axis.

**In this report, we are concerned with the spatial sampling of the MPS pulse for reconstruction purposes.

- Broadband spatio-temporal spectrum ([±20 m^{-1}, ±0.5 MHz])
- Circularly symmetric wave function [$f(x,y,z,t) = f(r,z,t)$].

In the next section, we will show how these properties can be used to design various sampling arrays.

Progress

With the constraints of symmetry and spectra in mind, we investigated the design of two-dimensional (2D) sampling arrays situated along the axis of propagation (z axis). Since the MPS pulse is circularly symmetric, sensors placed on different radii emanating from the origin will satisfy the symmetry constraint. More realistically, when noise is present, sensor redundancy is necessary to help enhance signal-to-noise ratio (SNR). We have used the following sampling array design procedure:

1. Select the candidate design
2. Determine its spatial bandwidth properties
3. Reconstruct the transient wave from its spatial samples
4. Compare the reconstructed wave with the "truth," which was generated by simulation of a line array with identical spacing.

For our reconstruction problem, we simulated the MPS pulse with a sampling interval of $\Delta t = 2 \times 10^{-7}$ sec and an aperture of $D = 0.1$ m using a line array of $L = 101$ sensors spaced at $d = 0.001$ m apart. The results are shown in **Fig. 2.** Here we have selected all of the physical parameters to match our experimental and fabrication capabilities. We will use the "truth" pulse to assess the reconstruction performance.

Practically, we are constrained to a $D = 0.1$ m aperture and are physically limited to $L = 121$ sensors for the design. We investigated various classes of arrays: square, hexagonal, coincentric circular, and the broadband designs that match these constraints. Arrays that theoretically have no bandwidth limitation are called *frequency independent*. The low frequency limit is determined by the size of the array, and the upper limit is bound by the precision in fabrication. These arrays are called *log-periodic* because they respond periodically with the logarithm of frequency. As an example, consider the so-called *log-spiral* design shown in **Fig. 1.** In this design, we distribute the sensors logarithmi-

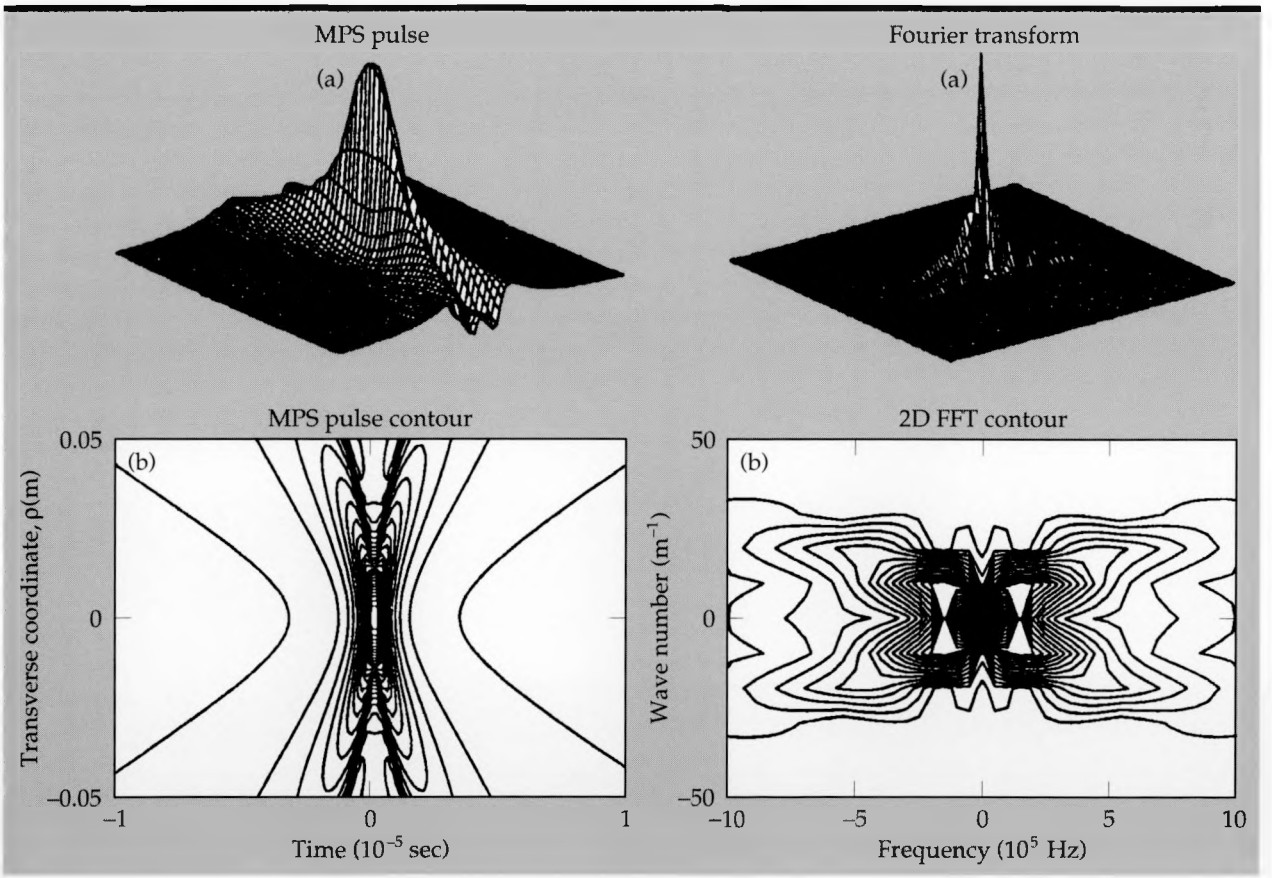


Figure 2. MPS pulse truth simulation for $L = 101$ element line array with 1 mm spacing; (a) spatial-temporal wave, (b) wave number-frequency spectrum.

cally on a spiral—the initial sensor location based on the minimum Nyquist spacing and the largest determined by the allowable aperture D. This design offers some desirable features based on the MPS properties. Besides the broadband response shown by the pattern in **Fig. 1a,** it enables unique sampling of the transient wave, which is very useful in a high SNR environment. Many sensors are located around the origin, providing a large number of samples of the Gaussian-like center region of the MPS pulse. Fewer sensors are located along the array extremities to sample the low-frequency tails. In **Fig. 1b,** we see the transient wave sampled by the log-spiral design. Note the large number of samples of the Gaussian pulse-like portion of the MPS pulse.

The problem of reconstructing the MPS pulse from a set of discrete spatial samples is the prime motivation in the array design. In the ideal case (e.g., sampling theorem) the signal is "exactly" reconstructed from its sampled values using cardinal interpolation as long as Nyquist sampling can be achieved. Unfortunately, this solution requires an infinite aperture or a spatially band-limited signal, which the theoretical MPS pulse does not satisfy. Our reconstruction approach is to develop an interpolator that sorts through the available samples for the pre-selected interval and linearly interpolates (fills in values) between the missing samples. In this scheme, a desired spatial sampling interval is pre-selected over the entire aperture length; those samples satisfying it are selected, and missing samples are linearly interpolated. In this way, we achieve a uniformly sampled spatial-temporal function from the non-uniform array samples.

We show the reconstruction and corresponding spectra for the log-spiral design in **Fig. 3.** Statistically, we quantify the reconstruction error by its mean and root-mean-square values, which are estimated using standard statistical sampling techniques. The error is calculated by comparing the reconstructed pulse to the "truth" pulse generated at the same spatial sampling interval used by the interpolator. As expected, since the log-spiral has many unique spatial samples of the Gaussian central region of the MPS pulse, most of its reconstruction error is located in the tails, which primarily

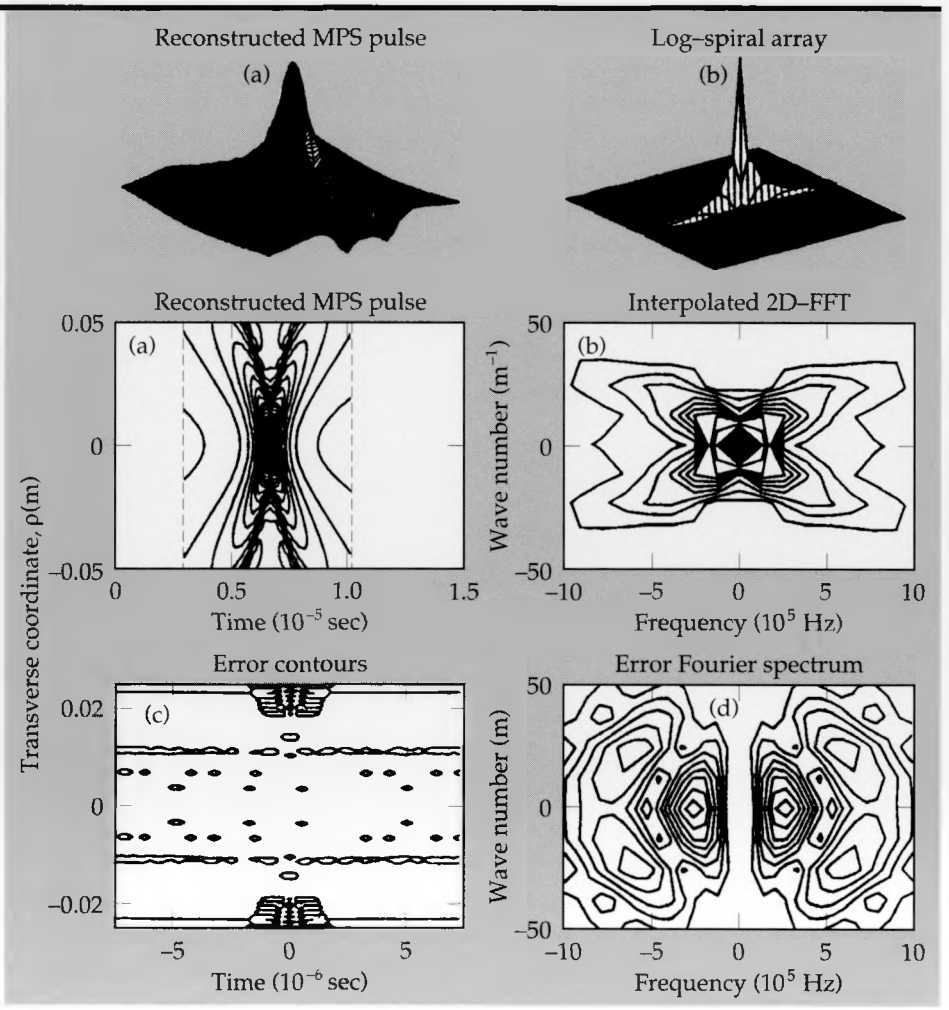


Figure 3. MPS pulse reconstruction from a log-spiral sampling array; (a) reconstructed spatial-temporal wave, (b) reconstructed wavenumber-frequency spectrum, (c) difference between truth and reconstructed spatial-temporal wave, (d) difference between truth and reconstructed wavenumber-frequency spectrum.

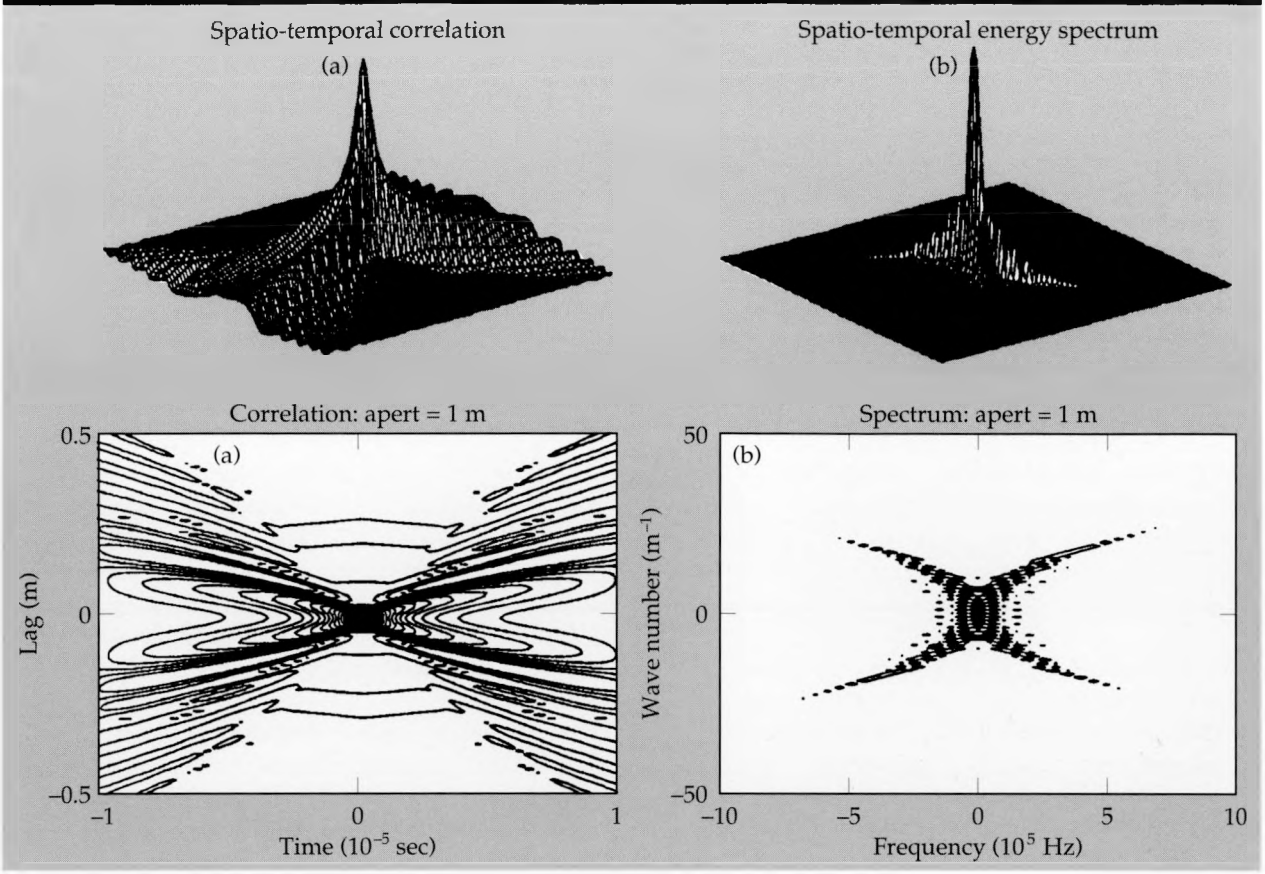


Figure 4. MPS pulse correlation/spectrum for L = 101 element line array with 1 cm spacing; (a) spatio-temporal correlation, (b) wave number-frequency energy spectrum.

occupy the lower frequency part of the wave number spectrum (central region). Statistically comparing this design to the standard 11 × 11 square array, we find that the mean and rms reconstruction error statistics indicate superior performance by the log-spiral array because m_ε: ($3 \times 10^{-6} < 200 \times 10^{-6}$) and σ_ε: ($8.5 \times 10^{-4} < 95 \times 10^{-4}$), respectively. Noisy data can also be processed using the arrays mentioned, and we have shown that the redundant (not unique in radius) sensors enable one to perform *signal averaging*, which results in enhancing performance of the arrays (see **Fig. 3**).

Future Work

Our plans for future designs will take into account the correlation properties of the transient and noise field. We will use this correlation information to optimize the design of the sampling array.

Initially, we will investigate larger apertures and their corresponding spatial correlations and spectra. The MPS "truth" pulse observed through a uniformly spaced line array of 101 sensors with a 1-m aperture at $z = 0$ with 101 temporal samples at $\Delta t = 0.2$ μsec is shown in **Fig. 4a.** Since the line array is uniformly spaced, we can use 2D Fourier transform techniques to estimate the spatio-temporal correlation function and the corresponding energy wave number-frequency spectrum. The results of this run are shown in **Fig. 4b.** Most of the correlated transient energy of the MPS pulse (at this aperture) is concentrated in the Gaussian region. As expected, high correlation exists primarily between ±0.1 m and the corresponding wave number-frequency spectrum lying between [±20 m^{-1}, 0.5 MHz]. We plan to estimate the acoustic noise field and then develop a design procedure that uses subarrays located at the regions of high spatial correlation while minimizing the noise correlation.

1. R. W. Ziolkowski, D. K. Lewis, and W. D. Cook, "Evidence of Localized Wave Transmission," *Phys. Rev. Lett.* **62**(2), 147 (1989), and *LLNL Report UCRL-99034* (1988).
2. R. W. Ziolkowski, "Exact solutions of the wave equation with complex source locations," *J. Math. Phys.* **26**(4), 861–863 (1985).
3. J. V. Candy, R. W. Ziolkowski, and D. K. Lewis. "Transient Wave Reconstruction and Processing," *LLNL Report UCRL-103451* (1990).

Electronic Circuit Analysis by Computer Simulation

Waldo G. Magnuson, Jr.
Nuclear Energy Systems Division
Electronics Engineering

Computer simulation is essential in evaluating the performance of large and complex electronic circuit designs. Our circuit modeling project has had the continuing goal of providing LLNL engineers with advanced computer-simulation capabilities for studying the behavior of electronic systems. In FY 89, the popular and powerful NET-2 program was brought over from the phased-out CDC-7600 computers, on which it had run for years, to the Cray-1 and Cray-X/MP computers of the Octopus and NMFECC systems. Thus NET-2 joins two other simulation programs already running on the Cray computers, SCEPTRE2 and SPICE2, to add significantly to our capabilities for simulating electronic circuit behavior.

Introduction

The circuit modeling project in the Computational Electronics and Electromagnetics Thrust Area has been providing Laboratory engineers with an electronic circuit simulation capability which uses computer models to analyze complex analog and digital systems. For the past several years we have had a strong capability based on the SCEPTRE2, NET-2, SPICE2, and SILOS computer programs. We have been actively involved in maintaining this capability. With the phasing out of the CDC-7600 computers, which was completed in 1988, a major part of the effort in maintaining our capability has consisted of moving the circuit simulation programs from the CDC-7600s to the Cray computers.

Four years ago we adapted the SCEPTRE2 program to the Cray-1 computer. This program is used frequently by both mechanical and electrical engineers at the Laboratory. Likewise, the NET-2 program has been heavily used by engineers since it began running on the CDC-7600s more than ten years ago. One of the more robust analysis programs available, NET-2 has been in use at LLNL and at other government laboratories as well. But with the CDC-7600s being phased out, it also had to be adapted to the Cray computers.

All three of our circuit-simulation programs now running on the Crays—SCEPTRE2, Net-2, and SPICE2—can solve very general nonlinear networks that may include complex descriptions (models) of electronic devices, magnetic devices, and transmission lines as well as linear and nonlinear resistors, capacitors, and inductors. However, SCEPTRE2 and NET-2 are considerably more flexible in their modeling capabilities than SPICE2. Both SCEPTRE2 and NET-2 are programs that have been developed on large computer systems, which makes them powerful but also makes them hard to move to personal computers. SPICE2, on the other hand, is a more modest program which has already been moved to the Apple and IBM personal computers.

SILOS is a digital logic simulation program that runs on the Engineering Department's VAX computer network.

Progress

Over the past two years we have been engaged in a project to move the NET-2 program from the CDC-7600 computers—the last one of which was phased out late in 1988—to the Cray computers. With the aid of Allan Malmberg, the developer of NET-2, we have successfully adapted the program to the Cray-1 and Cray-X/MP computer systems, and it is now operational on both the Octopus and NMFECC networks.

More than 400 program changes were made at LLNL in the conversion effort. Some 45 test problems were used to verify the program, and it was released for use in the summer of 1989. After the Cray-1 version of NET-2 was tested, the program was also adapted to the Cray-X/MP. Our Cray versions of the NET-2 programs have been transferred to Los Alamos National Laboratory, where they are currently being installed on the LANL computers.

NET-2, SCEPTRE2, SPICE2, and SILOS have

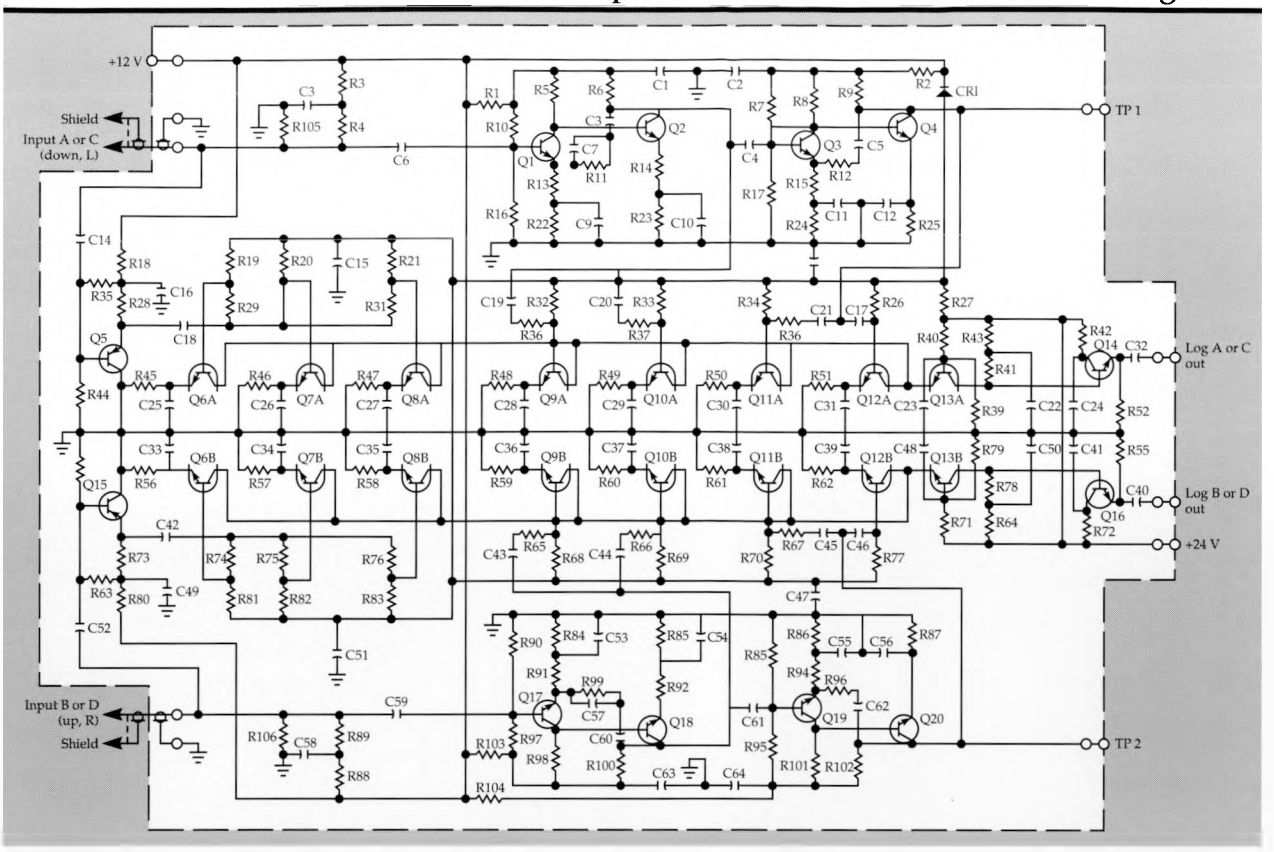


Figure 1. Schematic diagram of a complex electronic logarithmic amplifier circuit . This is typical of circuits modeled with the NET-2 computer program to analyze their response to high-power microwave pulses.

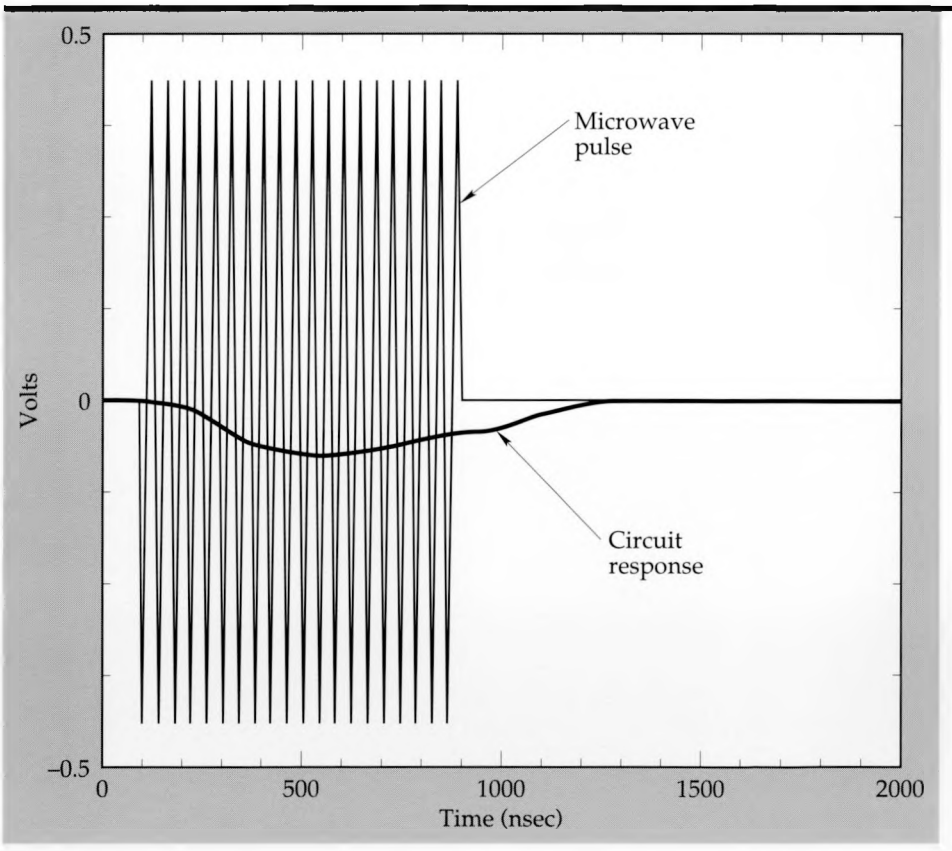


Figure 2. NET-2 analysis showing response of the circuit in Fig. 1 to a high-power microwave pulse.

been heavily used at LLNL during the past year to study the effects of high power microwaves on electronic components and circuits. The computer simulation studies have been done in conjunction with experimental measurements on the upset and burnout of electronic components and subsystems. Portions of a typical analysis are shown in **Figs. 1** and **2**. **Figure 1** is a schematic diagram of a complex logarithmic amplifier circuit. **Figure 2** is a composite of the analysis showing an out-of-band microwave signal input to the circuit and the resulting in-band response of the circuit, which is used as input to following system modules whose responses are analyzed in turn to determine if a system error results.

Future Work

In FY 89 we completed the transfer of NET-2 from the CDC-7600 computers to the Cray-1 and Cray-X/MP computers. NET-2 is finding heavy use in hardness evaluations of weapon system electronics. With the transfer of NET-2 to the Cray machines, LLNL now has one of the more powerful capabilities for simulating the behavior of large, nonlinear electronic systems. In the first quarter of FY 90 we plan to develop a short course or seminar for NET-2 users, which should be ready for presentation in late spring of 1990.

Miniature Vacuum Tubes for Harsh Environments

William J. Orvis, Dino R. Ciarlo, Charles F. McConaghy, Jick H. Yee, and Ed W. Hee
Engineering Research Division
Electronics Engineering

We are establishing the technology and modeling necessary to build and analyze micron-size, integrated vacuum tubes for use in high-radiation and high-temperature environments.

Introduction

Miniature vacuum tubes are vacuum switching and control devices fabricated on a silicon wafer, using the same methods as are used to make integrated circuits. While these miniature vacuum tubes will operate much the same as do conventional vacuum tubes, they are different in that they are micron-size devices, they will use field emission instead of thermionic emission as the electron source, and they will be integrable with existing integrated circuit technology.

The usefulness of vacuum microelectronics will derive from their expected extreme radiation hardness and temperature insensitivity. Additionally, they may be extremely fast devices. These characteristics, if realized, will make them strong candidates for use in power and control applications in nuclear and EMP environments (for example, weapons systems, nuclear reactors, fusion reactors, and satellites).

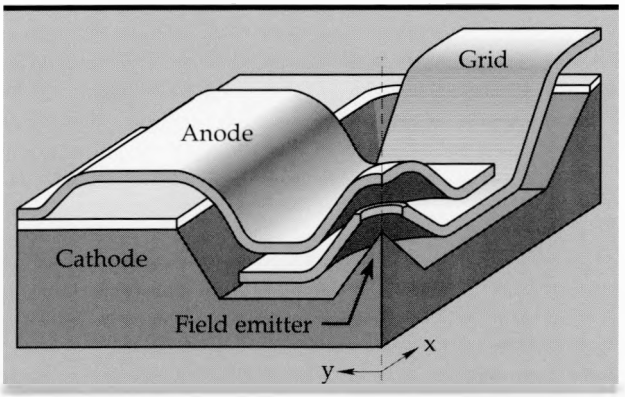


Figure 1. Cutaway of the structure of the vacuum triode shown in Fig. 2. The anode and grid are free-standing structures, supported only at the edges and hanging over the field emitter at the center of the device. The grid has a hole in it directly above the field emitter through which the field-emitted electrons can flow when a voltage is applied to the anode. Voltages applied to the grid will then control that flow of electrons from the field emitter to the anode.

According to research at Los Alamos National Laboratory, these devices could be capable of absorbing up to 10^{11} rad(Si)/s without upset, and more than 10^{17} neutrons/cm^2 and 10^8 rad(Si) of gammas without sustaining permanent damage. These radiation levels are three to four orders of magnitude greater than those that present silicon devices can withstand. The devices will be radiation hard because their active volumes will be a vacuum instead of being filled with silicon. Vacuum, unlike silicon, cannot be damaged and cannot be a source of photoionized charge carriers, which are the source of the large currents that cause upset and damage in silicon devices.

Vacuum microelectronics should also be able to withstand temperatures in excess of 775 K, as compared to a maximum of 650 K for the best silicon devices. The reason for this temperature insensitivity is the same as that for the radiation hardness; that is, the active volume is a vacuum, and thus the device cannot be shorted by thermally generated charge carriers.

Because the vacuum microelectronics are fabricated using integrated circuit technology, the potential exists to integrate a number of these devices on a single substrate to build circuits. This potential opens up whole new areas in electronics.

Progress

Progress in the last year has been in three areas—device design, tip sharpening, and modeling.

Device Design

The basic device structure is shown in **Fig. 1.** First, the surface of the silicon wafer is coated with

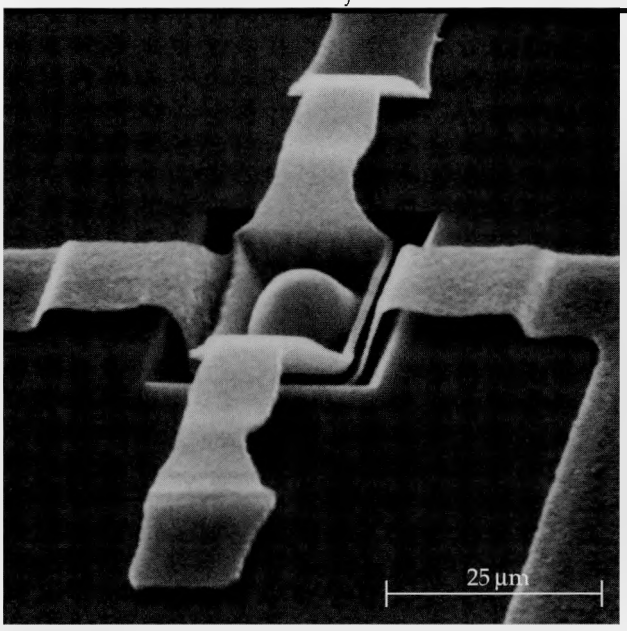


Figure 2. A miniature vacuum triode. The vertical structure on top is the anode, the horizontal structure that passes under the anode is the grid, and the field emitter is under the hemispherical bump on the anode. (See Fig. 1 for a cutaway of this structure.)

an isolation layer consisting of 0.5 µm of silicon dioxide and 0.4 µm of silicon nitride. Next, field emitters are created by etching around and behind a 5-µm-square mask of silicon nitride, using an isotropic etch of nitric acid, hydrofluoric acid, and acetic acid. The tips are sharpened by growing and removing a 500 Å to 1000 Å thermal oxide layer. We then bury the field emitter with glass, and apply a doped polysilicon grid. This structure is then buried again with glass and topped with a doped polysilicon anode. The final step is to remove the glass with hydrofluoric acid, leaving the free-standing anode and grid. A completed vacuum triode is shown in **Fig. 2** and a vacuum diode in **Fig. 3**; the latter is similar in structure to the triode just described, but lacks the grid.

Field Emitter Sharpening

The field emitters used in the initial fabrication of these devices produced insufficient emission to yield a useful device. To remedy this, we researched different methods of fabricating and sharpening these field emitters. Enlisting the aid of Prof. Charles Hunt of the University of California, Davis, and his student Johann Trujillo, we studied several different isotropic and anisotropic etches for creating field emission tips. The result of these studies was the isotropic etch mentioned above. This etch produced the field emitters shown in **Fig. 4**, having a radius of curvature on the tip of between

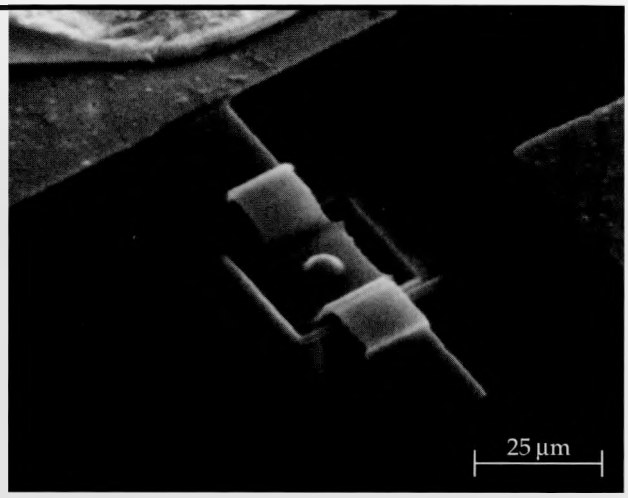


Figure 3. A miniature vacuum diode. The structure is essentially the same as the triode (Fig.2), but the device lacks the grid.

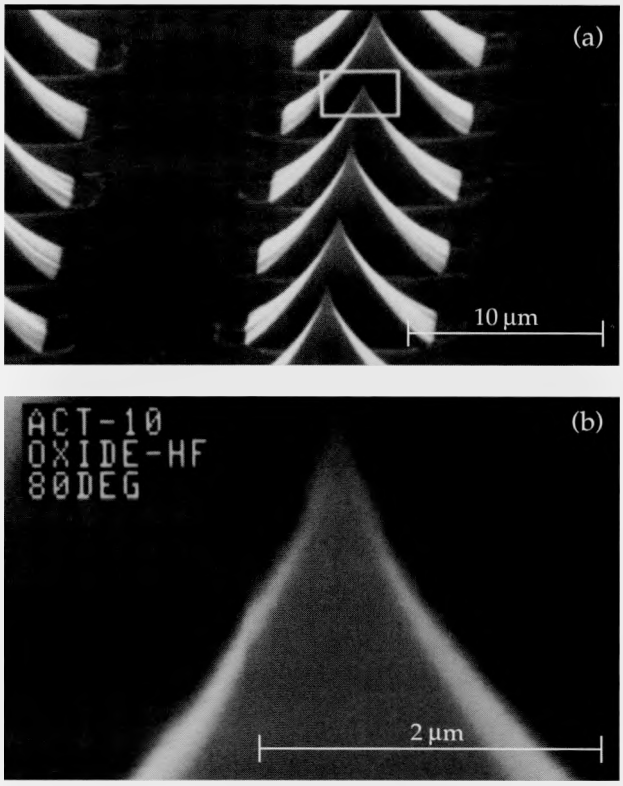


Figure 4. Isotropically etched field emitters. Using the nitric-hydrofluoric-acetic acid etch, we have produced field emitters with points that have a tip radius of from 400 Å to 1500 Å. View (b) is an enlargement of the tip enclosed in the rectangle in view (a).

250 Å and 1500 Å. A field of 10,000 to 15,000 of these tips gives the field emission results shown in **Fig. 5**.

While these isotropically etched tips give reasonable field emission, sharper ones would both give better emission and lessen the voltage requirements for the device. Thus, in a cooperative re-

search effort with Bob Marcus of Bellcore (Red Bank, New Jersey), we used an oxide growth technique to sharpen the tips. The resulting tips, shown in **Fig. 6**, have a radius of curvature at the point of less than 10 Å. We have imported this capability and are in the process of creating a miniature vacuum tube incorporating it.

Modeling

Using the combination of a two-dimensional static-field modeling code (STAT2D) and the theoretical Fowler-Nordheim equation, we have estimated the operating characteristics of our vacuum triode (**Fig. 7**). The curves were created by fitting the calculated field at the tip of the field emitter to the voltages on the anode and grid, inserting that fit into the Fowler-Nordheim equation, and then normalizing the results to some experimental field-emission data. The resulting equation gives the current for different voltages applied to the anode and grid. By taking the appropriate derivatives of this equation, we obtain the tube parameters. Note that these tube parameter values are not typical of those expected from more traditional vacuum tubes. The plate resistance is higher, and the transconductance is lower. These differences must be taken into account when designing circuits for these devices.

Future Work

Future work on this project will run in several directions. In the device-fabrication work, we are in the process of building a complete vacuum triode utilizing the device structure shown in **Fig. 2** and the sharpened tips shown in **Fig. 6**. As soon as an operating device is available, we will test the thermal and radiation hardness. We are also building a lateral device in which the field emitter points horizontally (in the plane of the surface of the silicon wafer) rather than vertically out of the surface. The device is formed by first coating the surface of the silicon with an insulator and a metal layer, and then forming a point in the metal by etching to form the field emitter. The anode is formed by a metal strip just beyond the tip of the field emitter. While we cannot control the anode-to-cathode spacing as well in a lateral device, such devices are much simpler to build. We will continue to work on field-emitter design and sharpening, especially focusing on the uniformity of our field emitters. Because of our new field-emitter design, the modeling results need to be redone to take the new shape and sharpness into account. The extremely

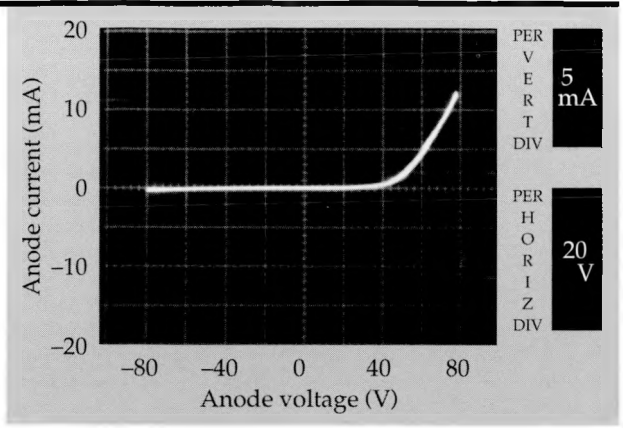


Figure 5. Current vs voltage curves for a field of 10,000 to 15,000 field emitters spaced 2 μm from a removable anode.

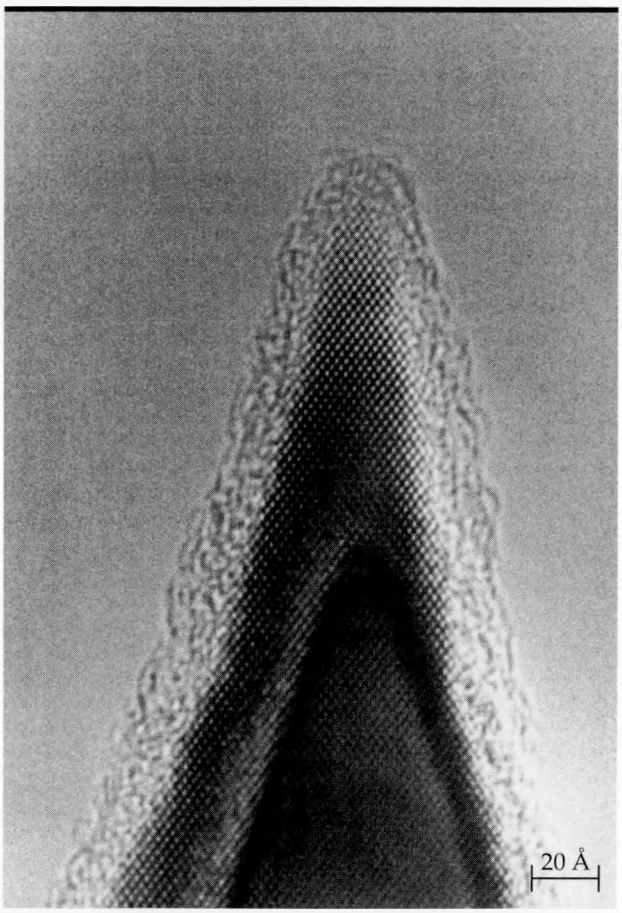


Figure 6. TEM (transmission electron microscope) photo of a sharpened silicon field emitter with a radius of curvature of less than 10 Å. The field emitter is coated with a 10- to 20-Å layer of silicon dioxide. This *native* oxide layer is formed whenever silicon is exposed to air. The fringing in the silicon is caused by scattering of the electron beam by the {111} crystal planes, which are spaced 3.13 Å apart. The fringing disappears near the tip because there are too few atoms left to scatter the incident electron beam. Thus, the point is less than four or five atoms across.

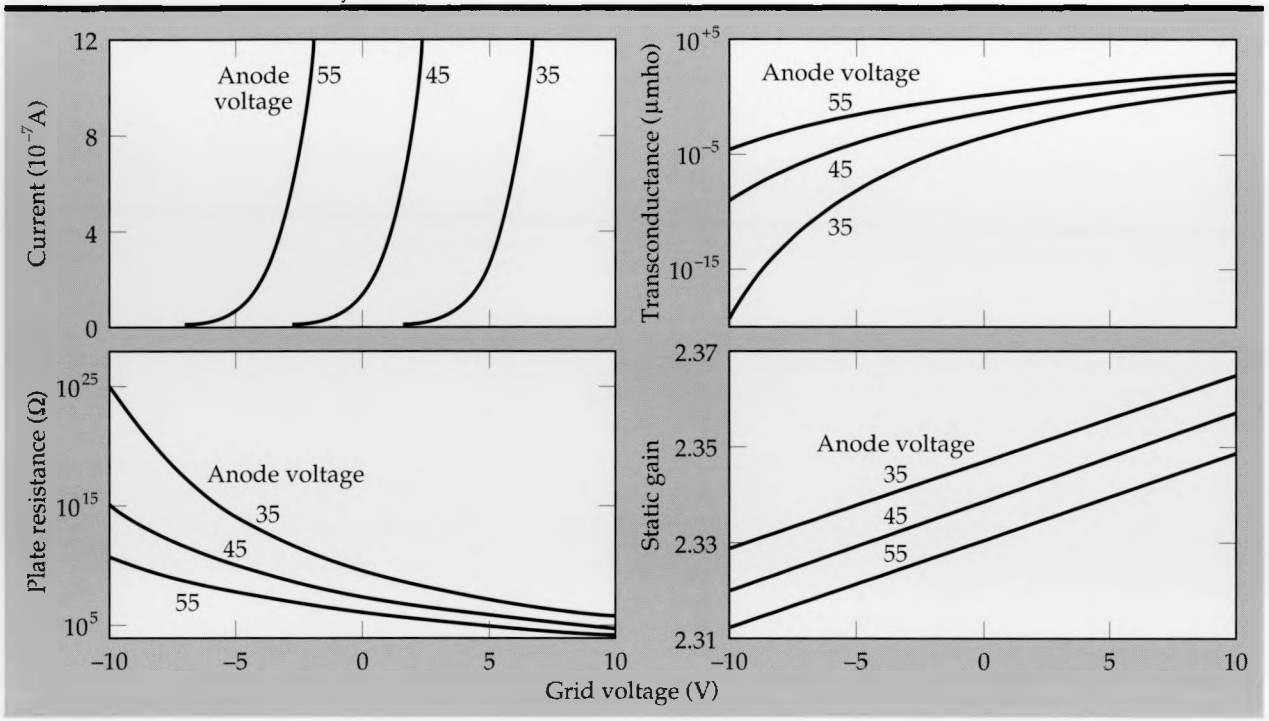


Figure 7. Theoretical current-voltage, plate resistance, transconductance, and static gain curves for the miniature vacuum triode. The calculations were performed assuming a 1.5-μm anode-to-cathode distance, with the tip of the cathode centered in a 1-μm-radius hole in the grid. The data was then normalized to some experimental field-emission data.

productive collaborations with U.C. Davis and Bellcore will continue with more studies on tip sharpening and design. At the completion of this project, we should have the capability to design and fabricate miniature vacuum switching and control devices and be able to predict their operational characteristics.

Modeling the Performance of Photoconductors

W. Travis White III, Clyde G. Dease, and Gizzing H. Khanaka
Engineering Research Division
Electronics Engineering

To help develop fast, optically triggered high-voltage pulsers in support of various LLNL programs, we are modeling photoconductive switches. Our goal is to provide the physical understanding necessary to design reliable and efficient switches. This year, we investigated two main issues: the effects of Gunn instabilities and the role of traps.

Introduction

The Laboratory is investigating gallium arsenide (GaAs) photoconductors for applications that require subnanosecond high-voltage switches. (A photoconductor is a semiconducting slab that becomes electrically conductive when exposed to light.) At low voltages, such switches can close on a time scale of picoseconds. However, because low-voltage switches are typically only a few microns thick, they suffer dielectric breakdown at high voltages. Our switches must be millimeters thick to hold off the high voltage (tens of kilovolts) in pulsers. The performance of these thicker switches cannot be extrapolated from data on conventional thin devices because interactions among electrical charges, electric field, and electron-hole traps are too complex.

A combined experimental and theoretical effort is required to develop reliable and efficient subnanosecond high-voltage photoconductive switches. The Pulsed Power thrust area leads such an effort at the Laboratory; the Computational Electronics and Electromagnetics thrust area provides modeling support.

One of the goals of the project is to identify which mechanisms dominate the switching behavior of our devices. From a knowledge of the dominant switching mechanisms, we want to establish the fundamental limits on the closing speed, recovery time, and switched-out voltages. Additionally, we need to identify the failure mechanisms in our devices and determine how device failure relates to the overall switching performance.

We have conducted experiments with 5-nsec diode-laser pulses illuminating semi-insulating GaAs switches. The devices show a delay of 2 to 3 nsec between the start of illumination and the onset of strong switching. During the delay, a comparatively constant, weak current occurs. At the end of this delay, the current rises suddenly by well over an order of magnitude. The rise in current, which we refer to as the closing of the switch, does not occur unless the spatially averaged electric field exceeds a critical threshold approximately equal to 4 to 5 MV/m. These results show that the mechanism of switching—whatever it is—is a function of the initial electric field and occurs with a voltage-dependent delay.

Additionally, we, Sandia National Laboratories, and others have observed a "lock-on" effect in high-voltage GaAs photoconductors.[1] It occurs at field strengths between 0.8 and 2.5 MV/m and may result from field-current instabilities. Instead of recovering to full voltage when the optical pulse ceases, switches recover only partially, locking onto a fixed voltage. The voltage persists until the external power supply is turned off. The details of lock-on are not understood, although Sandia speculates that the effect may be a result of impact ionization of traps.[2]

The correct theoretical model of switching should account for the magnitudes of the voltage threshold, the voltage-dependent delay, and the lock-on effect. We are considering the following mechanisms for switching: impact ionization, field-dependent trapping, and instabilities associated with negative differential resistance (**Fig. 1**). All of these mechanisms depend on the electric field strength; all of them can produce delayed, sudden large rises in current; and all of them can be associated with known device-failure mechanisms. Present experimental data do not discriminate among the various mechanisms. Hence, we must explore them theoretically.

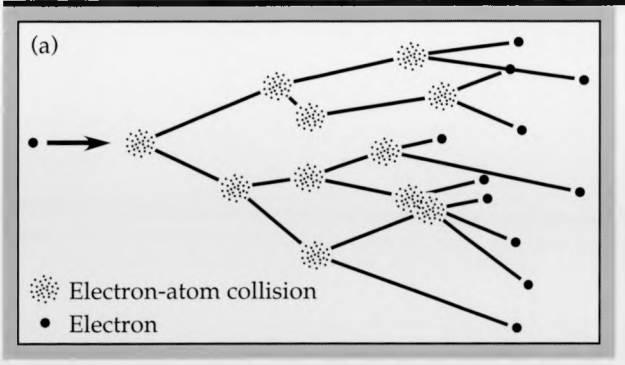


Figure 1. Three mechanisms for switching in photoconductors. (a) *Impact ionization.* A fast electron collides with a bound electron, freeing it to move through the lattice. An applied electric field accelerates the free electron as it moves, enabling it to knock loose more bound electrons, which accelerate and knock loose still more electrons. This cascade process rapidly multiplies the number of mobile electrons and produces a condition called *current avalanching.* The bound electrons may originate in lattice defects (traps) as well as in perfect bonds (valence states). The energy to ionize a trap is almost always much less than that required to ionize a valence state. Therefore, the electric fields necessary for impact ionization are correspondingly smaller for traps than for valence states. If the number of traps is large, which it is in the semi-insulating GaAs used for photoconductive switches, then a large number of trapped electrons (5×10^{22} m^{-3}) are available to contribute to current avalanching.

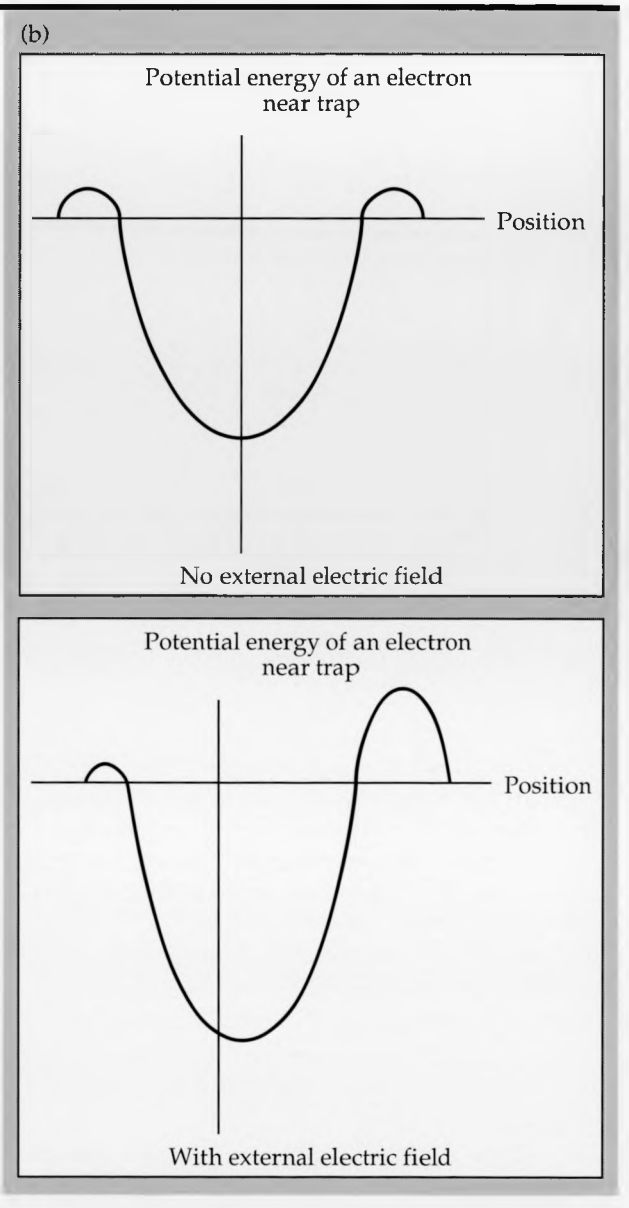


Figure 1 (continued) (b). *Field-dependent trapping or detrapping.* Externally applied electric fields distort the bonds near a lattice defect, raising or lowering the potential-energy well that is responsible for trapping electrons. If the well becomes shallower, then electrons can leave traps more easily, which raises the background current density and facilitates impact ionization. If the well becomes deeper, then the probability for trapping an electron increases. Although increased trapping reduces the photocurrent, the photocurrent will suddenly increase when all of the traps are occupied.

Progress

In FY 88, we showed with computer calculations that a nonuniform distribution of traps can significantly enhance the electric field across an illuminated photoconductor.[3] Photogenerated charges produce a large electric field that combines with the external field due to a battery or a power supply. This combination temporarily produces a high total field at certain locations inside the switch. We have dubbed the process *dynamic field enhancement.* The trapping of charge near an electrode further enhances the field. According to our low-field calculations, local field enhancements can exceed a factor of five when the switch contains a nonuniform distribution of traps. However, with either a uniform distribution of traps or no traps at all, the maximum low-field enhancement never exceeds a factor of 2.6. These results have provided the first theoretical evidence that intrinsic current avalanching might be possible at field levels five times below the expected threshold. While encouraging, these findings are only a first step in understanding the mechanics of photoswitching. They apply to situations in which the initial electric field is relatively weak. Since actual measurements occur at high fields, we began to study high-field switching effects in FY 89.

First, we examined effects of field-dependent carrier velocity, particularly NDR, and the formation of narrow zones of high charge density called

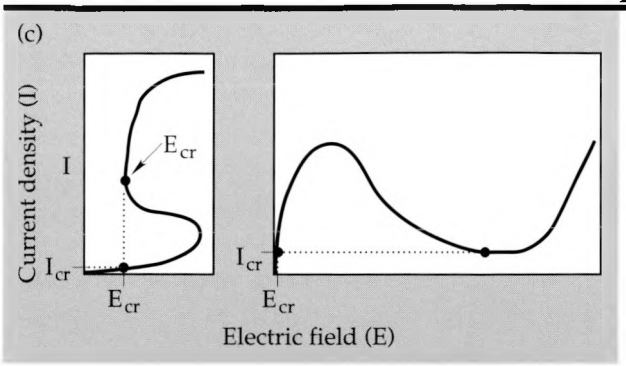


Figure 1 (continued) (c). ***Negative Differential Resistance (NDR).*** **NDR is not a single mechanism but a family of mechanisms. Any process that reduces the *current* when voltage is raised belongs to this family of mechanisms. The figure shows two classic current-vs-voltage curves with a region of NDR: an S-shaped curve and an N-shaped curve. Note that for current densities above a certain critical current on the S-shaped curve, one field can correspond to two or more different current densities. Consequently, the field in a given plane may be nearly constant; yet, the current density may assume any of several different values. This form of multi-stability leads to filamentary instabilities. In contrast, for fields above a certain *critical field* intensity on the N-shaped curve, one current can correspond to two or more different field strengths. As a result, the field may vary sharply as a function of position while the current density remains virtually constant. Associated with the N-shaped curve are charge-density waves called *Gunn domains*, thin sheets with large numbers of electrons that move towards the anode.**

Gunn domains. We tried to determine computationally if field enhancements due to NDR alone can produce bulk current avalanching in an ideal uniform device. Additionally, we tried to determine if the lock-on effect might be a result of the formation of Gunn domains.

Our model treated the GaAs material as semi-insulating by assuming that 10^{22} m^{-3} traps compensated 10^{13} m^{-3} p-type impurities. We assumed that the traps that did not compensate a hole were occupied (neutral) initially. Simple chemical kinetics governed the occupancy of the traps as the system evolved in time. We used room-temperature values for the cross sections and emission rates of the dominant trap, EL2. The one-dimensional code that we used has been previously described.[4] The photoconductor in our model was a 0.1-mm-thick slab of GaAs with two planar metallic boundaries. Electrically, the switch was in series with an ideal battery and a 100-Ω resistor. For the Gunn domain calculations, we assumed it to be illuminated uniformly between the electrodes. We chose an energy per photon just large enough to create one electron-hole pair for each absorbed photon.

Our calculations showed Gunn domains forming and moving across the device (**Fig. 2**). The computed growth and motion of the domains agreed with established theories of Gunn instabilities.[5] These results indicate that NDR can produce large local fields and that current avalanching can occur at the electrodes as a result of the field enhancement even if the initial electric field is as low as 5 MV/m. Despite the avalanching, however, the photoconductor in our model did not completely short circuit because we maintained a low enough optical intensity to isolate various physical effects. At a bias field of 1 MV/m, the voltage across the switch fell and rose roughly in proportion to the instantaneous optical intensity. Although 100 V was applied initially across the switch, only 48 V was switched out. At a bias field of 5 MV/m, the magnitude of the peak switched-out voltage reached only 60 V. The peak density of photogenerated electron-hole pairs was about 3×10^{22} m^{-3}, and the Gaussian optical pulse had a full width of 10 nsec between the $1/e$ points. Although the bias fields differed by a factor of five, the switched-out voltage was almost the same in both cases.

This simulated switching behavior illustrates an important practical problem of photoconductive switches—velocity saturation. As the electric field rises, the induced velocity of charge carriers (the *drift velocity*) reaches a limiting value. Further increases in the voltage across the switch at a fixed optical intensity do not increase the photocurrent once the drift velocity saturates. Switched-out voltage is proportional to the photocurrent. Consequently, the switch closes less efficiently as the initial voltage is raised. The only way to make the switch close completely is to force more current out of it. In the absence of charge multiplication, the optical intensity must be increased.

The computations showed two other interesting features. First, although the electric field and charge density both vary rapidly as a function of position when the Gunn domains form, the external voltage across the switch does not fluctuate noticeably. We observe this phenomenon even when Gunn domains reach the electrodes. Second, the electric field in the wake of the Gunn domains tends to lock onto one of two approximate values: 0.10 to 0.14 MV/m and 1.4 to 1.7 MV/m. Coincidentally, these values correspond roughly to the lowest electron and hole velocities, respectively, at which the field-vs-velocity curves in our model become double-valued. At such a value, the speed of a charge carrier (and hence, the current) remains constant as the carrier moves across a domain boundary from low to high fields. This phenome-

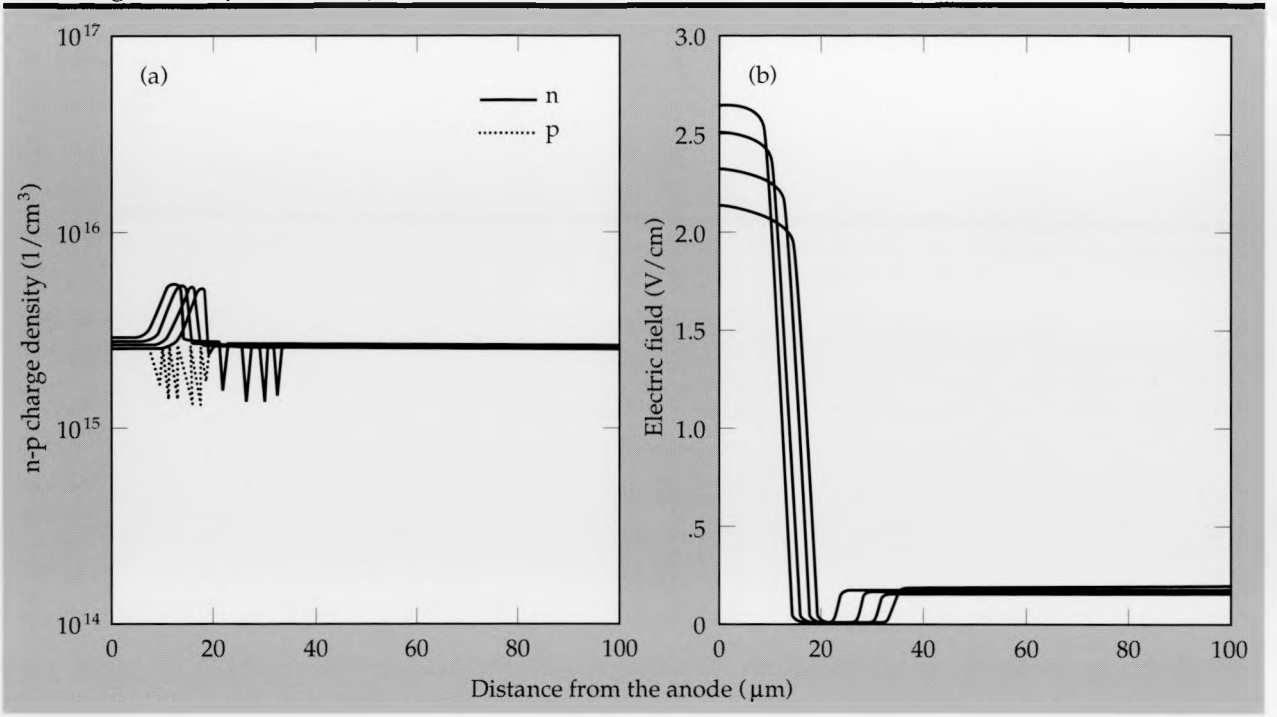


Figure 2. Field enhancement from Gunn domains. Shown are (a) plots of carrier density and (b) electric field strength vs position in the switch at four closely spaced times. At electric fields greater than 0.35 MV/m in GaAs, an increase in field strength reduces the velocity of electrons. Any irregularity in the field causes electrons to pile up where the field is greatest and to form a Gunn domain. Near the metallic contacts at the edges of a photoconductor, the electric field falls sharply to zero. Due to this sudden drop from high to low field strength, instabilities then naturally originate at the electrodes and propagate with the electrons toward the anode. As they move, the electrons pile up ever more sharply, raising the field in front of them and lowering it behind them. The strong gradient in the field reinforces the tendency of the electrons to pile up. This further increases the field enhancement until the Gunn domain reaches the anode. The domain can either exit the device at the anode or be sustained by impact ionization if the field is high enough.

non allows the fields to vary strongly across a device without significantly perturbing the local current density. Assuming the current varies smoothly with position, no large bursts of current will occur when the Gunn domains reach the electrodes. Thus, internal field lock-on helps produce a smooth current output. Our results do not show whether or not this field lock-on is tied to the experimentally observed voltage lock-on effect. More experiments are needed to resolve this issue.

In addition to effects caused by a field-dependent velocity, we also considered field-dependent trapping. We used a rate-equation analysis to show that field-dependent trapping can explain the electronic pulse-shape that we have measured on some of our switches. The rate-equation model presumes that spatial variations inside the photoconductor are unimportant and that the switch can be treated as a two-terminal nonlinear electronic component. The rate equations consider the time variation of the numbers of mobile charge carriers and empty traps in the switch. We ignored impact ionization of traps in our model.

To simplify the analysis, we assumed that the optical intensity was constant and that electrons had the same generation and recombination rates as holes. In this approximation, the numerical solution shows that the density of carriers has a two-step behavior (**Fig. 3**). We observe the density first rising until the rate of trapping matches the rate of photoabsorption. The density of electrons then remains constant until the traps eventually fill. Once all of the traps are occupied, the density rises again until the rate of recombination equals the rate of photoabsorption. We modeled the field dependence of trapping with a static approximation, choosing the trapping cross section as a function of initial voltage only and not of instantaneous voltage.

Current is proportional to carrier density times carrier velocity. Thus, the current also has a two-step response. If the first step is much smaller than the second, then the photoconductor appears to switch after a delay, as observed experimentally.

We chose the parameters for our model so that the computed pulse-shape resembled observed

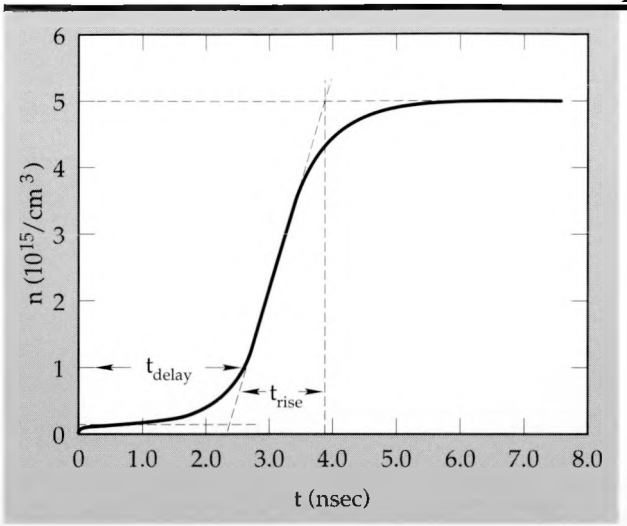


Figure 3. Two-step switching by trap filling. During illumination by a constant optical intensity, the density of photogenerated carriers, n, rises until the rate of trapping matches the rate of photoabsorption. The carrier density then remains constant. Eventually the traps fill, and the density rises again until the rate of recombination equals the rate of photoabsorption. Photocurrent is proportional to the density of carriers. Therefore, to the outside observer, the photocurrent appears to suffer a delay in switching due to the filling of traps. The length of the delay is the time required to fill the traps. It varies inversely with the optical intensity. The rise time is the time required for recombination to build to high levels and, in this model, varies inversely with the square root of optical intensity.

data. All parameters were within an order of magnitude of measured physical values, except the recombination rate constant. It had to be 1000 times stronger in our calculation than the measured radiative recombination rate constant of GaAs. Such a large recombination constant may suggest that a nonradiative process dominates recombination in our experiments.

The rate equations, although simplistic, have an important advantage over detailed numerical simulations. They can produce analytical solutions that help us identify physically important combinations of parameters. Thus, the simplified model is useful in guiding the interpretation of more elaborate calculations and also in suggesting meaningful experiments.

Future Work

We have identified several mechanisms that may account for the behavior of our GaAs high-voltage switches. By calculations and analysis, we have shown that dynamic field enhancement can occur as a result of both NDR effects and nonuniform distributions of traps in the material. In turn, the field enhancement can produce current avalanching and thereby short circuit (close) the switch. NDR effects can result from field-dependent trapping as well as from effects that are totally unrelated to traps. Current avalanching can result from impact ionization of traps or from intrinsic bulk avalanching at localized hot spots of the electric field. For sufficiently long optical pulses, trap filling can lead to a two-step switching similar to what has been observed experimentally. To identify which mechanisms actually dominate the performance of our switches now requires additional experiments.

In addition to modeling switching speeds, delays, and voltages in FY 90, we will study the effects of thin barrier layers at electrodes and field-dependent trapping parameters. This work will directly extend our present study of switch performance and will help guide future experiments.

Some future experiments suggested by this year's results include:

- Measurement of the density of traps as a function of position in our samples,
- Comparison of the instantaneous optical input to the instantaneous output voltage from the switch in an attempt to determine delay time, rise time, and output current as functions of optical parameters (e.g., intensity), and
- Search for evidence of dynamic field enhancement and Gunn domains in the switches.

By working with experimenters on these issues, we expect to identify the basic switching mechanisms in our high-voltage photoconductors this coming year.

1. G. M. Loubriel, M. W. O'Malley, and F. J. Zutavern, "Toward Pulsed Power Uses for Photoconductive Semiconductive Switches: Closing Switches," *Digest of Technical Papers, 6th IEEE Pulsed Power Conference,* 145–148 (1987).
2. H. J. Hjalmarson, Sandia National Laboratories, Albuquerque, N. M., private communication (1989).
3. W. T. White III, "Modeling the Performance of Photoconductors," *Engineering Research and Development Thrust Area Report FY 88,* UCRL 53868-88, Lawrence Livermore National Laboratory, 1-24–1-26 (1988).
4. W. J. Orvis, G. H. Khanaka, and J. H. Yee, "Development of a GaAs Solid-State Device Model for High-Power Applications," *Electrical Overstress/Electrostatic Discharge Symposium Proceedings, 1987,* 145–152 (Orlando, Florida, Sept. 29–Oct. 1, 1987).
5. S. M. Sze, *Physics of Semiconductor Devices, Second Edition* (John Wiley and Sons, New York, 1981).

Accelerator Cavity Modeling with TSAR

Robert R. McLeod, Steven T. Pennock, and John F. Deford
Engineering Research Division
Electronics Engineering

During FY 89, we adapted the three-dimensional electromagnetics code TSAR for use in modeling particle accelerator cavity problems of interest to the Laboratory's free-electron laser program. The TSAR code, in concert with the advanced pre-processing codes ANA and IMAGE, provides an advanced tool for investigating the behavior of accelerator cavities. Comparison of TSAR models with those produced with older, two-dimensional codes has shown that TSAR produces accurate results, and we are now applying TSAR to accelerator cells with fully three-dimensional geometries.

Introduction

Computer analysis is an important element in designing particle acceleration devices. Designers need detailed information about the behavior of components of the accelerator, especially the cells in which the acceleration of the particles occurs. In particular, information on the electromagnetic (EM) properties of those cells and how EM fields induced in them depend on the cell geometry is needed. However, the complicated and often novel cells designed at LLNL defy many of the traditional analyses. We have found that advanced numerical simulation of the accelerators can provide solutions to the problems encountered in traditional analyses.

Conventionally, simplifying the cavity geometry has been an initial step towards these solutions. Many accelerator cells are dominated by circularly symmetric features: the cylindrical beam pipe, radial induction gaps, and so forth. Thus, a common starting point with conventional analyses is to assume complete symmetry and reduce the problem to two dimensions.[1] However, this has the disadvantage of removing three-dimensional features that may strongly affect the behavior of the cell. A fully three-dimensional code such as TSAR (Temporal Scattering and Response) can work with the true, unsimplified cell as well as with complex cells that cannot be handled with standard two-dimensional methods.

In addition to the three-dimensional geometrical features of these devices, many cells designed at LLNL contain complex materials, such as ferrites. The presence of these materials necessitates a three-dimensional simulation even in symmetric devices. For instance, the nonlinear behavior of the ferrites can cause irregular, three-dimensional domains of magnetization to form. These features can be simulated in a code such as TSAR.

Progress

Progress during FY 89 was accomplished in several areas.

Modification of TSAR

TSAR was developed initially as a tool for calculating the electrical response of devices to external microwave radiation. For that purpose, objects were placed in the center of the computational volume, and special "radiating boundary conditions" were applied to the edges of the volume to simulate a free-space environment. The analysis of accelerator cells, however, is characterized by internal stimulation in a cavity with occasional transmission lines or ports leaving the problem. Though these applications are similar, the conditions imposed by the cell analysis required us to incorporate several new features into TSAR.

Primarily, a variety of new boundary conditions were needed in TSAR. First, an "impedance" condition was added to terminate accelerator ports, areas of the accelerator cells that connect the internal working volume to the outside world. The ports are modeled as transmission lines, and the impedance boundary condition allows the user to specify a particular field impedance (a specific relationship between the electric and magnetic fields) at the end of the transmission line. Next, symmetry plane boundary conditions were added. With these, the memory and time needed to model cells that have one or two planes of symmetry can

be halved or quartered, respectively. This is especially important for simulations of large cells that require massive amounts of memory to run.

Tools

A simulation code capable of processing these very large accelerator cell models is only one component of a successful modeling tool. Such a code is only as good as the models of the problem the user can create for it. The pre-processing tools for TSAR are very powerful, giving us the capability to rapidly create and modify large, complex models.

The core of these problem-generation tools is a solid modeler, MGED (Multi-Device Graphics Editor).[2] This modeler allows the analyst to construct a three-dimensional representation of the accelerator cell using combinations of simple geometric shapes. A discrete finite-difference model for TSAR is then created from the MGED solid model by ANA, a code developed by Gary Laguna at LLNL.[3] ANA has been sufficiently validated that a user can now have confidence that the mesh it produces is a good representation of the solid built in MGED. During previous forms of the mesh generator and during the validation of ANA, Brian Cabral and others at LLNL developed IMAGE,[4] an interactive graphics program that allows the user to view three-dimensional meshes for verification and modification purposes.

With these three codes—MGED, ANA, and IMAGE—researchers and designers at LLNL have a powerful and unique capability to create complex, three-dimensional computational models. The combination of these tools and the general TSAR simulation code has allowed us to investigate a number of different accelerator problems.

Accelerator Designs Modeled with TSAR

Most of the accelerator design problems to which TSAR has been applied have been of the "very asymmetric" variety. This means that these problems are tractable only in a three-dimensional environment such as the one provided by TSAR. Examples of the accelerator problems of interest to the free-electron laser program that have been simulated with TSAR include the free-electron laser side-coupler injector, the beam breakup study cell, the four-port measurement section, and the advanced test accelerator fast correction coil.

The fast correction coil is typical of our present work. This device was designed to examine the possibility of steering the electron beam in LLNL's advanced test accelerator (ATA) in real time. Designers needed to know how the structure of the cell would passively affect (through the beam's wakefields) the beam dynamics, as well as how the accelerator cell itself would function.

The first step in getting the needed information was, as outlined above, to construct an MGED solid model of the cell, as shown in **Fig. 1.** In this cutaway view, one can see the interior of the beam pipe as well as the current steering bars. This solid model was then used by ANA to produce a computational mesh for TSAR. A cross-section of this mesh, viewed with IMAGE, is shown in **Fig. 2.** In a typical TSAR run, a current beam was driven

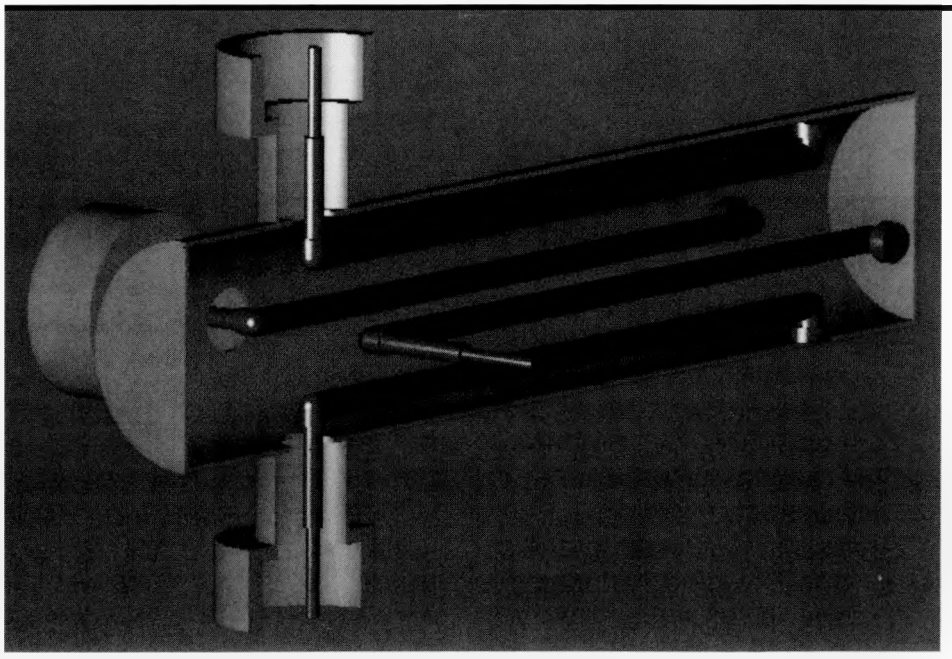


Figure 1. Ray-traced view of fast correction coil solid model constructed with MGED. In this cutaway view, the interior of the pipe beam and the current steering bars are seen.

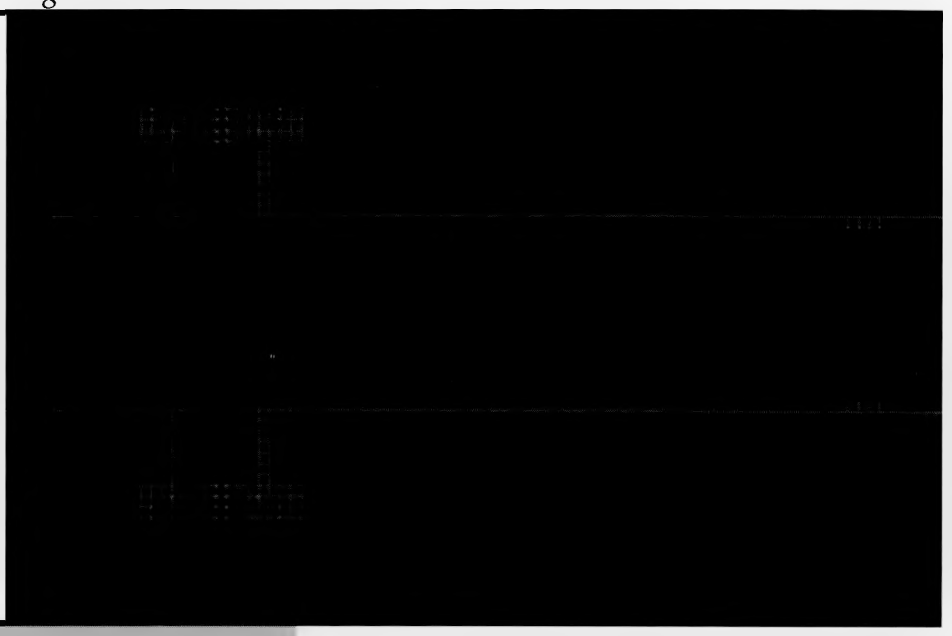


Figure 2. Slice of finite-difference mesh produced by ANA from model shown in Fig. 1 and viewed with IMAGE.

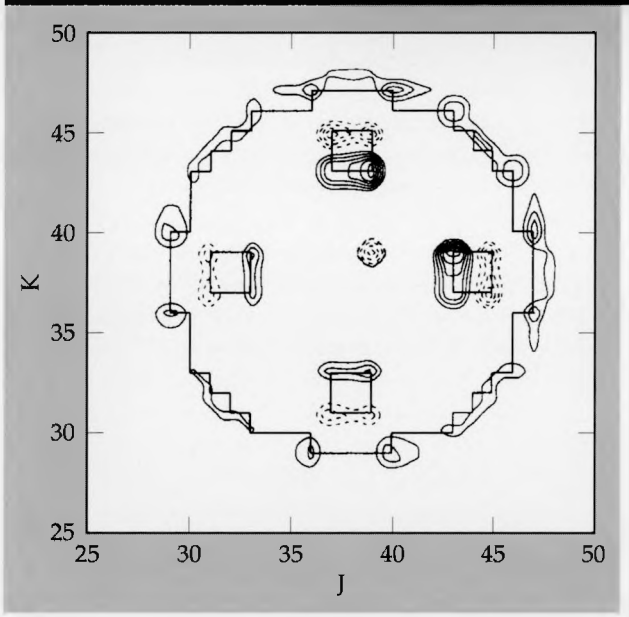


Figure 3. Contour plot of axial currents in fast correction coil with off-center stimulation. Solid contours are positive currents, and dashed lines are negative currents. The current beam can be seen above and right of center.

down the length of the cell, slightly off of the center line. This induced return currents on the correction bars and on the cell walls. In **Fig. 3,** both the beam current (above and right of center) and the return currents can be observed. An important operating mode revealed by this data is the presence of circulating currents on the outside of the bars that travel in the same direction as the beam. The discovery and understanding of these types of unexpected features is continuing computationally and experimentally. It is an example of the importance and utility of validated electromagnetics modeling codes like TSAR.

Future Work

Work in the coming year will focus on two areas: running larger, more detailed models and implementing a ferrite-simulation capability. The first task will be accomplished by conforming the computational grid to the size and shape of the cell. Currently, large amounts of memory are wasted outside of the cell where all of the EM fields are zero. The second task, ferrite simulation, involves the development and implementation of new materials subroutines that describe the ferrite physics. This work is underway and will greatly expand our simulation capabilities.

1. J.F. DeFord, R.R. McLeod, and G.D. Craig, "The AMOS Wakefield Codem," *Thrust Area Report FY 88,* Lawrence Livermore National Laboratory, Livermore, CA, UCRL-53868-88 (1988), pp. 1-27 to 1-30.
2. M.J. Muuss, et al., *Ballistic Research Laboratory CAD Package, Release 1.21,* Ballistic Research Laboratory Internal Publication (June 1987).
3. G.W. Laguna, "Solid Model FDTD Mesh Generator," *Thrust Area Report FY 88,* Lawrence Livermore National Laboratory, Livermore, CA, UCRL-53868-88 (1988), pp. 1-8 to 1-11.
4. G.W. Laguna, W.T. White, and B.K. Cabral, "Generating Meshes for Finite-Difference Analysis Using a Solid Modeler," Lawrence Livermore National Laboratory, Livermore, CA, UCID-21189 (1987).

Localized Transmission of Wave Energy

Richard W. Ziolkowski
and D. Kent Lewis
Engineering Research Division
Electronics Engineering

We are investigating, both theoretically and experimentally, a new mode of wave energy transport called localized wave transmission (LWT), in which pulses do not dissipate or diminish in amplitude when propagated over extreme distances. As a part of this research, we are studying the launching and propagation characteristics of LWT. Experiments substantiating the theory have been performed in an acoustic environment.

Progress

We have been developing specialized wave system solutions that are capable of synthesizing pulses which can give localized transmission of wave energy in a variety of physical systems.[1-19] These localized wave transmission (LWT) solutions describe a localized (highly directive), slowly-decaying transmission of energy in space-time. We have shown in theory that LWT can occur in a variety of wave systems, including scalar wave equation systems (acoustics), guided wave systems (wave equation with boundaries), electromagnetic systems (Maxwell's equations), particle systems (Klein-Gordon, Schrodinger, and Dirac equations), and plasma systems (Maxwell's equations combined with a fluid model of the plasma).

We have also developed a new representation of the solutions for these wave systems, based on a decomposition into base states consisting of bidirectional traveling plane-wave solutions (i.e., products rather than sums of forwards and backwards). This bidirectional representation is invertible and has been successful in reproducing standard time-dependent solutions as well as the new LWT formulations. It has proven more effective than conventional Fourier plane wave decompositions for describing LWT pulse solutions that can be tailored to give the LWT effect in a variety of physical systems.

These LWT pulses may impact several fundamental ideas of physics involving the wave-particle duality concept. The latter has been a main thread in the evolution of quantum mechanics. Wave-particle duality connects the wave behavior of a photon in Young's two-slit interference topology with the photon's particle behavior in photoelectric experiments. We have been investigating the use of the fundamental LWT solutions themselves as a model of a photon. These solutions exhibit both local and nonlocal characteristics usually associated with wave-particle duality notions. They can assume either a transverse plane wave or a particle-like character depending on the size of the defining parameters. We have modeled the scattering of a single LWT wavepacket in a Young's two-slit experiment to show how a photon can interfere with itself to recover the classical two-slit diffraction pattern. Similarly, we have uncovered LWT solutions of the Klein-Gordon and Dirac equations that are classical equations describing elementary particles such as electrons and protons. These solutions exhibit the standard billiard-ball characteristics associated with macroscopic entries together with the wave-like properties of microscopic ones. We are currently exploring the implications of the LWT solutions in ionization, Compton scattering, and photoelectric events.

Traditional wave solutions to linear systems are generally treated as monochromatic (single-frequency or CW). This separates the space and time characteristics of the waves and allows the immediate application of Fourier analysis. Laser-produced Gaussian beams, for instance, are studied in this fashion. The resulting solutions begin to spread after their Rayleigh distance is surpassed, a distance determined by the original size of the wavepacket and its frequency. Conventional antenna systems such as uniphase or phased arrays are driven essentially with monochromatic signals allowing only for spatial phasing. The resulting diffraction-limited signals begin to spread and decay when they reach the Rayleigh distance. In contrast, the LWT pulses are intrinsically space-time oriented, i.e., their space and time properties are nonseparable. This allows one to design pulses whose constituents remain highly correlated in space and in time.

The LWT effect is a byproduct of the extra degree of freedom offered by the full utilization of space-time. Moreover, LWT pulses are character-

ized by a high bandwidth, with the difference in hertz between high and low frequency components being as high as 10:1. This frequency range is much wider than that of traditional CW systems whose bandwidth may be less than 1.0%. Thus, one is paying for the enhanced localization by increasing the requisite interconnections of the spatial and temporal characteristics and the associated bandwidths of the pulses that are used to construct the LWT wavepacket. We have been faced with the difficult fact that there is no precise value for the Rayleigh distance in the LWT case because of the broadband character of its constituent pulses. The "goodness" of the pulses is thus difficult to characterize. However, one can design the LWT solutions to remain localized over a distance many times greater than the Rayleigh distance expected from traditional solutions. As a result, the LWT effect can be viewed as extending the near field to very large distances.

Having uncovered a very interesting class of theoretical pulses that describe localized transmission of wave energy, we have begun to determine if the LWT fields can be replicated by a finite set of sources, e.g., by a finite planar array of radiating elements. The hope is to produce fields that would be extremely close facsimiles of the exact solutions, and hence would share their properties. We are currently investigating potential LWT pulse launching mechanisms both numerically and experimentally. The type of system that is required may be characterized as an independently-addressable, pulse-driven array of radiating elements.

In particular, we have been studying the physical realizability of launching one of the LWT solutions, called the modified power spectrum (MPS) pulse, with numerical simulations and with acoustic experiments. The MPS pulse is so named because of the form of the weighting functions used to construct it from the basic LWT solutions. The physical characteristics of the MPS pulse are very appealing. This pulse can be optimized so that it is localized near the direction of propagation and its original amplitude is recovered out to extremely large distances from its initial location.

The current numerical model of an MPS launching system is a finite planar array of point sources. The array elements radiate spherical pulses that can be described by a causal, time-retarded Green's function. This construction is known as Huygens' representation. The essential idea is that with an array of point sources one is always in the far-field of each radiator and the overall field response of the array is readily obtained by superposition. The causal, time-retarded nature of the propagation insures that the wave is leaving the antenna in the desired direction at the correct time. We have also constructed a version of this simulator that decomposes the input drive functions into their Fourier components, propagates each Fourier component away from the radiators, and then recombines these constituent wavelets into the total field. This alternate model provides us with a tool to study the pulse reconstruction as a direct function of its frequency content.

In contrast to conventional arrays, the LWT driven arrays we are investigating consist of radiating elements that are driven by a specified spatial distribution of signals having broad bandwidth time signals. The resulting field is a sum of these individually radiated pulses. We have shown numerically that this combination of spatial and temporal distributions of driving elements generates pulses that maintain their localized shape at distances well beyond the conventional Rayleigh distance. Unlike the past attempts at enhanced localized transmission such as super-gain antennas, the LWT driven arrays appear to be very robust and insensitive to perturbations in the initial aperture distributions. Although they were not optimized, these results are sufficient to indicate that from a modest-size array, energy can be transmitted without spreading over interestingly large distances.

We have performed a number of acoustic experiments to test some of these claims for the MPS pulse. A series of water tank experiments have shown evidence of the LWT effect, and have demonstrated that the individual field components can indeed be generated by real sources. Moreover, an array of these sources recreated a reasonable facsimile of the MPS pulse away from the array. It was generated from a synthetic aperture consisting of a single ultrasonic transducer radiating the requisite pulses and a matched receiving transducer. The resulting waves were reconstructed by computer synthesis. Both linear and square arrays have been studied. Comparisons were made with the LWT pulse driven array and CW tone burst, uniformly-driven, and Gaussian-tapered arrays.

Typical results obtained from these synthetic square array, water tank experiments are shown in **Fig. 1.** The field generated by an LWT pulse driven array exhibits the predicted, localized pencil-beam characteristics. The experimental result and the simulations have compared extremely well; the analysis has predicted improvements that will be checked with further experimentation. Since our water tank is limited in size, a larger, more definitive experiment has been designed that would

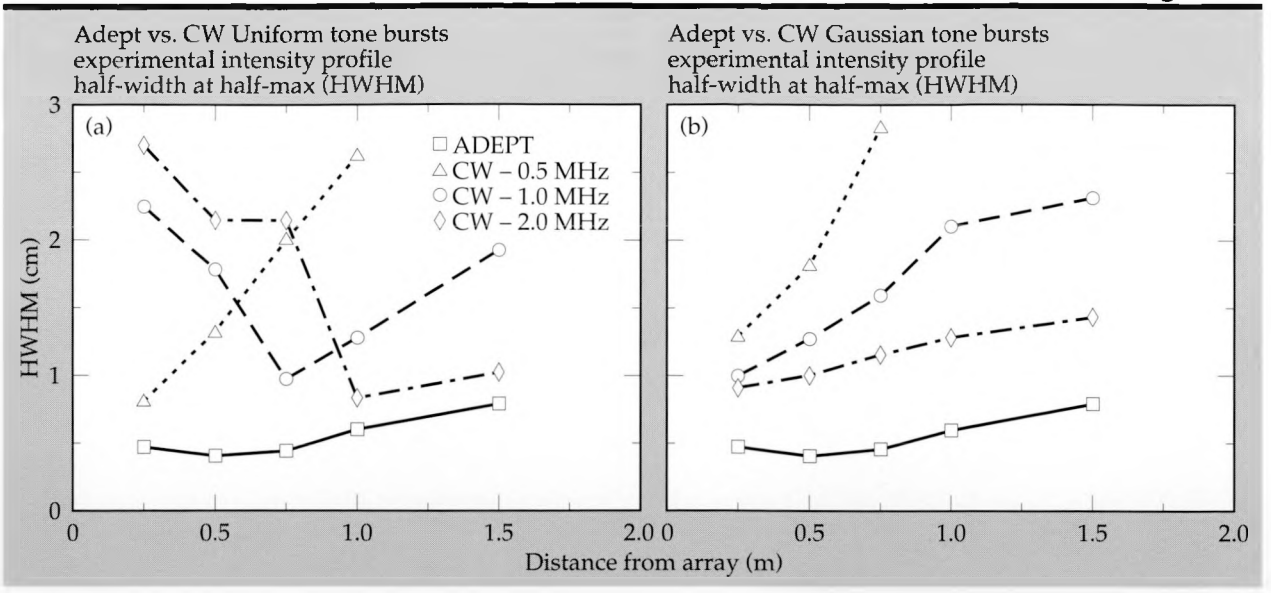


Figure 1. The waist (HWHM) of the intensity profile (radius at which the intensity has decreased to half its value) of the wave field is a standard measure of its localization. The figure shows the measured waists of the fields generated by a series of pulse driven, 3.0 cm radius, circular arrays. The MPS pulse driven array (ADEPT) is compared to (a) a uniform array driven with 0.5, 1.0, and 2.0 MHz tone bursts, and (b) a Gaussian tapered array driven with 0.5, 1.0, and 2.0 MHz tone bursts. The Gaussian waist was set equal to the MPS pulse waist, 1.5 cm. The spectra of the MPS drive functions have an amplitude e-folding point at 0.5 MHz. The Rayleigh distance for the Gaussian tapered array at 0.5 MHz was 25.0 cm. The focal distance for the uniform array at 2.0 MHz was 100.0 cm. The enhanced localization of the field generated by the MPS pulse driven array is apparent.

investigate the LWT effect over distances surpassing ten conventional Rayleigh distances. The array and the respective pulse drivers have been designed. The experiment will test how far away from the array the LWT effect can be maintained.

There are many potential applications for pulses designed to have LWT characteristics, including remote sensing, communications, power transmission, and directed energy weapons. This is particularly true for pulses that could be optimally designed for the particular environment in which they propagate. The LWT effect could significantly impact many areas of physics and engineering. Future experimental and theoretical efforts will concentrate on the propagation and scattering characteristics of the LWT pulse in more realistic environments, including the effects of reflection, refraction, diffraction, absorption, and nonlinearities. This will help determine the potential usefulness of the LWT effect in a variety of applications.

1. R. W. Ziolkowski, "Localized Transmission of Electromagnetic Energy," *Phys. Rev. A* **39**(4), 2005, (1989); see also Lawrence Livermore National Laboratory, Livermore, CA, UCRL-98507 (1989).
2. R. W. Ziolkowski, A. M. Shaarawi, and I. M. Besieris, "A Space-Time Representation of a Massive, Relativistic, Spin Zero Particle," *Nucl. Phys. B* **6**, 255–258, (1989); see also Lawrence Livermore National Laboratory, Livermore, CA, UCRL-98448 (1988).
3. I. M. Besieris, A. M. Shaarawi, and R. W. Ziolkowski, "A Bidirectional Traveling Plane Representation of Exact Solutions of the Scalar Wave Equation," *J. Math. Phys.* **30**(6), 1254–1269 (1989); see also Lawrence Livermore National Laboratory, Livermore, CA, UCRL-98870, June 1988.
4. A. M. Shaarawi, I. M. Besieris, and R. W. Ziolkowski, "Localized Energy Pulse Trains Launched from an Open, Semi-Infinite Circular Waveguide," *J. Appl. Phys.* **65**(2), 805–813 (1989); see also Lawrence Livermore National Laboratory, Livermore, CA, UCRL-98871, June 1988.
5. R. W. Ziolkowski, D. K. Lewis, and B. D. Cook, "Evidence of Localized Wave Transmission," *Phys. Rev. Lett.* **62**(2), 147–150 (1989); see also Lawrence Livermore National Laboratory, Livermore, CA, UCRL-99304, July 1988.
6. R. W. Ziolkowski, "Localized Transmission of Wave Energy," invited paper, SPIE Conference, Los Angeles, CA, January 1989; see *Microwave and Particle Beam Sources and Directed Energy Concepts, SPIE* **1061**, 395–402 (1989); see also Lawrence Livermore National Laboratory, UCRL-99304, January 1989.
7. A. M. Shaarawi, I. M. Besieris, and R. W. Ziolkowski, "A Novel Approach to the Synthesis of Nondispersive Wave Packet Solutions to the Klein-Gordon and the Dirac Equations," to appear in *J. Math. Phys.*, (1990); see also Lawrence Livermore National Laboratory, Livermore, CA, UCRL-100411 (1989).

8. A. M. Shaarawi, I. M. Besieris, and R. W. Ziolkowski (1989), "Diffraction of a Classical Wave Packet in a Two Slit Interference Experiment," submitted to *Phys. Rev. Lett.*; see also Lawrence Livermore National Laboratory, Livermore, CA, UCRL-100756, March 1989.
9. A. M. Shaarawi, I. M. Besieris, and R. W. Ziolkowski, "A Nondispersive Wavepacket Representation of Photons and the Wave-Particle Duality of Light," submitted to *Int. J. Theor. Phys.*; see also Lawrence Livermore National Laboratory, Livermore, CA, UCRL-101694 (1989).
10. R. W. Ziolkowski, D. K. Lewis, and B. D. Cook, "Propagation of Energy in Nonspreading Wave Packets," *Joint American-Japan Acoustic Society Meeting* (Honolulu, Hawaii, November 1988).
11. R. W. Ziolkowski, D. K. Lewis, and B. D. Cook, "An Experimental Search for the Localized Wave Packet," *Joint American-Japan Acoustic Society Meeting* (Honolulu, Hawaii, November 1988).
12. R. W. Ziolkowski, "Localized Transmission of Wave Energy," *SPIE Conference 1061: Microwave and Particle Beam Sources and Directed Energy Concepts* (Los Angeles, CA, January 1989).
13. R. W. Ziolkowski, "Generating Localized Packets of Wave Energy," invited paper, *National Radio Science Meeting* (URSI, San Jose, CA, June 1989).
14. I. M. Besieris, A. M. Shaarawi, and R. W. Ziolkowski, "Double Slit Diffraction of a Localized Wave Packet," *National Radio Science Meeting* (URSI, San Jose, CA, June 1989).
15. R. W. Ziolkowski, "Localized Transmission of Wave Energy," *Energy and Technology Review*, Lawrence Livermore National Laboratory, Livermore, CA, (November 1988), pp. 16–23.
16. R. W. Ziolkowski, "Evidence of Localized Wave Transmission," *Energy and Technology Review*, Lawrence Livermore National Laboratory, Livermore, CA, (November 1988), pp. 24–28.
17. I. Amato, "Making Waves that Travel Like Beams," *Sci. News* **135,** 38 (1989).
18. K. Davidson, "Science Boldly Goes into Sci Fi Arena," *San Francisco Examiner*, Sunday, January 19, 1989.
19. R. Ruthen, "Beating the Spread," *Sci. Amer.*, Citizen and Scientist, (April 1989), pp. 28–29.

Modeling Pulse Driven Antenna Systems

Marvin J. Barth and
Richard W. Ziolkowski
Engineering Research Division
Electronics Engineering

We have developed computer modeling codes to analyze the behavior of high-power, pulse driven microwave antenna systems. These codes enable us to model the far-field performance of complicated, pulse-driven antenna systems from their near-field behavior. Our approach employs discrete modeling of the near-field environment of the antenna system with standard finite difference and finite element time domain (FDTD and FETD) electromagnetic modeling codes, and then transforms the resultant near-field information to the far-field with a newly developed post-processor. This post-processor is termed the FAR code. The resultant code set will improve our ability to model a variety of high-power microwave (HPM) sources, including associated potential antenna designs and related HPM effects of current Laboratory and national interest. For example, we can now model the far-field distortions (antenna pattern modifications) and reduction of radiated power caused by air breakdown near HPM antennas and sources.

Introduction

Many microwave sources under development at the Laboratory and elsewhere produce pulsed fields, i.e., fields with time histories of finite duration that contain many frequencies. Many conventional electromagnetic modeling codes, such as those based upon integral equation approaches, generally deal only with single frequency or monochromatic (CW) fields. They can be used to model polychromatic pulsed fields by dealing with each frequency separately and then recombining the results through Fourier transform techniques. The field associated with each frequency is weighted by its amplitude in the spectrum of the source.

On the other hand, discrete methods such as the finite difference and finite element approaches can directly model the time evolution of fields. One models the source and antenna structures, and the surrounding environment by creating a mathematical representation called the mesh or grid. Specification of the appropriate boundary conditions at the locations in the grid defines the various structures and the interfaces between different materials. Finally, one runs the codes to produce time histories of the fields at all of the mesh points. The time evolution of the fields can then be viewed with a variety of methods. Moreover, if specific frequency domain information is desired, it can be readily obtained from the time histories of the fields with Fourier transform techniques. The distinct advantage of time domain modeling is the simultaneous acquisition of the behavior of all of the frequencies of interest at every point in the mesh.

Unfortunately, these methods have their limitations. CW methods such as the integral equation approach can model only limited physical environments. On the other hand, discrete methods can model very complicated environments, but the information they produce is limited to the size of the mesh that can be squeezed into any given computer system. They can be generally used to directly give near-field results. Nonetheless, if one could combine a method to extract far-field information from the near-field data with the ability of the discrete methods to model accurately the near-field environment, one would obtain an approach to modeling the fields generated from pulsed sources that would be superior to many conventional modeling techniques.

To accomplish our objective of using near-field information to model the fields far from any sources or antennas, we have developed a post-processor termed FAR which translates the near-field results of the finite difference time domain (FDTD) code TSAR[1] to the far field. TSAR is an LLNL-developed, general purpose FDTD code. It calculates the $\vec{E}$ and $\vec{H}$ field values over a specified closed surface within the mesh which includes the near-field geometries and sources. These values are stored for processing by FAR. We then use the equivalence principle as shown in **Fig. 1**. Accord-

ing to this principle, any electromagnetics problem may be transformed into one in which currents are specified on a closed surface that generate the original source field exterior to this surface and a null field (exactly canceling the original source field) in its interior.[2] The values of these surface currents are determined by the original source fields simply by the values of the cross product of the outward field component normal to this surface and the electric and magnetic fields. Because the TSAR computed values exist only on the FDTD grid, we do not have a complete description of the electric and magnetic fields at all points on the surface. We assume, however, that these point values are adequate samples of the continuous fields and surface currents, and treat the resulting equivalent surface currents as point electric and magnetic dipoles radiating into free space. FAR then calculates the resulting fields of each of these dipoles and sums all contributions to determine the total field.

Progress

There are now three varieties of the FAR code, which produce different descriptions of the same physical situation. The code FFAR computes the resulting fields as frequency domain far-field patterns. This allows us a direct comparison of our results with those generated with standard electromagnetics modeling codes. The code TFAR computes the actual time domain far-field time histories. This allows us to study directly the time evolution of the fields produced by a pulse driven array. The code NFAR computes the time domain field histories completely and includes the $1/r$, $1/r^2$, and $1/r^3$ terms of the point dipoles which model their near-field as well as their far-field behavior. Both FAR and FFAR only account for the traditional $1/r$ far-field term.

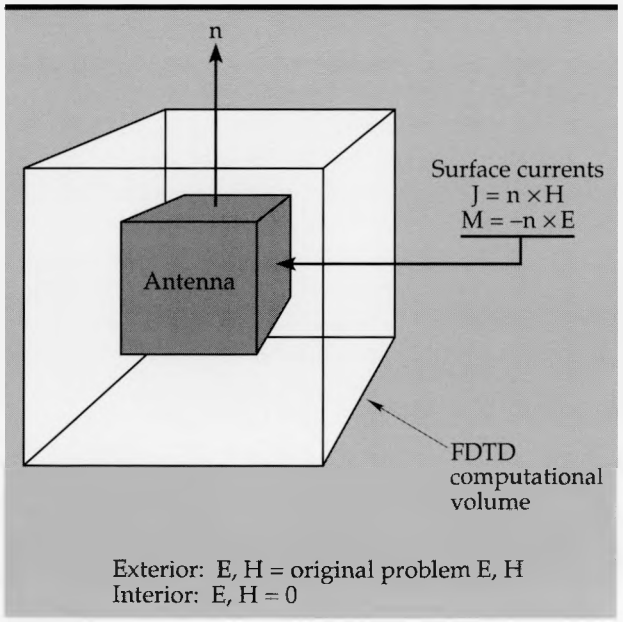


Figure 1. The near-to-far-field transform technique, based upon the electromagnetic equivalence principle. This technique requires a closed surface within the FDTD grid volume that encloses the antenna and the near-field region.

The FFAR code is both computationally faster and more accurate than the other codes. It simply uses Fourier transform properties which trade frequency multiplication for the time differentiation. The TFAR code is slower and less accurate because a numerical differentiation routine is used. The NFAR code is even slower because it also has a numerical integration routine in addition to the numerical differentiation routine and calculates more field components (the $1/r$ far-field terms model only the components of the field transverse to the direction of propagation, while the $1/r^2$ and $1/r^3$ terms also include its longitudinal components).

We have used primarily two test cases to validate these codes. In these test cases, the near field (FDTD) environment is composed of:

- Two dipoles radiating into free space with no blockages,
- A 3×5 dipole array with a left-T blockage.

The left-T blockage case uses a metal blockage in the shape of a "T" to simulate a plasma breakdown region that may be formed in a HPM antenna system and degrade its usefulness. In both cases, we compute the far-field frequency domain horizontal and vertical patterns. We compare those patterns with corresponding patterns obtained with the LLNL-developed Numerical Electromagnetics Code (NEC).[3] Since the NEC code is extremely accurate for this type of problem, we consider our error as the maximum error computed from the point-to-point difference between the patterns predicted by both codes.

These validation problems have enabled us to identify the three sources of inherent errors in the near-to-far-field transform process. The major error occurs because we are only sampling the surface currents. To isolate this error, we constructed a code called JOB. JOB uses analytical formulas to compute the surface field values, the values which TSAR would normally calculate. It then uses the FFAR code algorithms to calculate the far-field components. This eliminates from the far-field calculations the diffusion and dispersion

errors inherent in the FDTD codes. The error rate is determined by comparing the patterns obtained with different FDTD grid spacings with the NEC reference case. The results of this comparison are shown in **Fig. 2**, which indicates the linear regression of the error rate for the JOB code as a function of the inverse of the point density. As more grid points are included on the surface, hence, increasing the information or detail of the surface current structure, the pattern error decreases exponentially. The regression coefficient is 0.9989. Satisfactory results are obtained with grid spacing of at least $\lambda/8$, where λ is the wavelength of the principal frequency of interest in the pulses generated by the source. This result coincides with the generally recognized grid spacing criterion that finite difference code practitioners satisfy to avoid large errors in their simulation results.

The second error source we uncovered is the usual set of errors associated with numerically forming the discrete Fourier transform (DFT): aliasing, leakage, and picket-fence errors.[4] These errors always occur with a numerical DFT and are unavoidable. We mitigate the effects of these errors on the far-field calculations by properly choosing the excitation pulses and by artificially (within the DFT computer routine) extending the Fourier transform period (zero padding). We have found that the best results are obtained if one employs a transit pulse (a pulse with no DC component) whose duration is short and whose derivative is also a transit pulse.

The final error source is a collective effect composed of FDTD diffusion and dispersion errors, FDTD boundary errors, and interpolation errors. The diffusion and dispersion errors are mitigated by choosing a smaller grid spacing. The boundary errors are mitigated by placing the FDTD problem boundary at least $\lambda/8$ beyond any material or sources within the FDTD problem space. The interpolation errors arise within routines in the FAR code. Because the actual time histories are sampled only at a discrete number of points but are needed in a continuous sense, interpolation is required to obtain those additional time history values. The interpolation errors associated with FAR decrease linearly as more points are used to interpolate the field values between two points, up to a threshold where they actually begin to increase binomially.[5] We have found that a two- or five-point interpolation algorithm produces the best results.

To illustrate the efficacy of the TSAR-FAR code set, we show results for each of the two validation cases. **Figure 3** shows the two-dipole horizontal pattern (polar angle: $\theta = 90$; azimuthal angle: $-90 \leq \phi \leq 90$) of the E_θ field at a frequency of 1.0 GHz for TSAR-FFAR runs with grid spacings of $\lambda/10$ and $\lambda/20$ as well as the standard NEC pattern. The maximum error at any point over the range is 0.6528% for the $\lambda/20$ plot and 2.7277% for the $\lambda/10$ plot. **Figure 4** shows results for the 3×5 array with a left-T blockage horizontal pattern for the E_θ field and a corresponding NEC pattern again at a frequency of 1.0 GHz. The maximum error at any point over the range is 2.2550%.

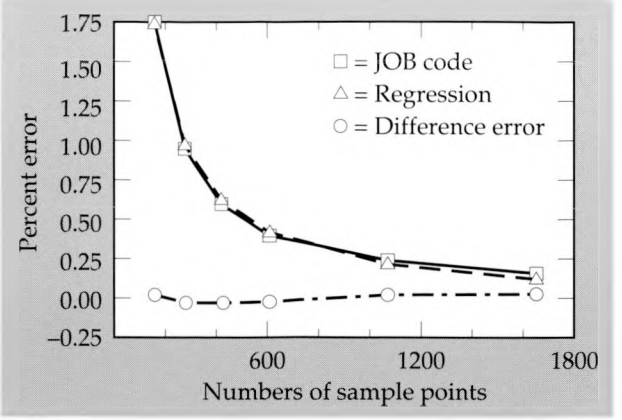


Figure 2. Linear regression of JOB maximum point error on the inverse of sampling point density, showing that the errors in the near-to-far-field transform are significantly decreased with an increased sampling density. For the case presented, frequency = 1.0 GHz; grid sizes = $\lambda/6$, $\lambda/8$, $\lambda/10$, $\lambda/12$, $\lambda/16$, and $\lambda/20$.

The corresponding errors over the frequency range from 0.5 to 1.5 GHz for both the horizontal and vertical patterns are shown in **Table 1**. In this table, the *Average error* column shows the errors averaged over 181 points in each pattern, and the *Maximum error* column shows the maximum error for either the horizontal or vertical pattern.

Future Work

During the course of our work, we have found that the number of data files created during a FDTD code run needed for processing by FAR can easily fill a large computer memory disk. To avoid this problem, we are presently incorporating the FAR near-to-far-field transform directly into TSAR. This effort involves rewriting the FAR algorithms so that instead of working on all the time histories at once (post-processing), the calculations are performed "on the fly" (i.e., as each time step occurs). At this time, all three codes (FFAR, TFAR, and NFAR) have been incorporated into TSAR and have been partially tested, and the results are very

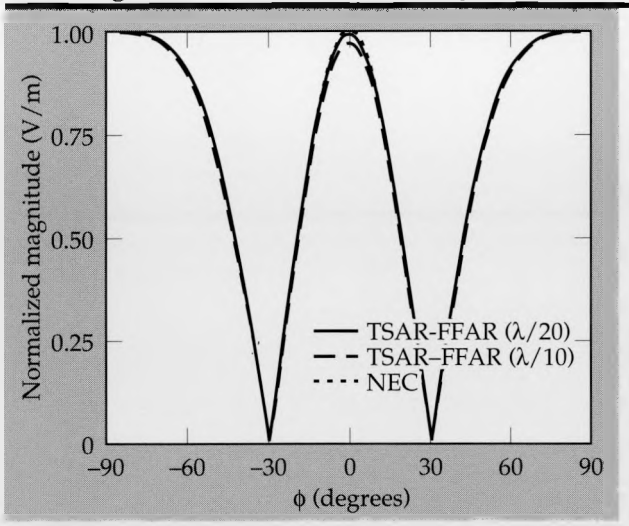


Figure 3. Comparison of the E_θ far-field patterns predicted by three codes for the problem of two z-directed dipoles radiating into free space at 1.0 GHz.

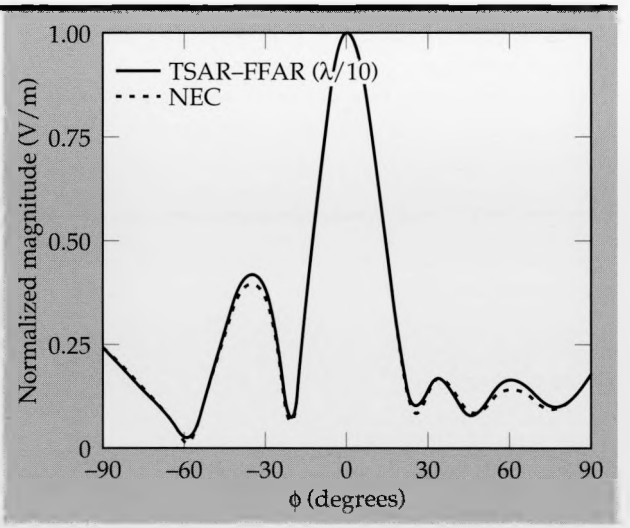


Figure 4. The E_θ far-field patterns predicted by two codes compared with a 3 × 5 array of z-directed electric dipoles at 1.0 GHz with a left-T blockage.

Table 1. TSAR-FFAR (E_θ) percent errors for the two-dipole (λ/20 and λ/10) and the 3 × 5 array left-T (λ/10) validation problems.

	Two dipole, no blockage				3 × 5 array, left-T	
	λ/20		λ/10		λ/10	
Freq (GHz)	Average error (%)	Maximum error (%)	Average error (%)	Maximum error (%)	Average error (%)	Maximum error (%)
0.5	0.0491	0.0977	0.3070	0.4541	0.9364	2.9694
0.6	0.1690	0.3057	0.1971	0.3927	0.5552	1.6840
0.7	0.2855	0.6760	0.3030	0.5721	0.2000	1.8647
0.8	0.3352	0.9093	0.2993	1.6302	0.2359	1.6726
0.9	0.3154	0.6610	0.7834	2.9523	0.4052	2.3505
1.0	0.1883	0.6562	1.0448	2.7277	0.5015	2.9078
1.1	0.4012	1.0569	0.4569	2.1848	0.1000	2.8616
1.2	0.9241	2.4561	0.3307	2.8069	0.8057	4.5043
1.3	0.8771	2.6649	2.4923	5.0393	1.1572	6.6254
1.4	0.1759	1.8902	5.3601	8.9404	0.8456	3.4841
1.5	1.2704	3.8062	4.1853	14.9557	1.7136	5.8389

promising. The final testing phase is in progress.

Because some of the problems of interest to Laboratory programs are too large to include within currently available computer memories (problems which require an enormous three-dimensional TSAR grid), we feel that it is important to develop a two-dimensional FAR code that can work with all our two-dimensional discrete codes. The equations for this work are complete, but the coding and testing remains to be carried out. Since all the error sources are known from our three-dimensional work, the remaining two-dimensional coding and testing should require minimal effort.

1. R. R. McLeod, *Temporal Scattering and Reflection Software User's Manual*, Vers. 1.1, Lawrence Livermore National Laboratory, Livermore, CA, UCID–21285 (1987).
2. R. F. Harrington, *Time-Harmonic Electromagnetic Fields*, (McGraw-Hill, New York, NY, 1961) p. 106.
3. G. J. Burke and A. J. Poggio, *Numerical Electromagnetics Code (NEC)—Method of Moments*, Lawrence Livermore National Laboratory, Livermore, CA, UCID-18834 (1981).
4. A. Ralston, *A First Course In Numerical Analysis*, (McGraw-Hill, New York, NY, 1965) p. 51.
5. G. D. Bergland, "A Guided Tour Of The Fast Fourier Transform," *IEEE Spectrum*, July, 1969, p. 41.

A 3D Modified Finite Volume Technique for Maxwell's Equations

Niel K. Madsen and
Richard W. Ziolkowski
Engineering Research Division
Electronics Engineering

A solution of Maxwell's curl equations with high-speed computors has greatly facilitated the analysis of electromagnetic scattering, coupling, and propagation problems. We present here a modified finite volume method for solving Maxwell's equations in the time domain. This method, which allows the use of general nonorthogonal mixed-polyhedral grids, is a direct generalization of the canonical staggered-grid finite difference method. Employing mixed polyhedral cells (hexahedral, tetrahedral, etc.), this method allows more accurate modeling of nonrectangular structures. The traditional "stair-stepped" boundary approximations associated with the orthogonal grid based finite difference methods are avoided. Numerical results demonstrating the accuracy of this new method are presented.

Introduction

One of the most important tools in the analysis of electromagnetic scattering, coupling, and propagation problems is the computerized direct solution of Maxwell's equations. Designers of modern aircraft, weapons systems, radars, and communication systems have a need to analyze how their systems respond to electromagnetic pulses and radiation. Cost-effective computer analyses of these problems are highly useful. Such analyses require a software package that can represent the objects and environment being studied in the computer, discretize Maxwell's equations by converting them into matrix equations, and then determine the required solution parameters. Such computer programs for three-dimensional problems require a great number of individual calculations, and an important goal of the software designer is to minimize the running time required to solve the problems of interest while maintaining the desired accuracy.

The conventional approach for numerically solving Maxwell's equations in the time domain has been the use of finite difference methods in conjunction with orthogonal grids.[1-5] It is well known that the use of such methods can produce very accurate results, particularly when the domain is rectangular or almost rectangular. For problems involving general nonrectangular domains, two approaches have been used: a) "stair-stepped" orthogonal approximations to irregular boundaries, and b) transformations which map the original irregular domain into a rectangle in the transformed coordinates. The first approach may give rather poor approximations to the boundary shapes unless very fine discretizations are used, and the second results in more complicated equations to solve.

We will present a method which is derived using a modified finite volume technique. This new method actually reduces to be the conventional finite difference method when applied on an orthogonal grid. We have primarily used this technique with irregular hexahedral grids. However, this technique is easily applied using grids which are composed of tetrahedrons, or mixtures of hexahedrons and tetrahedrons, or other cell types. This work generalizes to three dimensions our previous two-dimensional algorithm.[6]

We begin by assuming that we wish to solve Maxwell's curl equations on an irregular three-dimensional domain R which has a boundary surface denoted by S. We will also assume that the domain R has been discretized into hexahedrons. **Figure 1** shows a discretized twisted waveguide, which is typical of the problem complexity we wish to solve.

Modified Finite Volume Method

Finite volume techniques are usually derived by integrating the basic equations to be solved over a grid cell and forming discrete approximations for the resulting integral quantities. We will describe our modified finite volume technique using a basic hexahedral grid and its associated "dual" grid. For

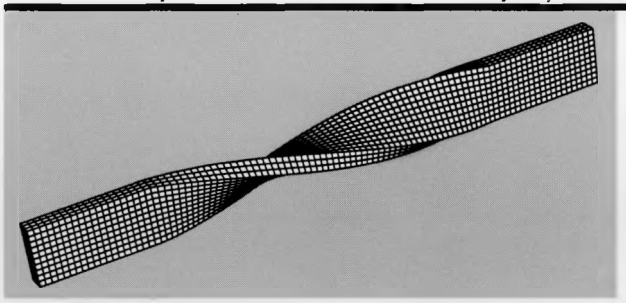


Figure 1. Twisted waveguide discretized into hexahedral cells.

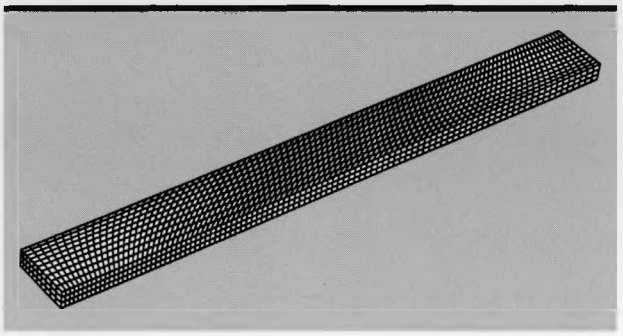


Figure 2. Skewed 10 × 3 × 100 nonorthogonal waveguide grid.

each hexahedral cell, we define its barycenter to be the point which is located at the average of the coordinates of the nodes which define the cell. We construct the dual grid by connecting barycenters of adjacent cells with straight lines passing through each of the interior cell surfaces of the original grid. Our discrete solution components will be associated with the edges of the original grid and also with the edges of the dual grid. The quantity associated with a cell edge is the projection of the electric field vector onto that edge, i.e., $\mathbf{E} \cdot \mathbf{s}$, where $\mathbf{s}$ is a unit vector in the direction of the edge. The magnetic field projection $\mathbf{H} \cdot \mathbf{s}^*$ is associated with a dual cell edge where $\mathbf{s}^*$ is a unit vector in the direction of the dual edge. We will denote with an asterisk, *, geometric quantities associated with the dual grid.

We begin by decomposing the electric field vector $\mathbf{E}$ into three orthogonal components which are determined from the dual grid. We have

$$\mathbf{E} = (\mathbf{E} \cdot \mathbf{s}_1^*)\mathbf{s}_1^* + (\mathbf{E} \cdot \mathbf{s}_2^*)\mathbf{s}_2^* + (\mathbf{E} \cdot \mathbf{s}_3^*)\mathbf{s}_3^* \qquad (1)$$

where $\mathbf{s}_1^*$ is a unit vector which is the average normal to the dual cell face which cuts the edge determining a particular $\mathbf{s}$. The vectors $\mathbf{s}_2^*$ and $\mathbf{s}_3^*$ are arbitrary mutually orthogonal unit vectors orthogonal to $\mathbf{s}_1^*$ so that $(\mathbf{s}_1^*, \mathbf{s}_2^*, \mathbf{s}_3^*)$ form a right-handed system. For the particular cell edge defining $\mathbf{s}$, the quantity to be computed is $\mathbf{E} \cdot \mathbf{s}$. Using (1) we have

$$\frac{d(\mathbf{E} \cdot \mathbf{s})}{dt} = \left(\frac{d\mathbf{E}}{dt} \cdot \mathbf{s}_1^*\right)(\mathbf{s}_1^* \cdot \mathbf{s}) + \left(\frac{d\mathbf{E}}{dt} \cdot \mathbf{s}_2^*\right)(\mathbf{s}_2^* \cdot \mathbf{s}) + \left(\frac{d\mathbf{E}}{dt} \cdot \mathbf{s}_3^*\right)(\mathbf{s}_3^* \cdot \mathbf{s}) \; . \qquad (2)$$

To form an approximation for the first term on the right side of (2) we integrate Maxwell's electric field equation over the face of the dual grid which is penetrated by the edge defining the vector $\mathbf{s}$. Assuming that $\mathbf{E}$ is constant over this face we have

$$\varepsilon A^* \frac{d\mathbf{E}}{dt} \cdot \mathbf{s}_1^* = \int (\nabla \times \mathbf{H}) \cdot \mathbf{s}_1^* dA^* = \oint \mathbf{H} \cdot d\mathbf{l}^*, \qquad (3)$$

where ε represents an average permittivity value, A^* is the area of this dual grid face, and the line integral is performed around the perimeter of this dual face in the positive direction relative to the normal direction $\mathbf{s}_1^*$. It is the computation of this particular term that "modifies" our approach from a conventional finite volume method.

To compute the other terms on the right side of (2), we simply integrate or average Maxwell's electric field equation over the two dual grid cells which contain the edge defining $\mathbf{s}$ (in the conventional finite volume manner). That is,

$$\varepsilon V^* \frac{d\mathbf{E}}{dt} \cdot \mathbf{s}_2^* = \int (\nabla \times \mathbf{H}) \cdot \mathbf{s}_2^* dV^* = \int (\mathbf{n}^* \times \mathbf{H}) \cdot \mathbf{s}_2^* dA^*, \qquad (4)$$

$$\varepsilon V^* \frac{d\mathbf{E}}{dt} \cdot \mathbf{s}_3^* = \int (\nabla \times \mathbf{H}) \cdot \mathbf{s}_3^* dV^* = \int (\mathbf{n}^* \times \mathbf{H}) \cdot \mathbf{s}_3^* dA^*, \qquad (5)$$

where ε is the average value, $\mathbf{n}^*$ a dual surface normal, and V^* is the volume of the two dual grid cells.

Since $\mathbf{n}^* \times \mathbf{H}$ is tangent to the integration surface, we can form approximations to the surface integrals in **Eqs. (4)** and **(5)** using the projections of $\mathbf{H}$ onto the dual edges which define the surfaces. That is,

$$\mathbf{n}^* \times \mathbf{H} = -\mathbf{H} \times (\mathbf{n}^*) = -\mathbf{H} \times (\mathbf{a}^* \times \mathbf{b}^*) = -(\mathbf{H} \cdot \mathbf{b}^*)\mathbf{a}^* + (\mathbf{H} \cdot \mathbf{a}^*)\mathbf{b}^*,$$

where $\mathbf{a}^*$ and $\mathbf{b}^*$ are the directions of two dual edges which meet at a dual face corner.

To compute the magnetic field values, which are associated with the edges of the dual grid cells, we simply proceed in a completely dual manner. **Equations (2)–(5)**, together with the analogous dual equations for the magnetic field quantities, constitute the Modified Finite Volume (MFV) approximation method.

The total number of unknown approximate field quantities is clearly equal to the number of cell edges in the original grid plus the number of cell edges in the dual grid. For a nonorthogonal logically regular hexahedral grid, the computation of a typical interior $\mathbf{E} \cdot \mathbf{s}$ will involve the use of 20 surrounding magnetic field values. When the MFV algorithm is applied using an orthogonal grid, the number of dual cell edge values used to compute a original grid field value reduces to 4, a considerable savings.

Exterior (perfect electric conductor) boundary conditions are very easily handled by simply setting $E_{tan} = \mathbf{E} \cdot \mathbf{s} = 0$ for those edges of the grid which form the boundary. As implied above, interior material interfaces are also easily handled by using appropriately averaged values for the discontinuous permittivity, ε, and the permeability, μ, as is customary with the conventional orthogonal grid finite difference method.

It is also important to note that the modified finite volume method as defined above numerically preserves (to within computer round-off) the divergence of the fields (e.g., charge will be conserved). This is easily seen to be true by recognizing that the left side of **Eq. (3)** is essentially the time derivative of the integral of the field divergence over a cell face. Since the right side of **Eq. (3)** is the line integral around a cell face, the integral of the time derivative of the field divergence over a closed grid volume will be zero as each cell edge will be traverse exactly twice—once in each direction. Thus the integral of the field divergence over any closed grid volume remains constant in time. To perform the time integration, we use the explicit leap-frog time integration algorithm. We have found that the efficiency and second order accuracy of this method are completely adequate.

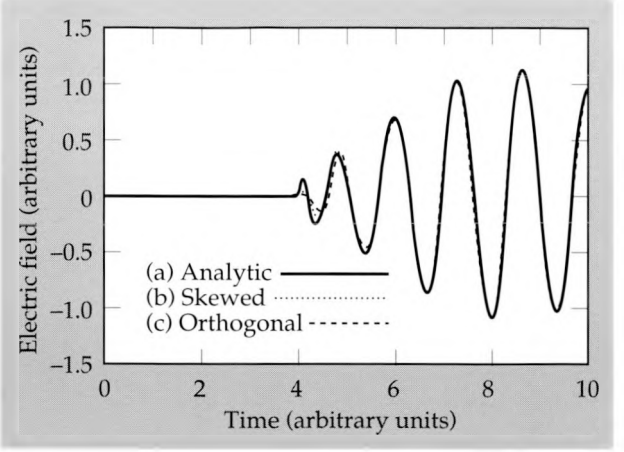


Figure 3. Numerical TE_{10} waveguide solutions compared with the analytic solution for a point 4. 0 m down the guide: a) analytic, b) skewed grid, and c) orthogonal grid.

Numerical Results

We first consider a simple problem which is designed to provide some assessment of the wave propagation characteristics of the MFV algorithm. We consider the propagation of a TE_{10} signal in a rectangular waveguide which is 1. 0 m in width and height, and 10. 0 m in length. The signal is initiated in the waveguide by specifying the tangential field, E_y, at the left boundary ($z = 0.0$). The function used to drive this signal is

$$f(t, x, y, 0) = sin(1.5\,\pi t)\, sin(\pi x) \ .$$

We assume that all of the waveguide walls are perfect electric conductors, i.e., $E_{tan} = 0$. Normalized values for the material parameters ε and μ are used: $\varepsilon = 1.0$, and $\mu = 1.0$. The signal is allowed to propagate until $t = 10$. The analytic solution for

Grid type	Max error	Run time (%)	Steps	Edges (sec)	Dual edges
Hexahedral orthogonal	0.173	145	160	11773	7670
Hexahedral skewed	0.093	183	204	11773	7670
Tetrahedral orthogonal	0.150	841	370	22103	27340
Tetrahedral skewed	0.191	1048	463	22103	27340
Triangular prism orthogonal	0.222	439	339	15773	12670
Triangular prism skewed	0.233	440	339	15773	12670

Table 1. Comparison of TE_{10} waveguide solution errors, run times, and problems sizes for various discrete MFV grids.

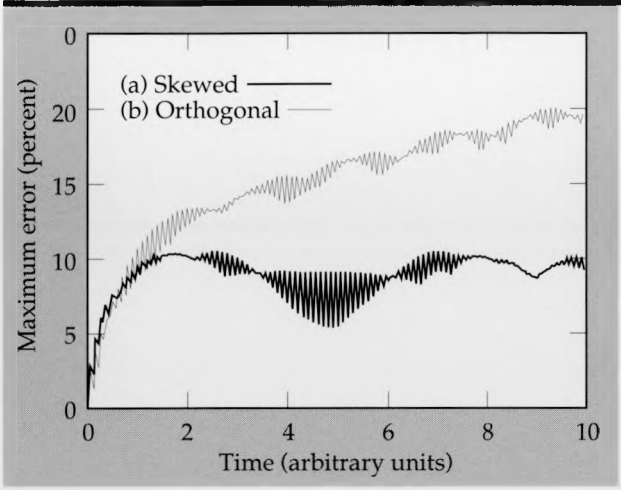


Figure 4. Comparison of maximum errors as a function of time for the skewed (a)and orthogonal (b) waveguide grids.

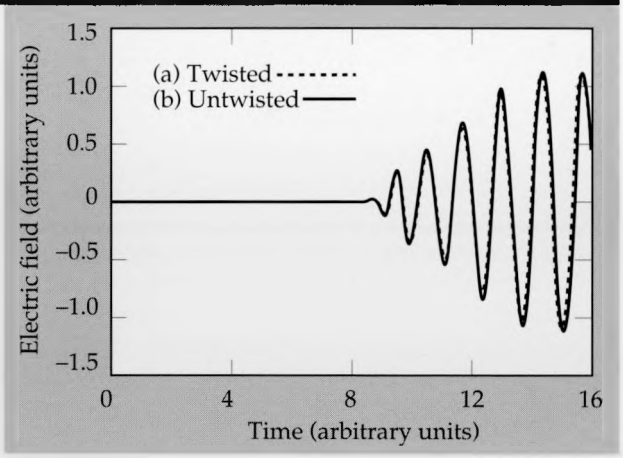


Figure 5. Comparison of the time history for the twisted waveguide MFV numerical solution (a)with a nontwisted waveguide MFV numerical solution.(b)

this problem has been derived and computed so we may examine the errors precisely.

We have used a variety of grids to numerically solve this problem. Clearly the simplest grid for this problem is an orthogonal hexahedral grid. However, to demonstrate the ability to use nonorthogonal grids, we deliberately skew the $10 \times 3 \times 100$ grid so that it is composed of nonorthogonal hexahedrons as shown in **Fig. 2.** The skewing for this grid is such that halfway down the length of the guide the cells have angles of about 45° and 135°. **Figure 3** compares the time histories of the numerical solutions for the skewed and orthogonal grids, respectively, with the analytic solution for a point 4.0 m down the guide. We notice that the significant error occurs as the signal first reaches the observation point. Both solutions exhibit typical dispersion errors; however, the skewed grid results are somewhat better. To more precisely compare the two solutions, **Fig. 4** shows the maximum errors over the entire waveguide for both solutions as a function of time. Surprisingly, for all but the earliest times, the skewed grid produces a uniformly better solution. To demonstrate that other grid cell types can produce acceptable solutions, **Table 1** compares the maximum errors, run times, and problem sizes for several different grids. The grids all have the same number of nodes, but differ in the number of elements and edges. The nonhexahedral grids were all formed from the orthogonal and skewed hexahedral grids by subdividing each hexahedral cell into 5 tetrahedrons or 2 triangular prisms. The solution times are in units of seconds using a Floating Point System 264 computer. It is obvious that this problem is most efficiently solved using hexahedral cells.

As a final example we consider the problem of computing a TE_{10} waveguide solution for the rectangular waveguide (shown in **Fig. 1**) which goes through a 180° twist. This problem cannot be effectively solved using a stair-stepped orthogonal grid finite difference algorithm unless a tremendously large number of cells are used. The analytic solution for this problem is not known. **Figure 1** shows the $10 \times 3 \times 100$ cell MFV grid used and **Fig. 5** compares the time history for a point after the twist with a TE_{10} solution for a similarly discretized waveguide with no twist. We notice that the waves transition the twist with little distortion. Conversion of the TE_{10} to other modes does not occur because of the thinness of the waveguide.

1. K. S. Yee, "Numerical Solution of initial boundary value problems in isotropic media", IEEE Trans. Antennas Propagag. AP-14(3), 302–307 (May 1966).
2. A. Taflove and M. E. Brodwin, "Numerical solution of steady-state electromagnetic scattering problems using the time-dependent Maxwell's equations," IEEE Trans. Microwave Theory Tech. MTT-23, 623–630 (Aug. 1975).
3. R. Holland, "THREDE: A free-field EMP coupling and scattering code," IEEE Trans. Nucl. Sci. NS-24, 2416–2421 (Dec. 1977).
4. K. S. Kunz and K. M. Lee, "A three-dimensional finite-difference solution of the external response of an aircraft to a complex transient EM environment: Part I—The method and its implementation," IEEE Trans. Electromag. Compat. EMC-20, 328–333 (May 1978).
5. A. Taflove and K. R. Umashankar, "The finite-difference time-domain (FD-TD) method for electromagnetic scattering and interaction problems," Elec. Waves Appl. 1(3), 243–267 (1987).
6. N. K. Madsen and R. W. Ziolkowski, "Numerical Solution of Maxwell's Equations in the Time Domain Using Irregular Nonorthogonal Grids," Wave Motion 10, 583–596 (1988).

Calculations of Optical and Transport Properties of Solid-State Superlattice and Quantum-Well Structure Materials

Jick H. Yee, Gizzing H. Khanaka, W. Travis White, and William J. Orvis
Engineering Research Division
Electronics Engineering

Solid-state superlattice and quantum-well materials are among the most important materials that have been recently developed for electronic device applications. These materials are made with molecular beam epitaxy (MBE) machines and consist of atomically thin layers of one material embedded within another. By adjusting the thickness and the atomic composition of the layers, one can change their optical and electronic properties. The Engineering Department at LLNL is now developing the necessary capabilities to make superlattice and quantum-well layer devices and materials. In order to provide the necessary theoretical support to complement LLNL's emerging fabrication capabilities, we must also be able to effectively model quantum-engineered devices and materials. Therefore, the goal of the present project is to develop quantum energy band and transport codes to help select superlattice and quantum-well materials with properties applicable to devices fabricated for Laboratory programs. Recently, we have completed a nonrelativistic energy band pseudopotential code, and we are in the process of completing a Monte Carlo transport code for superlattice materials. Using the energy band code, we have investigated the optical properties of some superlattice materials, and the results obtained from our code agree very well with experiments. Currently, we are investigating the transport properties of superlattice materials. In addition, we are modeling electron devices made with quantum-well and superlattice materials that are important to Laboratory programs.

Introduction

The advent of molecular beam epitaxy (MBE) has given rise to a new class of devices called superlattices or quantum-well structures.[1–3] Such devices are made by depositing two different layers of material, one on top of the other, with an MBE machine. Changing the thickness and atomic composition of the superlattice layers can tailor the optical and transport properties of the finished device to exhibit a desired electronic characteristic. Such devices could be used, for example, as quantum-well lasers, where photon emission takes place with the transition of an electron from a bound state in the conduction band to a bound state in the valence-band well. By adjusting the thickness and composition of the layers making up the quantum wells, one can select the wavelength of the emitted photons and maximize the electrical efficiency of the laser.

To date, most theoretical studies of the optical and electronic properties of superlattice and quantum-well materials have been based on simple potential energy models for the electrons.[4] Only recently have attempts been made to use more realistic models.[5] For these attempts, calculations are usually performed on only one kind of superlattice material.[5–8] In addition, most are based on perturbation techniques.[6–8] In that approach, the energy bands of the individual materials are calculated first, and then the energy bands of the superlattice material are calculated using the difference in the potential energy of the individual materials as the perturbation Hamiltonian. The weakness of this approach is the nonuniqueness of the reference energy of the superlattice material. Its value has to be adjusted with respect to experimental data. We do not use perturbation techniques in our calculations; hence, we do not need to adjust any reference energy values.

Band Theory of Superlattice and Quantum-Well Structure Materials

To construct a realistic value for the potential energy of the electrons in superlattice or quantum-

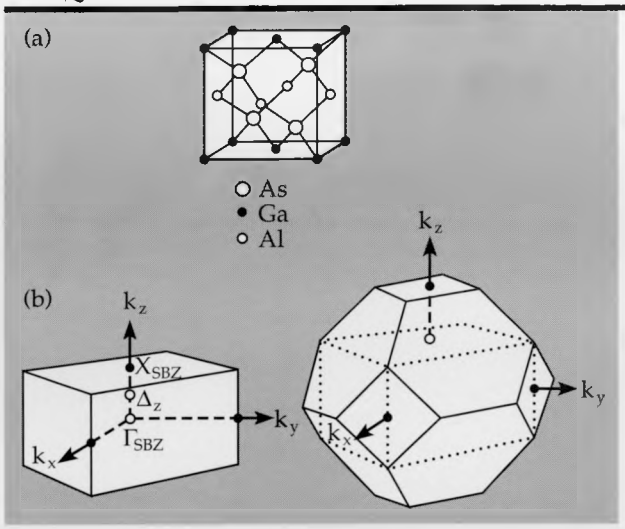


Figure 1. A typical simple quantum-well unit cell (a) used in our calculations, and (b) a typical Brillouin zone used in our calculations.

well structures, we use pseudopotential energy-calculation techniques. Considering the superlattice shown in **Fig. 1a**, which consists of two different materials, GaAs and AlAs, we can write its one-electron wave equation as

$$\frac{\hbar^2}{2m}\nabla^2\psi + U(\mathbf{r})\psi = E\psi \quad , \qquad (1)$$

where $\hbar$ is Planck's constant, m is the electron mass, ψ is the electron wave function, and E is the electron energy. The potential energy,U**r**, may be written as a superposition of effective atomic potentials Uj;

$$U(\mathbf{r}) = \sum_n \sum_j U_j(\mathbf{r} - \mathbf{R}_n - \mathbf{p}_j) \, , \qquad (2)$$

where $\mathbf{R}_n$ is the position of the center of the σth superlattice unit cell and $\mathbf{P}_j$ is the position of the given atomic core relative to the center of a cell. The function U(**r**) is translationally invariant from one superlattice unit cell to the next. A natural set of basis functions for expanding it is a set of plane waves of the form exp(i**K** · **r**).

By taking the inner product between the Schrödinger wave equation and every possible plane-wave basis function, exp(i**K** · **r**), we obtain an infinite-order matrix eigenvalue problem, the secular equation of which is Bloch's determinant form of the Schrödinger wave equation:

$$0 = \det\left| \left(\frac{\hbar^2(\mathbf{k}-\mathbf{K})^2}{2m} - E \right)\delta_{KK'} + \sum_j V_j(\mathbf{K},\mathbf{K}')S_j(\mathbf{K}-\mathbf{K}') \right| \, , \qquad (3)$$

where the quantity between the vertical lines is an infinite Hermitian matrix whose rows and columns are indexed by the reciprocal lattice vectors **K** and **K′**. The wave vector **k** is restricted to the first Brillouin zone [**Fig. 1(b)**]. The symbol $\delta_{kk'}$ is a Kronecker delta function, and E is the energy eigenvalue, which depends on **k**. The quantities $V_j(\mathbf{K},\mathbf{K}')$ and $S_j(\mathbf{K}-\mathbf{K}')$ are the atomic form factor and structure factor of the j^{th} type atom in a unit cell. They are defined as

$$V_j(\mathbf{K},\mathbf{K}') = \frac{n_j}{\Omega}\iiint d^3r \cdot \exp(-i\mathbf{K}\cdot\mathbf{r})U_j(\mathbf{r})\exp(i\mathbf{K}'\cdot\mathbf{r}) \, , \qquad (4)$$

and

$$S_j(\mathbf{K}'-\mathbf{K}) = \frac{1}{n_j}\sum_{p_i}\exp[i(\mathbf{K}'-\mathbf{K})\cdot\mathbf{p}_i] \, , \qquad (5)$$

where n_j denotes the number of atoms of type j per superlattice unit cell, Ω. Here, j indexes each different kind of atom and i indexes all atoms within type j in a superlattice unit cell.

Progress of the Project

During FY 89, we have completed and tested a pseudopotential code for superlattice and quantum-well materials based on **Eq. (3)**. Using this code and the data in **Table 1**, we have investigated the energy bands of some superlattice and quantum-well materials made from the III–V semiconductor compounds AlAs, GaAs, InAs, and GaSb.

Figures 2 to **5** show some of the results obtained from our calculations. **Figure 2** shows our code calculations of the band-gap of a superlattice made with GaAs and AlAs as a function of layer thickness and compared with experimental data. As one can see from this figure, the calculations agree very well with experimental data.

Figures 3(a) and **3(b)** show how our code results

Table 1. The symmetric and antisymmetric pseudopotential form factors used in the calculations. The values are given in Rydberg and the normalization volume is the atomic volume.

From Gell[6]				
	AlAs		GaAs	
q^2	Sym	Asym	Sym	Asym
3	–0.2307	0.0725	–0.2396	0.0700
4	———	0.0625	———	0.0500
8	0.0254	———	0.0126	———
11	0.0700	–0.0075	0.0600	0.0100
12	0.0000	0.0000	0.0000	0.0000

for the refractive index compare with experimental data for $Al_{0.5}Ga_{0.5}As$ alloy. Both the real and the imaginary parts show good agreement. It should be pointed out that our calculations were carried out at a temperature (0 K) which was different than the temperature (300 K) at which the experimental data was measured.[9] As a result, one may argue that the comparison between our calculations and the measured data is not conclusive. This, however, is not true. If we take the effect of temperature shifts into account in our calculations, the change of the energy bands is at most a fraction of an electron volt. We still expect our calculations to be in good agreement with the experimental data even if finite temperature effects were incorporated into our model.

Figures 4(a) and **4(b)** show how the imaginary and the real parts of the complex dielectric function vary as a function of layer thickness of the superlattice. We have not found experimental data for this case. However, for thin layer superlattice materials, we and others[7] have found the energy band structures of alloys consisting of the same materials as the thin layer superlattice to be approximately

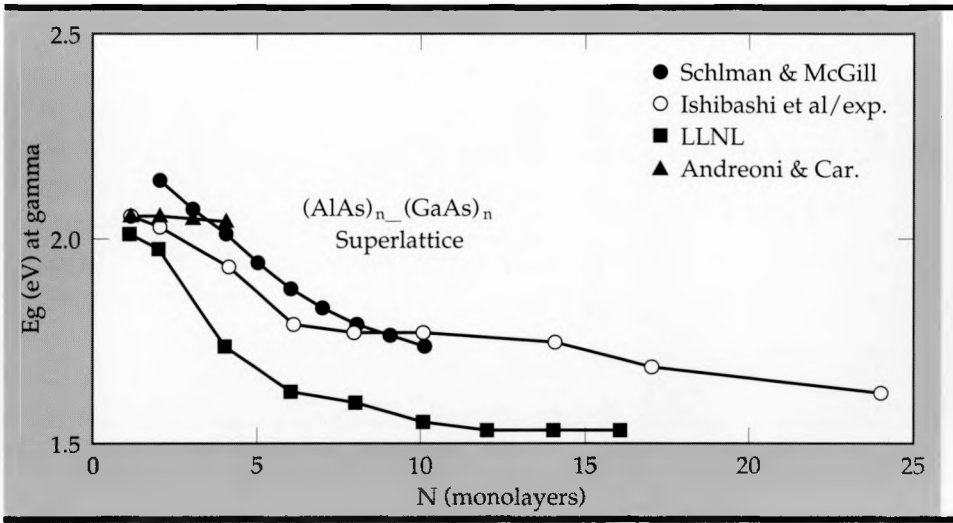


Figure 2. Superlattice band gaps calculated with our code, compared to experimental data. The curves show the energy gap of the gamma point vs number of monolayers of superlattice at 0 K. Our calculations, based on the empirical pseudopotential method, are plotted alongside other theoretical and experimental values.

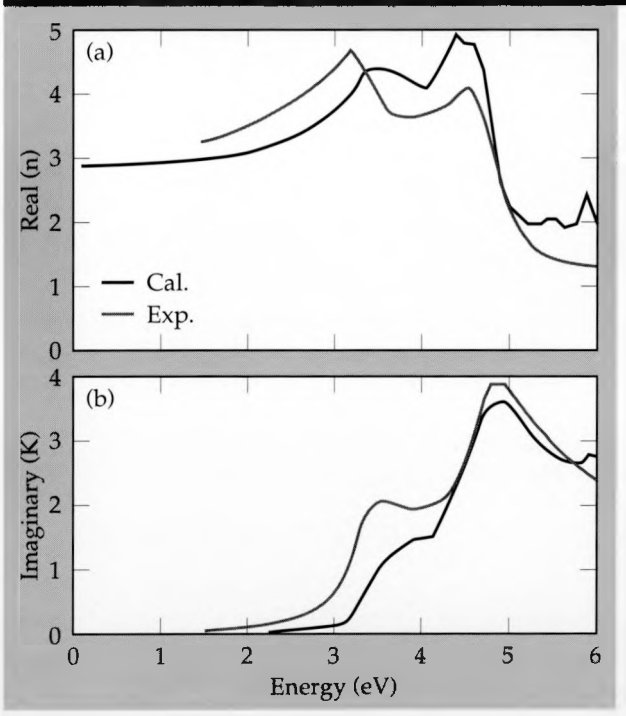


Figure 3. Refractive index calculated with our code at 0 K, compared with reported values of Aspnes at 300 K for $Al_{0.5}Ga_{0.5}As$ alloy. Both real (a) and imaginary (b) parts show good agreement.

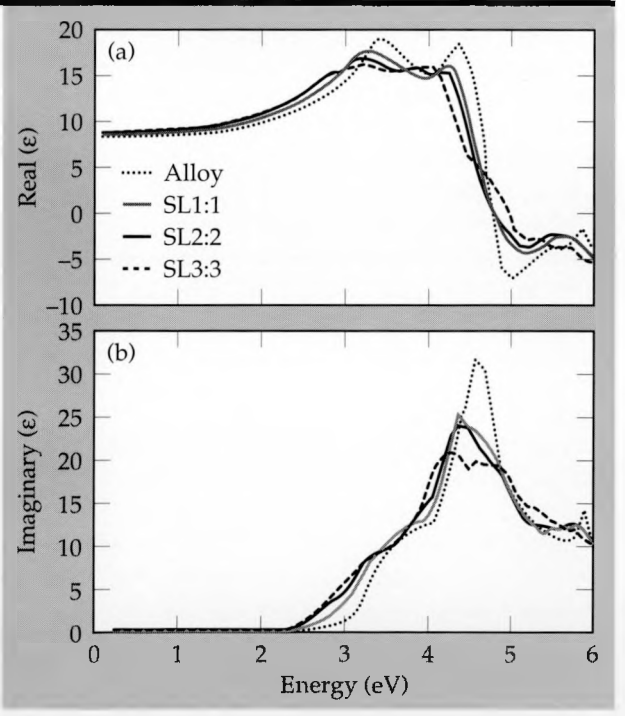


Figure 4. Variation of real (a) and imaginary (b) values of dielectric function as a function of layer thickness. Values are calculated at 0 K for $Al_{0.5}Ga_{0.5}As$ alloy, and also for three superlattice structures, $(AlAs)_1$—$(AlGa)_1$, $(AlAs)_2$—$(AlGa)_2$, and $(AlAs)_3$—$(AlGa)_3$.

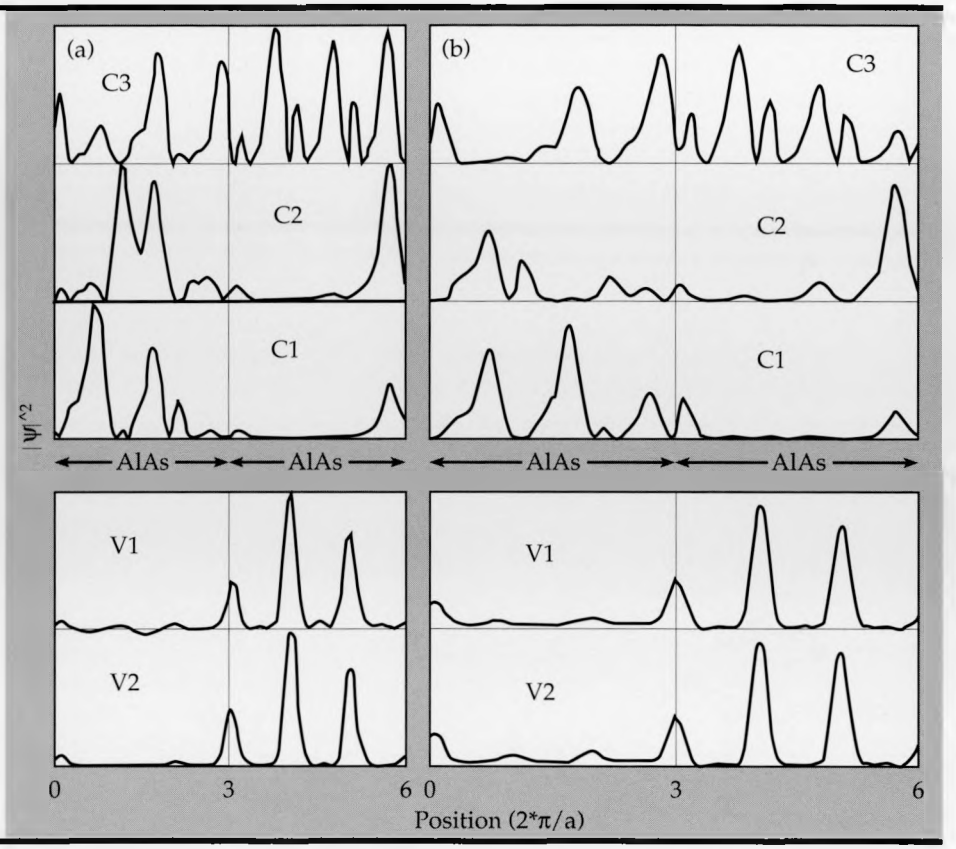


Figure 5. The square of the wavefunction for different energy bands at the center of Brillouin zone, calculated (a) with our code and (b) with a perturbation method.[6] The two sets of calculations are in good agreement.

the same. Consequently, we expect that the refractive index of both materials will be approximately the same.

Figure 5(a) shows how the square of the wavefunction of different energy bands varies as a function of position for a superlattice. **Figure 5(b)** shows the quantity obtained with a different energy band calculational technique.[6] As one can see from a comparison of the curves, our calculations and the others agree very well.

Future Work

During FY 89, we have successfully completed and tested a quantum mechanical superlattice (quantum-well) computer code. In addition, we nearly completed a Monte Carlo transport code. Comparisons of our code results with other calculations and experimental data have shown good agreement.[8] During FY 90, we will investigate with our computer codes electronic devices such as quantum-well lasers and optical modulators that are important to Laboratory programs. In addition, we plan to finalize the inclusion of spin orbit effects in our model, an effort also started in FY 89. Finally, we plan to make our energy band computer code self-consistent.

1. I. Esaki, "History and Perspectives of Semiconductor Superlattices," *Synthetic Modulated Structures*, L. L. Chang, Ed. (Academic Press, Orlando, Fla., 1985), p. 20.
2. G.H. Dohler, "Solid-State Superlattices," *Sci. Am.* **24**(5), 144 (1983).
3. F. Capasso, "New Multilayer and Graded Gap Optoelectronic and High Speed Devices by Band Gap Engineering," *Surf. Sci.* **142**, 513 (1984).
4. D.D. Coon and R.P.G. Karunasiri, "New Mode of IR Detection using Quantum Wells," *Appl. Phys. Lett.* **45**(6), 649 (1984).
5. J. Ihm, P.K. Lam, and M.L. Cohn, *Phys. Rev. B* **20**, 4120 (1979).
6. M.A. Gell, D.Ninno, M.Jaros, and C.C. Herbert, *Phys. Rev.* **B 34**(4), 2416 (1986).
7. W. Andreoni and R. Carr, *Phys. Rev. B* **21**(8), 3334(1980).
8. G.A. Sai-Halasz and L. Esaki, *Phys. Rev. B* **18**(6), 2812 (1978).
9. D. E. Aspnes, S. M. Kelso, R.A. Logan and R. Bhaz, *J. Appl. Phys.* **60**(2), 754 (1986).

Transient Induced Surface Breakdown

Jick H. Yee, Karna J. Wahlstrand, Raymond A. Alvarez, Jr., David J. Mayhall, and Richard W. Ziolkowski
Engineering Research Division
Electronics Engineering

During FY 89, we have developed two computer codes to investigate surface breakdown near metal and dielectric surfaces. One is a Monte Carlo transport code; the other is a two-dimensional gas breakdown code. We have also performed experiments to determine the effects of the dielectric gas SF_6 on the breakdown threshold of air. Mixtures of air and SF_6 may prove to be important in preventing surface flashover. Preliminary results from our codes for the breakdown threshold in a mixture of air and SF_6 have been compared with experimental data. These comparisons have suggested future improvements in our models and the experiments.

Introduction

Pulsed power and high power microwave technology are important in a number of Laboratory and national R&D programs. The ability to handle and transport high electromagnetic fields is limited by breakdown processes at component surfaces. Our ability to improve voltage and power handling capability is limited by our poor understanding of the physical and/or chemical processes that occur at surfaces. Information that is available is largely empirical; existing models are crude and in some instances conflicting.

Among the principal limitations on the generation and transmission of extremely high-power microwave pulses are several electrical breakdown phenomena that occur in the generating apparatus. These include "flashover" discharges across dielectrics separating metallic electrodes (typically initiating at the "triple junction" where the dielectric joins a metallic electrode in a vacuum or insulating gas), "vacuum discharges" between two metallic electrodes in a vacuum (including the resonant phenomenon of multipactoring), and breakdown in insulating gases which are sometimes used to fill the volume between electrodes when high vacuum is not practical.

Typical examples that arise in many microwave systems are dielectric windows which are used to separate evacuated cavities or waveguide segments from system elements at higher pressure, as well as the pressurized or evacuated cavities and waveguides themselves. In some evacuated systems maintaining an extremely good vacuum is impossible or impractical; nor can high-pressure insulating gases be used. An example is the electron beam interaction region in some types of microwave generators. In such cases one must consider the role that possible ionization of residual gas molecules in the necessarily poor vacuum might play. Since fabricating and maintaining large and/or complex systems capable of operating at high vacuum (~10^{-6} or 10^{-7} Torr or better) is generally much more difficult and costly then building a system capable of containing pressurized insulating gases (e.g., freon or sulphur hexafluoride), knowledge of the insulating characteristics of such gases is important for the broad range of cases where SF_6-pressurized systems can be employed.

The objective of this project is to investigate both theoretically and experimentally the physics of surface breakdown induced by high electric field, and to construct realistic computer models for the breakdown processes near the surfaces of metals or insulators, based on the results obtained from these investigations.

Experimental Approach

The various breakdown phenomena are strongly interrelated in the experimental techniques and diagnostics required as well as in some aspects of the physics. Consequently, it is our intention to perform experiments in all three areas. We will study the phenomena over a broad range of parameters; i.e., with a variety of materials, surface conditions, geometries, microwave pulse parameters (length, risetime, repetition rate), vacuum qualities, and other ambient conditions (e.g., pres-

ence of x-rays, uv light, magnetic fields). We will use, where possible, apparatus developed primarily for ongoing microwave-induced air breakdown and propagation studies. The most fruitful options, with the greatest leverage from the use of existing apparatus, appear to be the following:

1. Use an existing pair of rectangular ceramic waveguide windows, one at 45° to the waveguide axis (see **Fig. 1**) and one at 90° to the axis in very similar fixtures, to investigate surface flashover phenomena at microwave frequencies. Comparison of the effects in these two geometries would provide information on the effect of the dielectic angle at the triple junction in high-frequency fields. In pulsed "dc" fields, it is known that the breakdown threshold can increase dramatically (see **Fig. 2**) if the angle between the dielectric and the negative electrode is positive, peaking in the neighborhood of 45°. The concept of negative or positive electrode is not well defined in the microwave case, since the top and bottom broad walls of the waveguide alternate as positive and negative electrodes roughly three times per nanosecond, so it might be argued that 90° must be the optimum angle. Some materials show a reduced breakdown threshold when the angle is negative, as would be the case at each of the triple junctions for half of each rf cycle. This is the case for the two glass dielectrics in **Fig. 2**. Other dielectrics appear to show a minimum breakdown threshold at or near 90° (e.g., Lexan, in **Fig. 2**). In any case, the data in **Fig. 2** indicate that the breakdown threshold should increase for sufficiently large triple-junction angles, either positive or negative. In these experiments, the waveguide on one side of each window is pressurized with an insulating gas, while the other side is evacuated. [**Figure 2** is taken from the paper by H.C. Miller in LLNL Report, CONF-8808171, August 1988.]

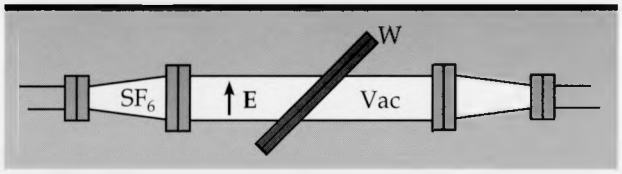


Figure 1. Microwave test fixture, showing ceramic waveguide windows at 45° and 90° to the direction of microwave propagation.

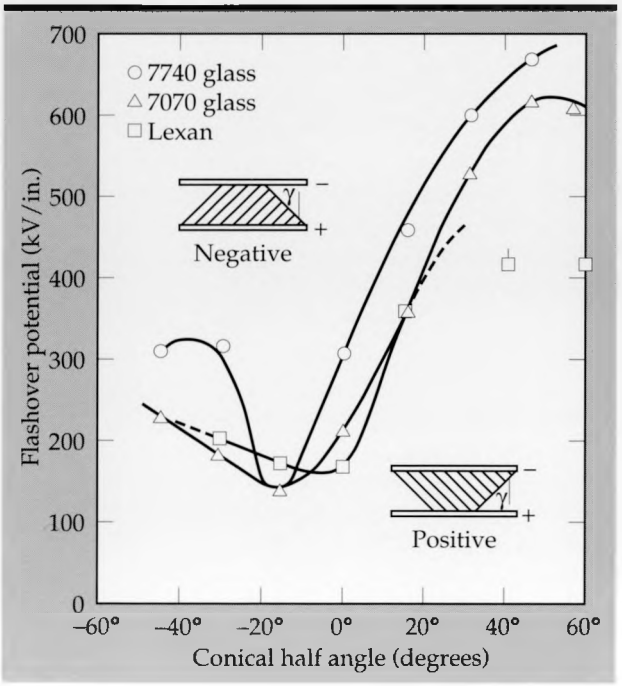


Figure 2. Dependence of dc flashover threshold on the angle between dielectric and negative electrode.

2. Study the insulating properties of SF_6, both pure and in mixtures with air and other gases, at microwave frequencies. Mixtures of air and SF_6 may be important in preventing surface flashover on the interior of windows at antenna feeds or on the exterior surfaces of small radomes in high-power microwave systems. Some study in this area was also required for our implementation of a plan to combine, coherently, the output power of four 20-MW klystrons to provide increased power levels for other experiments on microwave breakdown and propagation. In particular, it was not known whether commercially available waveguide hybrid power splitter/combiners could handle the proposed power levels at the nominal operating pressure or would have to be mechanically strengthened and operated at higher pressures.

Special resonant-ring waveguide structures using a magic-T hybrid for input and output coupling have been assembled for testing the power-handling capability of the commercial magic-T structure. The latter has vanes, irises, and a large indentation at the four-arm junction for matching purposes. Two configurations have been assembled, one with the resonant ring formed with E-plane bends and one with H-plane bends. WR-284 waveguide and components have been used throughout the hybrid/ring systems. The power circulating in either ring builds up to several times the drive power (the gain being slightly dependent on the gas mixture and pressure). This allows a single klystron driver to be used to test the hybrids at several times its input power. The circulating power can be directly monitored by a directional loop coupler incorporated into the ring. In addition to the enhancement of the input power in the hybrid due to the recirculation, one expects further enhancement of the electric fields at the matching vane and indentation. Although the actual fields in the magic-T junction have not been directly meas-

ured, they can be inferred, as a function of input and/or circulating power, from observations of breakdown in pure dry air at 300 Torr. The breakdown threshold field is known from previous investigations for that configuration.

3. Use the TM_{01}-mode circular waveguide system developed for measurements of free-electron lifetimes in air (see **Fig. 3**) to study vacuum breakdown and multipactoring with a variety of metallic surfaces. The field configuration at the ceramic window in this structure can be manipulated by changing the location of the end shorting plate. These changes vary the angle of electric field lines at the window W. In the shorted configuration shown in **Fig. 3**, it is possible to operate with tangential fields at the triple junction of the waveguide and window. It is also possible to move the pattern to produce a high normal field at the center of the window and essentially zero tangential field at the triple junction. During the electron lifetime experiments, breakdown at the window was observed in both configurations for certain operating conditions. The latter included instances where relatively high gas pressure—of air or N_2—existed in the waveguide rather than a high vacuum. Localized breakdown initiating at the metallic shorting plate was also observed in these experiments.

The cylindrical symmetry of the apparatus makes it relatively simple to modify the geometry and material of the shorting plate and window. The symmetry also makes it possible to try to model the experiments with two-dimensional codes. The system can also be operated in a traveling-wave mode when the shorting plate is replaced by an available matched load. It offers versatility for a variety of studies of vacuum breakdown and dielectric flashover.

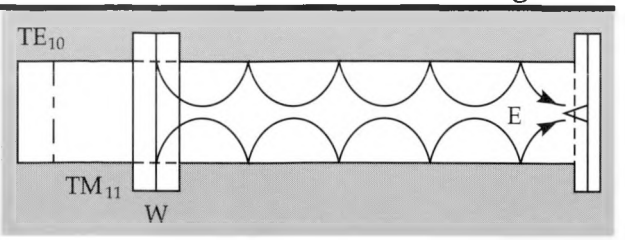


Figure 3. The TM_{01}-mode circular waveguide system for measurement of free-electron lifetimes in air. Changing the position of the end shorting plate varies the angle of the electric field lines at the window W. By replacing the shorting plate with a matched load, the system can be operated in a traveling-wave mode.

Progress

Quantitative experiments comparing surface flashover on the 45° and 90° windows in the rectangular waveguide fixture (Option 1) have been delayed by a small vacuum leak in the 45° window element. We believe that the leak is in the seal between the Macor window and the flanged holder, which should be relatively easy to fix. If the leak is due to a microscopic crack in the ceramic, a replacement window will have to be bought or fabricated.

We have made a limited series of measurements of breakdown thresholds for SF_6, air, and N_2 (and various mixtures) in the two resonant ring configurations (Option 2). The measurements thus far have been restricted to pressures at or below the nominal maximum operating pressure of the commercial magic-T hybrid: 30 psig, or approximately 3 atm absolute pressure. Since there is a slow pressure leak in the E-plane system, it was difficult to maintain a constant pressure except for the case of pure gases. Only a limited set of data were taken for that configuration, and only for various mixtures of dry N_2 and SF_6. No data was taken with dry air. The pressure was allowed to vary by approximately ±2 psi from the nominal value for measurements other than the pure N_2 or pure SF_6 cases. An additional uncertainty of about 0.5 psi in the pressure on the E-plane ring measurements arises from the accuracy of the pressure gauge used on those measurements. The gauge used for the H-plane measurements produces an uncertainty of about half that amount. There was no appreciable pressure drift during those measurements. All measurements used only bottled gases.

Although the location of the breakdown initiation was not directly observed visually on the measurements made thus far, it is reasonable to assume that it always occurred near the matching vane of the magic-T hybrid. Because the circulating high power is fed back into the hybrid in different arms in the E-plane and H-plane configurations, the exact location of initiation may have differed for the two cases. (The same hybrid unit has been used in both the E-plane and H-plane ring configurations.) Unless otherwise noted, all the measurements were made with a pulse length of 1 ms (approximately 50-ns rise time), a 1-pps repetition rate, and a microwave frequency of 2.85 GHz. No external preionization source has been used on the measurements thus far performed. In the case of pure nitrogen, one may expect on all pulses following an initial breakdown a reasonably high level of seed electrons from the plasma afterglow. Previous experience indicates that there are afterglow processes in the 300-Torr air case (and possibly at higher pressures) which also will pro-

vide seed electrons for a second or more following a breakdown event.

No quantitative measurements have yet been made using the TM_{01}-mode waveguide under high vacuum conditions (Option 3).

Results of SF_6 Breakdown Measurements

Generally, the breakdown power level is not well defined. For this reason, some of the following remarks bearing on this point are necessarily qualitative.

The phenomenon of rf processing undoubtedly plays some role. The E-plane system was operated for some time before systematic measurements were made, whereas the systematic accumulation of data on the H-plane configuration began very shortly after assembly of that system. The latter system may therefore show more time-dependence of the data than the former system. There is some evidence of this in the results that will be discussed below. A more systematic approach to this issue is still needed, however.

The typical sequence used in making the breakdown measurements was the following:

1. For each gas mixture, the ring was partially filled with SF_6. Nitrogen (or air) was added to the partial pressure required for the appropriate percentage mixture at 3 atm absolute pressure. Then additional SF_6 was added to bring the total absolute pressure to 3 atm.
2. After the 3-atm breakdown measurements, the pressure was lowered to 2 atm by release of some of the gas mixture, and the breakdown measurements were repeated.
3. The pressure was reduced to 1 atm (0 psig) and the measurements were repeated.
4. Before refilling the ring to 3 atm with a different mixture, the system, including gas fill lines, was pumped out to a pressure of <<1 Torr.

To the extent that processing time creates a bias in the data, the points at lower pressures for a given mixture may be anomalously high relative to the higher-pressure points. Further investigation of this effect is needed. Chemical effects due to breakdown processes may also cause a pressure-dependent bias. The system was generally not refilled between pressure changes for a given gas mixture, except in a few cases where a specific "repeat" run is specified. It is not known at this point whether a chemistry-effect bias would move the data upward or downward. An effort was made to minimize the number of breakdowns on each gas fill to reduce the uncertainty due to chemical effects.

In addition to a long-term processing effect, which may be mainly due to modification of surfaces due to repeated rf discharges, there appeared to be a short-term "processing" effect noticeable after each new gas fill. Typically, a few random breakdowns would appear at relatively low power levels as the power in the ring was being raised. After a few such events, much higher power levels could be reached with few or no breakdowns. Further increase of the power levels would result in frequent breakdowns. Generally it is the latter phenomenon that we have taken as a breakdown limit.

In a more detailed sense, more than one level could be defined as a breakdown threshold or "limit." Specifically, there is a level at which there is a breakdown only rarely, say on one pulse in 10

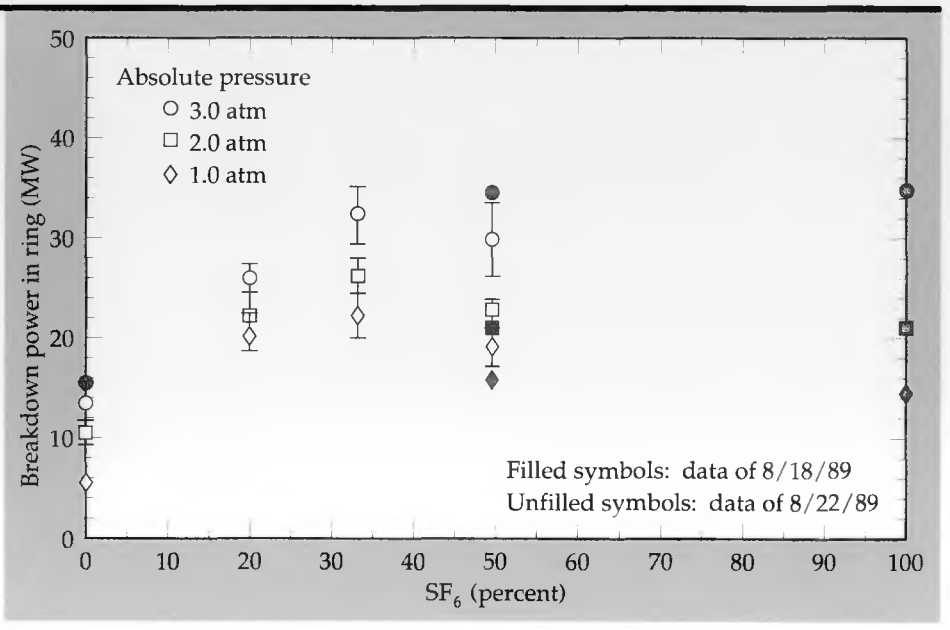


Figure 4. Microwave breakdown threshold data for dry N_2–SF_6 mixture.

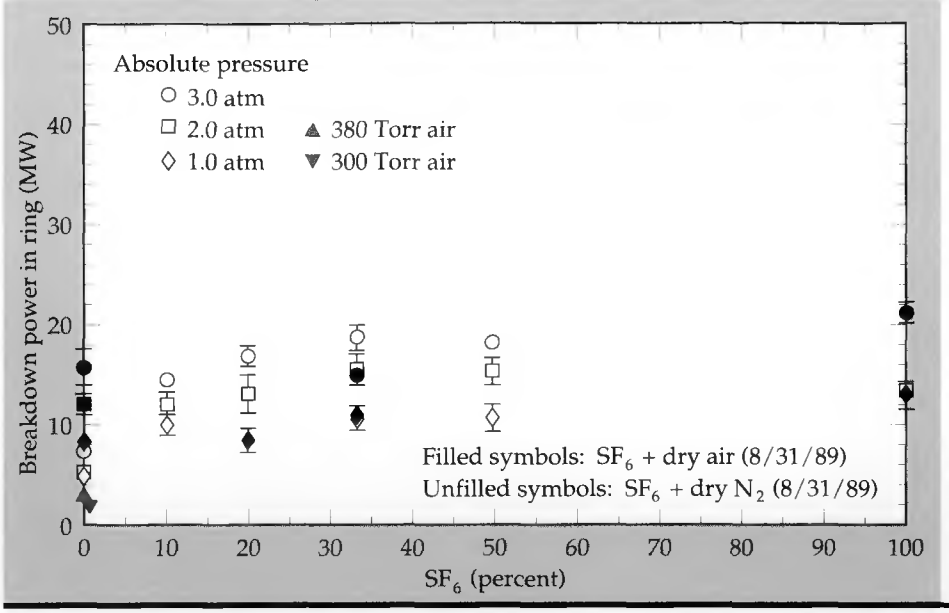


Figure 5. Microwave breakdown threshold data for mixtures of SF_6 with dry air and SF_6 with dry N_2.

or more. After a breakdown occurred, however, it was found that there were generally "pulse-to-pulse" effects, i.e., in order to prevent breakdown continuing on all further pulses the power level had to be reduced. On the initial day of data collection with the E-plane ring, a subjective average between these two levels (each of which is itself somewhat subjective) was recorded. These are shown on the data plots for the E-plane ring as filled figures without error bars. On subsequent days, both levels were recorded. These data are shown as data points with error bars: the point represents the arithmetic mean of the two values; the top of the error bar represents the level at which breakdown is initially observed on approximately 10% of the pulses; the bottom of the error bar represents the reduced power level at which the continuing pulse-to-pulse breakdown is extinguished. The latter, therefore, represents a reasonable electric field limit in applications where no breakdown can be tolerated. The tops of the error bars represent a rough operational limit for single-pulse or low-rep-rate operation where occasional breakdown can be tolerated. Where no error bars are shown for those measurements, the difference between the two limits was smaller than the size of the plotted data point.

At some point during the H-plane ring operation, it was realized that both limits could apparently be raised by "forced processing," i.e., overdriving the ring beyond the initial breakdown limits. In some of the H-plane plots, two sets of data points with error bars are plotted for those sets of measurements: the one plotted with a slight displacement to the left represents the first set of breakdown limits; the one plotted with a slight displacement to the right of the nominal mixture percentage represents limits achieved after the forced processing. In one forced-processing case the limits actually decreased slightly.

E-Plane Ring

Figure 4 shows relative breakdown threshold data for SF_6-N_2 mixtures measured in the E-plane ring. The key feature is the expected increase in breakdown threshold with the addition of SF_6. The addition of 20% SF_6 raises the breakdown level achievable with pure N_2. Increasing the percentage of SF_6 to 33% indicates some further increase in the insulating effect, but a further increase in the SF_6 concentration achieves little or no improvement. Parenthetically, with pure N_2, in addition to having low voltage hold-off capability, the recovery time following a breakdown is quite long, since diffusion is the only seed-electron clearing mechanism.

One set of measurements (not plotted) at 2 pps repetition rate and 25% SF_6 suggests a reduction in the breakdown threshold at the higher repetition rate. Further investigation of any rep-rate effect is needed. Other systematic effects can not be ruled out for the single measurement made thus far.

H-Plane Ring

Figure 5 shows data for mixtures of SF_6 with both dry air and dry N_2 taken on the same day. The improvement in insulation with addition of SF_6 to nitrogen is again evident, as in **Fig. 4**; saturation in the improvement again occurs at about 30%. The power levels are generally about a factor of two lower overall, however, than in the E-plane ring. This probably reflects the difference in field enhancement in the magic-T for the two configura-

tions. **Figure 5** indicates that pure air is a much better insulator than pure nitrogen. This is to be expected; attachment of electrons to oxygen constitutes an electron loss mechanism during breakdown. The improvement in insulation from adding SF_6 to air is therefore less dramatic than in the N_2 case. The chronology of the data points, however, may have produced some upward bias in the pure-air points (see **Table 1** and above discussion). The two low-pressure breakdown points for pure air provide a normalization for obtaining absolute field levels [see, e.g., D. Byrne, UCRL-53764, Oct. 1986]; the measured pressure on the 300-Torr point is more accurate than on the 380-Torr point.

Figure 6 gives an indication of the reproducibility of points (SF6-air mixtures) taken on two successive days. The offset pairs of points (filled symbols) for 50% and 100% SF_6 data of the later date give an idea of the effect of "forced processing" discussed above. The 2-atm data for the 50% mixture show that the process is not unfailingly successful.

Figure 7 indicates the gain in apparent insulating efficiency with repetition of data on the same date. The two sets of points indicate an increase in the breakdown levels, which could be due to additional processing time. Since only two conditions were repeated, however, other systematic or random effects cannot be ruled out.

It is clear that further study of the effect of rf processing is needed. The systematic effect of pressure and SF_6 concentration should be examined after thorough processing of the apparatus has taken place. The magic-T and other ring compo-

Figure 6. Example of the reproducibility of data, giving microwave breakdown data for dry air and SF_6 on two successive days.

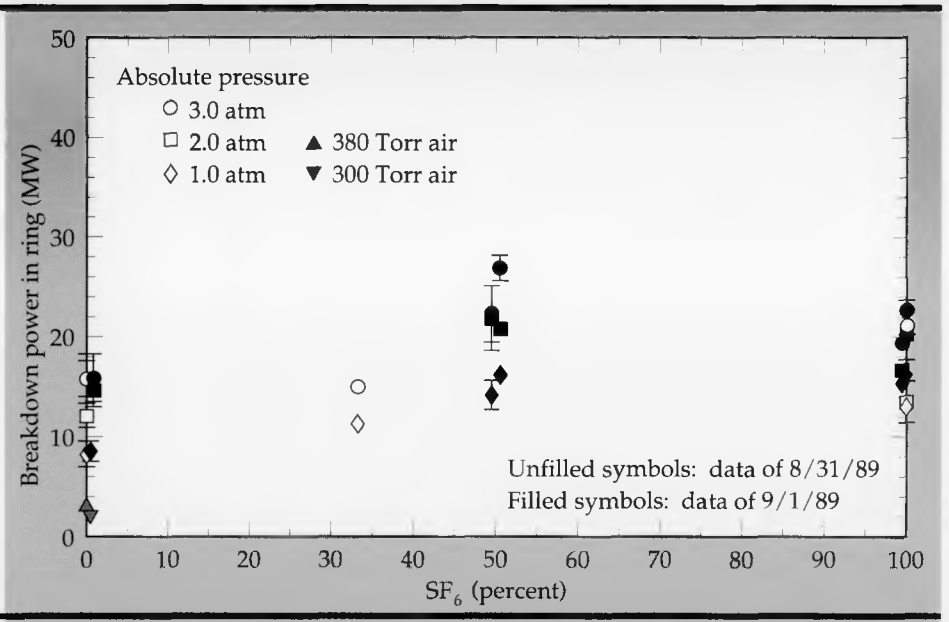
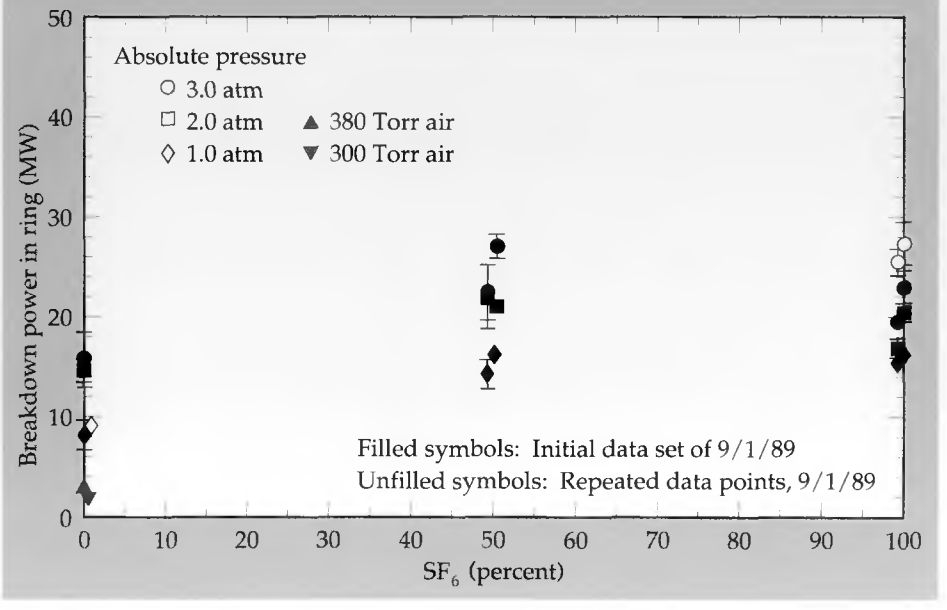


Figure 7. Example of gain in apparent insulating efficiency shown through repeating experiment later in the same day.



Average values from simple analytical model					
		(cgs units): F is electron field in kV/cm			
Quality	Average value	F = 5.0 Fawcett	F = 5.0 Code	F = 15.0 Fawcett	F = 15.0 Code
(100) Valley population fraction	f_{200}	0.25	0.267	0.65	0.667
(000) Valley drift velocity (cm/sec)	$\langle V_d \rangle_{000}$	2.3×10^7	2.4×10^7	2.1×10^7	2.2×10^7
(000) Valley mobility, (cm^2/Vsec)	$\langle \mu \rangle_{000}$	5×10^3	4.8×10^3	1.5×10^3	1.47×10^3
(000) Valley average energy (eV)	$\langle \epsilon \rangle_{000}$	0.15–0.20	0.17	0.25	0.29

Table 1. Comparison of average values obtained with our model and values obtained by Fawcett et al.,[2,3] showing good agreement between our model and well-established Monte Carlo methodology.

nents should be mechanically strengthened to allow the extension of these measurements to much higher pressures.

Theoretical Approach

It has been found experimentally that a thin dielectric layer covering a sharp point on a metal surface greatly reduces the threshold for field emission. Therefore, to understand breakdown processes near surfaces, one has to understand how electrons are transported from the metal through the dielectric and then to the surface. For this reason, during FY 89 we have developed two Monte Carlo transport codes to investigate the transport phenomena in dielectric materials. In one of these codes, the conduction bands of the material are represented by analytical equations. The other code uses the conduction bands calculated with a pseudopotential energy band code. The latter is more accurate for very high field transport phenomena when the effect of ionization is important. We tested these codes using a GaAs test crystal.

The Monte Carlo method[1,2] uses random numbers to sample probability distributions that describe physical systems in order to obtain their physical properties. In semiconductors the physical system is an electron, modeled as a particle with an effective mass m^*, drifting in a uniform applied electric field through a crystalline lattice. The lattice acts as a scattering agent. The relevant probability distributions include free flight time, scattering rates, and the angular distribution of the final wavevectors.

The equation of motion for the wavevector describing an electron's free flight in an electric field is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{K} = e\mathbf{E}/h \quad , \tag{1}$$

where the time distribution is given by

$$P(t) = \frac{1}{\tau_0}\exp(-t/\tau_0) \quad . \tag{2}$$

In the latter formula τ_0 is the inverse of a maximum scattering rate. It includes self-scattering events where the free flight continues and no actual collision takes place. Other scattering processes are selected by examining partial sums of scattering probabilities compared with a random number times τ_0^{-1} and then choosing the process whose partial sum just exceeds the latter. These scattering probabilities are calculated from Fermi's Golden Rule and are functions of the initial and final energies of the electron being scattered. We tabulate a histogram of time spent by the electron in a given energy range. After dividing by the density of states, this gives the distribution function as a function of energy. We compute average values $\langle a \rangle$ of transport quantities by summing integrals over the free flight time of the given quantity divided by the total time spent by the electron in the crystal:

$$\langle a \rangle = \frac{1}{T}\sum_i \int_0^{t_i} a(t)\,\mathrm{d}t \quad .$$

In the first code, we represent the conduction band by a band-structure similar to that of GaAs. The conduction bands consist of the central (000) valley and three equivalent (100) valleys. The (100) valley conduction bands are separated from the (000) conduction band by δ. For GaAs, this is 0.33

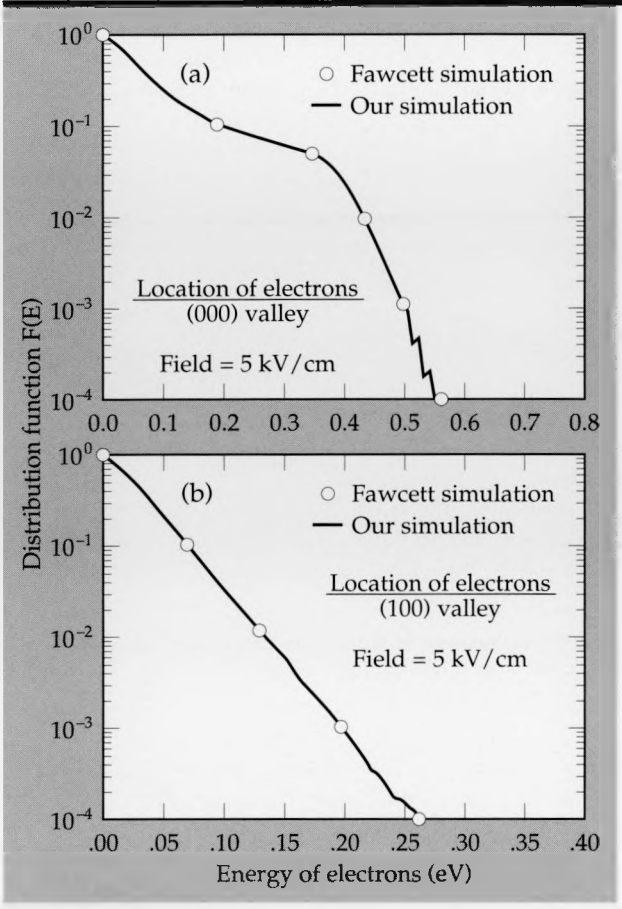


Figure 8. Comparison of distribution function computed with our model and distribution function obtained by Fawcett.[1,2]

eV. Near $k = 0$, the energy band for the (100) valley is represented analytically by

$$E = \frac{\hbar^2 k^2}{2m^*}$$

and for the (000) valley it is represented by

$$E(1 + \alpha E) = \frac{\hbar^2 k^2}{2m^*} .$$

The scattering processes include polar optical scattering and acoustic scattering in both valleys, equivalent intervalley scattering in the (100) valleys, and nonequivalent intervalley scattering between the (000) and (100) valleys. Both phonon absorption and emission can occur in all cases. Acoustic scattering, however, is treated as elastic. The angular distribution of the final states is random in intervalley scattering. On the other hand, for both polar optical and acoustic scattering it is a complicated function of initial and final energies and the angle β between initial and final wavevectors. The details of all scattering cross-sections will be presented in a future publication.

This model is identical to that used in 1970 by Fawcett and co-workers.[2,3] We have obtained distribution functions which agree exactly with theirs (see **Fig. 8**) for our test GaAs crystal. Since our average values are also very similar to theirs (see **Table 1**), we consider this code validated in its present form.

There are some shortcomings to the simple model, however. Later studies of GaAs[4] showed that this simple model is only valid near minima in the bandstructure. In practice, this means at low fields, $F \leq 30$ kV/cm. If this model is used at higher fields, the energy and drift velocity increase without limit (see **Fig. 9**), which disagrees with experimental results for GaAs.

Since we are primarily interested in high field transport, it became necessary to write a second Monte Carlo code which included impact ionization and the exact bandstructure of the crystal. This code uses only the first conduction band, obtained from an empirical pseudopotential band-

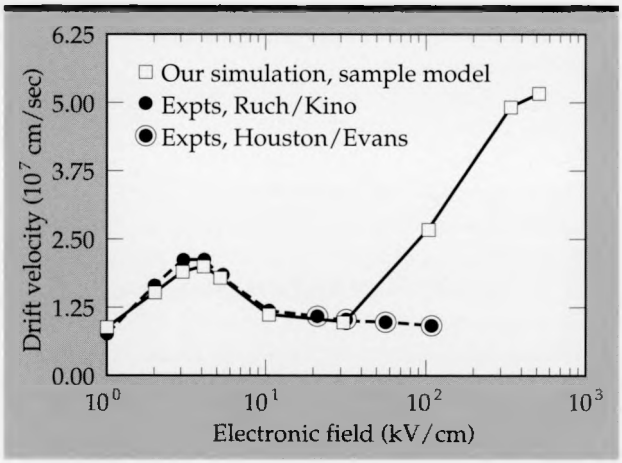


Figure 9. Average values from simple analytical model as a function of electric field, showing disagreement with experimental results at higher fields.

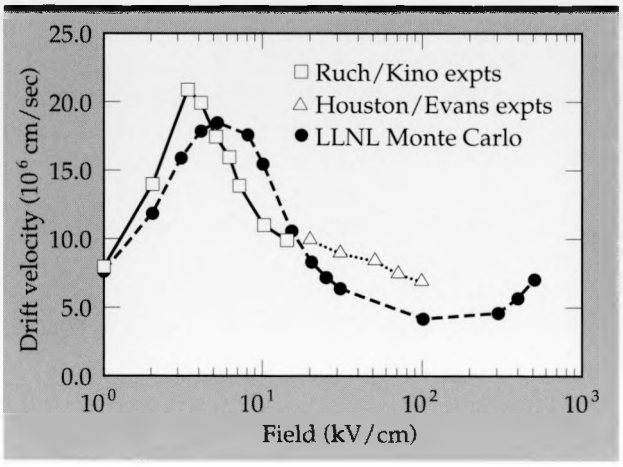


Figure 10. Drift velocity obtained from our more complex, exact bandstructure model as a function of electric field, compared with experiments.

structure code and stored in a table for interpolation. Impact ionization, the creation of an electron-hole pair above a threshold energy E_T, is modeled as another scattering process whose probability is given by the Keldysh formula,[5]

$$\lambda_{II}(E) = \lambda_{phn}(E_T) P \left(\frac{E - E_T}{E_T} \right)^2 , \qquad (7)$$

where E_T = 2.0 eV and P is a constant much greater than one (taken as 400 here). This process is found to occur only at high fields, $F \geq 100$ kV/cm. Other scattering processes (omitting acoustic scattering) are modeled as in the simple analytical code used by Littlejohn et al.[4] Different parameter values are used and the (111) valleys are included.

We compare our code's results with two similar Monte Carlo code calculations for GaAs in the literature. The first one,[5] done by Shichijo and Hess in 1981, is very much like our code, except that they do not distinguish between satellite valleys in the bandstructure. We define valleys separated by symmetry line energy maxima to tie the scattering rates to the bandstructure explicitly. In addition, they use the (000) scattering rates in all valleys rather than the Littlejohn parameters. Finally, they use time steps rather than sample the time distribution **[Eq. (2)]** directly.

The other exact bandstructure Monte Carlo code for GaAs in the literature was developed by Fischetti and Laux of IBM in 1988.[6] This is a more sophisticated code, using exact scattering rates taken from the bandstructure and including electron-electron interaction. Space charge effects are also included using a Poisson solution. Moreover, they include the first five conduction bands, not just the lowest one.

In **Fig. 10**, we compare the average drift velocity values as a function of electric field obtained with our simulations with experiments where available data exists. The drift velocity is computed at low fields from analytical fits of the bandstructure according to

$$V_d = \frac{1}{\hbar} \frac{\partial E}{\partial k} \qquad (8)$$

and at high fields is calculated by direct interpolation of energy gradient tables. Considering the fact that some numerical error occurs in the interpolation and fitting of the energy gradients, we have very good agreement with the available experimental drift velocity data. We are now in a position to proceed with studies of electron transport and dielectric breakdown.

In addition to the development of the Monte Carlo transport codes, we have developed a two-dimensional electron fluid gas breakdown code which includes the effects of field emission from sharp tips on the metal surfaces. **Figure 11** shows the geometry used in our calculations and **Fig. 12** shows the results of our calculations for the spatial variation of the electron density near a sharp discontinuity in the cylindrical waveguide. As expected from field emission theory, the electron density has the highest value near a sharp corner where the field is highest.

Future Work

During FY 89, we have investigated the effect of SF_6 gas on the breakdown threshold fields of air and nitrogen gas. The results obtained from our investigation confirmed the results obtained by others,i.e., that only a small impurity fraction, in the vicinity of 10%, has a dramatic effect on the breakdown threshold. After this level has been reached, further increase of the SF_6 fraction above

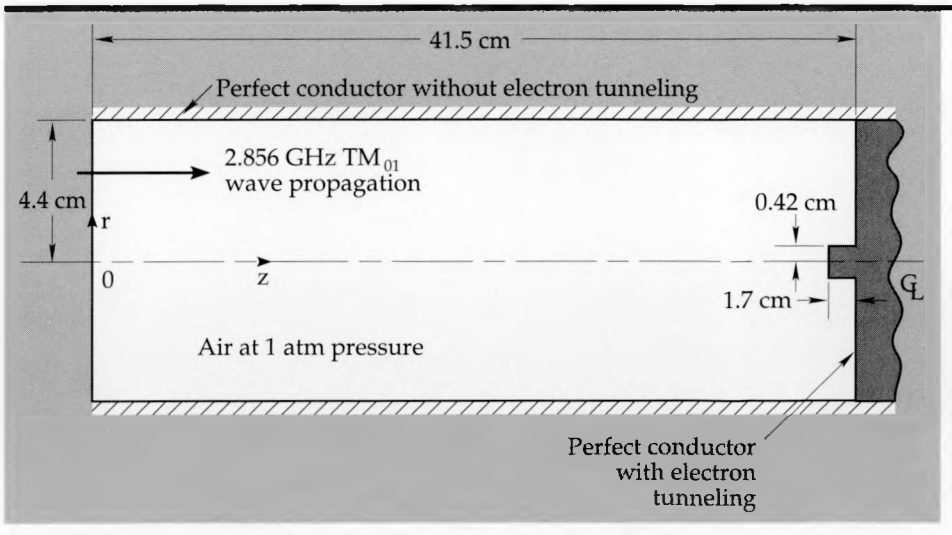


Figure 11. Computational geometry for our two-dimensional, electron-fluid gas breakdown code.

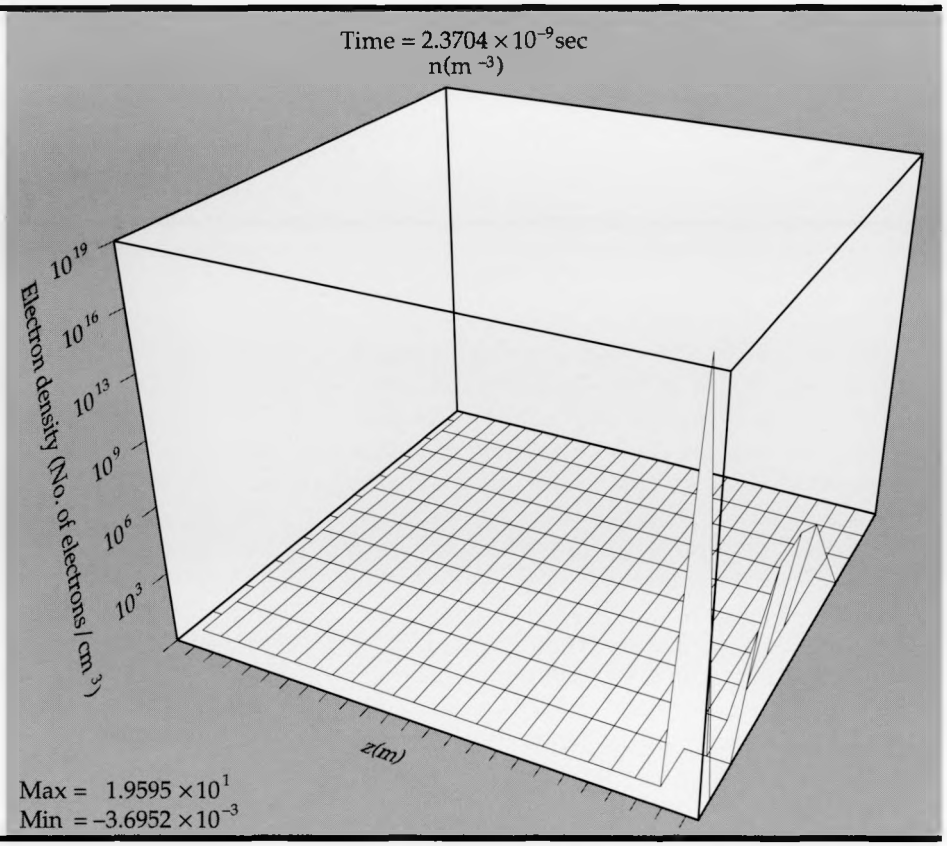


Figure 12. Results of electron fluid code calculations, showing spatial variation of electron density near a discontinuity in waveguide. The electron density is highest near a sharp corner where the electric field is also highest.

10% has only a slight effect.

SF_6 has often been used in waveguides and microwave systems to prevent the occurrence of high field breakdown in high power applications. The understanding of the physics of the breakdown in SF_6, when subjected to high fields, will help ultimately in the understanding of surface breakdown when microwave systems are immersed in mixtures of SF_6 and air.

Three codes were developed in FY 89: two Monte Carlo transport codes and a two-dimensional nonlinear electron fluid gas breakdown code that takes into account the effect of field emission from the metal surface. The Monte Carlo codes have been validated with existing data for GaAs crystals.

None of the codes developed to date has taken into consideration the effects of scattering near the surface where breakdown occurs. Therefore, our existing codes are still not adequate for modeling surface breakdown phenomena. As a result, future efforts will concentrate on the development of a physical model that will take the scattering processes near the breakdown surface into consideration.

1. C. Jacoboni and L. Reggiani, *Rev. Mod. Phys.* 55, 645 (1983).
2. A.D. Boardman, *Physics Programs* (John Wiley and Sons, New York, 1980), p. 355.
3. W. Fawcett, A.D. Boardman, and S. Swain, *J. Phys. Chem. Solids* **31**, 1963 (1970).
4. M.A. Littlejohn, J.R. Hauser, and T.H. Glisson, *J. Appl. Phys.* **48**, 4587 (1977).
5. H. Shichijo and K. Hess, *Phys. Rev. B* **23**, 4197 (1981).
6. M.V. Fischetti and S.E. Laux, *Phys. Rev. B* **38**, 9721 (1988).

AMOS Development and Applications

John F. DeFord, Robert R. McLeod, and Ronald F. Schmucker
Engineering Research Division
Electronics Engineering

George D. Craig and Bernhard Kulke
F-Division

AMOS is an electromagnetic-simulation computer code that has been designed specifically to study the rf properties of highly complex, rotationally symmetric cavities (or cells). The code has attracted significant interest from the particle-accelerator community at large, although to date it has been used primaily to design induction cells for high-current particle accelerators at LLNL. The most recent improvements made to the code, and several interesting applications of AMOS to actual problems are discussed in this article.

Introduction

AMOS (Azimuthal MOde Simulator)[1] is a "two-and-a-half-dimensional (2.5D)", finite-difference time-domain (FDTD), electromagnetic simulation computer code. It has been designed specifically to study the rf properties of rotationally symmetric cavities (or cells), especially those of highly complex geometric form with a variety of dielectric and magnetic inclusions. The code has been used primarily in the design of induction cells for high-current particle accelerators at LLNL.

Figure 1 is a simplified cross-sectional diagram of a portion of a "typical" induction accelerator, which shows a serial grouping of two induction cells and identifies the cell's internal regions and components, including empty (vacuum) spaces and materials. The diagram is of a full azimuthal cross section, rotationally symmetric about the axis, except for the pulse-power feed line.

The electron beam travels through the beampipe on axis, and experiences an accelerating voltage as it traverses the gap. This voltage is induced across the gap by a pulse delivered by the pulse-power feed line. The large ferrite toroid serves both as a magnetic switch and as an inductive load to the pulse-power drive. (It is required to match the cell impedance with the impedance of the pulse-power feed line.) The ferrite pucks are placed in the cell at strategic points to damp out undesirable rf resonances in the cavity.

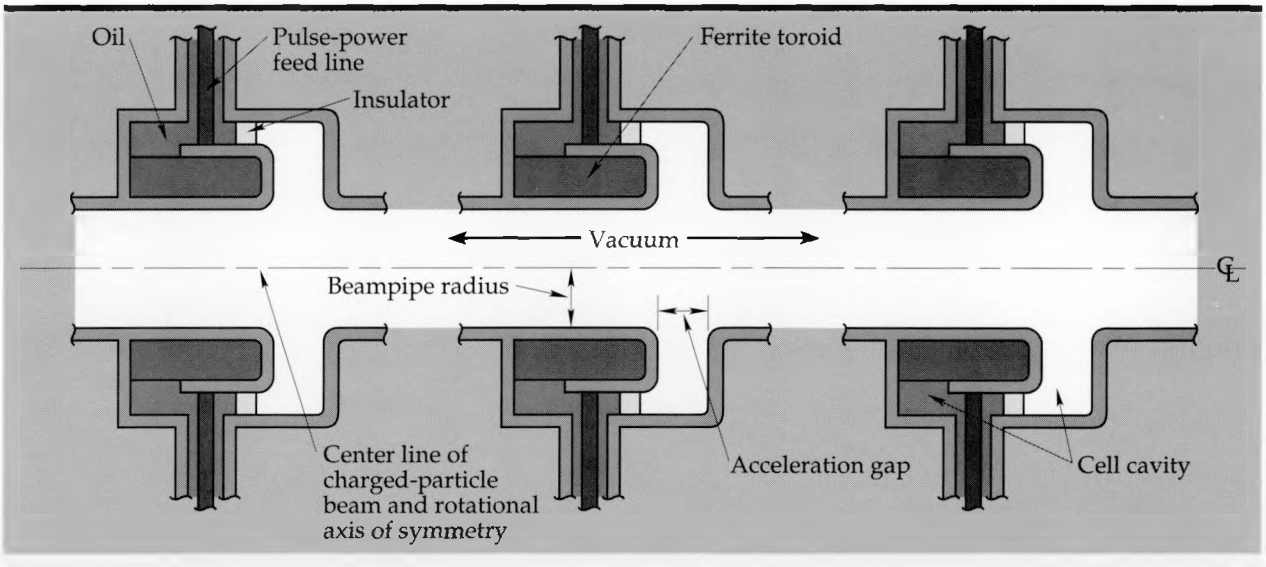


Figure 1. Simplified diagram of a full azimuthal cross section of a three-cell portion of a charged-particle induction linear accelerator, showing typical features. Actual cell is rotationally symmetric about the axis, except for the pulse-power feed line, that delivers a beam-accelerating pulse to the gap. The electron beam travels the beampipe axis and is accelerated as it traverses each gap. The large ferrite toroid acts as an inductive load to the pulse-power drive; it is required to match the cell impedance with the impedance of the feed line. Ferrite pucks are placed in the cell at strategic points to damp out undesirable rf resonances (see Fig. 4).

AMOS computes the temporal evolution of electromagnetic fields in rotationally symmetric simulation volumes (commonly called "bodies of revolution") by using the FDTD algorithm on fields distributed on a two-dimensional (2D) field grid with radial and axial coordinates, r and z, respectively. A highly simplified diagram of an AMOS simulation geometry and associated FDTD grid is shown in **Fig. 2**. The projection of the fields in the volume onto the 2D grid is accomplished by expanding the fields in a Fourier series in the third cylindrical (azimuthal) coordinate, ϕ. The mth term, or multipole mode, in the series varies as $\varepsilon^{(im\phi)}$; AMOS is used to solve for the temporal evolution of the mth mode via user specification of m. AMOS allows the fields to vary in a predefined (harmonic) fashion in the ϕ coordinate; it is for this reason that AMOS is called a 2.5D code to distinguish it from a purely 2D code in which the fields do not vary with ϕ.

The simulation grid is composed of a set of rectangular subgrids, each having a single set of electrical properties that apply to every element within the subgrid; the electrical properties are ε, σ, and μ, where ε is permittivity, σ is conductivity, and μ is permeability. Radiation or impedance boundary conditions can be applied on the boundaries of the simulation grid as necessary. Both electric and magnetic currents (**J** and **K**) may be used as sources for the electromagnetic fields, or the values of **E** and **H** may be set directly over a portion of the grid. The user can request output on any field component at any specified time interval at any position within the simulation volume. Desired output positions that do not correspond to a node on the grid or to a particular time step are interpolated from adjacent values.

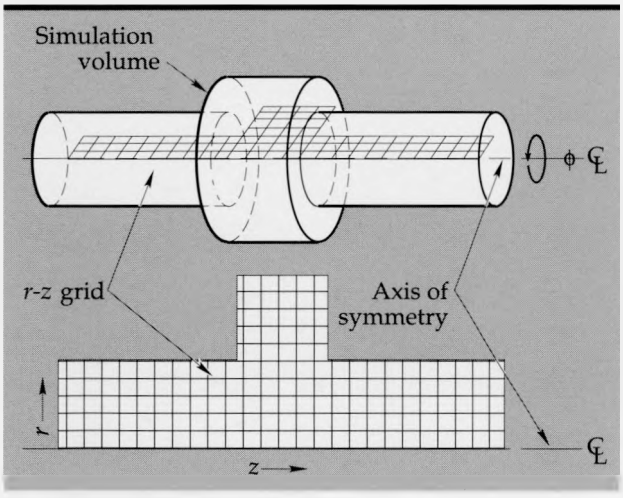


Figure 2. An example of an AMOS simulation volume and corresponding field grid. AMOS models the fields in a cylindrical volume by distributing the electric and magnetic field components (E_r, E_ϕ, E_z, and H_r, H_ϕ, H_z) in each rectangular element in an r-z grid, and by assuming a known ϕ-dependence of the form $\exp(im\phi)$. If a ferrite model is installed, then the vector components of the magnetization, M, are also represented on the grid.

To simplify the user definition of the simulation grid, the computer code DRAGON was developed. DRAGON is an interactive, window-based tool. It allows the user to define the problem geometry by entering a series of points, either by moving a cursor around on the graphics screen, or via an input file, or even directly from a computer-aided-design (CAD) system (see **Fig. 3**). The program connects the points with line segments, and these collectively define the problem geometry. DRAGON then allows the user to edit the geometry through a variety of operations such as moving points, moving line segments (collections of points), rotating line segments, adding and deleting points, etc. Once the model is defined, the user specifies the type of material in each closed polygon making up the problem geometry. DRAGON then constructs an FDTD grid that closely approximates the geometry. This grid is subsequently divided into subgrids, and DRAGON writes the descriptions of these subgrids out to a file that becomes part of the AMOS input specification.

DRAGON is classified as a "preprocessor," because it allows processing of input data before the simulation is performed by AMOS. A postprocessor has also been developed, called ANDY, that allows the user to carry out a variety of operations on the field data generated by AMOS. A simple example of the operations performed by ANDY is reading the time series of a particular field component from the AMOS output file, and plotting a time history from these data.

A time series may also be manipulated before being displayed. The available manipulations include Fourier transforms and algebraic operations. The latter include multiplication by a constant or by another series, addition and subtraction of series, etc. More sophisticated operations may be performed involving several time series at once, including the calculation of wake potentials and impedances; these two quantities are important for understanding the interaction of a charged particle beam with a structure such as a linac induction cell.[2,3]

In this article, we discuss our progress in recent development efforts in the areas of code documentation, AMOS and ANDY integration, and ferrite modeling. Also discussed are several applications of AMOS to specific design and diagnostic problems involving induction cells from three different

LLNL accelerators: ATA (Advanced Test Accelerator), SNOMAD, and DARHT (Dual Access Radiographic Hydrodynamics Test facility).

Progress

During FY 89, AMOS development efforts were focused in three main areas:

- Code documentation,
- Integration of ANDY and AMOS into a single code to simplify use and maintenance, and
- Development of magnetic material models for a broad spectrum of frequencies. The progress in these areas is discussed below.

Code Documentation

With any large computer program, dedication to well defined coding standards, as well as to concise and complete documentation, is critical to ease of maintenance and future development. During the past year, we have largely completed the internal documentation of the FORTRAN-77 that comprises all of AMOS and ANDY. This has been an extensive effort, involving the writing of some 25,000 lines of documentation for the approximately 12,000 lines of FORTRAN executable statements in these codes. In addition, some restructuring of the code was carried out to simplify the conceptual framework of the code and to improve its modularity. DRAGON is also written in FORTRAN and is extensively documented. It brings the total size of the DRAGON/AMOS/ANDY code set, including comments, to approximately 50,000 lines.

We have developed the rudiments of a user's manual and a programmer's guide for the code set. This part of the documentation effort will continue as more users access the codes and as others undertake the routine maintenance duties.

We anticipate that our work in documentation and attention to design details will be of significant benefit as other physics modules are added and the code becomes more heavily used.

AMOS and ANDY Integration

AMOS and its postprocessor, ANDY, have been integrated into a single code during the past year to simplify both the use and maintenance of the codes. (The code package retains the name AMOS.) Besides simplifying the actual simulation process (now only AMOS needs to be initiated, and ANDY is run by AMOS as a subroutine), this change also reduces the number of files that need to be written out to disk storage. In addition, all of the external files that are necessary to run the simulator are now uniquely named according to the particular problem that is being solved. This allows multiple, independent copies of AMOS to operate simultaneously in the same file space.

Future work will likely include the integration of DRAGON into AMOS; the entire DRAGON/AMOS/ANDY code set would then become a single package. We have not yet performed this final integration, because DRAGON is an interactive tool requiring a certain minimum level of interaction with the user via a graphical interface. The computer hardware necessary to support this graphical interface is not available everywhere.

It should be noted that while the component codes are being integrated into a single package for

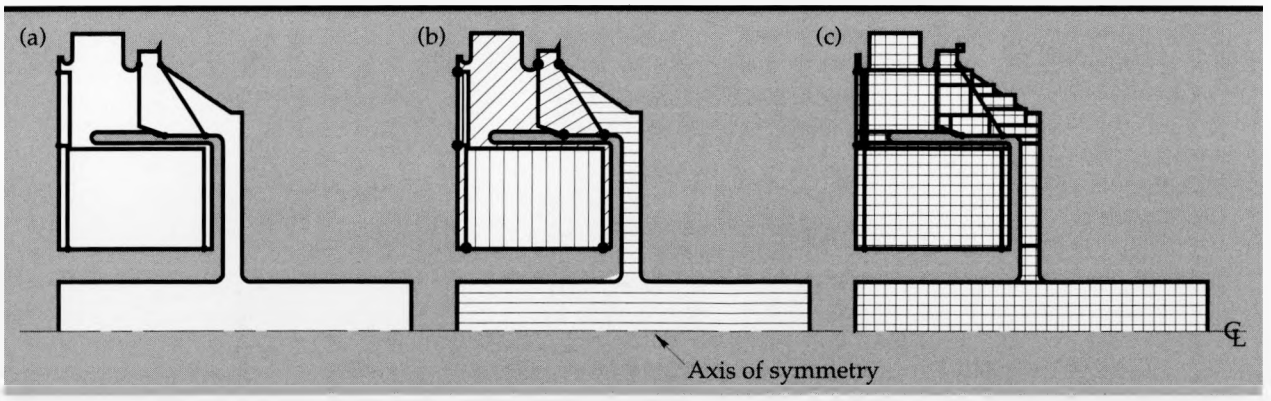


Figure 3. Example of definition of problem geometry using DRAGON to diagram a one-half azimuthal cross section of a "typical" induction cell. (a) Line segment definition of the model boundaries. The points are entered via mouse, keyboard, or input file. The resulting figure may be edited as necessary. (b) Material specification is made by placing the graphics cursor in each region and entering the material type for that region. (c) Finally, the FDTD grid is generated to fill the simulation area. The grid is divided into a set of rectangular subgrids, each associated with a region of different material. The specifications for the grid are then written into a data file, and this file is included as part of the AMOS input file.

ease of use and maintenance, the components themselves remain as modules. They may be removed and replaced with other processors that have the appropriate "hooks" or interfaces to work with the AMOS kernel. The modules may also be run as "stand alone" codes, separate from the AMOS kernel.

Ferrite Modeling

Ferrite modeling is important for linac induction cell design. There is typically a large quantity of the magnetic material in an induction cell, which serves as both a load to the pulse-power drive to a cell and as a microwave absorber to damp any electromagnetic resonances in the cavity (see, for example, **Figs. 1** and **4**).

In typical linacs, the ferrite loads in each induction cell are biased to near saturation by a charging current. In saturation, the magnetic dipoles in the ferrite are all pointing in a particular direction, and we say that the material is completely magnetized. The magnetization in the ferrite material reverses direction during the application of the pulse-power drive to the cell. This rotation of the dipoles occurs with a characteristic time of approximately100 ns in high-current, short-pulse accelerators such as the ATA and ETA-II (Experimental Test Accelerator II) at LLNL, and thus is characterized by frequencies of less than approximately 10 MHz. This large-amplitude, low-frequency rotation of the magnetization, **M**, may be modeled with a nonlinear equation relating $d\mathbf{M}/dt$ (time derivative of **M**) to **M** and to the magnetic field, **H**. A model suitable for a limited class of ferrites and field strengths has been developed by Avery, et al.;[4] it has been implemented in AMOS and is being tested.

The absorption of microwaves by the ferrite material at frequencies corresponding to resonances in induction cells results in relatively small oscillations in **M** about its equilibrium value. The frequencies of interest (in the range of 200 to 5000 MHz) are significantly higher than those associated with the pulse-power response. To model the rf absorption, a linear-tensor permeability model is appropriate. This model specifies a frequency-dependent matrix (tensor) relation between **M** and **H**. An experimental program was initiated during the summer of 1989 to determine the microwave properties of the ferrites used in the ATA and ETA-II accelerators. The experimental data was used to determine the coefficients in the tensor for these materials.

In general the implementation of a frequency-dependent constitutive parameter—such as the permittivity, ε, or the permeability, μ—in a time-domain code requires the use of a temporal convolution over the fields. From the standpoint of computer resources this is an expensive proposition in a time-domain simulation code. In fact, it is quite impractical for large problems involving many FDTD elements. However, it has recently been shown that for a certain class of frequency dependencies, the convolution may be done in a simple

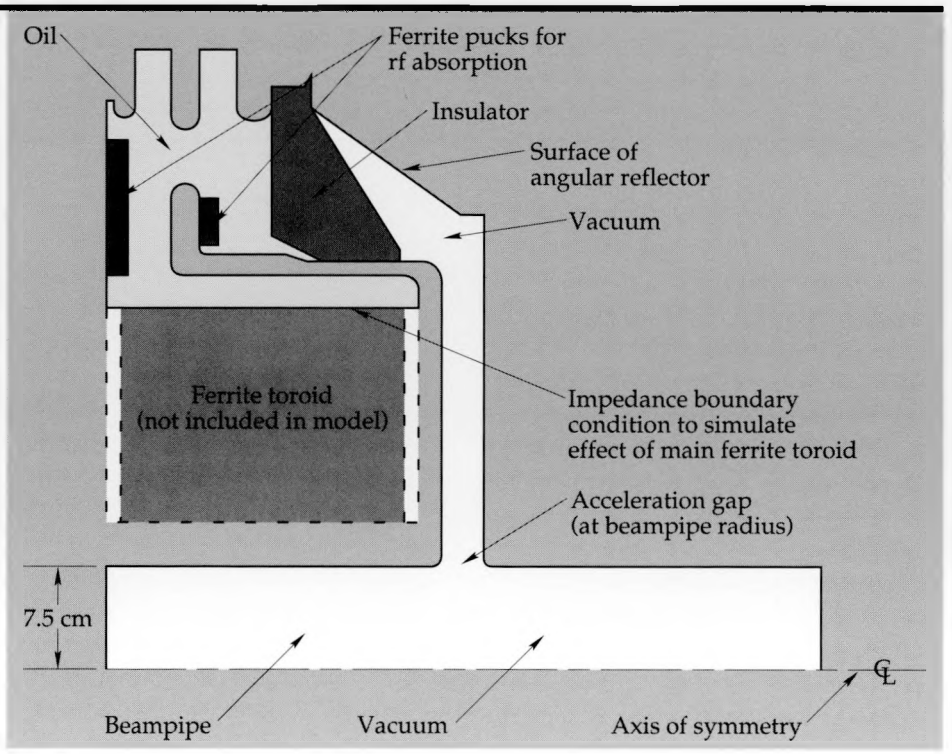


Figure 4. DRAGON rendering of azimuthal cross section of ATA cell. This cell was one of the first studied using AMOS. The primary focus in this early investigation was to determine which of the cavity resonances could be expected to couple strongly enough to the beam to cause beam oscillations and breakup of the beam. Ferrite serves as a microwave absorber to damp any electromagnetic resonances in the cavity and as a load to the pulse-power drive. Ferrite modeling is, therefore, very important for linac induction cell design.

fashion. In these cases, no significant increases in CPU time or memory are necessary over and above that required for dispersionless (nonfrequency-dependent) materials.[5] The tensor permeability suitable for the soft ferrite in most induction linacs may be approximated in this special form.

It is expected that the high-frequency portion of the overall ferrite model will be installed during the first quarter of 1990. Following our overall design philosophy, both the low-frequency and high-frequency ferrite models will be set up as software modules that may be inserted into AMOS whenever necessary. This will provide for simple updates to the models for different ferrites. It will also eliminate the need to perform the extra calculations necessary for ferrite simulation in those cases where there is no ferrite present.

Applications of AMOS to Specific Problems

During the past year, AMOS has been applied to a variety of design and diagnostic problems involving linac cavities of interest to both the Laboratory and to the national particle-accelerator community. The typical questions that AMOS is used to answer concern primarily the frequencies of cavity resonances and the design variations that would reduce beam/resonance coupling. In the following, we discuss the work on induction cells for three different accelerators: ATA, SNOMAD, and DARHT.

ATA Induction Cell

The ATA induction cell (**Fig. 4**) was one of the first cells studied using AMOS. The primary focus in this investigation was to determine which of the cavity resonances could be expected to couple strongly enough to the beam to cause beam oscillations and breakup of the beam. These beam breakup modes can lead to beam loss on the walls of the accelerator beampipe.

It was possible to determine which cavity modes would be excited by an electron beam transiting the acceleration gap by simulating the drive of a short pulse of current across the gap at the beampipe radius (see **Fig. 4**) and then taking the Fourier transform of the resulting axial electric fields in the gap. Also, by computing the wake potential and the cavity impedance, we were able to determine the strength of this beam/cavity coupling.

This procedure yielded the discovery of two modes, both quite near in frequency to the TE_{11} cutoff frequency of the beampipe. These modes were not appreciably damped by the ferrite in the cavity. In the absence of a high-frequency ferrite model, the ferrite was approximated by an impedance boundary condition. One of these modes may be expected in any perpendicular joint between a circular pipe and a radial transmission line; in this case, it is located in the vicinity of the acceleration gap at the beampipe wall. A field plot for this mode is shown in **Fig. 5(a)**. Since this mode is localized near the joint, and is evanescent into both the beampipe and the radial line, it follows that the placement of ferrite pucks in the ATA cell shown in **Fig. 4** will not damp this mode. Fortunately, because of the asymmetry in the axial field about

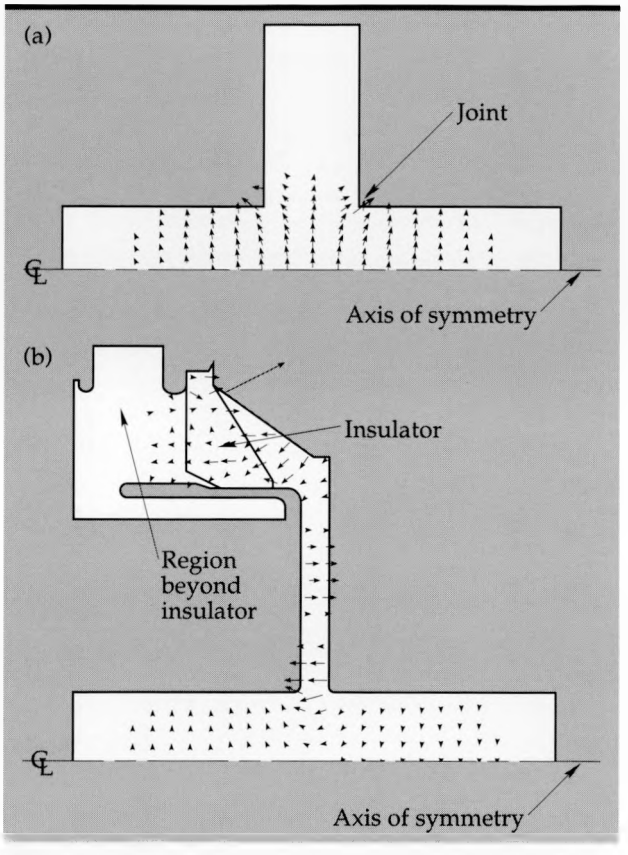


Figure. 5. Plots of two high-Q modes produced by AMOS in simulations of an ATA induction cell: (a) mode localized near the perpendicular joint between the beampipe wall and the acceleration gap, and (b) a mode evanescent in the oil behind the insulator. The plots show the direction and magnitude (length of vector tail) of the electric field in an azimuthal cross section of the cell. Note that the mode in (a) does not couple strongly to a transiting electron beam because the sign reversal of the axial electric field in the gap causes whatever energy that is put into the mode, while the beam is moving toward the gap, to be removed again as the beam passes and moves away from the gap. By the same argument the mode in (b) should be strongly excited by the beam.

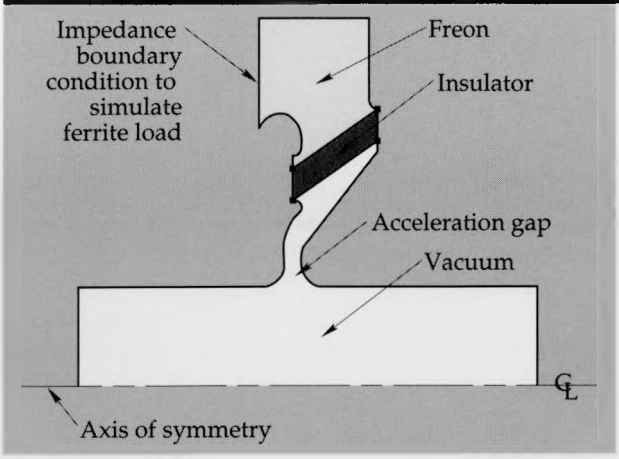


Figure 6. DRAGON rendering of azimuthal cross section of SNOMAD induction cell. Experimental measurements of the resonant mode frequencies were made for this cell, and these values were compared to AMOS simulations (see Table 1).

the center of the gap, this mode does not couple significantly to a transiting charged-particle beam, and it will only be weakly excited.

The second mode of interest, shown in **Fig. 5(b)**, is possibly unique to the ATA cell geometry. This mode is also not damped by the ferrite in the cell, as it is evanescent in the region beyond the insulator. In principle this mode should couple strongly to the beam because the axial field is of one sign (symmetric) in the gap; however, the experimental evidence for this mode is inconclusive. It may be that the proximity of the mode frequency to the beampipe cutoff frequency is allowing some of the mode energy to be carried out of the cavity in the form of traveling waves in the pipe so that the mode amplitude cannot reach an appreciable level.

AMOS has been used to suggest modifications to the ATA cell geometry and changes in the ferrite position that will reduce the effects of all modes that may interact with the beam.

SNOMAD Induction Cell

Another application that has yielded interesting results relates to the SNOMAD accelerator cell (**Fig. 6**). This cell is significantly smaller than the ATA cell. Since there was an isolated cell available for testing, we have been able to compare AMOS predictions of mode locations directly with experimental measurements. Modes up to $m = 4$ [modes that have a ϕ-dependence of the form $\exp(i4\phi)$] have been computed and compared with measurements, and these results are shown in **Table 1**. In this case, wherein the comparison is made with the insulator and oil removed and the ferrite is masked with a conductor, agreement with experiment is excellent. When the ferrite is unmasked, agreement with measurement is more qualitative. This is due to the use of an impedance boundary condition to simulate the effects of the ferrite. Results of this type are important benchmarks for AMOS—they have provided some of the earliest tests of the code against experimental measurements on a complex cell geometry.

DARHT Induction Cell

A third application concerns the induction cell of the DARHT accelerator being built at Los Alamos National Laboratory. The accelerator produces an electron beam that will be focused on a tantalum target to produce x rays for examination of extremely dense materials. AMOS was used to assist in the design of the prototype induction cell.

The preliminary design is shown in **Fig. 7(a)**. Several of the resonant modes exhibited unacceptably high coupling impedances with the electron beam, and AMOS was used to study the effects that geometric and material modifications would have on these resonances. The optimal design determined by AMOS is shown in **Fig. 7(b)**. This final design differs slightly from the optimal design suggested by AMOS because of pulse power and other considerations. For instance, sharp corners were rounded off to optimize the cell performance at high drive voltages.

The corresponding longitudinal and transverse coupling impedances for the preliminary and AMOS designs are compared in **Fig. 8**. The impedance of an induction cell is directly related to the strength of the interaction between the electron

Table 1. Comparison of experimental measurements and AMOS simulation results for the identification of resonant modes in SNOMAD induction cell.*

TM mode	Resonant frequency (GHz) Calculated	Measured
010	0.66	0.674
020	2.52	2.539
030	3.72	3.736
110	1.35	1.341
120	2.66	2.696
120	2.66	2.721
210	2.20	2.185
220	3.10	3.130
310	2.82	2.833
320	3.73	3.756
410	3.35	3.360

*For the data shown here, the ferrite was masked with a conductor, and the insulator and oil were removed.

beam and the cell. The impedance is a measure of the coupling between the electron beam and cell modes. The longitudinal impedance is a measure of the coupling of the axial beam motion to cavity modes, and the transverse impedance is a measure of the coupling of the transverse beam motion (that is, transverse to the beampipe axis) to cavity modes. In general, one would like the coupling impedances to be as small as possible to minimize the effects of cavity resonances on the beam.

During the process of optimizing the cavity design using AMOS, it was discovered that the coupling impedances, especially for the high-frequency modes, were affected strongly by both the positioning and the dielectric constant of the insulator, as well as the shape of the angular reflector at the top of the cell. Several parameters were varied over a range constrained by the fabrication requirements until the corresponding impedances were minimized. This cavity is currently being fabricated at Los Alamos and will undergo cold testing (no electron beam) in the near future. This will provide another test of AMOS against experimental measurements for a complex cell geometry.

Other Applications Studied

In addition to the applications detailed above, AMOS was also used to study various aspects of other accelerators of importance to Laboratory programs. These include the SLIA (Spiral Line Induction Accelerator), the ETA-II, the relativistic klystron, HGA (High Gradient Accelerator), the so-called "dielectric accelerator," and a proposed linac for heavy-ion fusion. The AMOS simulation code set may also be used in design work for the SSC (Superconducting Super Collider).

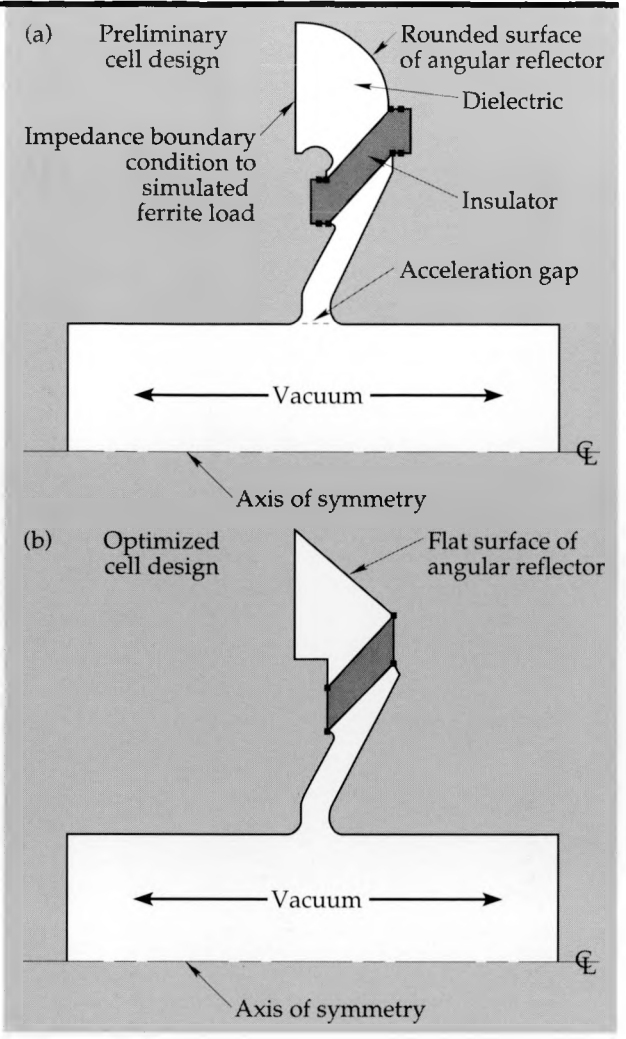


Figure 7. (a) Preliminary design of DARHT induction cell, and (b) DARHTdesign after optimization using AMOS. Design parameters, such as shape of the angular reflector and the shape and material properties of the insulator, were varied to optimize the rf characteristics of the DARHT cell for accelerator applications.

Future Work

Development efforts on AMOS during the coming year will continue the work begun this year. Work on the kernel of AMOS will primarily address implementation of a finite-element (FE) algorithm that can be used on irregularly shaped mesh elments. This will provide improved geometric modeling accuracy for those problems that require it. The faster finite-difference (FD) version of the code will remain for those problems that are insensitive to exact boundary representation. We are currently in the process of identifying an appropriate algorithm for the field updates on a mixed FD/FE grid; implementation of the algorithm as a module for AMOS will occur during 1990.

Work on ferrite modeling will continue in the upcoming year, with most of the effort focusing on the high-frequency model. Most of the high-frequency experiments were finished by the end of 1989, with a high-frequency ferrite module available at about the same time. The low-frequency ferrite switching model that has been installed in AMOS will be undergoing extensive testing against experimental data taken on small cores of ATA and ETA-II ferrite that were obtained during the previous year. Both the low- and high-frequency models will be in module form.

Various support software is being developed for AMOS, including output graphics and a more complete front end, or preprocessor. The completed user interface to AMOS will simplify the generation of the entire input file (not just the geometry section as is currently handled by DRAGON), and

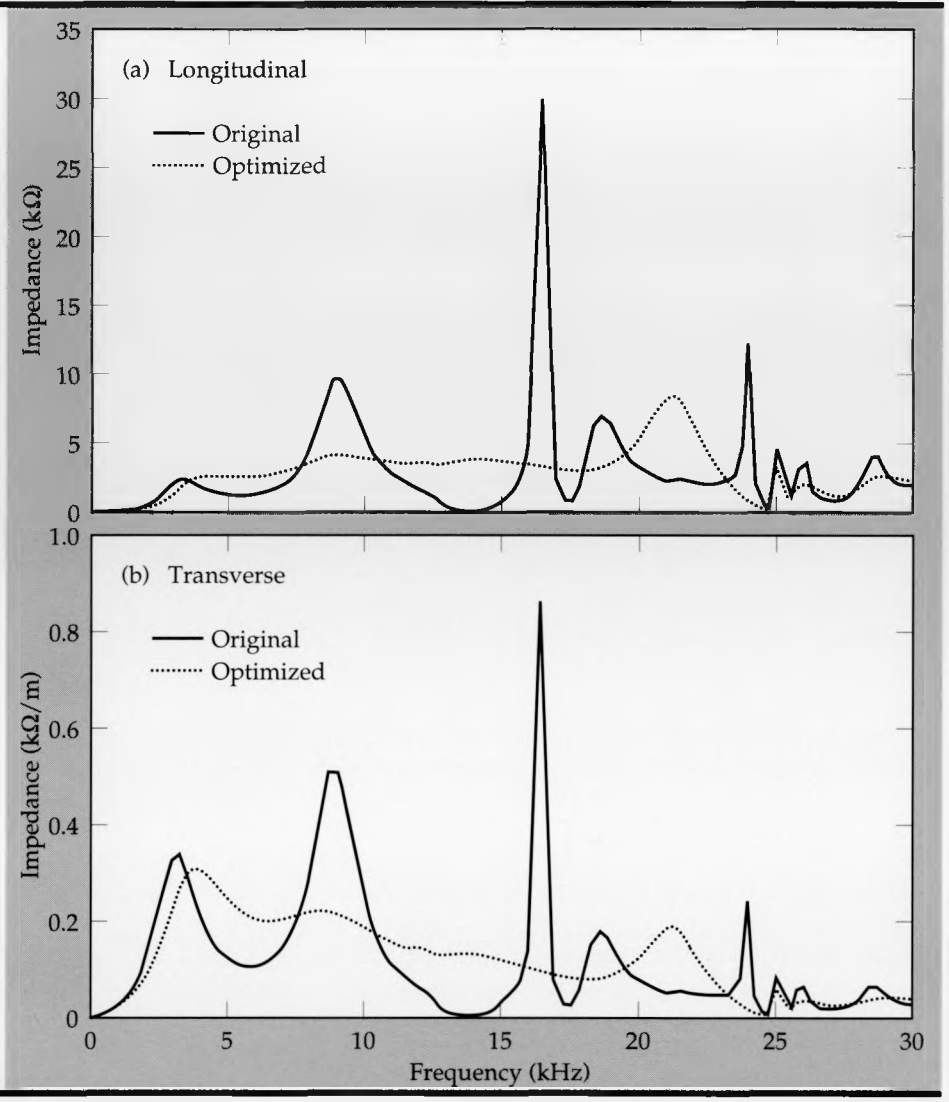


Figure. 8. (a) Longitudinal and (b) transverse coupling impedances are compared for the original and optimized DARHT induction cells. The impedance is a measure of the coupling between the electron beam and cell modes. It is desirable to have all components of the impedance as low as possible over a broad frequency range.

will also automate the concatenation of the necessary AMOS modules to construct an executable version of AMOS that is suitable for any given problem.

1. J. F. DeFord, G. D. Craig, and R. R. McLeod, "The AMOS (Azimuthal MOde Simulator) Code," *Proceeding of the 1989 IEEE Particle Accelerator Conference* (in press).
2. P. B. Wilson, *Introduction to Wakefields and Wakefield Potentials*, SLAC-PUB-4547, Stanford Linear Accelerator Center, Stanford, Calif. (1989).
3. K. L. F. Bane, *Wakefield Effects in a Linear Collider*, SLAC-PUB-4169, Stanford Linear Accelerator Center, Stanford, Calif. (1986).
4. R. K. Avery, J. F. DeFord, G. Kamin, W. C. Turner, and J. M. Zentler, "A Simple Domain Model of Flux Reversal in Ferrites," presented at the Conference on Magnetism and Magnetic Materials, Boston, Mass., November 27–30, 1989.
5. R. J. Luebbers, F. P. Hunsberger, and K. S. Kunz, "FDTD Formulation for Frequency Dependent Permittivity," *Proceedings of the 1989 IEEE AP-S International Symposium*, Vol. 1, pp. 50–53.

ENGINEERING RESEARCH AND DEVELOPMENT

Thrust Area Report FY 89

Computational Mechanics

The technical staff in the Computational Mechanics thrust area create and maintain software tools to solve complex design problems on supercomputers. In addition to our work on analysis codes, we also support the two- and three-dimensional mesh generators and graphic post processors, both on the super computers and on the workstations of various local area networks. Our suite of codes incorporates the governing mathematical equations of the general laws of mechanics and the particular phenomenology of a given materials problem. These codes can be used to solve heat transfer problems and to model physical phenomena in solid, structural, and fluid mechanics. The ability of the codes to accurately solve problems of large deformation and of inelastic material behavior makes them world-wide leaders in the domain of engineering software. Our codes are being acquired by industry as more organizations purchase powerful, high-speed computers.

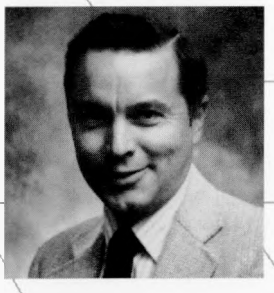
This year we sponsored two research projects:

- Discrete meshing of large-distortion Lagrangian problems
- Elastoplastic shell analysis in the DYNA3D code.

We also sponsored two development projects:

- The development of nonlinear solution algorithms for the TOPAZ heat transfer code
- The development of codes that solve fluid flow and heat transfer problems.

Gerald L. Goudreau
Thrust Area Leader



Using the NIKE2D Code to Model Severely Strained Ductile Metals

Bruce Engelmann
Nuclear Explosives Engineering Division
Mechanical Engineering

We have developed a new rezoning algorithm for inclusion in the NIKE2D finite-element code. Adding the algorithm to the code has allowed us to periodically remesh a problem, thereby increasing the accuracy of the results and preventing the premature termination of the simulation. We are currently applying the code to problems involving severe strain in ductile metals.

Introduction

In the past when we have used a Lagrangian finite-element method (see the article on computer modeling below) to analyze severe strain (>50%) in ductile metals, the model has become severely distorted, inducing numerical problems that can halt the analysis prematurely and degrade the accuracy of the results. The advent of more powerful computers, like the Cray XMP and Sun workstations, and more sophisticated codes, like LLNL's nonlinear finite-element code NIKE2D,[1] has increased our ability to analyze severe strain in ductile metals. NIKE2D can now simulate the changes that take place in the shape of a complex metal

Computer Modeling: Some Terms and Basic Concepts

In seeking to understand the physical phenomena in complex mechanical or electronic systems, engineers have recourse to empirical testing and modeling (simulating) the phenomena on a computer. Although each method excels (and, indeed, may be indispensable) in a given situation, the unique advantages of computer modeling increasingly make it the method of choice.

Computer simulation is still young, and, like other young disciplines, it has given rise to new terms and concepts or molded familiar ones to its own peculiar uses. When a design engineer first considers how to model an object or process (referred to as the "problem"), he or she begins with geometric modeling: the problem is thought of as a geometrical object that will change its shape with time or some other physics parameter(s). The engineer decides, on the basis of various requirements and constraints, whether to model the problem in two or three dimensions. A two-dimensional model represents a surface (which may be flat or curved) or a slice through the object. A three-dimensional model represents the volume of the object. Often a problem has some axis of symmetry, so only a part of it need be modeled. The elaborate computer programs or codes used to represent these problems are written as either two- or three-dimensional codes.

For either type of problem, the engineer begins by dividing (discretizing) the object into many small cells (elements or zones) that are the basic units on which the computer operates. Special codes have been written to generate the lines (mesh) that define the elements. The lines and intersection points (nodes) are important because they are associated with and behave according to instructions and parametric values specified by the engineer.

There are two main categories of mesh, according to movement: the rigid Eulerian mesh can model turbulent flow, but it is inappropriate for curved shapes; the flexible Lagrangian mesh conforms to any shape, but it can become hopelessly distorted when modeling extreme turbulence.

Meshes can also differ according to the modeling approach used. A finite-difference mesh usually, although not invariably, defines regular (orthogonal) cubes; the smaller these cubes are, the more precisely they can approximate an irregular geometry, but the longer the calculation will take. A finite-element mesh can have irregularly shaped elements, which means that they can conform to the shape of any object.

These two meshes derive from two different modeling techniques. Finite-difference meshes

characteristically result when the terms of the governing differential equations are approximated by their finite-difference formulas. The finite-element technique is very similar, but it functions by more global approximations; by this technique, discrete equations couple only adjacent nodes, but the equations are derived from global balance laws or variational principles.

Codes can also be described as being either implicit or explicit. Both finite-element and finite-difference methods transform the spatial dimension of the problem from differential to algebraic form, thus discretizing it. Time, however, is still expressed in the differential equations. These equations in the time domain can be integrated by explicit methods in which the unknown mesh-point variables are directly given as functions of known, previously existing data, or by implicit methods in which each time step requires the solving of simultaneous equations.

Another distinction modelers make is between linear and nonlinear solutions, which means that the code solves linear or nonlinear equations. Most engineering disciplines allow linearization of nonlinear phenomena, which permits substantially less numerical computation but restricts the validity of the solution to a narrow range of variables. The challenge for today's design engineers is to exploit the added power of supercomputers and new numerical methods to solve nonlinear equations and thereby increase the model's fidelity to the underlying phenomena.

structure as well as changes in its material properties (such as elastoplasticity) when a structure is under stress. Moreover, the code can model such changes within one specific part of a complex structure, an important attribute because the ultimate strength of a structure depends on the strength of its individual parts. Thus NIKE2D is suited to predict the ultimate strength of a complex structure, particularly when component failures tend to precede the failure of the entire structure.

Note that NIKE is suited for static or low-rate dynamic problems, such as analyses of bridges, earthquakes, or metal-forming processes; it is *not* meant to characterize results that are measured in microseconds. As an example of the kind of behavior NIKE *does* simulate, we recently used the code to model metal forming processes such as extrusion, hydroforming, rolling and drawing - all methods used to alter the size and shape of metal or to enhance its material properties. Again, simulating the behavior induced by these methods requires the ability to model severe distortion, as a metal workpiece may completely change shape during the process.

Lagrangian Formulation

The NIKE2D code uses a Lagrangian finite-element description of the equations that govern solid mechanics. In a Lagrangian description individual elements represent the same material points throughout the step-by-step analysis. The problem is that when a structure severely distorts, the finite elements also distort, resulting in a loss of accuracy. This distortion induces numerical problems that may eventually cause premature termination of the analysis. One way to reduce the amount of distortion is to smooth the mesh (that is, take the finite elements that have become streched or curved and make them more uniform and boxlike, as in the original mesh). The same logical grid is used before and after smoothing to reduce changes to the code's architecture and data base. Although smoothing increases the code's ability to handle severe strain problems, many analyses still encounter such excessive mesh distortion before the simulation is complete that smoothing is not enough. Thus, another way has been developed to reduce severe distortion: we periodically remesh the problem at appropriate times throughout the step-by-step analysis. This process is called rezoning and it involves freezing a numerical calculation at a point in time and changing the finite-element mesh to reflect the changes in geometry. The difference between smoothing and rezoning is that in the latter case we use an entirely new mesh, allowing us to apply the code to problems of severe distortion that simply do not yield to smoothing techniques.

Our objectives in FY 89 were to study various algorithmic aspects of rezoning, develop a prototype rezoner to be used in conjunction with the code and to evaluate the code's ability to analyze problems involving ductile metals. Our ultimate goal is to increase the Laboratory's ability to solve engineering problems complicated by severe distortion.

Progress

A rezoning algorithm has been added to the NIKE2D code, significantly expanding the code's ability to accurately simulate severe strain in ductile metals. Henceforth, we will call this algorithm

a "newzoner" because an entirely new mesh may be used at any point in the analysis. The newzoner is robust, accurate, and easy to use. Moreover, we have devised ways for users to store and manipulate the data accumulated while the problem is running, as well as a way to interrupt the problem at a given point to remesh. In the following sections we discuss some algorithmic highlights and illustrate the performance of the prototype by applying it to the metal forming process of forward extrusion.

Newzoner

In a finite-element analysis, we transform the differential equations that govern solid mechanics problems into algebraic equations that can then be solved on our computer systems. The method of transformation involves discretizing the problem in space (dividing a complex object into a number of finite elements) and time (dividing the analysis into time steps). Throughout the analysis kinematic variables such as displacement are tracked at the nodes and state variables such as stress are calculated at points inside the finite element. At the begining of the analysis as well as at the end of each time step, each of the kinematic and state variables are known, so the data necessary to rezone is available at each step. Typically, a quasi-static analysis may involve on the order of a hundred time steps, and depending on the amount of distortion, may require several rezoning phases. Once it is determined that a model should be remeshed, a rezoning phase would consist of constructing a new mesh, mapping the state variables from the old mesh to the new, and restarting the analysis at that time step. Note that rezoning is not a smoothing operation in which a variable calculated inside a finite element remains within that element. Rather, that variable may have migrated to a finite element in the new (rezoned) mesh that is at some distance from the finite element in the old mesh. Consequently for NIKE2D we had to make provision for tracking the data through the various rezoning phases.

When to Rezone

Like many aspects of analysis, there is a trade-off between interactivity (a user can stop the analysis because of severe distortion and remesh) and automation. We believe there is no substitute for an analyst's judgement, so it is necessary to have an interactive option available. As experience is gained, steps can be taken to automate the rezoning process. Rezoning should be performed whenever elements become severely distorted or when aspect ratios become high (an aspect ratio is a comparison of the characteristic length and width of an element). Severe element distortion will result in a degradation of accuracy and may introduce numerical problems that halt the analysis. The analyst can decide when to rezone either by monitoring the analysis in progress or by letting the analysis run its course and then examining the results with the aid of a post processor. If the examination shows that the finite-element mesh has become distorted at a given time step, the analysis can be backed up to that point, the mesh rezoned, and the simulation restarted.

During a newzone phase, a new finite-element mesh is constructed that minimizes element distortion. The mathematical model does not change and, therefore, the governing equations and boundary conditions remain in force; the only thing that is modified is the mesh. As the first step in the remeshing, the analyst defines four-sided regions that may be subdivided into quadrilateral elements. Although arbitrary regions may be harder to discretize by quadrilaterals than by triangles, NIKE2D uses four-node quadrilaterals exclusively because of their superior performance in modeling metal plasticity. Before the analysis can be restarted stress variables and variables describing material history must be mapped onto the new mesh. The process **(Fig. 1)** involves tracking the variables at numerical integration points (called Gauss points) located within a finite-element quadrilateral. To begin the remapping, Gauss point values are mapped onto the nodes of the old mesh using a least squares algorithm. Next the values are mapped to the new mesh nodes and then to the new element interiors using bilinear interpolation functions consistent with the finite-element approximation. After the remapping phase, the analysis is continued using the new mesh topology and associated stress and history variables.

One of the problems of rezoning is that when a new mesh is introduced, the number of nodes and elements changes. This poses a problem because the stress and history variables derived during an analysis are stored according to their physical position on the old mesh. That is, a given node or element interior will have values associated with it. The problem then becomes one of mapping the values from the old mesh to the new one. Our solution is to write all the old mesh values to a hard disk, thereby cleaning out the storage area, introducing the new mesh into storage, and then mapping the old values to the new mesh.

Once the simulation is completed a post processor interprets the results graphically. Again, rezon-

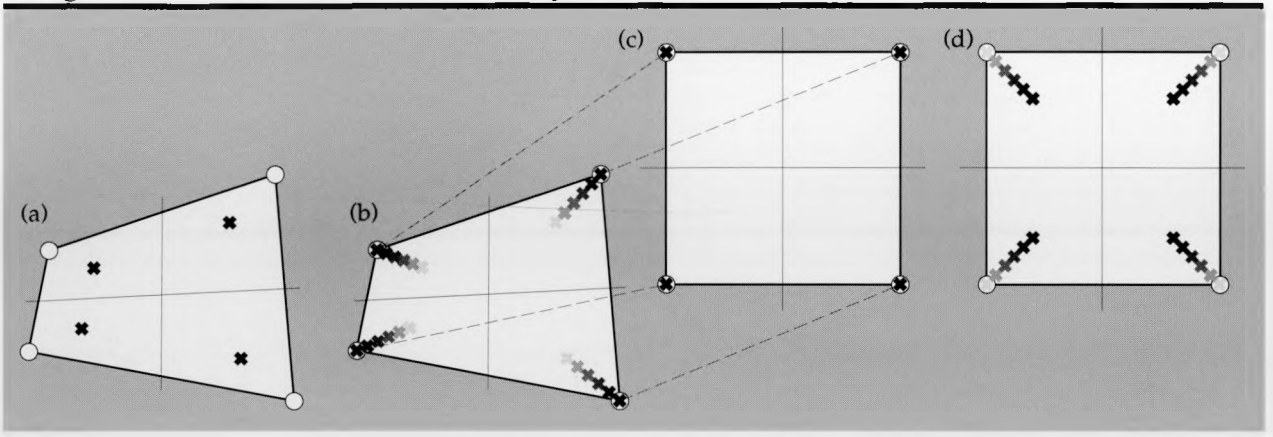


Figure 1. Mapping variables from the old to the new mesh. (a) Gauss points within a distorted finite-element quadrilateral. (b) Gauss points mapped onto the nodes of the old finite element. (c) Values at the old nodes mapped to the nodes of a new finite element. (d) Points mapped to the interior of the new finite element.

ing may introduce a problem when analyzing the results if we want to track the position or properties of a particular material point over time. The problem is that we have changed the mesh each time we rezoned and therefore the point we are trying to track may have migrated from one element to another. This means that we can't use standard finite element techniques to track material points. What we must do is track such points through the various rezoning phases. After analyzing the results of such a tracking procedure, the analyst may determine that the problem required rezoning at a previous time step. The mesh may then be rezoned and the analysis restarted.

Application to Forward Extrusion Analysis

To evaluate the performance of the rezoning algorithm, we have used the newzoner to analyze a forward extrusion process. **Figure 2** shows the axisymmetric finite-element discretization of a billet and die. We used an elastoplastic model to simulate the material behavior of the copper billet; the die is assumed to be elastic. This is a difficult problem for a Lagrangian code since the area reduction of the billet is 96% and strains will become extremely large as the billet is pressed through the die. Additionally, modeling the contact between the billet and die is complicated by a "dead zone" of material and a corner singularity.

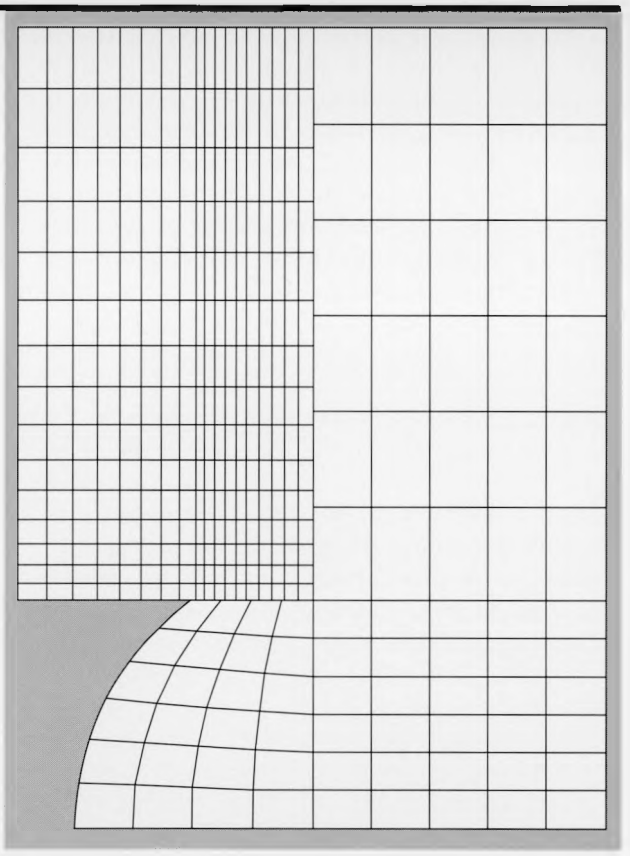


Figure 2. Initial finite-element discretization of a billet and die. An elastoplastic model simulates the material behavior of the copper billet; the die is assumed to be elastic.

Figures 3 and 4 depict the extrusion process, which was simulated for a total of 1.4 sec (including two intermediate rezoning phases). **Figure 3a** shows the deformed finite-element mesh after 0.6 sec, illustrating the dead zone that has developed in the corner. The most severe strains are just outside the dead zone. The mesh is rezoned in **Fig. 3b**, using more elements than the previous mesh to reduce element distortion. Strains reach several hundred percent of the original by 1 sec **(Fig. 4a)**, and the problem is again rezoned **(Fig. 4b)**. The deformed mesh of **Fig. 5** shows the slight expansion of the billet as it exits the die opening. If desired, the simulation can be continued by successive analysis/rezone phases.

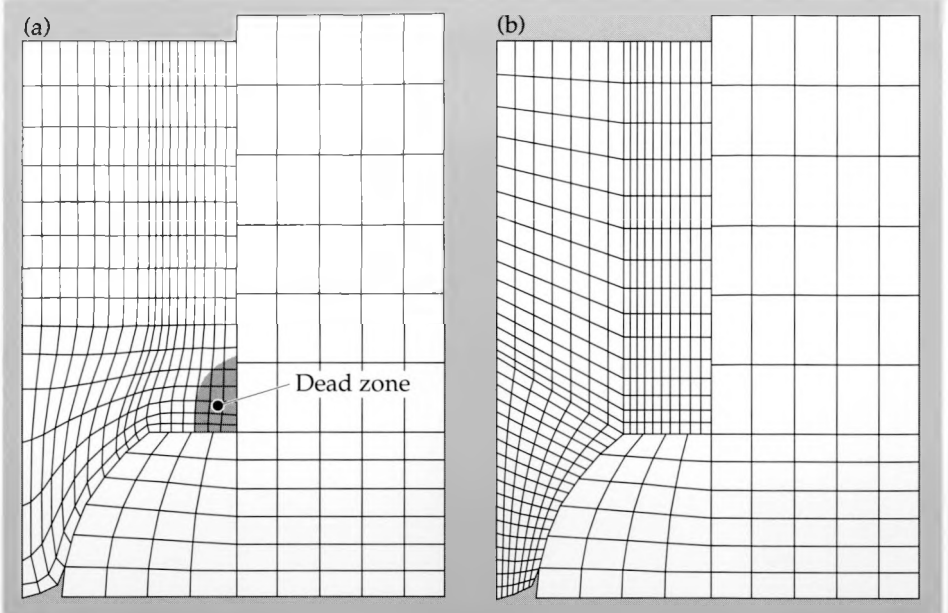


Figure 3. Beginning of the simulation of the extrusion process. (a) Mesh distortion at 0.6 sec; note that a dead zone has developed in the corner. (b) Rezoned mesh at 0.6 sec.

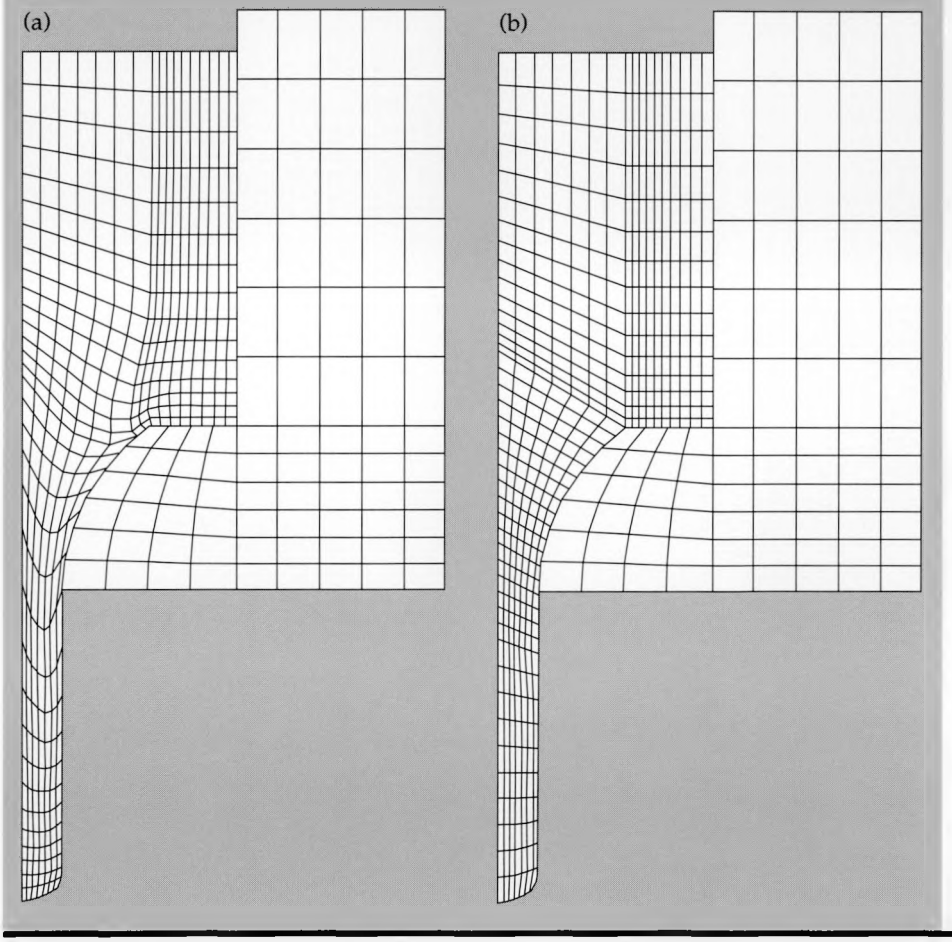


Figure 4. Continued simulation of the extrusion process. (a) Mesh distortion at 1 sec; at this time, strains have reached several hundred percent of the original. (b) Mesh rezoned at 1 sec.

Summary

The analysis described in the previous paragraph illustrates the increased ability of LLNL's NIKE2D code to simulate severe strain in ductile metals. The key to the improvement lies in our ability to rezone whenever the finite-element mesh becomes too distorted. Thus the newzoner option allows the engineering analyst to go much further in the analysis of a problem than would be the case if smoothing techniques were used. This, in turn, expands the number of problems that the NIKE2D code can solve successfully.

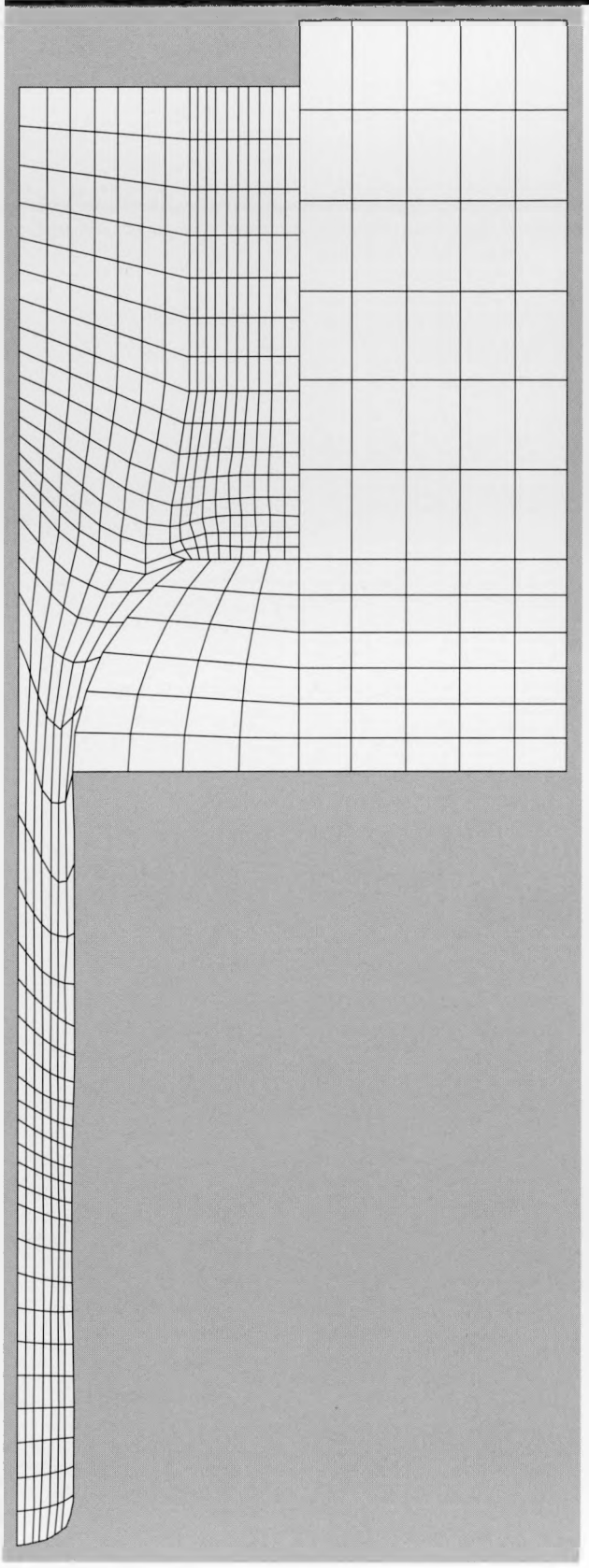


Figure 5. Final mesh configuration at 1.4 sec.

Acknowledgement

The author gratefully acknowledges the help of Dr. Peter Raboin during the course of this research.

1. J. O. Hallquist, "NIKE2D - A Vectorized Implicit, Finite-Deformation Finite Element Code for Analyzing the Static and Dynamic Response of 2-D Solids with Interactive Rezoning and Graphics," Lawrence Livermore National Laboratory, Livermore, CA, Report UCID-19677 (December, 1986).

Code Acquisition and Development Project of the Thermo-Fluid Mechanics Group

Werner Stein
Nuclear Test Engineering Division
Mechanical Engineering

The Thermo-Fluid Mechanics Group identifies, acquires, and develops computer codes that solve fluid flow and heat transfer problems. Current work includes:

•Developing a time-dependent code that uses finite-volume methods to model incompressible fluid flow and heat transfer equations in three dimensions. The code can already solve laminar flow problems and is now being expanded to solve turbulent flow problems.

•Modifying KIVA-II, a time-dependent code that uses finite-volume methods to model compressible flow and heat transfer equations in three dimensions. The modifications will expand the code's ability to handle boundary conditions, thereby enabling the code to solve a wider range of problems.

Introduction

The Thermo-Fluid Mechanics Group supports Laboratory programs that need sophisticated computer codes to solve their fluid mechanics and heat transfer problems. In many cases the applicable codes either need to be modified or are not available at the Laboratory. It then becomes the Group's task to modify Laboratory codes or to identify off-site codes that will meet programmatic needs. For example in FY 89 we helped the LIS Program with problems involving thermal radiation heat transfer and thermal stress. Since the Group's inception, we have provided the Laboratory with a number of fluid mechanics and heat transfer codes, including:

- A one-dimensional code that calculates the flow of compressible fluids through a network of pipes and the resulting heat transfer[1,2]
- A two-dimensional code that calculates the effects of heat transfer on curved and straight boundaries[3]
- A three-dimensional code that solves fluid flow and heat transfer problems in which the flow is both incompressible and laminar[4,5]
- A two-dimensional hydrodynamics code that calculates strong shock effects resulting from explosions[6]
- A Monte Carlo code that calculates thermal radiation exchange factors.[7,8] These exchange factors are used as input to the Laboratory's heat transfer codes.

In FY 89 we focused on three new code development projects. We continued to fund the development of a three-dimensional code that uses a finite-volume method to calculate laminar flow, turbulent flow, and heat transfer problems. Although the code is being developed at the University of California at Davis, it is being written with the Laboratory's needs in mind: the code accepts a computational mesh generated by a Laboratory mesh generation code, and has post-processing capabilities compatible with Laboratory graphics libraries. This finite-difference code is written in an easily modified modular format because the Group is frequently presented with problems of such a specific nature that the code must be modified before it can be successfully applied.

We are also working on KIVA-II,[9] a three-dimensional, time-dependent code that solves problems in fluid flow (both laminar and turbulent) and heat transfer using the finite-volume method. The code also models the chemical reactions that take place when injection sprays are added (injected) into a flow. Currently we are modifying the code to increase the number of options available to the user. Specifically, KIVA-II is being modified so that it can accept meshes from a mesh generator. By the same token, we are expanding the variety of boundary conditions that a user can specify, enabling the code to handle more diverse problems.

Also, we helped the Laser Program acquire and apply FIDAP, a finite-element code that solves incompressible fluid flow and heat transfer problems in three dimensions.[10,11]

Progress

University of California at Davis Incompressible Flow Code

We are solving a variety of fluid flow problems using a computer code that models fluid flow and heat transfer with the three-dimensional Navier-Stokes equations and an energy equation. Our philosophical approach has been to write systems of equations that have a direct relationship to the physics of the systems being modeled. Thus, it is clear in our code which equations apply to the physics of fluids and which to boundary conditions. The equations are generated using finite-volume principles, a practice that allows us to model irregular and curved boundaries. A predictor-corrector method is used to solve the resulting simultaneous equations. The code has been written in a modular format so that we can readily modify a particular subroutine.

The code has been written to solve incompressible flow problems (low Mach numbers) of interest to the R Program and the LIS Program. Incompressible flow problems differ from compressible flow problems because the equations of state cannot be readily applied. Consequently, when we are calculating a pressure field (across a cell boundary) we have recourse to a pressure correction technique based on a mass continuity equation. The technique uses two velocity variables, which allows for the continuity equation to be converged to machine accuracy on the cell boundaries. Currently we have solved the following types of problems using this code:

- External flow over three-dimensional objects in nonuniform flows
- Free convective flows over three-dimensional objects
- Unsteady flow in curved pipes of changing cross section
- Three-dimensional rotating flows in closed containers
- External flow over objects with surface mass transfer
- External gas flow over a deforming liquid droplet. The conditions that are being modeled are the surface tension of the droplet and the heat and mass transfer between the droplet and the gas.

All the above flows have been calculated assuming the flow to be laminar; however, we are now modifying the code so that it can also solve turbulent flow problems. **Figures 1 and 2** show typical code output.

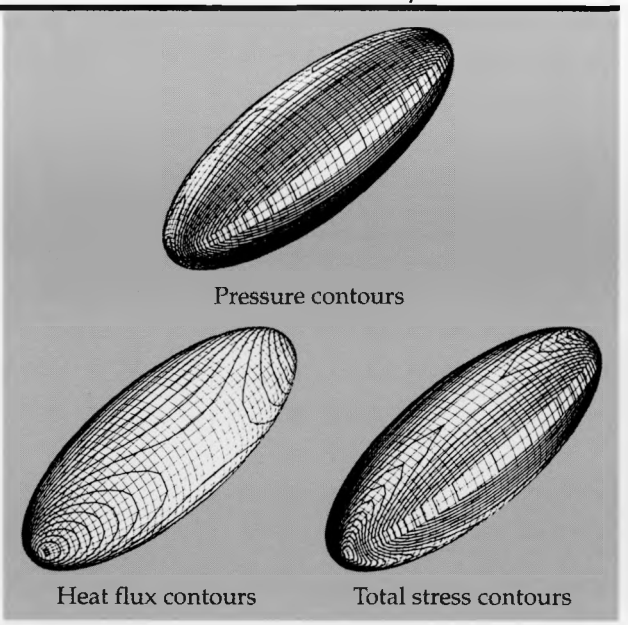


Figure 1. Contours of pressure, heat flux, and total surface stress for laminar flow (Reynolds number 66) over an ellipse. The results are in good agreement with on-going experiments.

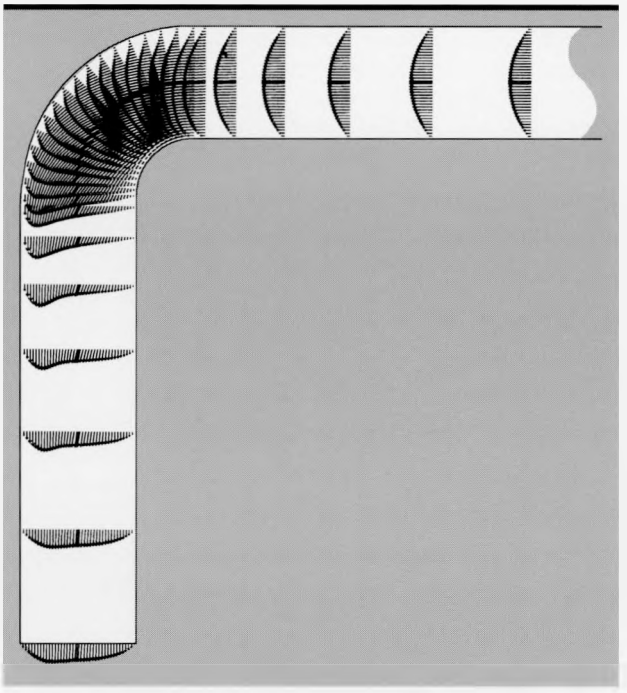


Figure 2. Velocity vectors in a curved pipe for laminar flow (Reynolds number 200). The problem is a complex one in that the model must solve a curved boundary problem in three dimensions.

KIVA-II Compressible Flow Code

The KIVA-II code calculates fluid flow and heat transfer occurring in an automobile engine cylinder. The injection of fuel, spark ignition, and chemical reactions in the cylinder are also mod-

eled. Presently the code accepts two- and three-dimensional cylindrical geometries and a three-dimensional rectangular mesh. We are now modifying the code so that it will accept a more general three-dimensional mesh to be generated by a Laboratory mesh generation code. The code presently allows the user to specify a constant velocity inflow condition over one face of the mesh. Again, we are modifying the code so that the user can specify an inflow velocity condition over an arbitrary region on the surface of the mesh. Similarly, we are modifying the code to allow a variety of outflow boundary conditions. All these modifications will enable the code to handle more diverse problems.

FIDAP Finite-Element Code

The Laser Program has recently purchased the FIDAP code, which calculates fluid flow and heat transfer problems in three dimensions. The code operates on the LLNL Cray and Sun workstations and has been aplied to a variety of fluid flow and heat transfer problems. Thermo-Fluid Group members have been directly involved in applying and modifying the code.

Future Plans

In FY 90 we will complete modeling turbulence and modifying the input and output routines for the U. C. Davis code. We will also complete modifying the mesh and boundary conditions for the KIVA-II code. We will begin research for two new codes that Laboratory programs have expressed an interest in: a three-dimensional hydrodynamics code and a two-dimensional fluid/structure interaction code.

1. W. S. Winters, *TOPAZ—A Computer Code for Modeling Heat Transfer and Fluid Flow in Arbitrary Networks of Pipes, Flow Branches, and Vessels*, Sandia National Laboratories, Albuquerque, NM. SAND83-8215 (July 1985).
2. R. W. Martin, *TOPP: A Post-Processor for TOPAZ, the One Dimensional Pipe Flow Analysis Code*, Lawrence Livermore National Laboratory, Livermore, California, UCID-21251 (July 1987).
3. M. E. Crawford and K. A. Thole, *TEXSTAN—A Two-Dimensional Boundary Layer Code*, Lawrence Livermore National Laboratory, Livermore, California (unpublished personal communication).
4. S. T. Chan, *FEM3—A Finite Element Model for the Simulation of Gas Transport and Dispersion: User's Manual*, Lawrence Livermore National Laboratory, Livermore, California, UCRL-21043 (April 1988).
5. R. C. McCallen, *An Investigation of the Finite Element Incompressible Flow Code FEM3*, Lawrence Livermore National Laboratory, Livermore, California, UCID–21527 (September 1988).
6. R. Tipton, *CALE Users Manual, Version 890601*, Lawrence Livermore National Laboratory, Livermore, California (June 1989).
7. P. Burns and J. Maltby, *MONTE User's Manual for "MONTE2D" and "MONTE3D,"* Colorado State University, Fort Collins, Colorado (September 1988).
8. T. A. Reitter, *An Investigation of the MONTE Radiative Exchange Factor Code*, Lawrence Livermore National Laboratory, Livermore, California, ENN 88-14 (August 1988).
9. A. A. Amsden, P. J. O'Rourke, and T. D. Butler, *KIVA-II A Computer Program for Chemically Reactive FLows with Sprays*, Los Alamos National Laboratory, Los Alamos, New Mexico, LA-11560-MS (1989).
10. Fluid Dynamics International, Inc., *FIDAP General Introduction Manual, Revision 4.0*, Fluid Dynamics International, Inc., Evanston, Illinois (September 1987).
11. R. Martin, Lawrence Livermore National Laboratory, Livermore, California, private communication, *FIDAP Availability* (March 1989).

New Nonlinear Solution Algorithm for the TOPAZ2D Heat Transfer Code

Arthur B. Shapiro
Nuclear Explosive Engineering Division
Mechanical Engineering

The computer code TOPAZ2D currently uses a functional iteration method to solve nonlinear equations resulting from the use of finite element methods in nonlinear heat transfer problems. The functional iteration method works and gives acceptable results, but is very slow in solving recent problems with highly nonlinear material properties (e.g., phase change) and difficult boundary conditions (e.g., radiation, boiling). We have found that Newton's method is more robust for solving this new class of problems. Fewer iterations are required per time step, and larger time steps can be taken during transient analysis.

Introduction

As TOPAZ2D[1] is applied to larger and more complex heat transfer problems, it becomes increasingly important that the solution process remain economical as well as accurate. The computer solution time for an analysis depends to a great extent on the numerical algorithms used for the solution (by iteration) of the systems of nonlinear algebraic equations resulting from the application of the finite element method.

The functional iteration method[2] is the most popular and the easiest algorithm to implement in nonlinear heat transfer computer codes. For the iteration process, we select an initial temperature vector $\{\theta_0\}$, and generate the sequence of vectors $\{\theta_i\}$ by

$$\{\theta_i\} = [H(\theta_{i-1})]^{-1}\{\mathbf{f}(\theta_{i-1})\} \text{ for each } i = 1, 2, \ldots, n$$

until $\| \theta_i - \theta_{i-1} \| < \varepsilon$.

The $[H]$ matrix contains terms for the material properties and the vector $\{\mathbf{f}\}$ contains terms for thermal generation and boundary conditions. An advantage of this method is that the matrix $[H]$ is banded, positive definite, and symmetric. A disadvantage of this method is that its rate of convergence is linear. Several convergence acceleration algorithms (e.g., Aitken's Δ^2 process, Steffensen's iteration) have been applied, but do not attain the quadratic convergence rate of Newton's method.[2]

For the Newton iteration process, we select an initial temperature vector $\{\theta_0\}$ and generate the sequence of vectors $\{\theta_i\}$ by

$$\{\theta_i\} = \{\theta_{i-1}\} - [H_T(\theta_{i-1})]^{-1}\{\mathbf{f}(\theta_{i-1})\} \text{ for each } i = 1, 2, \ldots, n$$

until $\| \theta_i - \theta_{i-1} \| < \varepsilon$.

Newton's method has the advantage of quadratic convergence. A disadvantage is having to calculate the tangent matrix $[H_T]$, which involves temperature derivatives of the variables that are a function of temperature.

Progress

During FY 89 we implemented Newton's method within TOPAZ2D for nonlinear material properties and convection boundary conditions. The tangent matrix is

$$[H_T] = [H] + [A] + [B] + [C]$$

where

$$H_{i,j} = \int_\Omega \nabla N_i^T \mathbf{k} \nabla N_j d\Omega + \int_\Omega N_i \rho c N_j d\Omega \ ,$$

$$A_{i,j} = \int_\Omega \nabla N_i^T (\nabla \mathbf{N}\Theta) N_j k' d\Omega \ ,$$

$$B_{i,j} = \int_\Omega N_i (\mathbf{N}\Theta) N_j \rho c' d\Omega \ ,$$

$C_{i,j}$ = boundary condition contributions.

The N terms are the finite element shape functions, ρ is the density, and Ω is the volume of the finite element. The $[H]$ matrix is the standard thermal stiffness matrix formulation for a linear heat transfer problem. The $[A]$ and $[B]$ matrices are

contributions to the stiffness matrix accounting for nonlinear thermal conductivity and heat capacity, respectively. The [C] matrix accounts for nonlinear convection boundary conditions.

The [A] matrix is nonsymmetric. This generates a full nonsymmetric tangent matrix [H_T] and thereby increases computer memory requirements and solution time per iteration. We would like to omit the [A] matrix from the formulation, noting that its presence can accelerate convergence but is not required to obtain a correct answer. To verify this approach, we formulated a mathematical definition for a test problem used to investigate convergence rate. The problem represents steady-state, one-dimensional heat transfer in a stainless steel (type 304) slab with one surface maintained at 298 K and the other surface exchanging heat by turbulent natural convection with an environment at 500 K. This nonlinear heat transfer problem may be written as

$$\frac{d}{dx}\left(k\,\frac{d\theta}{dx}\right)=0\ ,$$

where $0 \le x \le 0.1$ and $\theta = 298$ K at $x = 0$. Additionally,

$$-k\,\frac{d\theta}{dx} = h\,(\theta_s - \theta_\infty) \text{ at } x = 0.1\ ,$$

$$k = 8.116 + 0.01618^{*}\,\theta\left[\frac{W}{mK}\right]\ ,$$

and

$$h = 1.31\,(\theta_\infty - \theta_s)\,\frac{1}{3}\left[\frac{W}{m^2K}\right]\ ,$$

with

$$\theta_\infty = 500\ K\ .$$

As shown in **Table 1**, the rate of convergence is reduced from quadratic to super-linear by not including the [A] matrix in the formulation of the tangent matrix [H_T]. The functional iteration method is seen to converge linearly. For most real materials the thermal conductivity is a very weak function of temperature (i.e., $dk/dq << 1$). The magnitudes of the terms in the [A] matrix are much smaller than the terms in the [H], [B], and [C] matrices. Therefore, omitting the [A] matrix from the formulation in the presence of other highly nonlinear conditions has little effect on the overall convergence rate.

Table 1. The rate of convergence is reduced from quadratic to super-linear by not including the nonlinear thermal conductivity contribution matrix [A] in the tangential matrix [H_T]. The functional iteration method is seen to converge linearly.

Iteration	Newton's method with [A]	Newton's method without [A]	Functional iteration
1	3.1 e+01	3.1 e+01	3.1 e+01
2	1.4 e–01	1.4 e–01	7.3 e–01
3	4.8 e–06	4.5 e–04	1.7 e–02
4	1.8 e–11	1.0 e–06	4.2 e–04
5		1.5 e–09	1.0 e–05
6			2.5 e–07
7			6.1 e–09

Future Work

In FY 90, we will be implementing Newton's method for other TOPAZ2D features, including slide line contact surfaces, material phase changes, radiation boundary conditions, enclosure radiation, bulk nodes, special interior elements, and chemical kinetics. As each feature is added, the convergence rate of the code will be monitored for quadratic convergence and the accuracy will be validated against analytical answers. After Newton's method has been implemented for all TOPAZ2D features, quasi-Newton methods (e.g., Broydens method, BFGS) will be reviewed for inclusion into TOPAZ2D.

1. A. B. Shapiro, *TOPAZ2D—A Two-dimensional Finite Element Code for Heat Transfer Analysis, Electrostatics, and Magnetostatics Problems*, Lawrence Livermore National Laboratory, Livermore, CA, UCID-20824 (1986).
2. R. L. Burden and J. D. Faires, *Numerical Analysis* (Prindle, Weber & Schmidt, Boston) 1985.

Elastoplastic Shell Analysis in DYNA3D

Robert G. Whirley
Nuclear Explosives Engineering Division
Mechanical Engineering

The objective of this investigation was to examine and improve upon the elastoplastic shell modeling capability in the Laboratory's finite element code, DYNA3D. The finite element analysis of thin shell structures under transient, dynamic loads is an important capability to many Laboratory programs. Essential features of this capability are speed, accuracy, and robustness for both the shell element formulation and the constitutive model.

This article summarizes (1) the development of a new four-node, quadrilateral, finite element shell formulation, and (2) the evaluation of two basic methods employed in elastoplastic constitutive algorithms: stress resultant and thickness integration. The results of this research have been implemented in DYNA3D and offer a significant improvement in our elastoplastic shell modeling capability.

Introduction

The computer simulation of the elastoplastic behavior of thin shell structures under transient, dynamic loads plays an important role in many LLNL programs. Often the loads are severe, and the structure undergoes plastic (or permanent) deformation. These simulations are effectively performed using DYNA3D,[1] which is an explicit, nonlinear, finite element code developed at LLNL for simulating and analyzing the large-deformation dynamic response of solids and structures. It is generally applicable to problems where the loading and response are of short duration and contain significant high-frequency components. Typical problems of this type include the contact of two impacting bodies and the resulting elastoplastic structural behavior.

Several of the Laboratory's programs rely heavily on our ability to accurately model the large-deformation, elastoplastic behavior of thin shell structures. Three of the many programs that routinely use DYNA3D for this purpose include Nuclear Systems Safety, Nuclear Design, and Enhanced Safety:

The Nuclear Systems Safety Program is concerned with transportation safety of radioactive waste. There, DYNA3D is used to analyze waste-canister damage under projectile impact and drop loads.

In the Nuclear Design Program, DYNA3D is used to analyze shipping containers under simulated drop conditions to verify structural integrity. To model the shipping containers, the simulations rely heavily on shell element formulations capable of representing elastoplastic behavior.

In the Enhanced Safety Program, shell elements with elastoplastic behavior are used in DYNA3D to simulate an aircraft crashing into a stationary U.S. Air Force bomber, the B1-B. The objective of the simulation is to determine the loads that such an event would impart to the weapons carried by the B1-B. **Figure 1(a)** is a graphic depiction (cutaway view) of the finite element model used for this analysis. To simulate the impact, the model is given an initial velocity toward the crash barrier on the right in the figure. **Figure 1(b)** shows the post-impact geometry.

The objective of this investigation was to examine and improve upon the elastoplastic shell modeling capability in DYNA3D. This included a basic review of both the shell element formulations and the computational methods employed in the elastoplastic constitutive algorithms. Special attention was to be paid to accuracy/speed tradeoffs between the various algorithms. The outcome of these studies provides guidance for the analyst in choosing the most cost effective set of approximations to solve a particular problem to a given level of accuracy.

This article summarizes:

- The development of a new four-node, quadrilateral, finite element shell formulation, designated the "YASE shell."
- The comparative evaluation of two basic methods employed in elastoplastic constitutive algorithms for shell structure modeling: the *stress-resultant method* and the *thickness-integration method.*

In the thickness-integration method, plane-stress constitutive algorithms are applied at a number of integration points through the thickness of the shell. We evaluated seven algorithms, each with a

different combination of plane-stress constitutive integration algorithm and number of thickness-integration points.

Progress

A significant improvement has been made in our elastoplastic shell modeling capability. In the element technology area, the new shell element formulation has been implemented in DYNA3D.[2,3] In the structures modeling area, the performance of the various algorithms has been analyzed employing the new element, and accurate stabilization procedures have been devised for elastoplastic shell analysis. The results of this research have also been implemented in DYNA3D.

In addition to the research and development work, guidelines have been written that are useful to the engineering analyst.

Shell Element Formulation

The finite element shell formulation plays a key role in the elastoplastic modeling of thin shell structures. Accuracy, speed, and robustness are heavily influenced by the assumptions embedded in the element formulation. Accuracy requires a formulation capable of representing large rigid-body motions. Speed in an explicit analysis context requires one-point shell element integration. Robustness requires an effective stabilization procedure to suppress zero-energy modes, or "hourglass modes," introduced by the one-point element integration procedure.

We have developed a new shell element formulation that satisfies these requirements and offers more accuracy and robustness than previously existing formulations, while retaining comparable speed. In conjunction with the new element formulation, we have derived a new, consistent stabilization procedure.

The newly developed shell element (the YASE shell) is a four-node quadrilateral based on a three-field weak form of the governing equations. The three-field form of the governing equations allows independent interpolation of displacement, strain, and stress fields within a finite element. The associated Euler equations are: equilibrium, strain-displacement, and the constitutive relations. In-plane accuracy is attained through a three-field adaptation of the stress fields proposed in a two-field variational context by Pian and Sumihara.[4] The transverse shear stress fields are those proposed by Bathe and Dvorkin.[5]

All one-point shell element procedures require

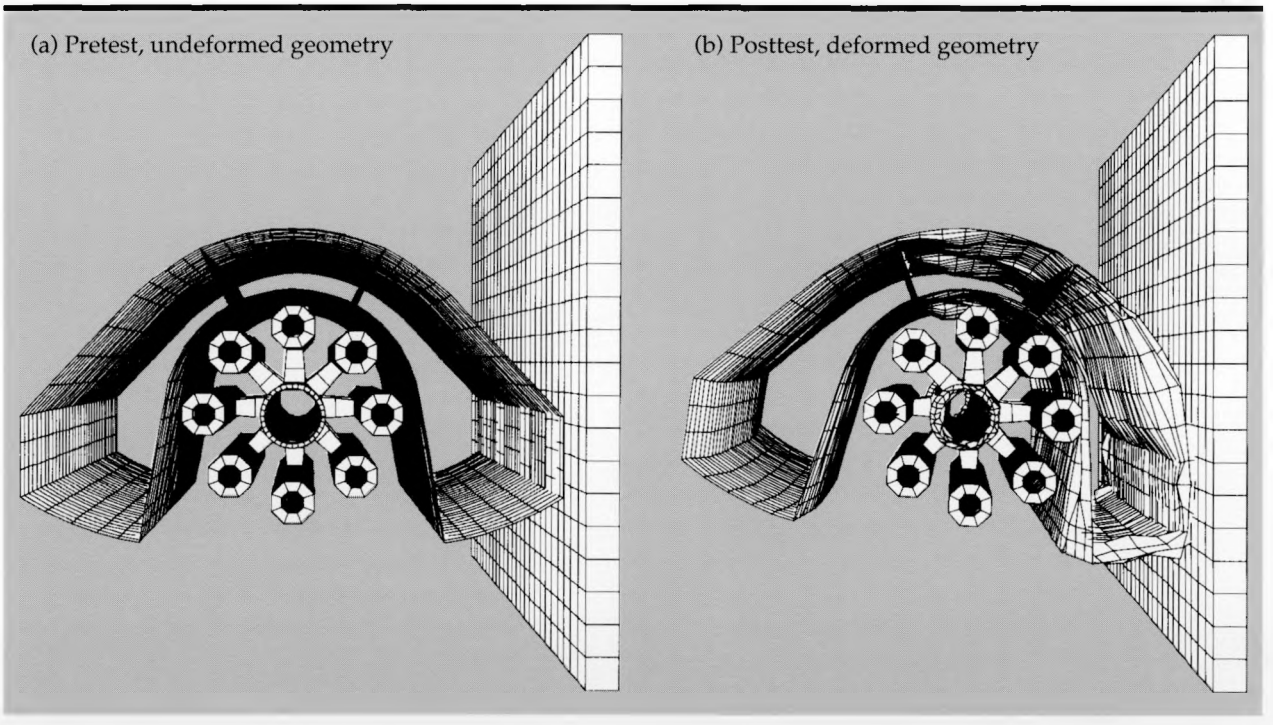


Figure 1. A finite element shell formulation with elastoplastic behavior is used in DYNA3D to simulate an aircraft crashing into the bomb bay of a stationary B1-B bomber. (a) Cutaway view of shell element geometry used for crash-analysis modeling of a body section of the B1-B. (b) Final, deformed geometry of body section resulting from crash-analysis modeling. The objective is to determine the loads that would be imparted to the B1-B weapons.

stabilization to suppress spurious zero-energy modes, often called "hourglass" modes. Failure of the stabilization algorithm to control these zero-energy modes results in a nonphysical solution, as shown in **Fig. 2(a).** This example problem is a corner-supported plate subjected to a point load at the center. the presence of hourglass modes is manifested by the characteristic checkerboard pattern of displacements, wherein adjacent nodes are displaced in opposite directions. The correct solution to this simple problem is shown in **Fig. 2(b).** Note the smooth curves in the correct solution. Stabilization procedures that require the user to select values for hourglass control parameters frequently have accuracy characteristics that depend heavily on the choice of values. In practice, it is difficult to choose appropriate values for hourglass stabilization parameters that yield an accurate solution.

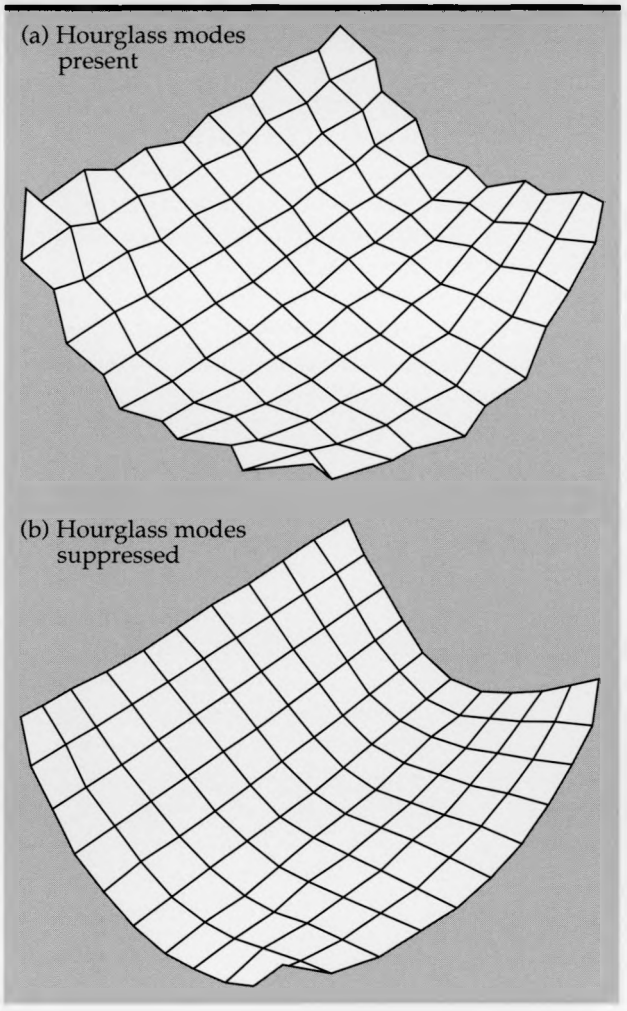
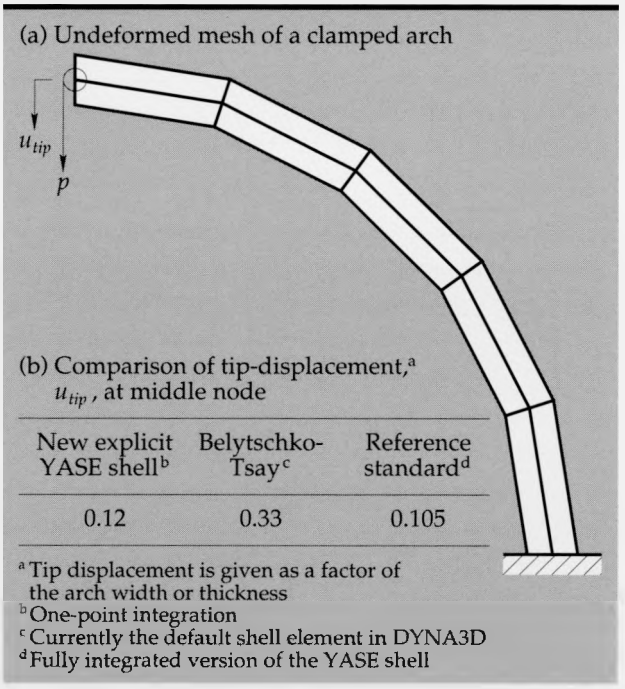


Figure 2. All one-point shell element procedures require stabilization to suppress spurious zero-energy modes, or "hourglass" modes. Failure to control these zero-energy modes results in a nonphysical solution. The example here is a corner-supported plate deformed by a point load at the center: (a) A nonphysical solution, with hourglass modes present. (b) The correct solution, with hourglass modes suppressed.

Our new stabilization procedure does not require the user to supply values for the control parameters. In the case of elastic material behavior, the new procedure with one-point shell element integration yields results identical to the fully integrated element. This represents a substantial improvement in accuracy over previously existing one-point shell element formulations.

We have devised a stabilization procedure for elastoplastic shell analysis that accounts for the effects of plasticity on the higher-order modes of the element. This procedure properly treats plasticity due to membrane deformations as well as both in-plane and out-of-plane bending deformations.

A simple example illustrates the improved behavior of the new explicit YASE shell element. **Figure 3(a)** shows an undeformed mesh of a



(b) Comparison of tip-displacement,[a] u_{tip}, at middle node

New explicit YASE shell[b]	Belytschko-Tsay[c]	Reference standard[d]
0.12	0.33	0.105

[a] Tip displacement is given as a factor of the arch width or thickness
[b] One-point integration
[c] Currently the default shell element in DYNA3D
[d] Fully integrated version of the YASE shell

Figure 3. Since it is difficult for a user to choose appropriate values for hourglass stabilization parameters that yield accurate solution, we derived a consistent stabilization procedure that requires no user input. For elastoplastic analysis, the procedure properly treats plasticity due to membrane deformations as well as both in-plane and out-of-plane bending deformations. A simple example illustrates the improved behavior of our new explicit, one-point-integrated YASE shell element: (a) Undeformed mesh of a clamped arch subjected to in-plane bending by a point load, p, at the tip. (b) Tabulated results of final displacements under the load. While achieving nearly a factor of three increase in accuracy over the Belytschko-Tsay element, the YASE element required only 10% more computer time.

clamped arch subjected to in-plane bending by a point load at the tip. Final displacements under the load are listed in **Fig. 3(b)** for three specific shell element formulations: (1) the new explicit YASE shell, (2) the Belytschko-Tsay shell (currently the default shell element in DYNA3D), and (3) a fully integrated version of the YASE shell that serves as the reference standard. The result using the YASE shell with one-point integration is within 15% of the reference (fully integrated) result, while the result using the Belytschko-Tsay element exceeds the reference value by more than a factor of three. While achieving this increase in accuracy, the explicit YASE element (with one-point integration) required only 10% more computer time than the Belytschko-Tsay element. Thus, the slight increase in cost appears well justified.

Constitutive Algorithms

Thickness-Integration Method. Most methods for elastoplastic shell analysis use continuum-based material models (i.e., stress is related to strain). One of the basic assumptions in the formulations of many shell elements, including the YASE shell, is that the "through-the-thickness" normal stress is negligibly small. As far as numerical constitutive evaluations are concerned, each shell element is composed of a number of lamina, with each lamina satisfying the postulates of plane stress. (Plane stress theory assumes that the out-of-plane normal stress is negligibly small.) These lamina stresses are used to evaluate the integrals of stress through the element thickness appearing in the computation of the element moment and force resultants. The advantage of this thickness-integration method is accurate representation of both partial yielding of a section and complex loading states, but at the expense of speed. The extent of approximation and level of cost can be partially controlled through the choice of the number of numerical integration points used in the thickness integration.

Stress-resultant Method. An alternative method uses material models formulated directly in terms of shell stress resultants (i.e., forces and moments are directly related to displacements and curvatures). This stress-resultant method offers potentially greater speed, since no thickness integration is performed and the material model is evaluated only once per shell element, but at the price of accuracy. Inaccuracies in the stress-resultant method occur from two principal causes:

- The assumption of full section yielding, which is basic to most stress resultant methods. Physically, yielding initiates at a point in the cross section and then spreads until the entire section has yielded. This causes a gradual reduction in the stiffness of the shell. Thus, the assumption of full section yield replaces this gradual stiffness reduction by an abrupt change from the elastic stiffness to the post-yield stiffness.
- The approximate method used to integrate the constitutive equation. Stress-resultant plasticity theory involves multiple yield surfaces that intersect in a nonsmooth fashion, and that generates considerable complexity in the associated constitutive integration algorithm.[6] Speed requirements in an explicit-analysis environment require simplifications to be made in these sophisticated procedures, and the modifications introduce errors into the integration.

Evaluation of Modeling Methods. Present efforts in constitutive modeling have focused on evaluation of the performance/cost (i.e., accuracy/speed) tradeoffs between the thickness-integration plasticity methods and the stress-resultant plasticity method for constitutive modeling in shells. Our studies of stress update algorithms employed in the thickness-integration methods are essentially complete.[7]

In order to assess the accuracy/speed tradeoffs between the various methods for constitutive modeling, a model problem was chosen for study. The selected problem is a cantilevered plate with a point load applied at one of the free corners. This load induces bending, shear, and normal forces in the plate. A comparison of the out-of-plane displacement and required computer time was made for:

- The stress-resultant plasticity method.
- Seven different thickness-integration shell plasticity methods, each with a different combination of plane stress plasticity algorithm and number of thickness-integration points. Results of the comparison are given in **Table 1.**

All computation times are based on a Cray Y/MP computer, although similar results would be expected on any vector processing machine. This test problem tends to highlight the influence of the constitutive algorithm on the overall execution time, since the problem does not contain other expensive numerical procedures such as those necessary to detect and treat the contact of two initially separate bodies. "Contact" algorithms contain search algorithms that are inherently scalar (i.e., "nonvector"), and this causes these procedures to be expensive on vector-processing computers such as the Cray. As a result, the effect of the constitutive model on the execution time of

most practical calculations is somewhat less dramatic than is indicated in this example.

As shown in **Table 1,** the stress-resultant method is the fastest and also the least accurate, introducing a displacement error of 45%.

Next in speed is thickness integration with two points through the thickness. If the stress scaling algorithm (SS-2) is used, then the error is 31%, but if one uses the fixed three-iteration algorithm (FTI-2) (default in DYNA3D), the displacement error is down to 18% with only a slight increase in computation time.

The results using thickness integration with three points, and either the fixed three-iteration algorithm (FTI-3) or the plane stress subspace algorithm (PSS-3), show a slight increase in error over the comparable results using two thickness integration points. This is somewhat unusual, since using more integration points typically yields better accuracy.

However, the expected trend in improved accuracy is again seen in the results using five thickness integration points. While somewhat slower, thickness integration with five points and the fixed three-iteration algorthm (FTI-5) yields an accurate result, with only 6% error.

The two calculations using the (scalar) iterative radial return algorithm (IRR-5 and IRR-10) are included for comparison purposes only, and are too inefficient to be used in practice.

In order to compare their effects on a large-scale analysis, we used four different elastoplastic constitutive algorithms to simulate the axial crush of a box beam. Each algorithm employed one of the following four computational methods:

- The stress-resultant plasticity method.
- The thickness-integration shell plasticity method using the plane-stress fixed three-iteration algorithm with two, three, and five points through the thickness: FTI-2, FTI-3, and FTI-5.

The computed deformed geometry for each of the methods is shown in **Fig. 4** along with their respective computation times. As seen in **Fig. 4,** the simulation using the stress-resultant method predicts a different deformed shape than the thickness-integration method, regardless of the number of thickness integration points. The stress-resultant calculation produces a fold with the first hinge inward, while all of the thickness-integration calculations predict the first hinge will fold outward. Since the thickness-integration method more accurately represents the physics of the problem, more confidence is placed in the deformed geometry predicted by that method.

This simulation illustrates the dramatic effect that constitutive-model errors can have on the evolution of the solution to a nonlinear problem. In addition, **Fig. 4** shows that the CPU time savings with the stress-resultant plasticity method in this large analysis are less significant than in the simple problem summarized in **Table 1.** The thickness-integration shell plasticity method with two-points in the thickness integration uses only 2% more computer time than the stress-resultant method.

Table 1. Evaluation of methods for elastoplastic constitutive modeling.

Type of algorithm	Thickness integration points	CPU time (sec)	Out-of-plane displacement factor	Displacement error (%)
Stress-resultant method				
Resultant	—	49.33	0.68	45
Thickness-integration method				
SS[a]	2	56.1	0.32	31
FTI	2	60.2	0.38	18
FTI	3	68.0	0.59	25
PSS	3	79.6	0.59	25
FTI	5	84.1	0.5	6
IRR	5	135.0	0.5	6
IRR	10	223.0	0.47	—

[a]Legend for type of plane-stress, shell plasticity algorithm:
SS stress scaling.
FTI fixed three iteration.
PSS plane stress subspace.
IRR iterative radial return.
(Details of these algorithms are described in Ref. 7.)

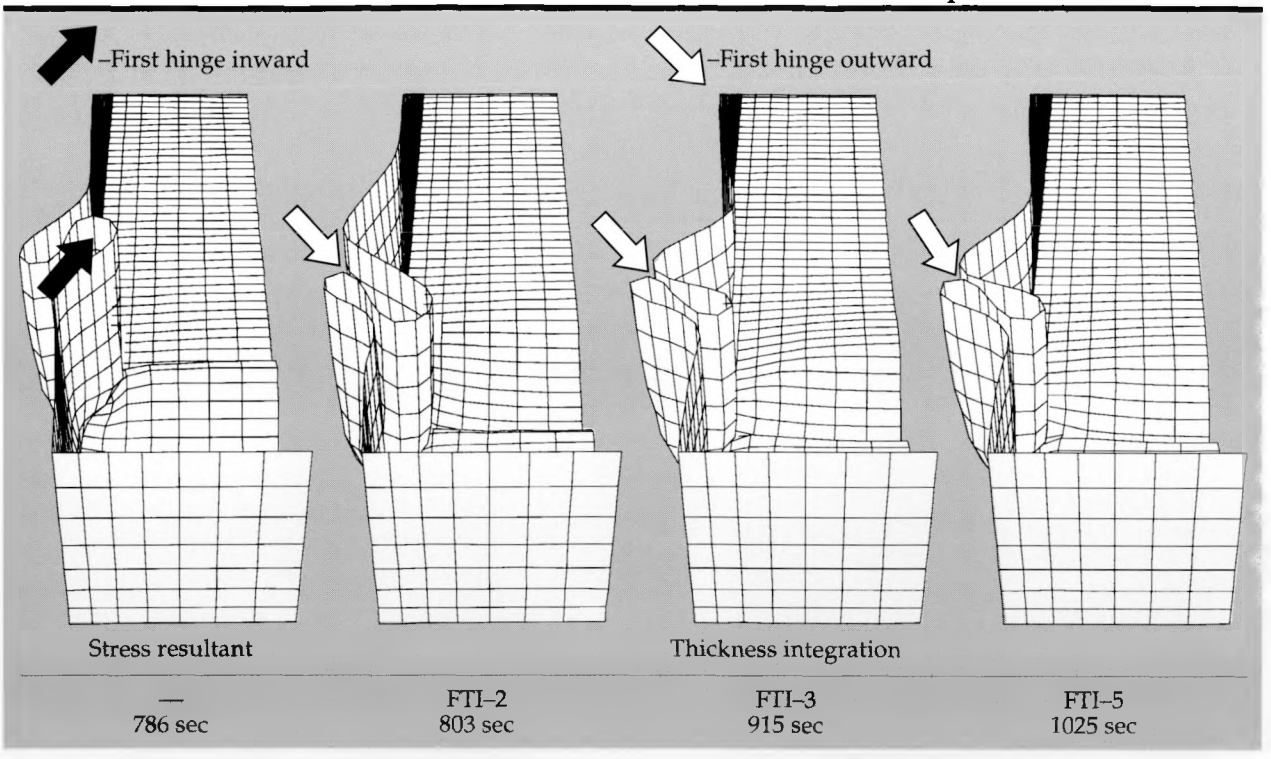


Figure 4. The effect of the constitutive algorithm on a large-scale analysis was demonstrated by the simulation of the axial crush of a box beam, using four computational methods for comparison: resultant plasticity; and thickness-integration shell plasticity using the fixed-three-iteration algorithm with two, three, and five points in the thickness integration. Shown here are the computed deformed geometries along with their respective computation times in seconds. Notice that the resultant calculation produces a fold with the first hinge inward, while all of the thickness-integration calculations predict the hinge will fold outward.

The above investigation indicates that the thickness-integration shell plasticity method is generally to be preferred over the stress-resultant plasticity method. The number of thickness integration points may be reduced to two for an approximate solution at minimum cost; this is generally more accurate than the stress-resultant method. In addition, the thickness-integration method generalizes more easily to complex constitutive models that do not lend themselves to representation in a stress-resultant form.

Conclusions and Future Work

A new quadrilateral shell element (the YASE shell) has been formulated with substantially improved accuracy properties in both elastic and elastoplastic applications. Concurrently, an associated stabilization procedure has been developed that controls spurious hourglass modes without the need for assigning ad-hoc values to the control parameters. The element is efficient and economical and has been implemented in the LLNL explicit finite element code, DYNA3D.

Thickness integration of plane stress constitutive equations has proven to be the method of choice for elastoplastic shell modeling. Due to the complicated structure of the yield surfaces, the stress-resultant plasticity method offers little speed increase over the thickness-integration shell plasticity method with two points through the thickness, and it suffers serious inaccuracies. The thickness-integration method using five points through the shell thickness offers an accurate solution to most elastoplastic shell problems.

Based on the above observations, it is recommended that thickness integration with two thickness points be used for preliminary calculations, and that thickness integration with five thickness points be used for final calculations to assure accurate results.

This concludes our two-year effort aimed at improving our modeling capability for thin shell structures undergoing elastoplastic deformation. Substantial progress has been made in increasing the accuracy and in decreasing the cost of large-scale simulations. These developments translate into faster turnaround times for analyses, allowing more design options to be investigated and ultimately leading to a better design.

Acknowledgements

Extensive collaboration by Dr. Bruce Engelmann of the Methods Development Group during the development of the YASE shell element is gratefully acknowledged.

1. J. O. Hallquist and R. G. Whirley, *DYNA3D: Nonlinear Dynamic Analysis of Structures in Three Dimensions, User's Manual,* Lawrence Livermore National Laboratory, Livermore, Calif., UCID-19592, Rev. 5 (1989).
2. B. E. Engelmann, R. G. Whirley and G. L. Goudreau, "A Simple Shell Element Formulation for Large-Scale Elastoplastic Analysis," *Proc. ASME Symp. on Comp. Methods in Shells,* December 1989.
3. B. E. Engelmann and R. G. Whirley, "An Elastoplastic Shell Element Formulation for Large-Scale Explicit Dynamic Analysis" (in preparation).
4. T. H. H. Pian and K. Sumihara, "Rational Approach for Assumed Stress Elements," *Int. J. Num. Methods Eng.* **20,** 1685-1695 (1984).
5. K. J. Bathe and E. N. Dvorkin, "A Continuum Mechanics Based Four-Node Shell Element for General Nonlinear Analysis," *Int. J. Computer-Aided Eng. Software* **1,** 77-88 (1984).
6. J. C. Simo and J. G. Kenney, "On a Stress Resultant Geometrically Exact Shell Model. Part IV. Nonlinear Plasticity: Formulation and Integration Algorithms" (to be published in *Computer Methods App. Mech. Eng.*).
7. R. G. Whirley, J. O. Hallquist and G. L. Goudreau, "An Assessment of Numerical Algorithms for Plane Stress and Shell Plasticity on Supercomputers," *Engineering Computations* **6**(2), 115-126 (1989).

ENGINEERING RESEARCH AND DEVELOPMENT

Thrust Area Report FY 89

Diagnostics and Microelectronics

The Diagnostics and Microelectronics thrust area researches and develops advanced microdevices for the Laboratory's high-speed diagnostic needs. We also support ways to improve the Laboratory's basic materials and device technology base. In the area of high-speed diagnostics, we are concentrating on three technologies:

- Semiconductor-based microfabrication.
- Semiconductor microelectronics.
- Integrated and quantum optical electronics.

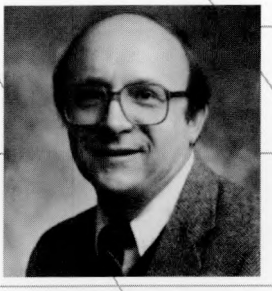
Our research in these areas combines the results of our own projects, projects conducted jointly with industry, and university research projects sponsored by our thrust area. In FY 89 we supported the following specific projects:

- Investigating the ability of charged-coupled devices made of gallium arsenide to capture and record high-speed transient signals.
- Developing a novel data transmission and recording system in which the time base of transient signals is encoded by optical wavelengths.
- Providing high-speed semiconductor laser diodes and electro-optic modulators that modulate optical signals transmitted over optical fibers.

In the area of basic materials and devices, we are supporting the following projects in FY 89:

- Upgrading our optical calibration facility.
- Establishing system-level models of complex diagnostic systems.
- Establishing computer control of our molecular beam epitaxy system such that complex graded bandgap materials and superlattice structures can be built.
- Developing a submicron etch technology based on chemically assisted ion beam etching.
- Making special diffraction gratings by developing advanced silicon anisotropic etching technology.
- Making complex sensors by developing silicon-based membrane technology.

Joseph W. Balch
Thrust Area Leader



Integrated Optics

Glen M. McWright
Engineering Research Division
Electronics Engineering

During FY 89, we continued to extend the performance parameters of lithium niobate integrated optical devices. Here, we discuss recent improvements in the fabrication and performance of Mach-Zehnder integrated optical modulators which now have > 20-GHz bandwidths, < 10-V switching voltages, and 3-dB optical insertion losses.

Introduction

Many diagnostic applications at the Laboratory require the transmission of a single analog signal from a remote radiation sensor. The simplest diagnostic system uses electrical cables. However, because electrical cable is lossy and dispersive, fast transient signals are badly distorted during transmission. For this reason, cable-based systems are generally not satisfactory for transmitting information faster than about 1 ns. Furthermore, electrical cables are very susceptible to interference from ionizing radiation, which is present in many of the environments in which diagnostic signals must be transmitted.

Both these problems with cable-based transmission systems can be circumvented with integrated optical systems in which the signal source is a semiconductor laser diode. The light signal is manipulated and controlled by a high-speed electro-optical modulator in much the same way that conventional electronic circuits manipulate electrical signals. We are now developing optical transmission systems using the integrated optics approach, which combines the microfabrication techniques of integrated electronic circuitry with the inherently high fidelity, broad bandwidth, and high information capacity that characterize optical media. Our goal is to build the optical equivalent of an integrated electrical circuit. Its advantages would include wider bandwidth, lower loss, smaller size and weight, and immunity to electromagnetic interference.

An example of an integrated optical transmission system suitable for diagnostic applications in underground nuclear tests is shown in **Fig. 1**. The voltage output from a radiation detector is used to externally modulate a laser, thus encoding the detector's electrical signal on the optical waveform produced by the laser. The resulting variations in the laser light intensity are recorded with a streak camera. A laser wavelength of 810 nm is used to accommodate the streak camera's photocathode, whose sensitivity is poor in the popular 1.3-μm region of the electromagnetic spectrum, which might otherwise have been chosen for the laser wavelength. The key component in this system is the integrated optical modulator. Development of this modulator has been the main focus of our research activities. We have chosen to concentrate on one particular type of integrated optical modulator—the Mach-Zehnder interferometer—because of its wide bandwidth and low switching voltage.

The Mach-Zehnder (**Figs. 2a** and **2b**) is an electro-optically tunable interferometer. It is formed by photolithographic techniques similar to those used in semiconductor chip manufacture.[1] First, optical waveguides are defined in the lithium niobate electro-optic crystal by the in-diffusion of titanium; then gold electrodes are deposited on the waveguides. The structure is designed so that the light splits at the input Y-branch into two beams that propagate independently for a few millimeters and then recombine at the output Y-branch. If no electric field is applied, the beams arrive at the

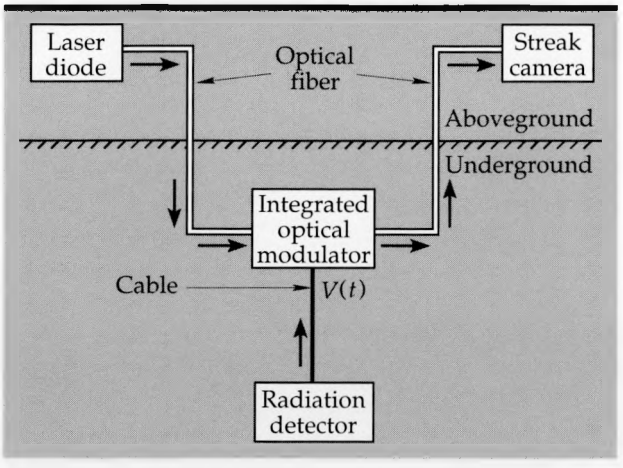
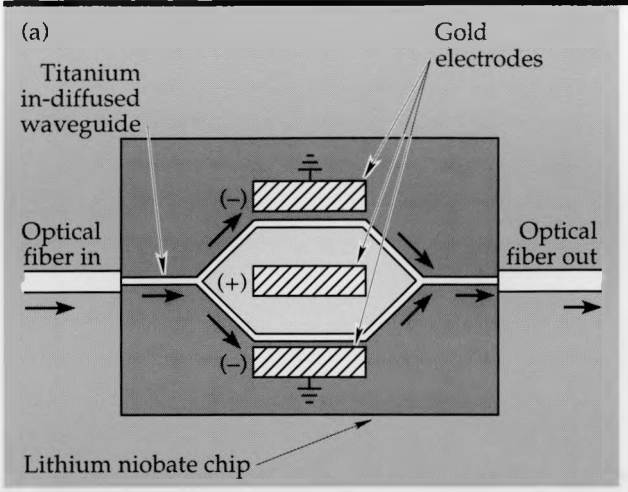


Figure 1. Functional block diagram of an integrated optical data transmission system configured for diagnostic applications in underground nuclear tests.



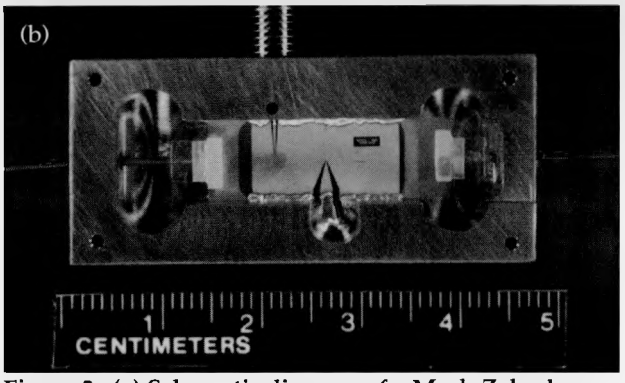


Figure 2. (a) Schematic diagram of a Mach-Zehnder interferometer fabricated in lithium niobate. (b) Photomicrograph of a packaged lithium niobate Mach-Zehnder modulator.

output Y-branch in phase and recombine without loss. However, application of an electric field introduces a path length difference between the two beams. This path length difference produces a phase difference between the two beams at the point of recombination, resulting in an intensity modulation of the recombined beam that is proportional to the applied electric field. Thus, the Mach-Zehnder operates as an electrically-driven optical modulator.

Integrated optical device designers are generally interested in three key device parameters: the bandwidth, the dc switching voltage (the voltage required to change the light transmission through the device from 100% to 0%), and the optical insertion loss. Device goals usually include wide bandwidth, low dc switching voltage, and low optical insertion loss. The interplay beween these three device parameters is quite complicated, and there is a trade-off between parameter areas. For example, higher bandwidth always implies higher drive voltage, and low drive voltage implies higher optical insertion loss. Nevertheless, we have been able to delicately balance the waveguide and electrode parameters to repeatably and reliably obtain wide bandwidth, low switching voltage, and low optical insertion loss within a given device. Thus, during FY 89, we improved the performance characteristics of lithium niobate modulators in these three specific areas. First, we extended the bandwidth to greater than 20 GHz. Second, we reduced the switching voltage to below 10 V. And third, we reduced the total optical insertion loss of the device to 3 dB.

Progress

The first of our three major FY 89 research efforts involving lithium niobate Mach-Zehnders led to an extension of their bandwidth to beyond 20 GHz. This improvement was accomplished through a combination of clever electrode and packaging designs. We measured the frequency response using the so-called swept frequency technique (**Fig. 3**). In this technique, an RF signal (0.01–26.5 GHz) is square-wave-modulated with a low frequency electrical signal (below 10 kHz). This square-wave-modulated RF signal is then applied to the Mach-Zehnder, thus modulating the optical carrier. The resulting optical intensity variations are monitored with a low-speed photodetector and lock-in detector. Most importantly, the optical modulator, because of its wide bandwidth design, responds to the high-frequency, square-wave-modulated RF signal. However, by design, the low-speed photodetector and lock-in detector respond only to the amplitude of the low-frequency envelope of the modulated optical signal. Indeed, the advantage of this technique is that the modulator bandwidth can be accurately measured without a high speed detection system. As the frequency of the RF signal is "swept" from 0.01–26.5 GHz, the amplitude of the envelope measured on the lock-in detector varies in direct proportion to the frequency response of the Mach-Zehnder. Thus, the bandwidth of the optical modulator is easily measured across an extremely wide range. We measured the frequency response of a Mach-Zehnder from 0.01–26.5 GHz using this swept frequency technique. As shown in **Fig. 4**, the bandwidth of this device is greater than 20 GHz. This bandwidth was subsequently confirmed with independent time response measurements from a streak camera.

Our second improvement to the lithium niobate Mach-Zehnders was to reduce the dc switching voltage for these devices to less than 10 V. This

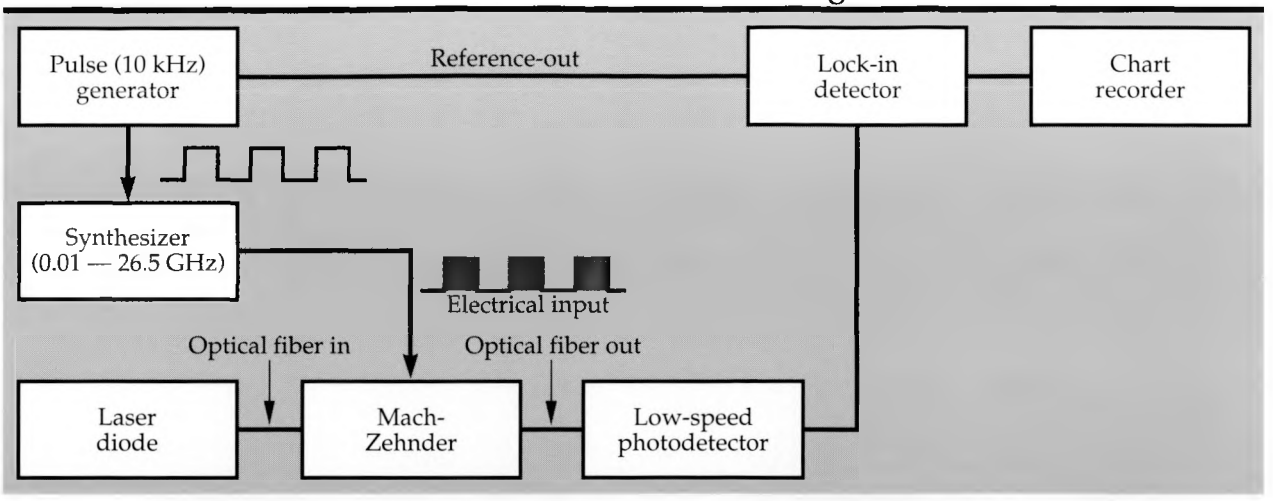


Figure 3. Schematic of swept-frequency setup used to measure Mach-Zehnder bandwidth. A square-wave-modulated RF signal is applied to the optical modulator, thereby modulating the optical carrier. A low-speed photodetector and lock-in respond only to the amplitude of the low-frequency (10 kHz) envelope of the modulated optical signal. As the frequency of the RF signal is swept from 0.01-26.5 GHz, the amplitude of the envelope signal as measured by the lock-in varies in direct proportion to the frequency response of the Mach-Zehnder.

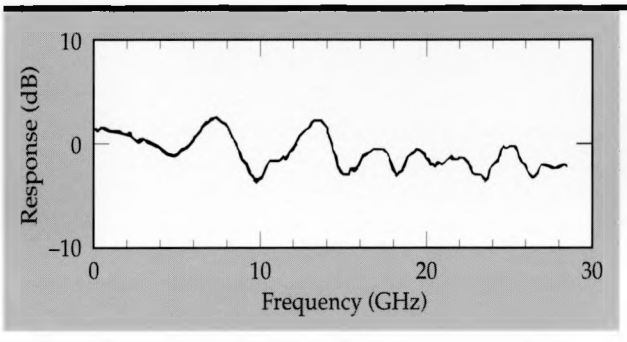


Figure 4. Swept-frequency measurement of electro-optic bandwidth of a typical Mach-Zehnder device, indicating that bandwidth is wider than 20 GHz.

advance was achieved by improving the match of the electro-optic overlap between the optical waveguide and the electrode.

Our third improvement to the lithium niobate Mach-Zehnder system was to reduce the total optical insertion loss to 3 dB. This advance was achieved by more effective polishing procedures and improvements in the waveguide design, which combined to give a better mode match between the optical fiber and the optical waveguide. Finally, it is important to emphasize that we can repeatedly and reliably obtain these superior performance characteristics—wide bandwidth, low switching, and low optical insertion loss—within any given device by delicately balancing the design parameters.

Future Directions

We have made important strides this year in lithium niobate integrated optical devices. Specifically, we have fabricated and evaluated Mach-Zehnders which have > 20-GHz bandwidths, < 10-V switching voltages, and 3-dB optical insertion losses. Our future plans include removing the resonances from the high-frequency, electro-optic characteristics of the modulators, and also ruggedizing the Mach-Zehnder package. This research and development program should enable us to realize the full potential of this integrated optics technology.

1. G. M. McWright and J. C. Koo, "Integrated Optics for Diagnostics," *Energy and Technology Review,* Lawrence Livermore National Laboratory, Livermore, CA, UCRL-52000-87-4 (April 1987), pp. 21–27.

Microfabrication Development

Dino R. Ciarlo, Wing C. Hui, and James A. Folta
Engineering Research Division
Electronics Engineering

During FY 89 we worked on a number of new processes to provide LLNL researchers with specialized microstructures for their experiments. Our development efforts included a new phase grating fabrication process for diffraction gratings that can be used in x-ray spectrometers, a method to fabricate laser alignment targets with variable reflectivity, a study of (110) silicon etching with applications to miniaturized sensor and heat sink design, and the electrochemical etching of silicon wafers to form thin silicon membranes for use as x-ray windows.

Introduction

We are using integrated circuit (IC) microfabrication technology to produce custom microstructures for several LLNL programs. These microstructures are specialized physical devices with dimensions of a few micrometers. They are used as diagnostic devices, as sensors, and as targets in LLNL experiments. Programmatic support for this activity has steadily increased over the last few years. During FY 89 we received programmatic support to deliver a number of microstructures, including the following:

- Micromachined silicon heat sinks for laser diode cooling
- Microelectrode chemical sensors
- Vacuum-integrated circuits for high-temperature, radiation-hard electronics
- Advanced optical sensors
- Laser alignment targets
- Thin x-ray transmissive windows

These types of microstructures are custom-designed for a specific application, and are not commercially available. Also, this past year we installed a third low-pressure chemical vapor deposition (LPCVD) system for the deposition of boron nitride thin films for x-ray transmissive windows. The other two LPCVD systems are used to deposit thin films of silicon nitride and polysilicon.

Progress

Phase Diffraction Gratings

We have developed a new way to fabricate diffraction gratings on a smooth silicon wafer substrate. In previous work, diffraction gratings were printed and etched into the surfaces of silicon wafers using chemical etchants such as potassium hydroxide. Lines and spaces as small as 0.75 μm were used. The reflection efficiency of these gratings tends to be low because the etched grooves have a rough texture and the incident energy can reflect only from the unetched tops of the grooves. Our new method of fabrication produces gratings with smooth surfaces in the etched grooves themselves as well as between the grooves. Such gratings are called phase gratings because the depth of the etched grooves is adjusted to get the desired phase angle shift between the energy reflected from the tops and the bottoms of the grooves. Since the collected energy is reflected from the entire grating surface instead of from just half the surface, the grating efficiency is higher.

These gratings are useful in x-ray diagnostic experiments where they serve as substrates for synthetic multilayers. Several groups at LLNL are conducting x-ray diagnostic experiments using such synthetic multilayers. They have an increased reflectivity to low-energy x rays because of Bragg diffraction of the low-energy x rays from the deposited multilayers. In some applications, new x ray spectrometers can be built by depositing the synthetic multilayers on top of an existing diffraction grating.

Laser Alignment Targets

We have developed a method to mass-produce laser alignment targets using our low-stress silicon nitride technology. To do this, we first deposit a 0.5-μm-thick, low-stress silicon nitride film on a 15-mil-thick, 3-in.-diam silicon wafer using LPCVD. We then evaporate a 0.5-μm-thick gold film on one side of this wafer. Using double-sided photolithography, an alignment target is printed on the surface of the gold film while a window etch pattern is simultaneously aligned and printed on

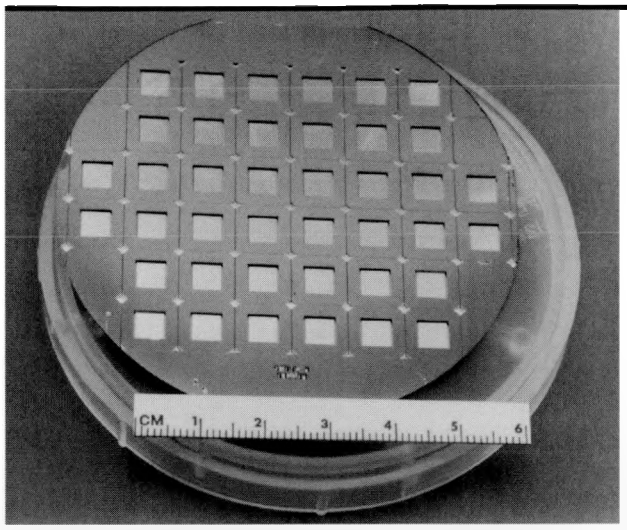


Figure 1. Silicon wafer, 3-in.-diam, with 40 laser alignment targets etched into its surface.

the opposite side of the wafer. The alignment target is then etched using a gold wet etch, and then the silicon nitride in the window pattern on the opposite side of the wafer is plasma etched. Finally, the silicon is etched away leaving several free-standing silicon nitride windows in the wafer, each window containing a 0.5-μm-thick gold target. **Figure 1** is a photograph of a 3-in.-thick silicon wafer containing 40 alignment targets. Scribe lines are also etched into this wafer so that the windows can be separated into individual parts without requiring any sawing. The yield on these parts is very high because of previous work on control of the parameters required to deposit low-tensile-stress silicon nitride. During a recent run, we fabricated these targets on six 3-in.-thick wafers and got a 100% yield to give $6 \times 40 = 240$ alignment targets in one run.

For some applications, the silicon nitride window supporting the gold alignment targets needs to be somewhat roughened so that it will scatter light. Such surfaces are made by depositing the LPCVD silicon nitride film on a silicon surface that has been roughened slightly by using a lapping grit. The resulting roughness of the surface is transferred to the deposited silicon nitride. **Figure 2** is a close-up photograph of one of the silicon nitride windows with the 0.5-μm-thick gold alignment target clearly visible. The width of these gold lines is 25 μm. The area of the silicon nitride in this window is 5 by 5 mm, and the overall chip size is 19 by 19 mm.

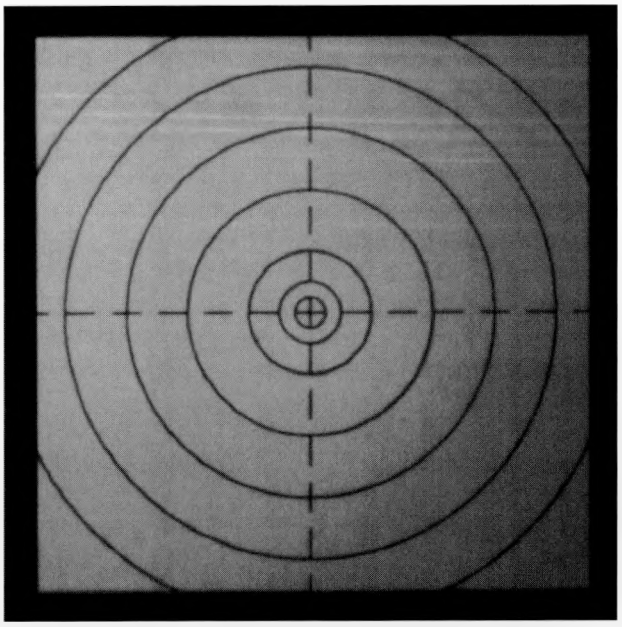


Figure 2. Close-up view of a laser alignment target showing the 25-μm-wide gold alignment lines.

Anisotropic Etching of Silicon

During FY 89 we continued to study the anisotropic or directional etching of silicon wafers in order to develop the techniques required to etch certain useful structures. Most of the interesting structures we etch are on (110) oriented silicon wafers. When we purchase these (110) oriented wafers we usually ask that an orientation flat be placed on a (111) crystal plane which is perpendicular to the (110) surface. If vertical deep grooves are to be etched, then the mask is aligned so that the grooves run parallel to this orientation flat. For other structures, it is necessary to intercept other (111) crystal planes which are also vertical to the (110) surface. There are two possibilities, as shown in **Fig. 3**. The second vertical (111) crystal plane makes an angle of 70.6° to the alignment flat, but it can be in either of two directions as shown. The only way to tell the direction of the second perpendicular (111) plane is to etch a small feature and then carefully inspect the edges. If one etches a small pit in the surface of such a wafer, the structure shown in **Fig. 4** will result. At the corners, there will be etched edges which make angles of 70.6° or 54.7°. The vertical (111) plane is the one that makes a 70.6° angle with the orientation flat.

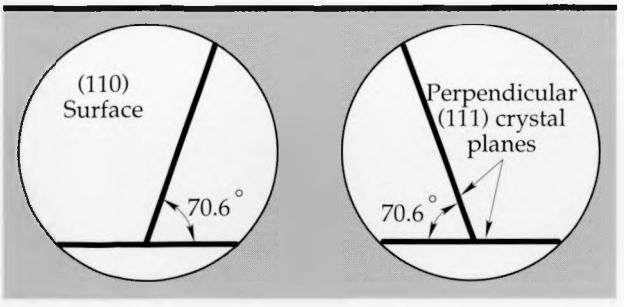


Figure 3. Possible orientations for the vertical (111) crystal planes on a (110) oriented silicon wafer.

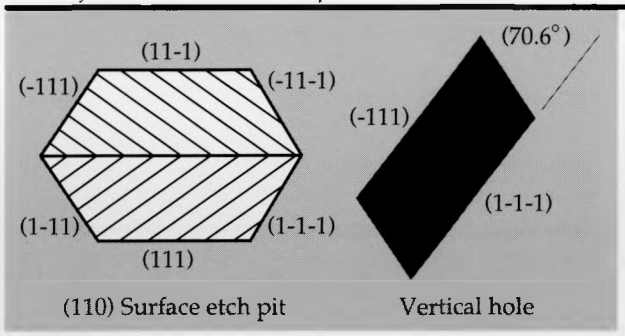


Figure 4. Etch pit in a (110) oriented silicon wafer showing the enclosed crystal planes.

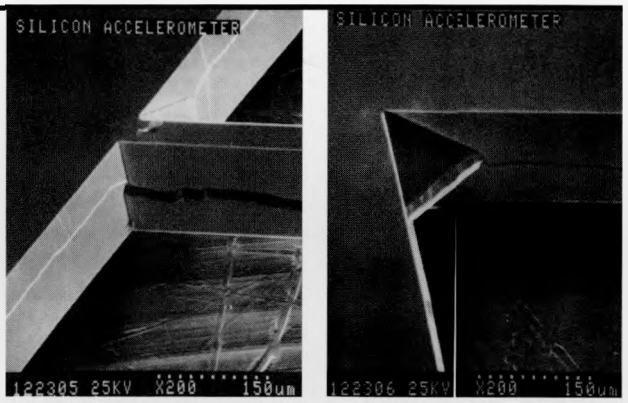


Figure 6. SEM images of the etched pattern using the mask shown in Fig. 5.

In general, when silicon is directionally etched with a wet etch such as potassium hydroxide, the shapes of the resulting inside angles in the etched cavities are well-behaved and easy to predict. Their shape approaches a terminal shape which can be predicted in advance. Outside corners etch more irregularly and are not easy to predict, since their shape depends on how long the part is etched. **Figure 5** is a diagram of a pattern etched into the surface of a (110) oriented silicon wafer. All the mask edges were aligned to vertical (111) crystal planes. **Figure 6** shows SEM images of a surface that has been etched with this pattern. As can be seen, the obtuse-angled corners come together in one sharply defined crease. The acute-angled corners contain a (111) crystal plane which makes a 35.3° angle to the surface of the wafer. This data is especially useful when designing structures which have to be etched through the entire thickness of a wafer, as is required in water manifolds for microchannel heat sinks.

Electrochemical Etching of Silicon

We are developing the capability to fabricate thin single crystal membranes of silicon by electrochemical etching.[1] Thicknesses between 1 to 3 μm are especially useful. Such membranes can be used as x-ray transmissive windows or as targets in spectroscopy experiments. The approach we are using is to purchase (100) oriented silicon wafers, which are p-type and which have a thin epitaxial n-type layer grown on their surface. The thickness of this n-type layer has to equal the thickness of the desired single-crystal silicon membrane. The wafers are placed in a potassium hydroxide solution and biased as shown in **Fig. 7**. Initially very little current flows, because for this bias arrangement, the p-n junction diode is reverse-biased and thus prevents current flow. In this configuration, the exposed p-type silicon is etched by the KOH. When the p-type silicon is completely etched, there is no longer a p-n junction to limit the current and it increases. This high current causes a thin oxide to form over the exposed n-type region and the etching stops. The resulting membrane has the thickness of the initial n-type epitaxial silicon layer. A special Teflon fixture was designed and built to hold the wafer during the etching. This fixture

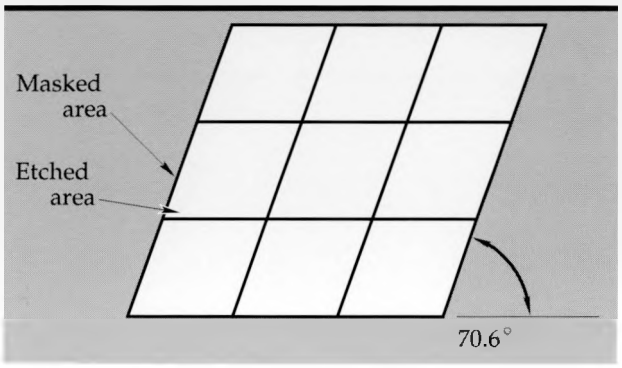


Figure 5. Pattern to be etched into a (110) oriented silicon wafer to study inside corner etching.

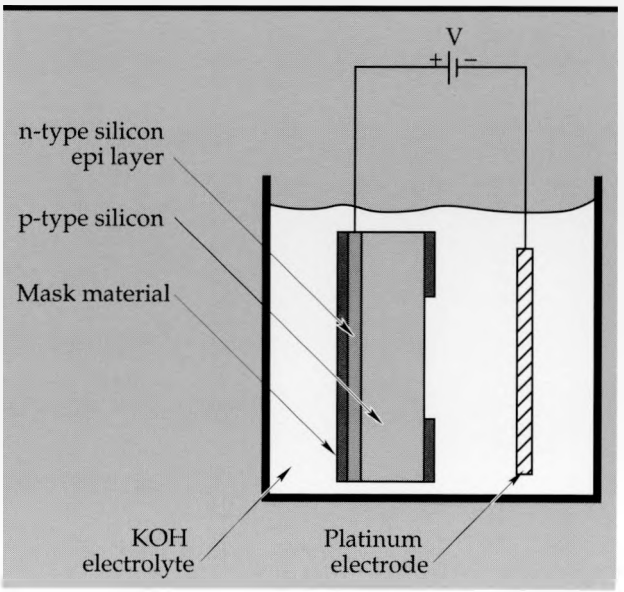


Figure 7. Experimental configuration for the electrochemical etching of silicon membranes.

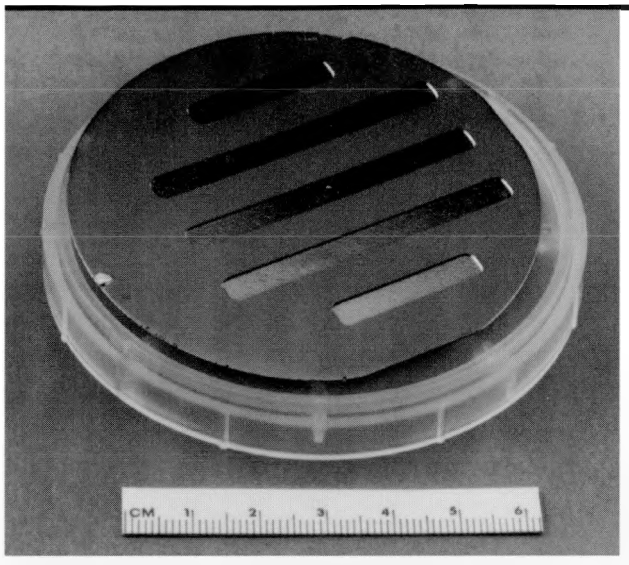


Figure 8. Silicon wafer, 3-in.-diam, that has been electrochemically etched to form five thin silicon membranes.

protects parts of the wafer from the etchant and also provides an electrical contact to the n-type region. A potentiostat was incorporated in the control circuit to control the working electrode potential and provide better process control. **Figure 8** is a photograph of a 3-in.-diam, 15-mil-thick silicon wafer that was etched using this method. This wafer contains five thin silicon membranes, 3 mm wide and 2 to 5 cm long.

Future Work

During FY 90 we plan to perform additional studies on several silicon wafer bonding techniques. In particular we want to measure the strength of several different wafer bonds and compare this data with acoustic microscope photographs of the same bonds. We also expect to continue refining our electrochemical etching technique for single crystal membrane formation. We would like to achieve silicon membranes as thin as 0.1 μm. Finally, we expect to work with a new process to achieve gallium arsenide (GaAs) thin membranes.[2] In this process, GaAs films are grown by molecular beam epitaxy (MBE) on a GaAs substrate that has been coated with a thin MBE aluminum arsenide (AlAs) sacrificial layer. The top thin GaAs layer can be floated off the thick GaAs substrate by dissolving the thin AlAs sacrificial layer in hydrofluoric acid. Such thin GaAs films have a number of useful applications. In particular, if laser diodes are fabricated in this thin GaAs layer, the layer can then be attached to a heat sink for the efficient removal of the heat generated during laser operation.

Acknowledgments

The authors acknowledge the valuable technical work of the microfabrication technicians including Conrad Yu, Glen Jameson, Ed Hee, and Rob Olson. The new phase grating fabrication method was suggested by Troy Barbee of the Chemistry and Materials Science Department.

1. B. Loeck, S. D. Scott, N. F. De Rooij, and R. L. Smith, "Study of Electrochemical Etch-Stop for High-Precision Thickness Control of Silicon Membranes," *IEEE Trans. Electron Devices*, **36**, 4 (April, 1989).
2. E. Yablanovitch, T. Gmitter, J. P. Harbison, and R. Bhat, "Extreme selectivity in the lift-off of epitaxial GaAs films," *Appl. Phys. Lett.* **51**, 26 (28 Dec. 1987).

Improvements in Molecular Beam Epitaxy Technology

Raymond P. Mariella Jr. and Gregory A.Cooper
Engineering Research Division
Electronics Engineering

Our objectives in FY 89 were to put our molecular beam epitaxy (MBE) machine's shutters and oven temperatures under full computer control, so that complex structures such as graded bandgap materials or superlattices could be grown reproducibly by "recipe." In addition, we intended to produce high quality films as characterized by electron mobility. We accomplished these goals, and now use our MBE to grow GaAs/$Ga_{1-x}Al_xAs$ thin films from which our coworkers have fabricated optical waveguides, ultra-high-speed photoconductors, GRINSCH SQW diode lasers, and charge-coupled devices (CCDs). We also developed the capability to grow InGaAs strained-layer quantum wells and superlattices by adding an indium oven and an arsenic cracker to our MBE system. We started growing these InGaAs structures at the end of FY 89.

Introduction

MBE is a technique by which substrate wafers of materials such as gallium arsenide (GaAs), silicon (Si), or indium phosphide (InP) are individually coated with single-crystal layers of $Ga_{1-y-x}Al_xIn_yAs$. The coating, which typically takes place at the rate of one monolayer per second, is primarily a result of the chemical reaction between arsenic (As) and one or more of the group III metals aluminum (Al), gallium (Ga), or indium (In). **Figure 1** shows a schematic drawing of an MBE system. Maintaining the substrate wafer in the temperature range of 200°C to 750°C, we expose it to molecular beams of As_4 (or As_2, from the cracker) and Ga, Al, and In, as well as trace quantities of selected dopants such as Si or beryllium (Be). Because the molecular beams can be shuttered off and on in a fraction of a second, and because the growth process tends to anneal the surface, it is possible to grow very high quality interfaces by MBE. **Figure 2** shows a transmission electron micrograph of an AlAs/GaAs superlattice which we grew under computer con-

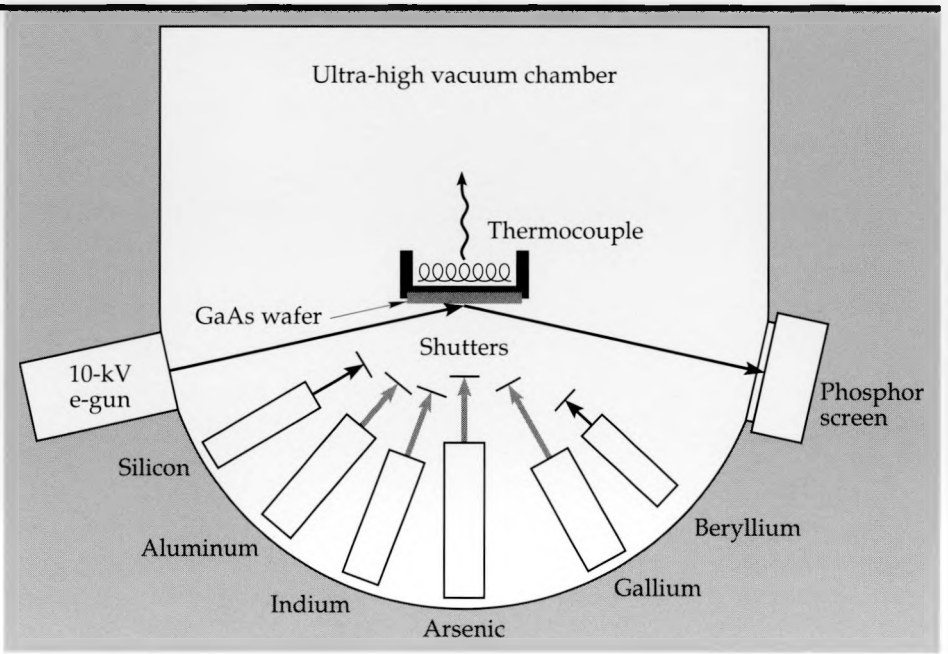


Figure 1. Schematic drawing of the molecular beam epitaxy (MBE) machine. The various elements are heated for evaporation and the associated elemental fluxes are turned off and on with shutters.

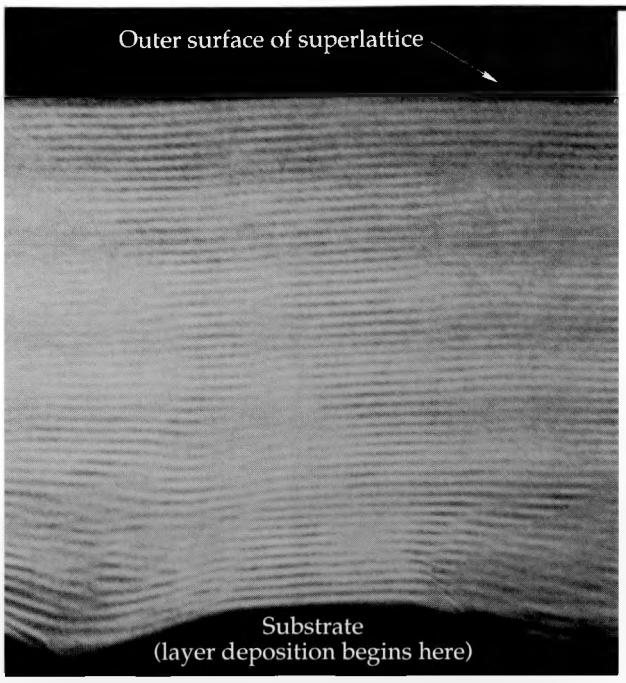


Figure 2. High resolution TEM micrograph of a superlattice with 50 pairs of GaAs/AlAs with each individual layer having a thickness of four monolayers or 11 Å. At the start of the growth some roughness of the interface can be seen, but this smooths out almost completely at the final layer. Total thickness of the superlattice is 550 Å.

trol. The roughness of the interface which is visible at the start of the superlattice is essentially smoothed away at the outer surface.

Progress

We grew more than 80 different films in FY 89. These have been processed into photoconductive light detectors with graded bandgap AlGaAs on low temperature GaAs which displayed enhanced sensitivity and response times under 50 ps; as well as GaAs CCDs with 128 cells, waveguide/light modulators with response times under 50 ps; and also GRINSCH-SQW diode lasers with a current threshold of 300 A/cm^2. **Figure 3** shows a plot of the electron mobility for a GaAs layer which we grew by MBE with Si doping at $7 \times 10^{14}/cm^3$. The room temperature mobility of 8,000 $cm^2/(V \cdot s)$ and the 90 K mobility greater than 70,000 $cm^2/(V \cdot s)$ indicate very high quality epitaxial growth.

One important factor in the successful film

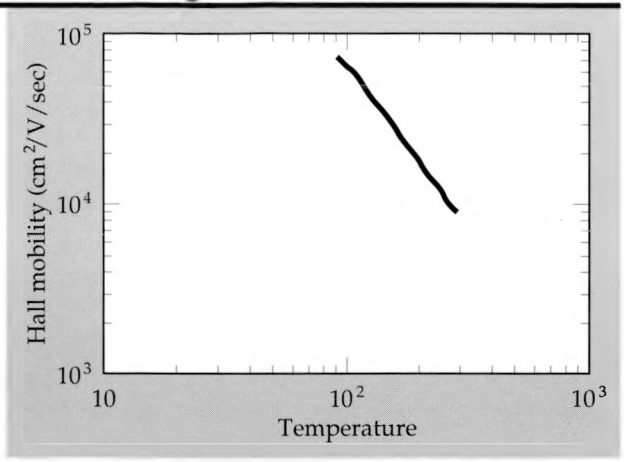


Figure 3. Log-log plot of electron mobility vs temperature for MBE GaAs doped to $7 \times 10^{14}/cm^3$, n-type. At 300 K, room temperature, the mobility is 8000 $cm^2/(V \cdot s)$, and this increases to more than 70,000 $cm^2/(V \cdot s)$ at 90 K.

growth for FY 89 has been the development of computer control of the MBE shutters and oven temperatures. We have accomplished this using a 16–MHz Everex 80386-based computer running Microsoft QuickBasic 4.0. Although the code can be run in the compiled form, the computer and program are fast enough so that the interactive mode is sufficient. The program constantly polls the keyboard for operator input, also polls the oven temperatures, and continuously displays the open/shut status of each shutter. During a growth by "recipe," the computer monitors the time of each individual growth step and can perform a change in oven temperature or shutter status with an accuracy of 0.1 sec or better. Growth steps can include superlattice structures such as the example shown in **Fig. 2** and also linear or parabolic band-gap grading, which is necessary for high-performance diode lasers.

Future Work

We will continue our epitaxial growth of thin films for use in the development of photoconductive light detectors, waveguide/light modulators, and GaAs CCDs. We will also grow GRINSCH SQW diode lasers with InGaAs strained-layer quantum wells to provide longer-wavelength laser emission. In addition, we will investigate the use of InGaAs strained-layer superlattices to build high-efficiency photocathodes for the near infrared.

Quantum Well Lasers

Raymond P. Mariella Jr.
Engineering Research Division
Electronics Engineering

In FY 89 we used our molecular beam epitaxy (MBE) machine with full computer control of shutters and oven temperatures for the reproducible growth of epilayers to fabricate GRINSCH SQW diode lasers based on gallium arsenide and aluminum gallium arsenide. We processed these wafers into broad-area lasers and observed threshold current densities as low as 300 A/cm^2. We also developed the capability to grow $In_yGa_{1-y}As$ strained-layer quantum wells by adding an indium oven and an arsenic cracker to our MBE system. We have recently (end of FY 89) begun to set up the apparatus required to characterize the emission wavelengths from these lasers.

Introduction

Graded-index separate-confinement heterostructure single-quantum-well diode (GRINSCH SQW) lasers based on gallium arsenide (GaAs) and aluminum gallium arsenide ($Al_xGa_{1-x}As$) have demonstrated the highest electrical efficiency of any laser, generating 12 W of constant power at 25% efficiency and 50 W average power under pulsed conditions at greater than 50% efficiency.[1,2] **Figure 1** shows the epilayers which comprise a typical GRINSCH SQW diode laser and **Figure 2** shows a typical broad-area device. The important features of a GRINSCH SQW laser are a high-quality waveguide structure, which is the parabolically-graded AlGaAs region, approximately 0.4 µm wide, for the confinement of the laser light; and a very narrow (60 to 120 Å) GaAs quantum well which effectively concentrates the electrons and holes, thereby providing the maximum optical gain for a given amount of injection current. The typical emission wavelength for a GaAs GRINSCH SQW laser is in the range between 820 nm to 850 nm; narrower quantum wells give shorter wavelengths. By using a quantum well of $In_yGa_{1-y}As$ instead of pure GaAs, the band gap of the quantum well is reduced with increasing indium mole fraction y, and the wavelength of the emitted light is correspondingly increased. Diode lasers of this type emitting wavelengths as long as 1 µm have been reported in the literature.[3] This wavelength range has a number of applications for LLNL programs such as satellite-submarine communications, and because of their extremely high efficiency, small size, and ruggedness these devices are ideal for the optical pumping of neodymium lasers or other solid state lasers.

Progress

In FY89 we grew and fabricated GRINSCH-SQW diode lasers with measured current thresholds of ≈300 A/cm^2. **Figure 3** is a plot of light power versus drive current for one of our lasers with a length of 350 µm and a width of 100 µm; the threshold current of 100 mA corresponds to a current density of 290 A/cm^2. Our best growth demonstrated threshold current and efficiency superior to that of the best wafers which we had purchased from commercial sources. We have also started growing our films with arsenic from an arsenic cracker, which is necessary for the growth of high-quality $In_yGa_{1-y}As$ strained layer quantum wells. The $In_yGa_{1-y}As$ crystal structure has a slightly larger lattice constant than that of pure GaAs, but

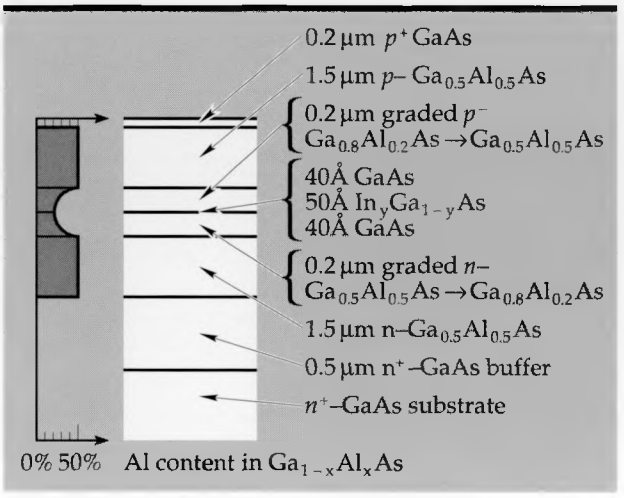


Figure 1. Schematic drawing of the layers which comprise a GRINSCH SQW $In_yGa_{1-y}As$ diode laser. The parabolically-graded AlGaAs region acts as a waveguide structure for the confinement of the laser light.

the quantum wells are thin enough so that the $In_yGa_{1-y}As$ undergoes a coherent strain to match the lattice of the rest of the crystal, hence, the term "strained layer." We have grown a number of films with $In_yGa_{1-y}As$ single or multiple quantum wells, and examinations of the films with high-resolution cross-sectional transmission electron microscopy show these layers to be of high crystalline quality.

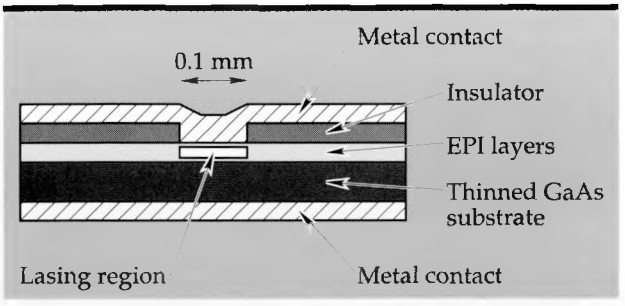


Figure 2. Typical broad-area diode laser. The lasing region corresponds to the graded bandgap layers in Fig. 1.

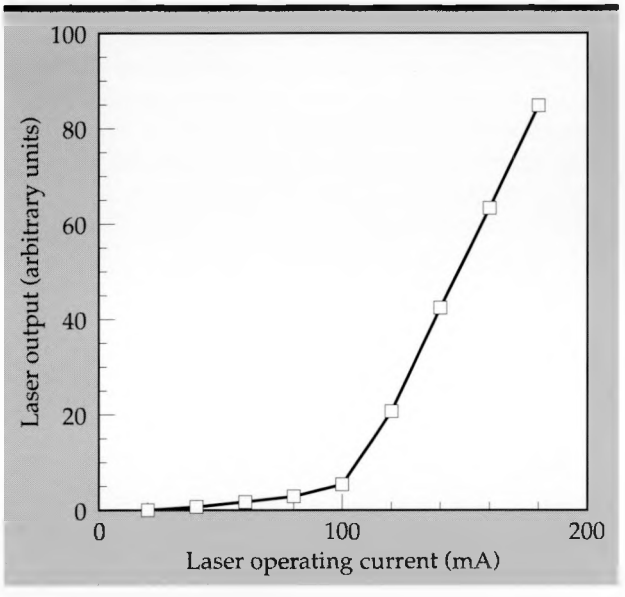


Figure 3. Plot of output power versus drive current for a laser made from our MBE grown material. The sharp "kink" at 100 mA corresponds to a lasing threshold of 290 A/cm².

Future Work

We will continue our work on GRINSCH SQW diode lasers with $In_yGa_{1-y}As$ strained layer quantum wells to provide laser emission wavelengths falling in the range from 0.9 μm to 1.0 μm. In particular, we will examine a variety of growth parameters such as growth temperature, reactant fluxes, and bandgap grading which affect the overall performance of these lasers. The bandgap grading is particularly interesting, since the present growth recipe generates a parabolic profile with 60% aluminum on the outside and 20% aluminum next to the quantum well. This structure has set world performance records for a GaAs quantum well, but may not be optimal for $In_yGa_{1-y}As$ quantum wells. Specifically, lowering the inner aluminum content to 15% or 10% may improve device performance with the smaller bandgap $In_yGa_{1-y}As$ material in the quantum well. Also, we will develop fabrication and characterization techniques so that we can support the other work being done at LLNL in diode lasers. Specifically, we will develop the metalization procedures necessary to produce bars of lasers which can be used by Y-Division in their high-power heat sinking program. Other specialized fabrication procedures to produce lasers for classified programs will be developed as well.

1. D. Mundinger, R. Beach, et al., *Appl. Phys Lett.*, **53,** 1030 (1988).
2. D. Mundinger, private communication.
3. L. E. Eng, et al., *Appl. Phys Lett.*, **55,** 1378 (1989)

Wavelength Deflection Technique for Recording Microwave Signals

Charles F. McConaghy
Engineering Research Division
Electronics Engineering

During FY 89 we have continued investigating an optical recording technique for single-shot electrical signals with bandwidths exceeding 10 GHz. The technique consists of optically encoding the electrical signal's time history into optical wavelengths, and then deflecting the optical signal across a readout array with a diffraction grating. During FY 88 we showed successful recording of repetitive signals having bandwidths up to 20 GHz with this technique. Our FY 89 goals were to characterize and minimize overall system losses, with the objective of recording single-shot signals. An additional goal has been research into chirped laser diode sources as a substitute for the large table-top dye laser used in our earlier experiments. Progress was made in both of these areas.

Introduction

We are exploring wavelength deflection as a method for recording fast electrical signals having very wide bandwidths. The basic principle is to encode each time point of an electrical waveform to a unique optical wavelength.[1] From a practical standpoint, the encoding is accomplished with an optical carrier wave provided by a chirped laser whose wavelength varies linearly over a time interval equivalent to the duration of the signal to be recorded. This chirped carrier is amplitude-modulated with the fast electrical signal by means of an integrated Mach-Zehnder modulator and then transported over a fiber. Readout and recovery are accomplished with a spectrometer which deflects the incoming light across an imaging array with a diffraction grating **(Fig. 1).**

Other optical encoding schemes that transport microwave-bandwidth electrical signals are based on either direct or indirect modulation of a laser diode. The direct modulation of a laser diode is limited to a bandwidth of about 10 GHz and must

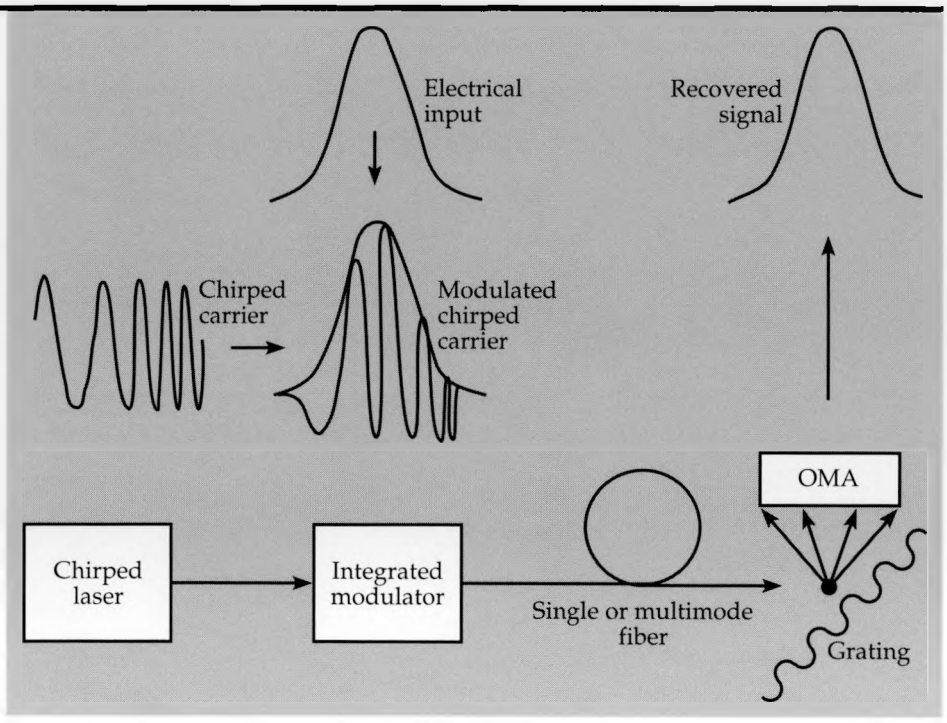


Figure 1. Wavelength-deflection recording process. A chirped laser and an integrated modulator provide a way to encode the electrical waveform. The signal is recovered by deflecting the incoming light with a diffraction grating across an optical multichannel analyzer (OMA).

Figure 2a shows a single-shot picture of a single pulse and **Fig. 2b** shows a much cleaner trace of a large number of these pulses. We have seen a large amount of variation in the noise spectrum from pulse to pulse when observing in the single-pulse mode. We believe that these variations are minimized when integrating a large train of pulses in a repetitive mode. At the present time, the noise spectrum and pulse-to-pulse spectrum variations of individual laser pulses are the main limitations in obtaining high quality single-shot data. In spite of these limitations, we were able to capture some single shots with the wavelength deflection system. **Figure 3** shows a single shot of an 8-GHz sine wave captured by the wavelength deflection system. Most of the noise on the sine wave is due to the spectral noise of the chirped carrier. Presently we are looking into techniques that we hope will minimize the spectral noise. One of these techniques is to narrow the laser spectrum by a birefringent filter, to increase its energy by cavity dumping or a dye amplifier, and then to use this higher peak power pulse to get increased bandwidth by self-phase modulation in a short fiber.

Energy Requirements and Losses

All photonic systems have optical losses. This requires us to provide enough energy in the original optical carrier to make up for the photons lost in the various couplers, modulators, gratings, fibers, mirrors, etc. that may be in an overall photonic system. The number of photons that are needed at the point in the system where they are converted back to electrons is a function of the desired signal-to-noise level. In the wavelength deflection technique, this conversion is accomplished with an OMA at the back end of the spectrometer. The OMA consists of a linear array imager preceded by a single or dual microchannel plate intensifier. The intensifier provides enough gain to achieve the shot noise limit and is capable of seeing down to the single photon level. In order to achieve the shot noise limit, it is important that other noise sources such as dark noise in the readout, clock noise in the readout, or dark noise in the intensifier are negligible. By gating and using a very high gain intensifier, it is possible to neglect these noise sources. Assuming the shot noise limit, the amount of energy needed in the optical carrier to obtain a specific signal to noise ratio is given by

$$E_{total} = \text{Losses} \times E_{photon} \times N_{points} \times (S/N)^2/QE \ .$$

E_{photon} is the photon energy (at 810 nm, E_{photon} is 1.53 eV), N_{points} is the number of data points, S/N is the desired signal to noise, and QE is the quantum efficiency of the device that converts the photons to electrons (in an intensified OMA this is a photocathode which might have a QE of 2% at 810 nm). To achieve a S/N of 100 with 100 data points and a system loss of 20 dB will require 1.2 nJ of laser energy. This number can be reduced by improving the QE or by reducing the system losses. Recently we acquired a microchannel plate with 10 % QE at 810 nm which should permit us to reduce the laser energy by a factor of ≈5.

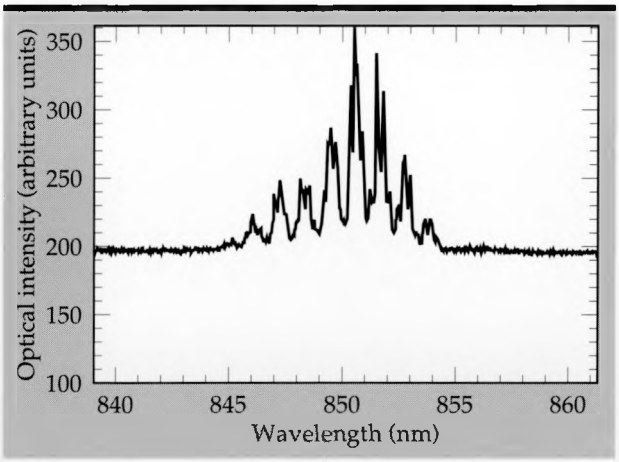


Figure 3. Single-shot picture of an 8-GHz sine wave using the synchronously-pumped dye laser for the chirped optical carrier.

Figure 4 shows a block diagram of the major components of the wavelength deflection system together with the optical loss of each component. The first type of optical loss is L_c, the coupling loss from the laser into the fiber. The second is L_{cfiber}, which is the loss in the chirping fiber. The third is L_{mod}, which is the loss in the integrated modulator. The fourth is L_{tfiber}, which is the loss in the fiber that transports the optical signal to the spectrometer. The fifth term is L_{spec}, which is the loss in the spectrometer and includes the grating,mirrors, and input coupling losses. Therefore, $L_{total} = L_c + L_{cfiber} + L_{mod} + L_{tfiber} + L_{spec}$.

We measured the losses of each component individually, and the results are shown in **Fig. 4.** Best and worst case numbers are dependent on several variables. For the chirp fiber and modulator, increased losses of at least 1 dB were observed if connectors were used. Another variable that affects the losses of the modulator and spectrometer is the polarization of the light, since both devices are polarization-sensitive. Currently the chirp fiber is Corning Glass FLEXCOR 850 with a total length of 1 km. A better choice would be to use a polarization-preserving fiber, but this was not obtain-

be done at optical wavelengths of 1300 or 1500 nm in order to circumvent fiber dispersion. Unfortunately, the photocathodes used in both streak cameras and microchannel plate (MCP) intensifiers are nonfunctional at these wavelengths, which consequently limits the recording of these signals to detector/oscilloscope arrangements. One advantage of the wavelength encoding technique presented here is the ability to maintain high bandwidths in fibers where dispersion might otherwise limit bandwidth.

Once the signal is wavelength-encoded at the Mach-Zehnder modulator, it can then be propagated over long fibers (1 km). In fact, the fiber can even be multimode. This has allowed us to propagate signals up to 20 GHz in bandwidth down 1 km of multimode fiber at a wavelength of 820 nm. Successful transmission of this bandwidth on such a long run of multimode fiber is unique. Another advantage of the wavelength-encoding technique is the potential to achieve better time resolution than a conventional streak camera. Subpicosecond resolution is possible with a trade-off in record length and number of data points. Yet a third advantage is a simplified detector arrangement consisting of a relatively slow-speed MCP and imager, without requiring either a streak camera or oscilloscope.

Progress

During FY 88 we demonstrated the feasibility of the wavelength-encoding scheme in our lab with a dye laser. These early experiments were done repetitively by synchronizing our chirped laser to the electrical signal to be recorded. This work was reported in the FY 88 Engineering Research and Development report.[2]

In FY 89 our goal has been to demonstrate wavelength-encoding technology in single-shot experiments. To achieve this goal, it was first necessary to isolate a single pulse from our repetitive laser. For most of our experiments, we have used a synchronously-pumped dye laser (Stryral 9 dye) which has a pulse width of 500 fsec at a nominal wavelength of 820 nm and a repetition rate of 82 MHz. A Pockel cell was used to isolate a single pulse. After some experimentation we found that the Pockel cell was not sufficient by itself to yield a single pulse. The Pockel cell has a rejection ratio of 1000:1 between the single switched pulse and the rest of the unswitched 82-MHz train of pulses. However, the residual unswitched pulses actually had considerably (1000 times) more total energy than the single switched-out pulse we were interested in using for our single shot measurements, because of the large number of unswitched pulses that occur during the optical multichannel analyzer (OMA) window time.

To minimize the effect of the unswitched background pulses, a gated microchannel plate intensifier was used in front of the OMA array to further attenuate the unswitched pulses. The gate time was not extremely critical as long as it was kept below 100 nsec, since the main purpose was to further reduce the 82-MHz signal the OMA would otherwise have integrated over its long window time of 10 msec. An additional benefit of the short gate pulse is a great reduction in the thermally-generated electrons which are emitted by the photocathode of the microchannel plate if it is left on for a long period of time. In fact, gating for single-shot events allows the noise level to be minimized to the shot noise limit where the noise is proportional to the square root of the signal. We found that the spectrum of individual laser pulses was considerably noisier than that of the repetitive pulse train that we had previously observed.

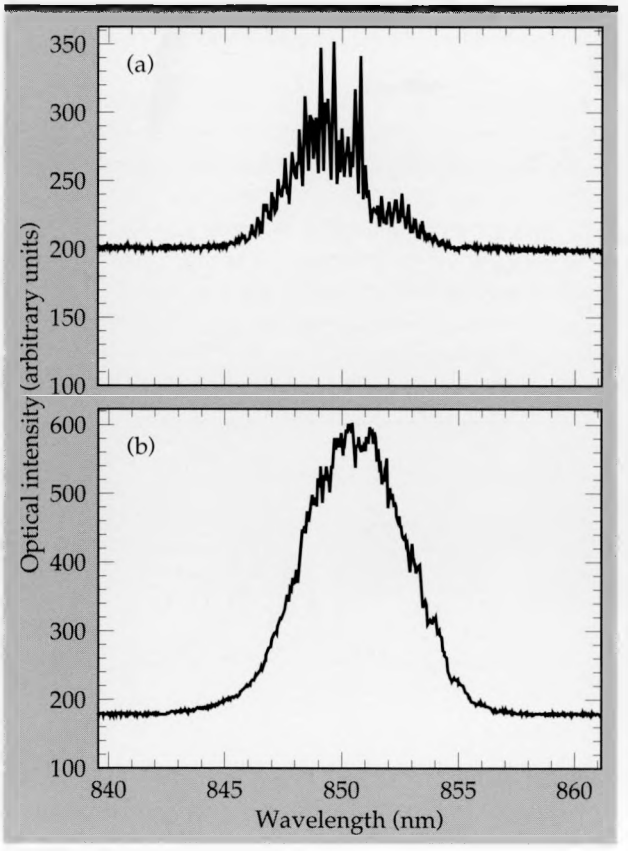


Figure 2. (a) Optical spectrum of a single pulse from a synchronously-pumped dye laser. (b) Optical spectrum of multiple pulses from a synchronously-pumped dye laser.

able. Some decrease in losses still might be possible at the modulator as better devices become available. Another possible loss term that could be decreased is the input coupling loss, which is highly dependent on the type of laser used as well as the optics selected to gather the light into the fiber.

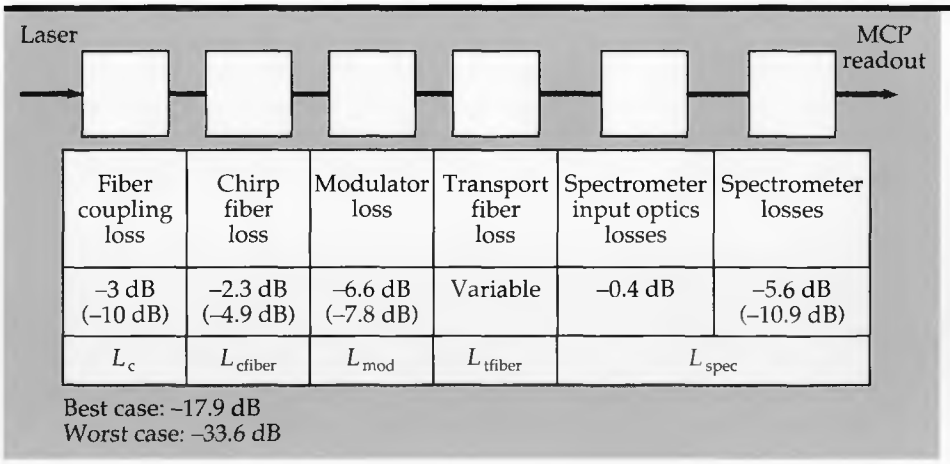
Chirped Diode Lasers

The key element in encoding time to optical wavelength is the chirped laser source. There are two ways to generate this chirped optical carrier. One way is to use a laser that can be electrically tuned in wavelength over the time period of interest (1 to 10 nsec in this case). An alternative approach is to start with a laser pulse that is sufficiently short and spectrally wide. This laser pulse is chirped by coupling it into a single-mode fiber which will give a linear chirp with respect to time based on the group velocity dispersion of the fiber. This technique makes use of the material dispersion of the fiber and does not require the laser pulse to be of high peak power. This latter approach is what we used exclusively in all the experiments with the wavelength deflection system.

Even though no experiments have been done with electrical wavelength-tunable lasers, it is worth commenting on this approach. There is no commercial source of wavelength-tunable diode lasers at this time. However, a great deal of research is going into these devices in Japan, Europe, and the U.S.[3] Most of the commercial interest in these devices is for wavelength division multiplexed (WDM) systems. Some of the results achieved show continuous tuning of over 5 nm by varying the current to the various tuning sections of the diodes. These devices, if available, could be used to provide a chirped source for our applications. One benefit of the wavelength tunable diodes would be an easy way of changing record lengths, simply by varying the time of the current tuning ramp. These diodes should also put out a relatively flat amplitude spectrum as they are tuned. Other advantages are that polarization can be maintained, and that the optical loss of the chirp fiber would be removed from the system. Tunable diodes probably will not achieve the time resolution that we have obtained with our short pulse/fiber chirped laser, but will probably meet the need for record lengths of 1 nsec or longer.

Although wavelength-tunable diodes have not yet reached the commercial market, we have been investigating a Q-switched diode laser as a possible substitute for our dye laser in the short pulse/fiber method of generating the chirped optical carrier. Soviet researchers have reported on a laser diode that has an integral saturable absorber within the laser cavity.[4] These diodes have produced pulses as short as 5 psec with peak power up to 10 W. The other interesting attribute of these diodes is broad optical spectrum (as wide as 20 nm). This type of laser, when coupled to a kilometer of fiber, could provide a fairly compact, easy-to-use chirped source.

During FY 89, development of the Q-switched diode was pursued from two directions. We have purchased a specimen of the only commercially available Q-switched diode. In addition to our evaluation of this diode, we have begun work on fabricating our own devices. The Q-switching is accomplished by holding off lasing by means of a saturable absorber. This allows a much larger population inversion to build up than in a conventional laser diode. Saturable absorbers have been built in two different ways. The best results have been obtained by building a diode with three separate sections. Two of these sections are pumped under forward bias, while the third section is reverse biased and acts as a voltage adjustable satu-



Fiber coupling loss	Chirp fiber loss	Modulator loss	Transport fiber loss	Spectrometer input optics losses	Spectrometer losses
–3 dB (–10 dB)	–2.3 dB (–4.9 dB)	–6.6 dB (–7.8 dB)	Variable	–0.4 dB	–5.6 dB (–10.9 dB)
L_c	L_{cfiber}	L_{mod}	L_{tfiber}	L_{spec}	

Best case: –17.9 dB
Worst case: –33.6 dB

Figure 4. Optical losses of the various components in the wavelength deflection system.

rable absorber. This is the type of device we have begun fabricating. The commercially available diode implements the saturable absorber by damaging one end of the diode with high energy ions (16-MeV oxygen ions). These ions form traps for some distance in from the end facet and this region of traps acts as the saturable absorber.

Experiments were done to measure the performance of the commercially-built laser diode. Streak camera traces showed that the diode has a 10-psec pulse width, and energy measurements indicated a 10-pJ output, which roughly corresponds to a 1-W peak power output. We used a negative-going 370-psec pulse of about 30 V to drive the diode. If the drive pulse is increased to higher amplitudes, the diode will pulse more than once during the drive time. One, two, or three pulses were observed, depending on the amplitude of the drive pulse. The interpulse time averaged about 100 psec but was dependent on the drive level.

The laser diode was coupled into a fiber to form a chirped carrier. The output was displayed on a spectrometer with streak camera readout so that both the spectral and temporal characteristics could be observed. **Figure 5** shows the results. The first thing to note is the straight-line dependence between time and wavelength. This is due to the material dispersion of the fiber and is the type of characteristic we need for the chirped source. Note also that the spectrum is quite broad (at least 10 nm). A third point of interest is that the spectrum is noncontinuous (holes are present across the 10 nm of bandwidth). The spectrum was also seen to change somewhat from pulse to pulse. Although we were able to record data up to about 12 GHz using this diode, it is not ideally suited for use as a chirped source because of the holes in the spectrum. These holes would translate to missed portions of the record when capturing data. However, the Soviet research indicates that it might be possible to obtain a continuous spectrum by using diodes where the saturable absorber is implemented as a reverse-biased section. We hope to verify this finding as we study our own multisection diodes.

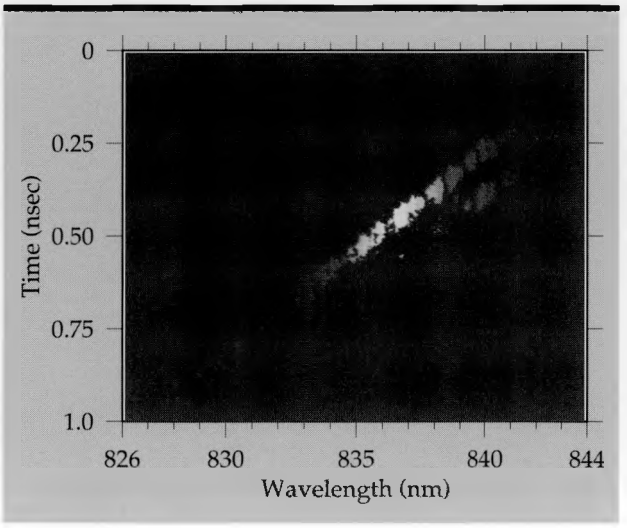


Figure 5. Single-shot recording of both time and wavelength for a Q-switched laser diode coupled to 1 km of single mode fiber.

Future Work

The FY 89 work has focused on establishing single-shot performance. To achieve this, we have measured and then minimized the optical losses of the various components of the wavelength deflection system. In addition, we were able to isolate single laser pulses from our sync-pumped dye laser and observe their optical spectrum. We now know that the present laser system lacks spectral stability on a single-pulse basis.

The main emphasis for future work will be to obtain a more stable spectrum for single-shot applications. For table top lasers such as our sync-pumped dye laser, we believe this can be accomplished by increasing the energy of the pulses by either cavity dumping or post-amplification. These higher power pulses could be used with a fiber to obtain a nonlinear self-phase modulation in a short fiber. This would increase the spectral bandwidth with the object of obtaining a cleaner spectrum than that available from the basic dye laser. Another possible way to obtain a more stable spectrum is to start with a more stable laser such as a colliding pulse laser (CPM). These lasers can generate pulses as short as 50 fsec. One disadvantage of the CPM is that the wavelength of operation is in the visible area where fiber losses are much higher. In addition, the Mach-Zehnder modulators would have to be specially fabricated for the visible range, since we have to date made devices only for 810 and 1300 nm. If the CPM laser were pursued, it might be worthwhile to try to build an infrared version, which has been shown to be possible by some researchers.[5]

Work is expected to continue on Q-switched laser diodes as a possible diode candidate for the chirped carrier. Currently a three-electrode device is in fabrication. There is still a fair amount of research to be done on these devices to explore how the saturable absorber works to generate both a short pulse and a wide spectrum.

Acknowledgments

The success of our efforts this year has been the combined result of a number of individual efforts. Glenn McWright and Don Lafaw provided the Mach-Zehnder integrated optic modulators. Jackson Koo provided a great deal of the design work on the Q-switched diodes, and Stacy Lehew has been responsible for the fabrication of these diodes. Gregg Lancaster and Mark Lowry provided the streak camera/spectrometer for the measurement of the data shown in **Fig. 5.** Finally, Brad Ross provided technical support to all aspects of our experiments.

1. J. A. Valdmanis, "Real-Time Picosecond Optical Oscilloscope," *Ultrafast Phenomena* **5**, 82 (1986)
2. C. F. McConaghy, "Wavelength-Deflection Technique for Recording Microwave Signals," Lawrence Livermore National Laboratory, Livermore, CA, *Thrust Area Report FY88*, UCRL-53868-88 (1988).
3. S. Murata, I. Mito, and K. Kobayashi, "Over 720 GHz frequency tuning by a 1.5 μm DBR laser with phase and Bragg wavelength control regions," *Electronics Letters* **23**(8), 403-405 (9 April 1987).
4. P. P. Vasil'ev, "Picosecond injection laser: A new technique for ultrafast Q-switching," *IEEE J. Quantum Electronics* **24**(12), 2386-2391 (1988).
5. W. H. Knox, "820nm colliding-pulse modelocked dye laser pumped by a cw Ti:sapphire laser," *Proc. of CLEO 1989, April 24-28, 1989* PD31-1 (Baltimore, MD, 1989).

Calibration of Optical Transient Detectors

Philip M. Armatis and
Sterling J. Andreason
Field Test Systems Division
Electronics Engineering

Many Laboratory programs require the calibration of devices that measure the energy content of optical pulses. To develop a calibration procedure, we studied techniques for the measurement of transient light sources and the calibration of suitable detectors. A method was developed to calibrate commercially-available energy (joule) meters with an accuracy traceable to the National Institute of Standards and Technology (NIST).

Introduction

Recent improvements in optical standards, detectors, equipment, and sources have created a need for greater accuracy in the Laboratory's radiometric techniques. For this reason, the Laboratory's Optoelectronic Calibration Facility (OCF) has expanded its activities to assist various Laboratory programs with radiometric and photometric calibration.

The accurate calibration of silicon detectors used to measure the energy of optical pulses is of particular interest to many Laboratory research activities. However, the manufacturer's calibration of commercially available silicon-detector joule meters widely used at the Laboratory has often been inadequate. We have therefore developed a method of optical pulse calibration for use at the Laboratory, and present here a progress report on the development of a reliable means for calibrating silicon detectors used for transient optical measurements. We discuss the traceability of the radiometric standard to NIST, the electronically-calibrated pyroelectric radiometer (ECPR) on which the procedure is based, and the operating principles of the joule meters for which calibration is required. Also described are the test setup, certain problems that may be encountered during the calibration procedure, and limitations of calibration accuracy.

ECPR Operating Principles

The ECPR (see **Fig. 1**) senses optical energy with a thermal detector which can be activated either by the heat from the absorption of an optical pulse, or the heat from an electrical current passing through the detector. The responses of the detector to these two different types of heating are nearly identical, enabling the direct comparison of incoming optical

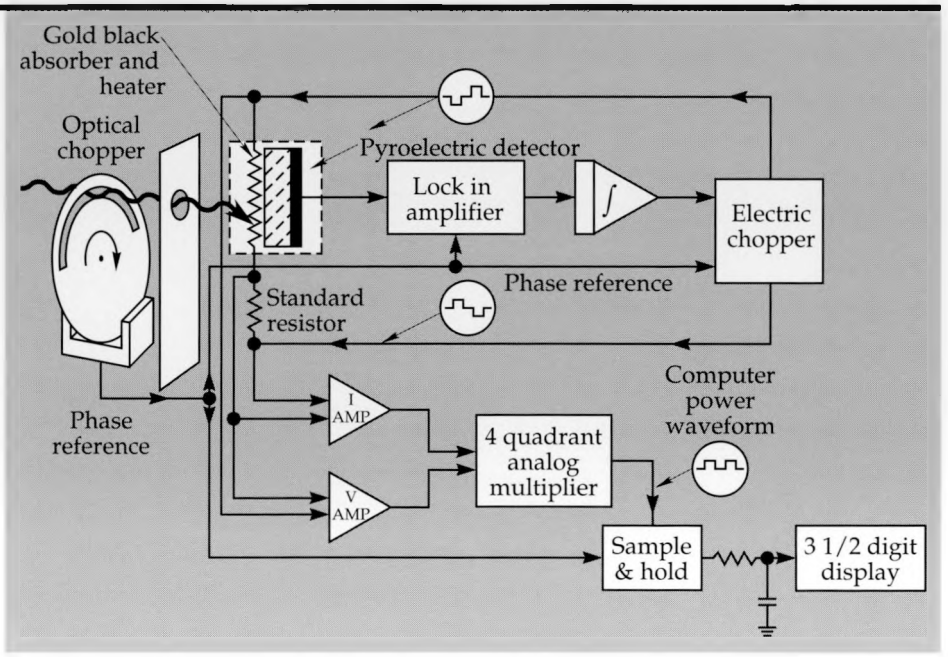


Fig. 1. Block diagram of the electrically calibrated pyroelectric radiometer (ECPR), showing the null balance servo loop (above standard resistor) and power computing circuitry.

power to units of electrical power. Since electrical standards are highly accurate and readily available, this type of radiometer can be accurately calibrated and readily traced back to NIST. The ECPR thus provides the radiometric standard on which our measurements are based.

The pyroelectric detector in this system is a lithium tantalate pyroelectric detector crystal with a black gold surface film. The incident radiation to be measured is modulated with an optical chopper before striking the surface of the detector. During the period that the incoming optical energy is blocked by the chopper, an electrical signal is applied across a resistive heater on the detector surface. The power of this electrical signal is kept equal to the incoming optical power by a closed servo loop, thus providing the basis for power measurement. Associated power computing circuitry drives a digital output display.

Energy Meter Operating Principles

The commercially-built energy meters for which calibration is required operate on a different principle than the ECPR. Instead of measuring incoming optical power indirectly by means of a servo-matched electrical signal, the energy detectors in these meters directly generate a voltage step function proportional to the total energy of a pulse of optical radiation. Both silicon and pyroelectric energy probes are available for these meters.

The pyroelectric detector consists of a thin slice of ferroelectric material with a permanent electric polarization that is highly dependent on temperature. As incoming optical energy is absorbed by the detector coating, an electrical current is generated that is proportional to the total temperature increase of the detector, and hence to the total energy of the optical pulse. The detector current is integrated, and the resultant step height of the signal is proportional to the input energy.

The silicon detector generates an electrical current through the formation of electron-hole pairs due to the absorption of photons. Again, this current is directly proportional to the number of incident photons, and therefore the incident power. As with the pyroelectric detector, the silicon detector current is integrated, and the resulting signal is proportional to the input energy.

Calibration Procedure

The test setup is shown in **Fig. 2**. A beamsplitter arrangement permits the output of a single laser to be directed at both the reference device (the ECPR) and the joule meter to be calibrated. The energy E entering the joule meter probe under test may be computed by

$$E = P\,(Rs/Rd)\,(d/f)$$

where P is the power measured at the ECPR, Rs/Rd is the beamsplitter ratio (which may not be precisely 50%), d is the joule meter chopper duty

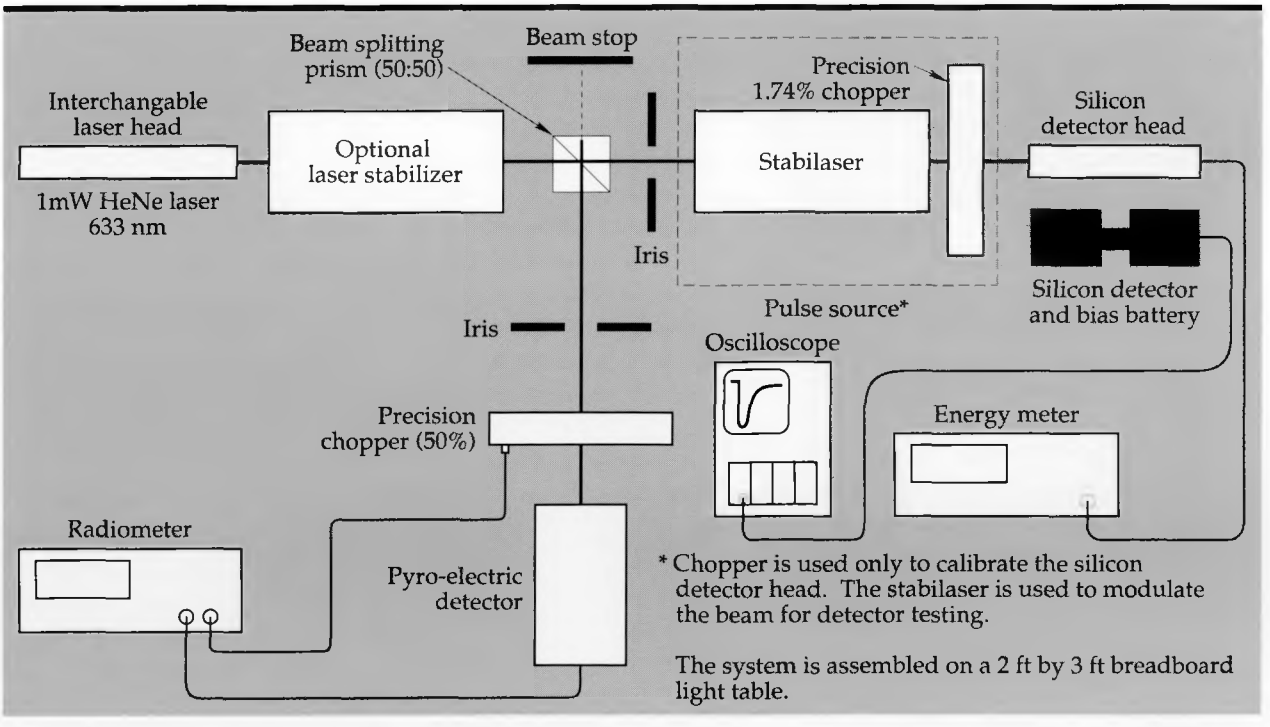


Fig. 2. Test configuration for calibration of joule meter.

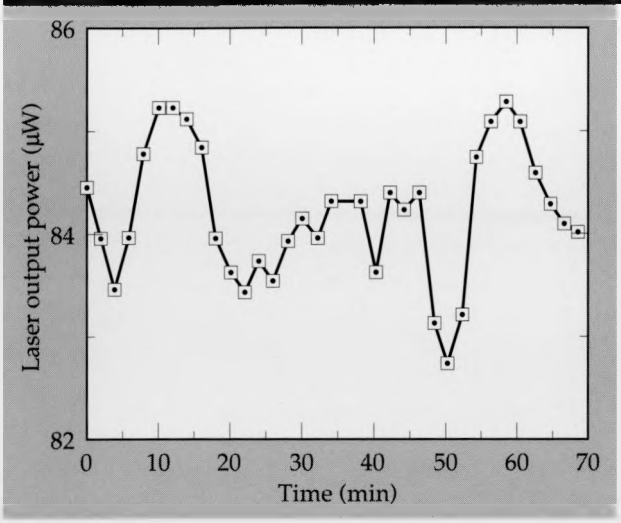


Fig. 3. Drift in calibration laser output power over time.

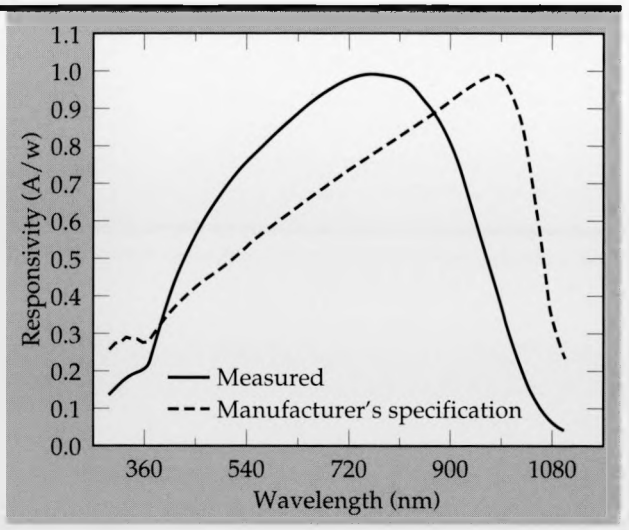


Fig. 4. Difference between actual measured spectral response of silicon detector and the manufacturer's specification.

cycle, and f is the joule meter chopper frequency. The probe gain control on the joule meter is then adjusted until the meter readout is correct.

This setup provides simultaneous measurements with the calibrated ECPR and the joule meter being tested. Source drift problems are thus avoided. Even after a laser has stabilized, we have found that output power can drift irregularly through a range of 2% (**Fig. 3**). We are aiming at an overall measurement accuracy of 5%, so source drift is a real issue.

Progress

Careful attention to detail is important for accurate calibration. All operating parameters must be checked for accuracy. For example, the ratio of the beamsplitter we used was determined by actual measurement to be 49.8/50.2 at 632 nm. The optical chopper at the input of the joule meter probe had a 1.74% duty cycle instead of the rated 2%, which if not corrected would have introduced an error of 13% into the calibration procedure.

The most significant source of error in the calibration procedure is the silicon probe itself. The response peak of one of the probes we calibrated is at 780 nm, not 940 nm as the manufacturer specified, and the spectral response curve is different in shape from that shown in the product literature (**Fig. 4**). This discrepancy could introduce calibration errors as high as 25%. Also, the meter's circuitry required modification to allow an acceptable range of probe gain control at HeNe wavelengths of 632, 612, 594, and 543 nm.

It is important to note that the calibration obtained with the procedure just described is valid only for the wavelength and energy levels used in the calibration procedure. The reason for the limited validity of the calibration is the varying spectral response of the silicon probe (**Fig. 4**). A typical pyroelectric probe has a flat response curve (%0.5%) from 400 nm to 3000 nm, and a calibration obtained at any frequency within this range will be valid over the entire flat portion of the response curve. However, the calibration of a silicon probe is valid only at the wavelength and gain used during the calibration procedure. The dynamic range of a silicon probe is also different from that of a pyroelectric probe such as that used in the ECPR. These problems currently limit the usefulness of our calibration procedure.

Future Work

Building on the results we have already achieved with a variety of laser wavelengths, we plan to expand and refine the OCF's pulse calibration techniques. Our goal is to obtain absolute irradiance or energy measurements over a silicon detector's entire spectral range, and to predict the errors introduced into our measurements from spectral simulations and approximations. The spectra to be measured range from individual spectral lines from lasers to broad-spectrum radiation similar to that emitted by fluors, phosphors, and flashlamps. In order to accomplish these measurements, we will use computer modeling, radiometric bandwidth normalization techniques, and experimental measurements for verifying predictive error analysis.

Ion Assisted Etching for Microelectronics and Microoptics

Dale E. Miller
Engineering Research Division
Electronics Engineering

The goal of this new, three-year project is to develop specific ion-beam-assisted etching (IBAE) and reactive-ion etching (RIE) processes to form precise, complex microstructures. These are needed to realize new microelectronic, optoelectronic, x-ray optic, and visible or near-infrared microoptic devices.

In FY 89 we acquired, installed, and customized a new vacuum system for etching (and deposition) with inert or reactive ion beams. We performed initial IBAE experiments on single-crystal surfaces of silicon and gallium arsenide. As required for microoptics and optoelectronics, we found conditions that preserve optical smoothness. Additionally, we modeled time-dependent heat transfer, due to ion bombardment, across layers of insulators and semiconductors typical of configurations used to make solid-state devices.

Introduction

Our prior engineering research and development effort on crystal-plane-selective, or anisotropic, wet etching has yielded techniques that have been very successfully used to make optical and electronic microstructures. Examples are patterned membranes of silicon or silicon nitride that function as laser-alignment targets and x-ray windows, high aspect-ratio grooves for linear and circular zone plates for spectroscopy and imaging, and sharp silicon pyramids for field-emitter electron sources. Prototype micromechanical devices, including a patented latching accelerometer that measures forces greater than 1,000 G, have also been made with wet-chemistry technology. This technology has been developed by LLNL and other labs for electronic integrated-circuit microfabrication.

However, while we are likely to continue to expand the types of silicon-based devices achievable through wet chemistry, we are reaching the limits of this technology for certain shapes and, more importantly, for whole classes of devices made with insulators and semiconductors other than silicon. Specifically, anisotropic etching produces silicon grooves with atomically smooth sidewalls but faceted, rough bottoms. Optical microstructures often utilize fused silica or other amorphous materials including polymers. Also, engineers in the Diagnostics and Microelectronics thrust area are developing semiconductor lasers and detectors made with gallium arsenide and related III-V materials.

Plasma-assisted "dry" etching, investigated by the semiconductor industry for more than a decade, has now replaced wet chemistry for many of the etching and pattern transfer operations in the manufacture of microelectronic circuits. For our current microelectronic device research and several programmatic development projects, there is also an increasing dependence on plasma-assisted etching. Such dry etching is done by ions and reactive neutral radicals from a plasma in which the device substrate is immersed. We use these chemically active ions and neutral species in two basic schemes. One, called plasma etching, uses a glow discharge without accelerating potentials to yield omnidirectional or isotropic etching. This process is similiar to most wet etching but is effective on a wider range of materials. The other, reactive-ion etching, achieves vertical sidewalls for grooves and holes by creating a potential of several hundred volts between the plasma and the substrate being chemically etched.

Although RIE has become the predominant method of materials processing for integrated circuits, the process is still poorly understood because its complex, coupled, and nonequilibrium plasma conditions are difficult to measure or to model. We seek specific reaction data on the effect of various plasmas and etching materials. Since the experimental parameters are independently controllable in the IBAE system, specific reactions are more easily studied than in an RIE apparatus. Consequently, from our IBAE studies, we expect to gain information that will benefit RIE operations. The

experimental reaction data will allow more precise use of our new IBAE system and three existing RIE systems.

IBAE is a newer form of dry etching that does not require the device substrate to be immersed in a plasma. The substrate is isolated from the plasma in a separate high-vacuum chamber. Etching is done by a collimated ion beam, extracted from a gridded plasma chamber called a Kaufman ion source. Consequently, among other advantages, substrate temperature and charging are more readily controlled.

Our IBAE system is more advanced than RIE systems in several respects: its dual beams and separate high-vacuum chamber enable greater flexibility in forming shapes. In its simplest mode of operation, called ion-beam etching (IBE) or ion milling, we use nonreactive (argon or xenon) ions alone at energies of a few hundred eV to physically remove (sputter) substrate atoms by collisional energy transfer in a billiard-ball-like process. In a plasma system, the same process is called sputter etching. An important advantage of IBE is that the substrate can be rotated relative to the direction of the etching ion beam to give well-defined oblique sputtering. This allows versatile control of slopes and angled notches that are needed, for instance, for some designs of microlenses and mirrors.

Research has shown the introduction of ions or neutral molecules of reactive gases dramatically improves the overall rate of ion-beam etching. Besides overall rate, anisotropy in etching is needed to realize complex shapes (or just smooth vertical sidewalls of a microgroove). It depends on both ion momentum and chemical reactivity. Moreover, experience with RIE suggests that, to also achieve the significant etching-rate selectivity required for many microstructures made from two or more materials, chemically active gases must be incorporated in the etching process. Quantitative selectivity must be experimentally determined. We have modified our IBAE system so that we can introduce one or more of eight gases, either as part of the ion beam, or, as reported to be even more effective, in the immediate vicinity of the substrate to be etched. For future LLNL programmatic applications, materials to be patterned by IBAE include silicon, silicon oxide, silicon nitride, boron nitride, carbon, gallium arsenide and other III-V compounds, plus selected metals.

This investigation is expected to yield process conditions that will allow even more complicated microstructures in silicon than have previously been possible. Moreover, new IBAE and better-understood RIE techniques should increase our ability to create new devices out of optoelectronic materials (primarily gallium arsenide and related compounds) using ion-beam processing. Examples of new devices include high-efficiency lasers and laser arrays.

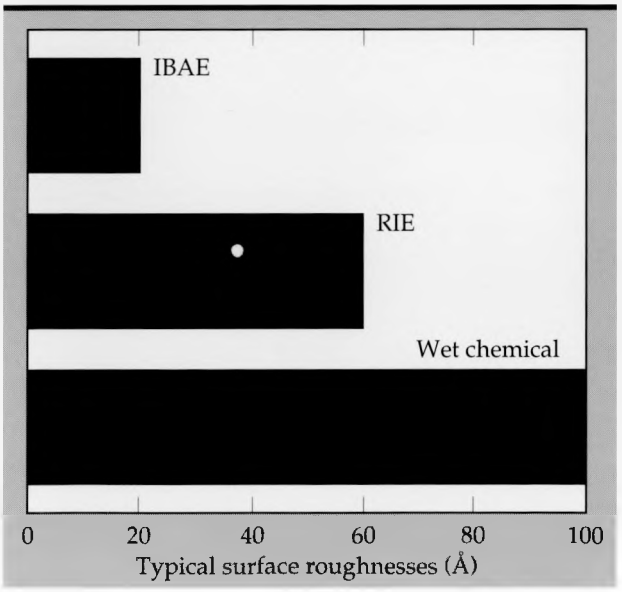


Figure 1. A comparison of typical surface roughnesses using three types of etching techniques: ion-beam assisted (IBAE), reactive-ion etching (RIE), and anisotropic wet chemical methods.

Progress

The IBAE system is now operational. It includes a special handling system, which we designed and installed, for eight toxic gases. Apparatus for advanced diagnostics including line-of-sight mass spectrometry and laser-induced fluorescence (LIF) were obtained.

Initial Ion-Beam Etching Experiments

Experimentally, we found conditions that preserve the subnanometer smoothness of single-crystal-silicon surfaces after noble-ion-beam removal of more than 100 nm of Si. We measured surface roughnesses of ion-beam, plasma, and anisotropic wet-chemical etched surfaces with a laser-interferometer system in the LLNL Optics Shop. **Figure 1** shows the comparison between these three methods. **Figure 2** shows that, although the optimal anisotropic wet chemical etch typically yields atomically smooth sidewalls, realization of smooth groove bottoms requires ion-assisted etching. Similarly, we demonstrated, as seen in the scanning electron microscope image of **Fig. 3**, that

trench bottoms more than 3 μm deep, which were formed in gallium arsenide by ion-beam etching, could maintain 10 nm smoothness. Furthermore, the morphology of the molecular beam epitaxy-grown GaAs sidewall exhibited both the desired slope for this laser-structure application and a direct dependence on the masking structures used.

Modeling of Ion Beam Heating

The common approach to controlling the temperature of an ion-assisted etching substrate, typically a wafer of Si or GaAs with oxide and polymer masking layers, is to mount the wafer on a water-cooled support made of a good heat conductor (e.g. copper). Yet only the temperature near the surface immediately adjacent to the support is effectively controlled this way because the temperature of the rest of the wafer and, especially, of the low thermal conductivity layers deposited on the wafer will be determined by the time-dependent heat flow from the opposite side bombarded by the ion beam. In a numerical analysis of the heating dynamics of multilayer structures with imperfect thermal contacts, most uncertainty derives from the quantitative model used for the contacts, at both the wafer-support interface and between deposited layers on the wafer. While the roughness of real surfaces can imply heat flow that is not necessarily one-dimensional near the interface, the lack of detailed knowledge of the microscopic geometry and thermal physics at a specific real interface precludes more elaborate modeling. In the absence of data, typical first approximations wrongly assume the contacts are atomically smooth. We estimated two one-dimensional parameters, roughness size and actual area of contact, as sketched in **Fig. 4,** to provide a more realistic model of the actual contacts. The purpose of the modeling was to outline critical conditions within which temperature control could be realized for a variety of wafer and layer thicknesses and materials and different contact integrities. Results are summarized in **Table 1.** The graph in **Fig. 5** is an example of calculated ion-beam heating of a thin wafer of fused silica coated with a high temperature polymer, as might be used in the formation of microoptics.

Future Work

Since laser-induced fluoresence diagnostics on the IBAE system would require variable short-wavelength laser sources for the IBAE non-plasma environment, we shall defer its implementation until we move into the new microelectronics facility. On the other hand, mass spectrometry of reactant molecules and background gases will be part of our experimentation in FY 90.

We will determine the physical and reactive conditions needed to form diffraction grating and microlens test structures, designed in collaboration with scientists in the Chemistry & Materials Science and Physics Departments, with fused silica and then with silicon. Fluorine-ion interactions are the principal study. In our experiments with

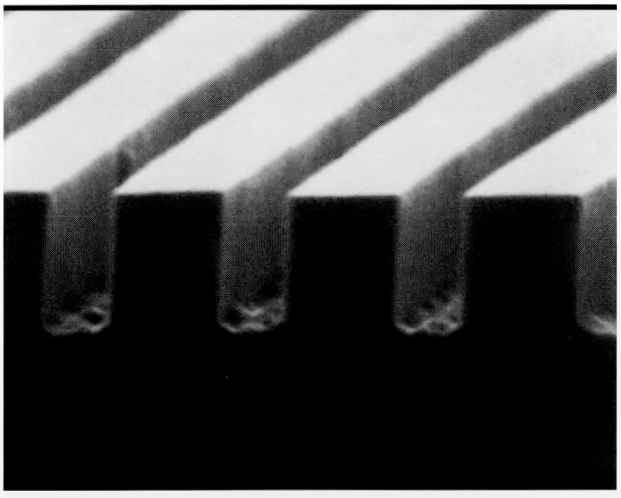


Figure 2. Scanning electron microscope image of grooves etched by anisotropic wet chemical etching and ion-beam etching illustrates the atomically smooth sidewalls and rough groove bottoms characteristic of microstructures formed by anisotropic wet chemical etching.

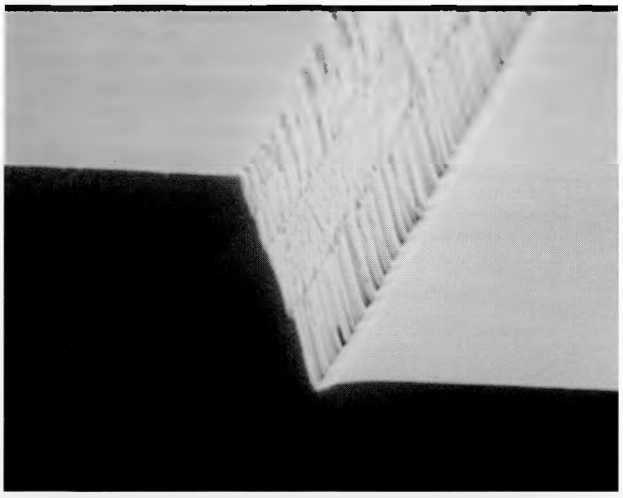


Figure 3. Scanning electron microscope image of gallium-arsenide trench bottoms more than 3 μm deep. The trench bottoms, formed by ion-beam etching, maintain 10 nm smoothness, and the morphology of the molecular beam epitaxy-grown sidewall exhibits both the desired slope for this laser-structure application and a direct dependence on the masking structures used.

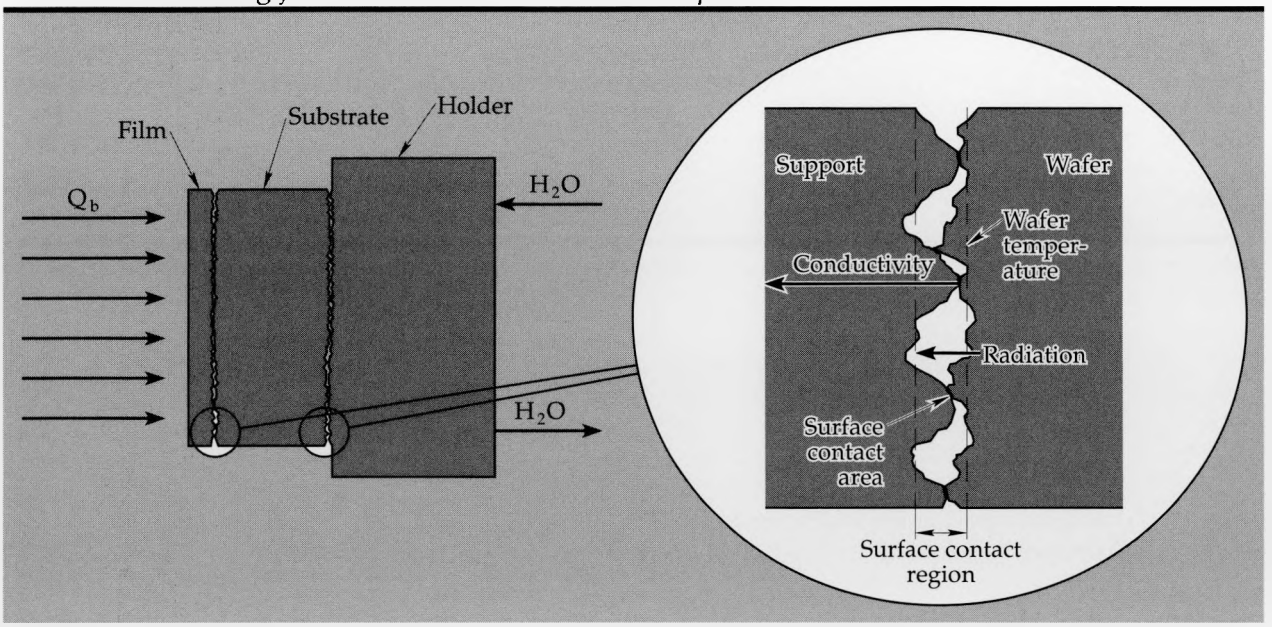


Figure 4. Temperature control of multlayer structures subject to ion bombardment. The control depends critically on thermal contacts between layers. At 1 mA/cm² (6.25×10^{15} ions/cm²) and 1,000 eV, the ion heat energy flux, Q_b = 1 W/cm².

plasma-based RIE processing on both silica and gallium arsenide, we will continue to seek masking techniques that will give better microstructures in both RIE and IBAE. Once basic techniques for making silica microoptics are established, we will commit the IBAE system to investigation of chlorine chemistry that will lead to formation of microstructures in GaAs-based compounds for optoelectronics. Specifically, we develop grooves, facets, gratings, and then refractive and reflective structures for surface emission from semiconductor lasers.

Acknowledgments

This first year was instrumentation intensive. Ken Vindelov was pivotal to getting the IBAE system to a functional status and determining the requirements for toxic gas handling. He also con-

Table 1. Relative sensitivity of wafer temperature to "downstream" contact parameters.

Wafer thermal conductivity / Wafer thickness	High	Low	
	Under 10 mm	Thick (>1 mm)	Thin (<1 mm)
Thermal conductivity of adjacent layer "downstream"	High sensitivity—layer controls composite contact conductivity	Low sensitivity—wafer controls contact conductivity	
Contact gap and absolute roughness	Low sensitivity with conductive adjacent layer, high sensitivity with resistive adjacent layer	Low sensitivity—wafer resistance far greater than contact resistance	High sensitivity—contact resistance far greater than wafer resistance
Wafer temperature control method	Direct with highly conductive support, temperature close to support temperature and stable	Indirect by adjusting beam intensity, temperature above support temperature, but stable	Depends on contact roughness. Temperature may fluctuate.

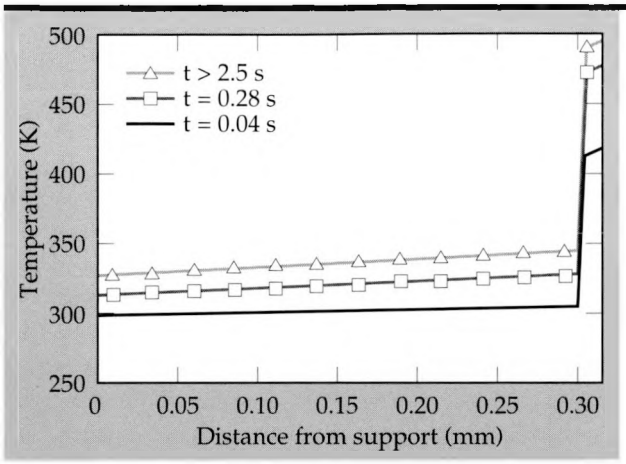


Figure 5. A graph of ion-beam heating of a 10-μm layer of polyimide on a 0.3 mm silicon-dioxide wafer with only 1% atomically intimate contact. Note the 200 K surface temperature increase. Parameters used to determine the necessary times were: heat flux = 8 W/cm², cooled support temperature of 300 K, the contact fraction at the support and at the coating = 0.01, absolute roughness at support = 10 μm, absolute roughness at coating = 5 μm. Note that the temperature remains steady over the distance from 0 to 0.3 mm and after hitting the surface contact region rises dramatically before leveling off.

ducted most of the experiments related to plasma-based etching, particularly on masking and selectivity issues with RIE. Jim Folta tackled the final design of the gas handling system, the safety documentation, and performed the ion-beam-etching smoothness experiments. Complementing the instrumentation and initial experiments efforts, Yury Zundelevich modeled the thermal responses of key types of substrates and film layers subjected to ion-beam bombardment.

1. D. E. Miller, "Thermally and Ion-Assisted Etching of Tungsten and Molydenum", Institutional Research and Development FY 88 , Lawrence Livermore National Laboratory, Livermore, Calif., UCRL-53689-88, pp. 161–162.
2. Y. Zundelevich, "Heat transfer across layered structures subject to ion bombardment", IAE internal project report (May 1989), available from the author.

Comparison of Electro-Optic Diagnostic Systems

Karla G. Hagans
Field Test Systems Division
Electronics Engineering

We have examined the component devices used in electro-optic transient recording systems and studied their contributions to system noise levels. This work includes the investigation of three system configurations: a 1300-nm laser diode system, an 810-nm Mach-Zehnder system, and a 1300-nm hybrid system. As a result of our work, we have developed a method of analysis for determining the theoretical noise sources in electro-optic transient recording systems.

Introduction

Currently, the design and construction of electro-optic transient recording systems is an imprecise exercise; that is, we assemble the various components into a diagnostic system and then accept the performance we get. A common assumption made in constructing such a system is that using components which individually exhibit the best performance will yield a system of higher performance. Many times, however, this assumption is not valid. For example, laser diodes are often a major source of noise in electro-optic transient diagnostic systems. Thus, for laser-noise-limited systems, it is very important to design the laser diode to have the lowest possible laser noise. This is not the case, though, for electro-optic diagnostic systems using the streak camera as the receiver. In these systems, the noise floor of the system is determined by the receiver; hence, improvements in the laser diode noise performance yield very little improvement in the system noise performance. One must look at the desired system performance and identify these system component tradeoffs.

A number of devices exist for performing the same functions in electro-optic transient recording systems, as shown in **Fig. 1**. For the modulator, one can choose either to modulate a laser diode directly or to modulate an integrated optics device. The laser diode has the advantage of being a simpler system to implement, but it suffers from wavelength instabilities in the 810-nm large-signal regime. The integrated optics modulator has the advantage of offering higher bandwidth than do laser diodes but at the cost of creating a more complex system that includes both a laser diode carrier and a modulator. There are also two choices for the wavelength of transmission. The 1300-nm systems operate at the optimum wavelength for the fiber optic cable, while the 810-nm systems are optimized to accommodate a streak camera as a receiver. Although the streak camera can record many data channels simultaneously, the quantum efficiency of the streak tube's photocathode falls off rapidly at wavelengths greater than 800 nm. (The

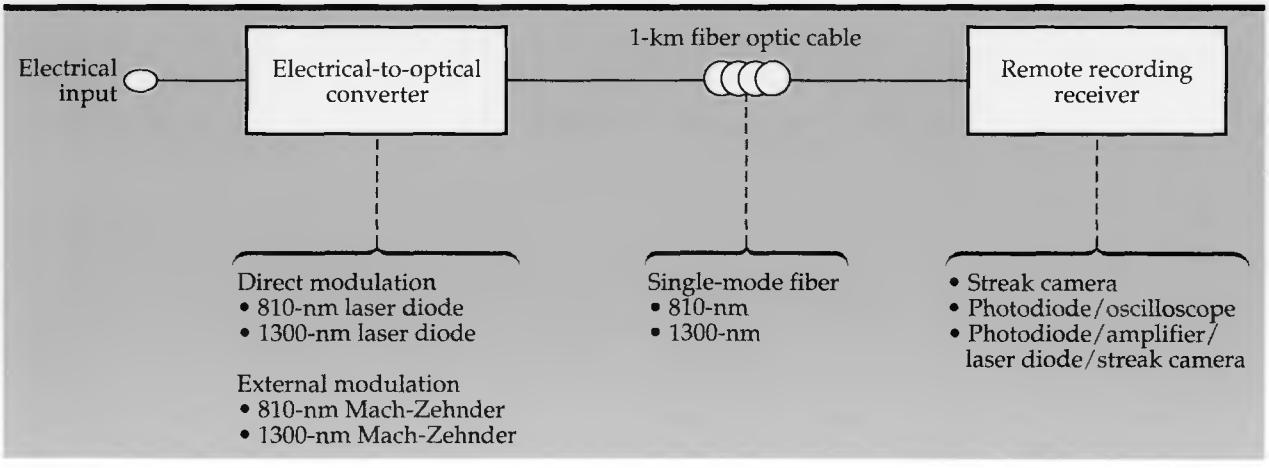


Figure 1. Component choices for electro-optics transient recording systems. Each choice impacts system complexity, cost, and signal fidelity.

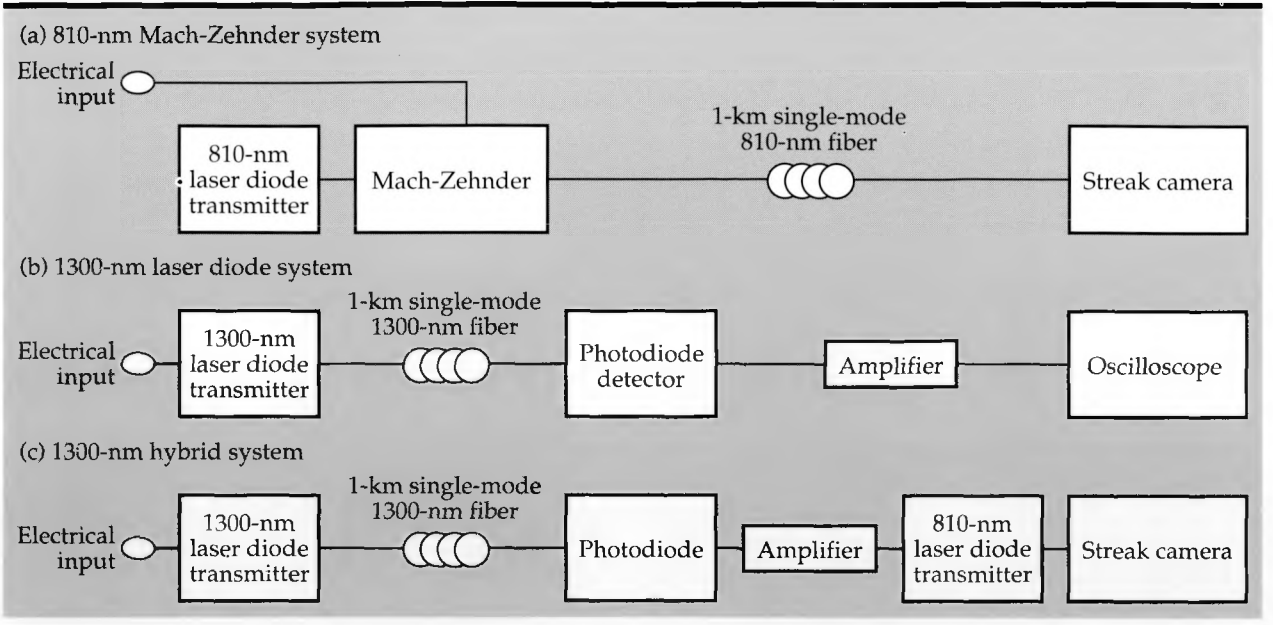


Figure 2. Electro-optic transient recording systems under investigation. The systems are (a) an 810-nm Mach-Zehnder system, (b) a 1300-nm laser diode system, and (c) a 1300-nm hybrid system. The noise sources for these systems are compared in Table 1.

zero-dispersion window for silica fibers is 1300 nm, but this wavelength is outside the operating range of the streak camera.) The 810-nm wavelength effectively accommodates the combined requirements of the laser diode source, the fiber, and the streak camera. Each of these choices for modulator, wavelength of operation, and receiver impacts system complexity, cost, and signal fidelity.

Progress

Systems Being Investigated

Figure 2 shows three configurations of electro-optic transient recording systems that we are now investigating. The first system **[Fig. 2(a)]** uses a lithium niobate Mach-Zehnder integrated optics device as the electrical-to-optical modulator. The intensity-modulated signal is then transmitted uphole via 810-nm single-mode fiber, where it is recorded by a streak camera receiver. The wavelength of operation has been chosen to accommodate the photocathode sensitivity of the multichannel streak camera receiver. The second system **[Fig. 2(b)]** trades off the multichannel capability of the streak camera receiver to operate in the 1300-nm zero-dispersion region of the fiber optic cable. This system directly modulates a laser diode to convert the electrical signal to an intensity-modulated light signal. The signal is received with a photodiode and then amplified to be recorded on an oscilloscope. The third system **[Fig. 3(c)]** is an attempt to provide the 1300-nm system with multichannel streak camera capability. In this system, the oscilloscope is replaced with a streak camera and a second laser diode operating at 810 nm.

Each of these systems offers a substantially higher bandwidth than do electrical coaxial cable systems, but at the cost of decreased dynamic range. The upper end of the dynamic range is difficult to improve, requiring either changing the laser diode structure to produce a more efficient, higher power laser diode or redesigning the lithium niobate modulator to make it more efficient. For this reason, we have chosen to focus on the theoretical limits of the lower end of the dynamic range (the noise floor) to identify which systems are performing at their theoretical limit as well as to make comparisons between systems. In addition, we can identify which system parameters must be improved to yield the greatest overall performance.

Sources of Noise in Each System

Each system has similar sources of noise to consider. The first such source is fluctuations in the optical signal generated by either the laser diode modulator or the laser diode carrier. For this analysis, we will assume that there is no additional optical carrier noise induced by external reflections and that the laser is limited only by its relative-intensity noise. Relative-intensity noise is defined as

$$\frac{\langle P_{noise} \rangle}{\langle P_{average} \rangle} = \frac{\text{rms noise power}}{\text{average power}}.$$

A typical laser diode relative-intensity noise figure is -150 dB/Hz for a laser diode operated away from its resonance peak. A second noise source is the shot noise of the detection system. Shot noise arises from fluctuations in the current that are due to the discreteness of individual electrons. Finally, we must consider the thermal noise in the amplification system. Thermal noise is the noise generated by the random thermal motion of electrons across a resistance.

Method of Analysis and Assumptions

For each system, we first calculate the gain required for a unity gain analog link. The link loss is determined by

$$\frac{I_{out}}{I_{in}} = \frac{(\alpha_m)(\alpha_L)(P_o)(\gamma_o)(R_s)(\pi)}{2V_\pi}$$

(for Mach-Zehnder),

and

$$\frac{I_{out}}{I_{in}} = (\eta_{LD})(\gamma_o)$$

(for direct-modulation links),

where α_m = 0.25 (modulator loss), α_L = 0.5 (fiber loss), P_o = 10 mW or 2.5 mW (power into Mach-Zehnder and laser diode, respectively), γ_o = 0.8 A/W and 0.02 A/W (detector and photocathode sensitivity), R_s = 50-Ω source resistance, V_π = 7 V (voltage for 100% modulation), and η_{LD} = 0.07 mW/mA and 0.35 mW/mA (modulation efficiency of the laser diode into the fiber for single-mode and multi-mode fibers, respectively).

We then calculate the noise due to the laser diode relative-intensity noise, detector shot noise, and amplification stage:[1]

$$I_{RIN}^2 = I_{DC}^2 B\, 10^{(G+RIN)/10}$$

(for laser noise),

$$I_{SDET}^2 = 2q I_{DC} B\, 10^{G/10}$$

(for detector shot noise),

$$I_{TH}^2 = \frac{KT B\, 10^{(G+NF)/10}}{R}$$

(for thermal noise),

and

$$I_{SMCP}^2 = 2q I_{DC} BF\, 10^{G/10}$$

(for photocathode shot noise with an internal MCP),

where $I_{DC} = P_o \gamma_o$ (for direct modulation) or I_{DC} = 0.15 $P_o \alpha_m \alpha_L \gamma_o$ (for indirect modulation), and F = 4 (MCP noise factor), B = 5-GHz bandwidth, G = gain for unity gain system, and NF = 6 dB (amplifier noise figure).

Table 1 compares the noise floors of the three investigated systems for a 5-GHz bandwidth. The 1300-nm direct-modulation-link noise floor is dominated by the relative-intensity noise of the laser. The 810-nm Mach-Zehnder link is dominated by the shot noise of the photocathode. Even though a second laser diode noise source is added to the 1300-nm/810-nm hybrid system, the total noise floor is still lower by a factor of four than that of the 810-nm Mach-Zehnder system. This is because electrical gain is added before the major streak-tube photocathode noise source. The streak camera receiver-based systems could all be improved by higher photocathode sensitivity. The 1300-nm system sacrifices the multichannel capabilities of the streak camera, yet because the photodiode has a higher sensitivity, its performance

Table 1. Source and magnitude of noise in three high-bandwidth fiber optics diagnostic systems. The 1300-nm direct-modulation-link noise floor is laser diode limited. The 810-nm Mach-Zehnder-link noise floor is limited by the streak camera receiver.

		1300-nm direct-modulation link	810-nm Mach-Zehnder link	1300-nm/810-nm hybrid link
1300 nm_{LD}	I_{RIN}^2	6.3×10^{-9} A²		6.3×10^{-9} A²
Dectector	I_s^2	1.0×10^{-9} A²		1.0×10^{-9} A²
Amplifier	I_{TH}^2	5.0×10^{-10} A²		5.0×10^{-10} A²
810 nm_{LD}	I_{RIN}^2		6.8×10^{-9} A²	6.4×10^{-10} A²
Photocathode	I_s^2		3.0×10^{-7} A²	1.0×10^{-8} A²
Total	I_N	0.09 mA	0.55 mA	0.14 mA

within a single channel provides an improved signal-to-noise ratio and increased dynamic range. The 1300-nm hybrid system offers higher dynamic range than does the Mach-Zehnder system but will have difficulty in achieving bandwidths greater than 5 GHz. In the laboratory, we were able to validate the noise floor model for the 1300-nm laser diode oscilloscope system by predicting 0.1 mA of noise and measuring 0.08 mA in 2.5 GHz.

Future Work

In the past year, we have determined the theoretical noise sources in electro-optic transient recording systems. Next year, we will expand upon this work to include the electrical transient and time division multiplexing of signals. System performance of the various implementations will be determined by bandwidth, signal-to-noise ratio, dynamic range, and number of channels. In addition, where it is appropriate, we will validate the system performance models that we develop. We will provide a road map for future component development work by identifying system tradeoffs for the device designers. By providing a focus to on-going research, we will identify the critical systems-level problems that must be solved in order to successfully implement novel systems still in the research phase.

1. W. E. Stephens and T. R. Joseph, "System Characteristics of Direct Modulated and Externally Modulated RF Fiber-Optic Links," *Journal of Lightwave Technology* **lt-5**(3), (March 1987).

GaAs CCDs for High-Speed Transient Recording

Gregory A. Cooper,
J. Courtney Davidson, and
Stefan P. Swierkowski
Engineering Research Division
Electronics Engineering

The goal of this project is to design and fabricate gallium arsenide (GaAs) charge coupled devices (CCDs) for imaging and transient recording applications where sample rates of 1–2 GHz are required. Based on published work[1,2] we have designed and fabricated functional CCDs up to 64 cells in length. We have established a modeling capability that we hope will help us to design and fabricate CCDs that will be faster than any CCDs yet made. Our resistive-film process, now under development, will increase the maximum operating frequency of the CCDs beyond 1 GHz. We have demonstrated that solid-state electronic instrumentation can be used in nuclear weapons tests at locations where the survival of such instrumentation was previously believed to be impossible. Our results confirm the feasibility of this approach for obtaining multi-channel, radiation-hard, and compact transient-recording systems.

Introduction

In many experiments conducted by Lawrence Livermore National Laboratory, the source of a transient signal is separated from the data recorders by as much as a kilometer. Transmitting high-bandwidth analog signals over such distances can result in serious signal degradation due to dispersion and loss in such transmission media as coaxial lines or fiber-optic cables. A recent trend has been to convert transient analog signals to digital form near the source of the signals, and then transmit the digital information to the permanent recorders at a much slower rate that is not affected by the restricted bandwidth of the transmission medium. One way of accomplishing this is to use very fast analog-to-digital converters (ADCs) and fast digital memory to convert each analog sample to its digital equivalent as fast as the samples are taken. Another technique is to store the analog samples in some form of analog storage device and then to digitize the samples at a slower rate using more conventional ADCs. There are a variety of trade-offs in performance, cost, size, and power requirement between these two techniques. The optimum solution depends on the specific application.

CCDs (see **Table 1**) are analog-signal charge-storage cell arrays in integrated circuit (IC) form that uniquely combine a signal sampling function and a signal storage function in one device. In a CCD, the value of an analog sample is represented

Table 1. Definition of terms.

CCD	Charge Coupled Device	A device in which packets of charge (usually electrons) are transported through a series of cells across the device under the control of isolated gate electrodes.
CTE	Charge Transfer Efficiency	A measure of the fraction of a charge packet that is transferred from one cell to the next cell of a CCD. Ideal transfer efficiency would yield a value of CTE = 1.0.
MBE	Molecular Beam Epitaxy	A process whereby oriented, single-crystal layers of a crystal (in this case GaAs) are grown on a single-crystal substrate in an ultra-high vacuum. Various electronic properties may be given to the layers as they are grown.

by the number of electrons stored in a potential well. A CCD is a linear array of closely spaced wells. The close proximity of one well to its neighbor enables the charge in one well to be transferred to a neighboring well. This is the meaning of the term charge-coupled. In the electrical transient recording mode, the analog signal is sampled at one end of the CCD linear array and a number of electrons representing the value of the sample is loaded into the first well. The packet of electrons representing the first sample is then transferred to the second well and the second sample is taken. In this "bucket brigade" fashion the wells of the CCD are loaded with electrons that represent the analog values of the transient signal at specific times. An alternate way of getting electrons into the storage well of a CCD is to illuminate the well with light. The electrons produced by the absorption of the light are collected and trapped in the well. The holes produced are swept away into the substrate below the channel. This ability to convert light into electrical signals makes CCDs very useful for one-dimensional and two-dimensional imaging applications.

Virtually all present CCD applications have been implemented in silicon using metal-oxide-semiconductor-field-effect-transistor (MOSFET) technology. In these structures the control electrodes are isolated from each other and from the channel (the location of the potential wells) by silicon dioxide. Examples of the most popular uses of silicon CCDs today are the video imagers inside VCR camcorders and the very-long-exposure imagers used in astronomy.

There have been previous efforts to apply silicon CCDs to very-high-speed transient recording applications,[1] but the MOS silicon CCD has several fundamental limitations. For high-speed operation, the electrons must be able to move quickly from one well to the next. There are two factors

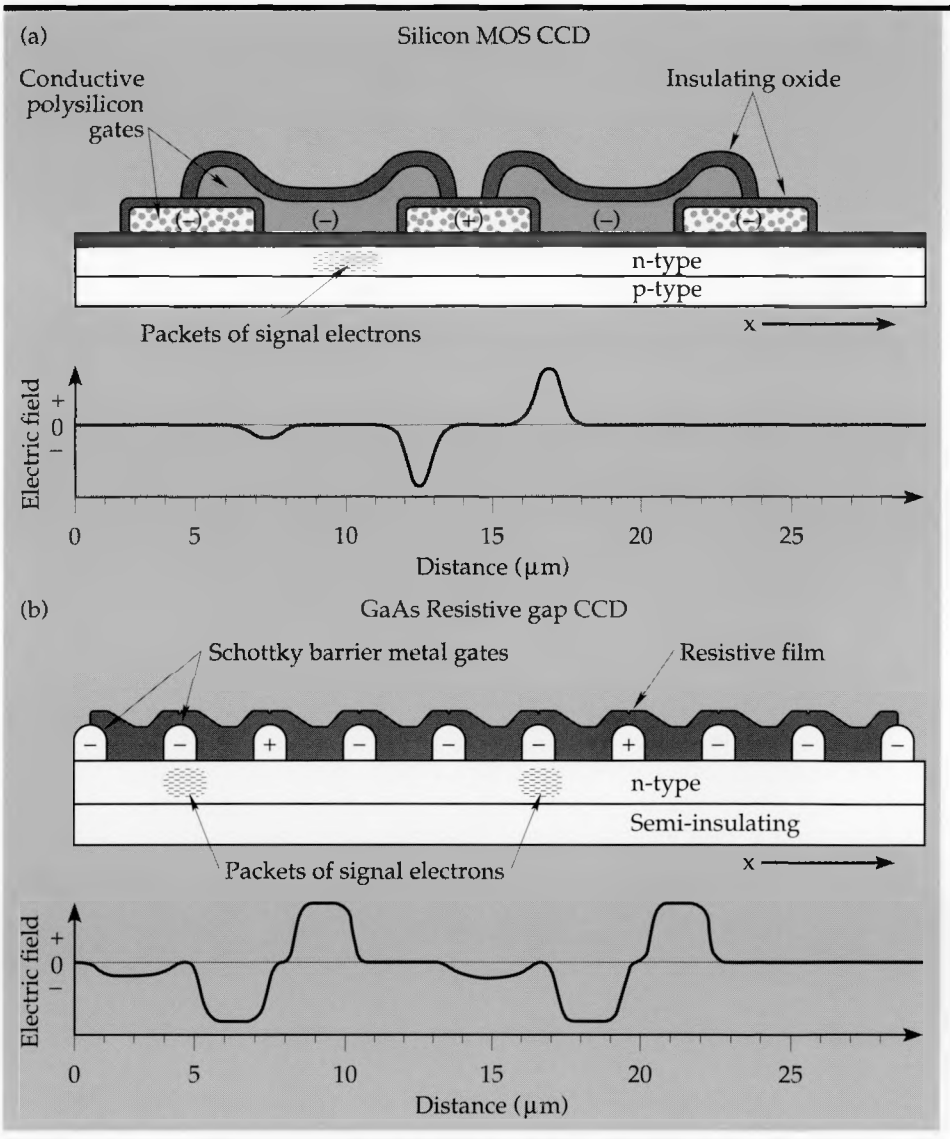


Figure 1. Cross sections of a typical three-phase silicon MOS CCD (a) and the four-phase structure used for GaAs CCDs (b). The electric field in the x-direction which controls the movement of the electron signal packets through the CCD is also shown for both types of CCD. A negative electric field tends to push the electrons to the right and a positive electric field pushes the electrons to the left. It can be seen that the resistive gap structure provides a more uniform negative electric field to propel the electrons from one gate to the next. The silicon CCD must rely on the slower process of diffusion to move the electrons from under the middle of the gates.

that control the transit time of the electrons from one well to the next, the electron velocity and the distance between the wells. **Figure 1a** shows a typical buried channel silicon CCD structure.

The minimum distance between the wells is limited by the MOS process used to fabricate Si CCDs. In order to minimize the amount of uncontrolled surface between the control electrodes, the electrodes of the Si CCD are overlapped. To prevent shorting, the electrodes are made from polysilicon whose surface is then oxidized. This process forms an insulator on the lower electrodes to isolate them from the upper, overlapping electrodes. Even though modern contact lithography can routinely produce structures that are 1 μm in width or less, the polysilicon electrodes must be made substantially wider to accommodate the surface oxidation process. This makes the unit Si CCD storage cell (which may contain 3 or 4 electrodes) quite wide (≈20 μm).

The electron velocity characteristics in silicon also limit overall device speed. Although the electron velocity at high electric fields is approximately the same for silicon and gallium arsenide, the electron velocity at lower fields is much higher for GaAs than for Si. To achieve high electron velocities in Si CCDs, two methods are typically used. First, the channel where the electrons flow is moved away from the surface to a depth where the electric fields generated at the electrode gaps penetrate under the electrodes. This is called "burying" the channel. The second technique, used in conjunction with the buried channel to increase the electric fields in the channel, is to use much higher electrode voltages (>10 V). Since the electrodes represent a capacitive load that is approximately the same for both Si and GaAs devices, the electrode drivers for Si CCDs must be much more powerful to deliver the larger currents required to rapidly generate the large required voltages. Because of these constraints, the operation of silicon CCDs has been limited to frequencies below 250 MHz.

CCDs based on the GaAs material system have been recently developed.[2,3] These devices are based on the same fabrication technology as MEtal-Semiconductor-Field-Effect-Transistors (MESFETs). MESFETs use a metal Schottky barrier gate to modulate the current through the device. A Schottky barrier gate is necessary for GaAs FETs because it is not possible to grow a high-quality, stable oxide on GaAs to make MOSFETs. As in the case of the silicon CCD, wide gates are not desirable because of the poor penetration of the electric fields under the gates. Metal gates can easily be made as narrow as 1 μm. Because metal gates are used for the electrodes of a GaAs CCD, it is not possible to minimize the gap between the gates by overlapping them. Since an uncontrolled surface between the gates can lead to very poor device performance, the gaps between the gates are filled with some resistive material to control the GaAs surface. It is important that the materials placed between the gates have similar electrical properties to the metal gates so that potential barriers or wells are not formed in the gap region which would interfere with the flow of electrons across the gap. A material that forms a Schottky barrier is required, and it must be sufficiently resistive so that it does not short out the electrodes on either side. A typical GaAs resistive gap structure is shown in **Fig. 1b**.

The concept of a resistive gap is not new. In fact, it was first proposed in the early days of silicon CCDs, but was discarded in favor of the simpler MOS process. Placing a resistive material in the gap between electrodes allows the voltage that controls the movement of the signal electrons to vary continuously from the potential on one electrode to the potential on the next electrode. Since the electric fields penetrate under the narrow gates even near the surface, it is not necessary to use large gate voltages to create large electric fields deep under the gates as is required with the silicon MOS CCD. Because electrons are transported at high velocity and the electrode-to-electrode distance can be made much shorter, the resistive gap technology allows much higher operating speeds for CCDs. GaAs resistive gap CCDs have been clocked up to 2 gigasamples per second with charge transfer efficiencies of 99.9%.

A question may be raised at this point: Why not use resistive gap technology to fabricate CCDs out of silicon, which is a much more mature technology than GaAs? There are several disadvantages to using silicon for resistive gap CCDs. The primary problem with silicon is that it has a smaller bandgap than GaAs, which makes the formation of Schottky barrier electrodes more difficult. Low-quality electrodes cannot operate at high voltages. This, as well as the observation made earlier that larger voltages are required to attain high electron velocities, makes it unlikely that high speed CCDs could be fabricated using resistive gap technology on silicon. Another advantage of GaAs is that very-high-resistivity material is available, which means that the CCD channel can be isolated very effectively from the substrate and neighboring devices. This can lead to improved charge transfer efficiency, since fewer electrons will be lost to the substrate.

Recent work on low temperature buffer layers

grown by molecular beam epitaxy (MBE) in GaAs[4] has led to improved isolation of devices fabricated on GaAs. Since silicon MBE is still in its infancy, this level of isolation is not yet possible in silicon devices. The availability of MBE layers for GaAs devices also gives a designer many more options for the construction of the doped layers of the CCD channel. A final reason for using GaAs instead of silicon for CCDs is the superior optical qualities of GaAs for imaging applications. The efficiency with which GaAs can convert light into electron-hole pairs is far superior to that of silicon, due primarily to the direct bandgap nature of GaAs.

Progress

Environmental Test Package

One of the primary goals of this project is to produce a transient recorder design that can be used in the harsh environments associated with underground nuclear testing. With this goal in mind, a portion of our efforts in FY 89 was devoted to demonstrating that solid state electronic packages can collect nuclear test data and can then survive long enough to transmit the data to the permanent recorders on the surface. Additional goals of this experiment were to determine which forms of energy (x rays, gamma rays, neutrons, electromagnetic pulse, etc.) would disrupt the function of our electronics, and what type and level of shielding is required to protect the electronics from this energy.

The design of the electronics package was made as simple as possible to reduce any ambiguity when analyzing the results of the test. The first simplification was to put the signal source inside the shielded portion of the package. This reduced the likelihood that stray energy could enter the electronic package in an unforeseen way and ruin the experiment. Because the effects of high intensity electromagnetic energy can destroy electronic circuits, the outputs of the package were optical signals carried on optical fibers. The only electric wires that penetrated the package were the doubly-shielded dc power lines. The primary shielding for the package consisted of an iron box with 12-mm-thick walls and multiple chambers through which the incoming dc power was fed and filtered. The idea behind the multiple chambers was to isolate the electronics from any large ground currents that might be generated on the cables outside the box. The lid of the box and the mating surface of the box were machined flat so that good electrical contact was made between the two pieces. A good, continuous electrical contact is required to eliminate any possible path to the interior of the box along which electromagnetic radiation could enter.

The electronics inside the package consisted of an oscillator driving a gallium arsenide logic device. The output of the oscillator was used to drive one commercial fiber-optic transmitter and the output of the logic device was used to drive a second fiber-optic transmitter. Both transmitters were located inside the iron box; only their optical output fibers penetrated the wall of the box. The main goal of the experiment was to test the hardness of a package containing electronic components fabricated at LLNL, but a backup package was also fabricated using a commercial oscillator and a commercial GaAs inverter. The primary design used an RCA gate array chip configured as a ring oscillator and a GaAs inverter chain. The final configuration of the gate array was done in our labs, as was the entire design and fabrication of the GaAs inverter chain.

In order to gain some knowledge about the level of radiation that the electronics package could tolerate, the fully assembled packages were subjected to intense x-ray pulses generated by a Febetron high energy electron beam pulser. Our intent was to subject the packages to radiation rates in excess of those expected in the field. Both the commercial and the in-house electronic packages were tested. The results of this testing led to several conclusions:

- The GaAs parts showed no indication of upset or failure for radiation rates up to two orders of magnitude higher that that expected in the field;
- The commercial oscillator was substantially upset at radiation rates at least one order of magnitude lower than the rate at which the in-house oscillator began to show some upset;
- The fiber-optic transmitters and the dc-to-dc power converter were not affected by the maximum radiation rate, 9×10^7 rad/sec; and
- None of the components of either the commercial or the in-house packages failed irreversibly at dose rates up to two orders of magnitude higher than that expected in the field. These results led to the decision to field two electronic packages using in-house parts, one package with additional neutron shielding and the other package without it. This configuration would give us information as to the extent of neutron shielding required.

The outputs of the transmitters were recorded during the actual experiment in the field along with timing fiducials so that the exact time of failures (if any) could be determined. This precise timing information is required to determine which

form of energy (EMP, neutrons, gamma rays, shock wave, etc.) caused the package to fail. After the experiment was completed, the data was analyzed to determine the time and cause of failure. The data revealed that both of the test packages continued to transmit data without upset until the shock wave arrived. This provided good evidence that the components and packaging techniques used were adequate for fielding electronic systems in locations where the radiation was previously believed to be too intense for solid-state electronic systems to survive. Since even the non-neutron-shielded package survived, we concluded that the additional neutron shielding was not required.

CCD Design

The design of the GaAs CCDs that we fabricated and tested is based on both our well-established technology for GaAs digital and analog circuits and our experience with the design, fabrication, and testing of silicon CCDs. Our initial design was a four-phase CCD structure. A four-phase CCD has four gates per cell to control the flow of electrons through the cell and pass them into the next cell. Four phases are desirable for very-high-speed operation because the signals on each of the four electrodes are simple sinusoids which are much easier to generate at very high frequencies than nonsinusoidal signals. The configuration of the CCDs we have fabricated is shown in **Fig. 2.**

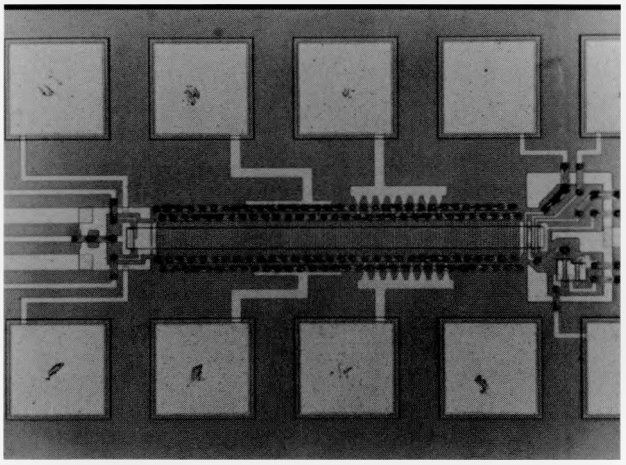


Figure 2. A 32-cell GaAs CCD designed and fabricated at LLNL. The input structure is shown at the left and the output structure is on the right. The charge storage cells are in the middle, and each cell is composed of four parallel, thin metal gates, one for each clock phase. The gates are connected to the bonding pads by air bridge interconnects that run perpendicular to the CCD gates above and below the CCD cells. Many crossovers are required to form the 123412341234... interconnect pattern of the four-phase gates in the CCD.

The input technique that we selected is known as partition sampling. In this technique the input signal is isolated from the input electron potential well by a control electrode that creates a barrier between them. The input well is formed in the GaAs under a suitably biased Schottky barrier gate next to the partition gate. When the barrier is lowered to take a sample, electrons from the signal electrode transfer into the input well. When the barrier is raised again, the electrons in the input well are isolated from the signal electrode and can now be transferred into the first cell of the CCD.

The gate-to-gate pitch of our CCDs was chosen as a compromise between maximum frequency of operation and ease of fabrication. A pitch of 3 μm was chosen with 1.25-μm gate metal and 1.75-μm gaps.

The output structure we used consists of a small electrode that is isolated on either side by potential barriers controlled by dc-biased Schottky barrier gates. The small output electrode is connected to the gate of an output MESFET. When the signal electrons are spilled over the first potential barrier into the output electrode, they cause the current through the output MESFET to be modulated. In this way a very small number of electrons can be used to create a large detectable signal using the very high power gain of the MESFET. After the signal has been detected, the electrons are dumped into another electrode and the output electrode is ready for the next packet of electrons.

Our CCD design incorporates two variations from previous designs that should improve the performance of our devices. One improvement is the use of air bridges to cross wiring traces over other traces. Air bridges, as the name implies, allow one metal conductor to cross over another conductor by leaving an air gap between them. A more common way of crossing a conductor over another is to have a solid dielectric fill the gap between them. Since all dielectrics have dielectric constants larger than air, the air gap will have the lowest parasitic capacitance of any crossover technique. Our design uses air bridges to distribute the four clock phase signals to the electrodes of the CCD. We hope that by reducing this source of stray capacitance, we will improve the speed performance of our devices. The technology to fabricate air bridges in our lab was developed for our digital process several years ago. The other variation is the use of MBE material and etched mesas. In order to isolate and define the active region of the CCD, we have selected to etch away all of the unwanted material around the active areas. This provides the highest possible degree of isolation, but does add slightly to the process complexity

because the active areas (mesas) are higher than the surrounding areas.

In anticipation of the future need to operate our CCDs at frequencies of 1 GHz and higher, we have designed a clocking generator and high-current drivers that will be able to cycle the large capacitive loads of the CCD gates. The clock generator has been designed using our buffered-FET-logic (BFL) technology and will be capable of generating all four of the clock phases as well as the input sampling pulse and the output reset pulse. The high-current driver circuits are designed to be driven by the clock generator, and will buffer the signal (i.e., provide more output current) and shift the output voltage to levels that are appropriate for the CCD. These circuits have been fabricated, and preliminary testing up to 700 MHz is now underway.

Experimental Results

Two different forms of our CCD were fabricated along with several test structures to aid in the evaluation of the CCDs. One of the two forms had wide gates and small gaps between them. This version was intended to be operated without resistive material between the gates. The other version had narrow gates and large gaps between them and was intended to be processed with resistive film layers between the gates. Our initial process run was completed without the use of the resistive layer (which is still under development) to verify the functionality of the design at low frequencies. The chips fabricated in this run contained 1-, 16-, 32-, and 64-cell CCD circuits. A simple structure without CCD gates was used to establish the operating conditions of the circuits and to test the input and output structures separately. Operation of these CCDs requires 6 dc-biased electrodes and 6 clocked electrodes.

Once the proper dc bias conditions were established, we tested the 16-, 32-, and 64-cell CCDs for charge transfer behavior at low frequencies. **Figure 3** shows the input signal, output signal, and the reset signal for 16-, 32-, and 64-cell CCDs on the same chip. One can observe the increasing time delay between the input and the output as the number of cells is increased. Also visible is the degradation of the leading and trailing edge sharpness as the number of cells is increased. This is the result of a nonideal charge transfer efficiency. As an electron packet moves from one cell to the next, it leaves behind a small number of its electrons and it also encounters in the next cell some of the electrons left behind by a previous packet. This small amount of mixing of neighboring packets degrades the fidelity of the signal and is undesirable.

Figure 4 shows a histogram generated by a simple model which predicts the shape of a CCD output given a nonideal charge transfer efficiency. By comparing the model to the data in **Fig. 4b**, one

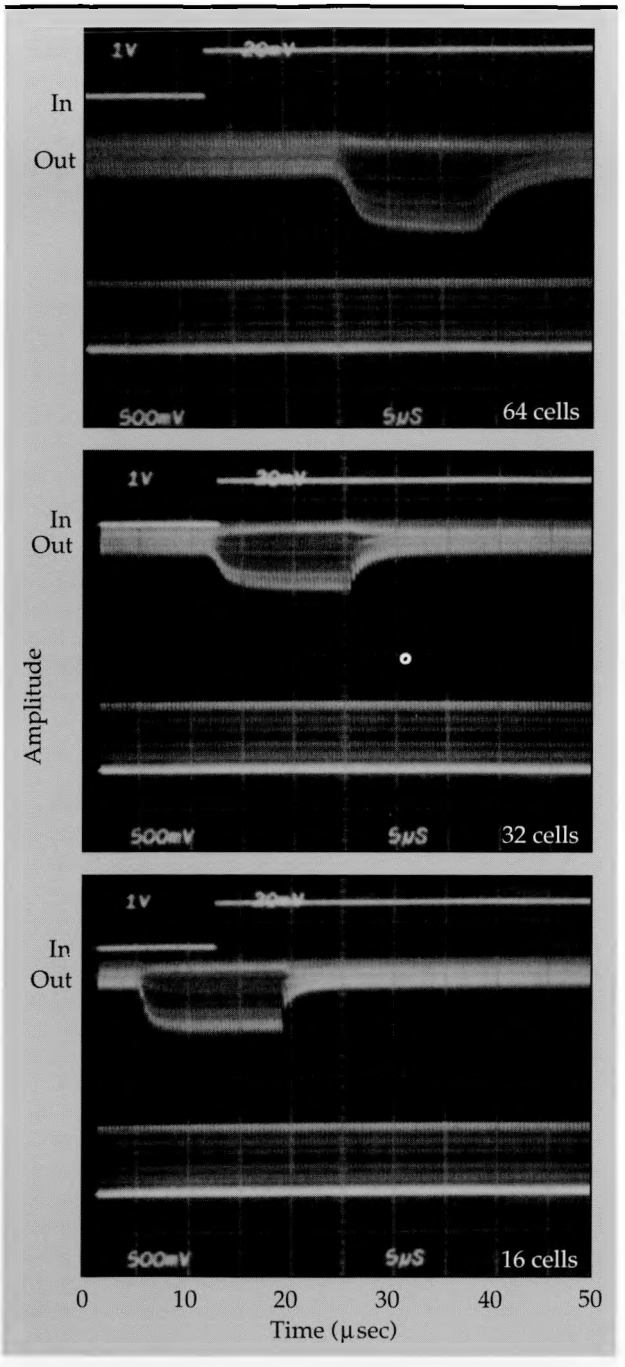


Figure 3. The input and output of three different sizes of GaAs CCDs showing the increased time delay between the input and the output as the number of cells is increased. Also visible is the increased degradation of the output pulse shape with a greater number of cells. The effects of imperfect charge transfer are more apparent as the number of transfers is increased.

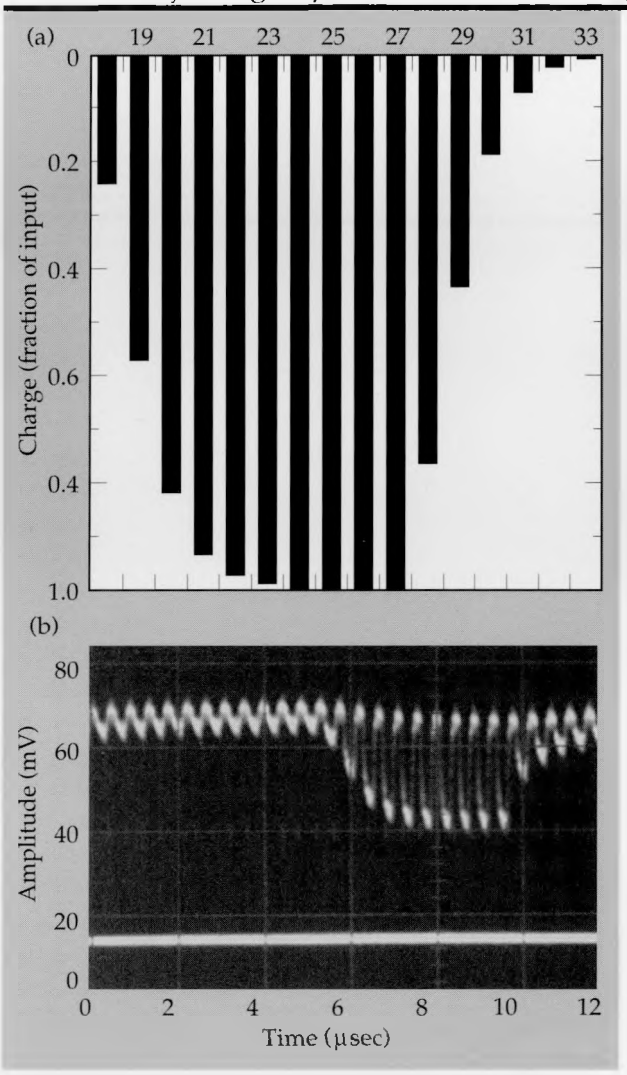


Figure 4. A comparison of a simple, proportional charge transfer model with a CTE set to 0.92 (a) and the actual data taken on one of our 16-cell GaAs CCDs (b). We thereby estimate the charge transfer efficiency to be about 0.92. We expected the CTE of this particular device to be as low as this, because the device was fabricated without the resistive material in the gate gaps that enhances the charge transport between potential wells.

can see that the charge transfer efficiency for our devices is approximately 0.92 at this frequency. This means that about 8% of the charge is left behind with each transfer. We were able to clock charge through these CCDs at frequencies up to about 15 MHz, where the CTE began to degrade due to the frequency limits of the probe cards we were using and the CCDs themselves. We had expected to observe this phenomenon in the first batch of chips because these chips did not have the resistive film between the gates. The uncontrolled surface between the gates creates electron traps or barriers that impede the flow of the electrons. We are encouraged by the fact that we were able to operate our CCDs at frequencies below 15 MHz because this is an indication that our design, materials, and processing were excellent. Poor design, materials, or processing can result in excessive amounts of thermally generated electrons which quickly fill the wells at low frequencies and render the CCDs useless. Achieving fully functional CCDs at any level of performance during the first year of this project was a major milestone.

To operate the CCDs at very high frequencies (>300 MHz) we are developing a process to deposit and pattern a stable resistive film in the gaps between the gates. The resistive film we are using is an amorphous mixture of Cr and SiO that is RF-sputtered from a target composed of 35% Cr and 65% SiO. This material is generally referred to as CERMET (CERamic-METal). There are many parameters that can be varied during the deposition of CERMET to control the resistivity and stability of the film. We have found a range of parameters that enable us to deposit CERMET films with sheet resistivities on the order of 300 kΩ/sq and thicknesses of the order of 3000 Å. The ability to produce films of the desired sheet resistivity and thickness was an important milestone during this first year because the resistive layer is crucial to the operation of these devices at high frequencies. Recently work has begun to develop a lift-off process for patterning the CERMET because this type of process would be more compatible with our CCD process than other patterning techniques such as ion milling. Our efforts are now focused on finding a combination of photoresist preparation and CERMET deposition rate that does not damage the photoresist. Once we have developed a suitable lift-off technique, we will begin fabricating CCDs with CERMET in the gaps.

Using some of the CCDs without CERMET, we have built a low-frequency, prototype transient recorder to demonstrate fast sampling and slow read-out recording. For the prototype, a gate-array was used to generate the four clock phases, the input sampling pulse, and the output reset pulse in response to externally applied arming and triggering signals. **Figure 5** shows the output of the CCD for two different triggering times. In **Fig. 5(a)**, the switch to the slower clock rate occurs after the input signal has already passed completely through the CCD, and in **Fig. 5(b)**, the clock rate slows down after the input signal has been sampled by the CCD but before the samples have passed through the CCD. In addition to the broader output signal when the clocks are slowed, one can also observe the increase in the amount of background signal coming from the CCD when the clock rate is

slowed. This increase is believed to be the result of an increased number of electrons generated in the wells by thermal processes. The rate of electron generation is constant, but at the slower clock rates the wells have more time to collect the thermal charges. This effect will be much less significant at the higher clock rates at which we intend to eventually operate these devices.

Modeling

This year we began an effort to model charge transport in CCDs and 2-D device design. We have contracted with UC Davis to take an existing two-dimensional, dynamic modeling code call BAMBI and modify it for use with GaAs CCD structures. BAMBI was originally developed to model silicon MOS devices, and several modifications have been made to allow the use of multiple Schottky barrier contacts and GaAs material. The goal of this modeling project is to examine the transient electron transport behavior and to optimize the design of the CCD topography and the MBE materials to give the best performance. The first phase of this project has been completed. The first phase included the following milestones:

- Adapt the FORTRAN source code for BAMBI to operate on a UNIX platform,
- Modify BAMBI to incorporate all of the structures we will need to simulate GaAs CCDs, and
- Develop (or acquire) an output package that will take the raw BAMBI output and produce graphical representations of the parameters we wish to observe.

The next phase of this project will be to use BAMBI to simulate various structures and materials to help guide us in the design of the next generation of CCDs.

Future Work

In order to meet the goal of designing a fieldable, high-speed transient recording system, our efforts in FY 90 will be applied in several areas. A major milestone for the next year will be to demonstrate CCD operation at clock rates in the range of 1 gigasample/second. Two tasks must be performed to achieve this milestone: (1) we must complete the development of the CERMET deposition and patterning process and combine it with the CCD fabrication process, and (2) we need to establish a high-frequency testing capability.

We intend to extend our modeling effort to study the transient behavior of charge carriers in CCDs in order to gain a better understanding of the processes that limit the CTE of CCDs. We hope to use the results of this modeling to help us design a second generation of CCDs that will have improved characteristics for use in transient recorders. We also plan to model and fabricate alternate CCD configurations such as two-phase and one-phase structures which could be superior to the four-phase structure and also easier to control.

Once we have established a fabrication process and a testing capability for high-speed CCDs, we also plan to investigate the sources of dark (thermal) charge generation processes, sources of noise, and the fundamental frequency limits of the devices.

We are also going to continue to work on developing a high-speed analog demultiplexing (ADMUX) circuit. A first generation of this device was designed and fabricated during the first phase of this project, and testing has recently begun. Using these devices we hope to be able to sample an analog signal and demultiplex the samples to sev-

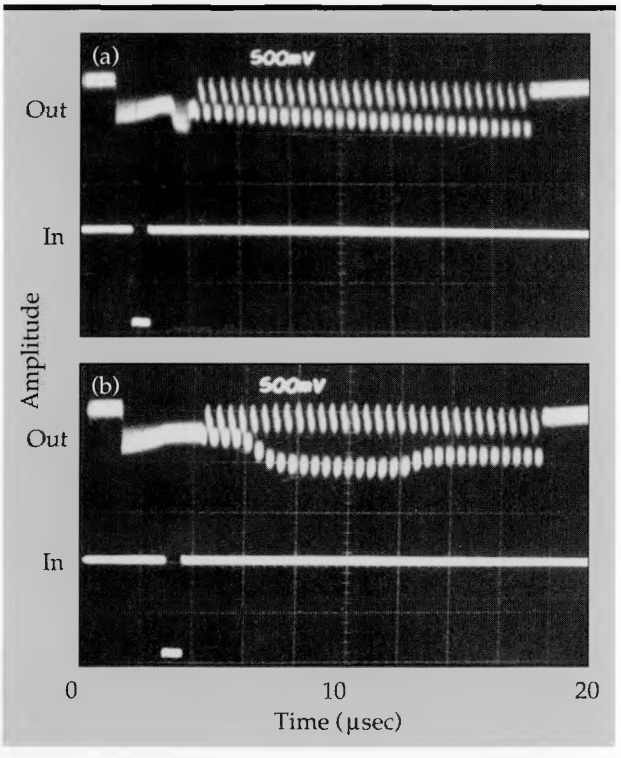


Figure 5. The input and output of a GaAs CCD being operated as a fast-in / slow-out analog recorder. The input pulse is being sampled by the CCD while the clocks are running fast. The sampled data can be seen arriving at the output of the CCD while the clocks are still running fast (a) and after the clocks have switched to a slower rate (b). The temporal width of the slow output is seen to be much greater than the original input signal width. Fast sampling and slow readout strategies such as this can be used to sample fast signals and then process the samples at a slower rate.

eral CCDs. This technique will improve the overall system bandwidth beyond the capability of an individual CCD.

1. T. E. Linnenbrink, et al., "A One Gigasample per Second Transient Recorder," *IEEE Trans. Nuc. Sci.* **26**(9), 4443 (1979)
2. J. A. Higgins, R. A. Milano, E. A. Soverno, and R. Sahai, "Resistive Gate GaAs Charge Coupled Devices," *Proc. IEEE GaAs IC Symp.* 1982, pp. 49–52.
3. E. A. Soverno, R. Sahai, W. A. Hill, and J. A. Higgins, "Microwave Frequency GaAs Charge Coupled Devices," *Proc. IEEE GaAs IC Symp.* 1984, pp. 101–104.
4. C. L. Chen, F. W. Smith, A. R. Calawa, L. J. Mahoney, and M. J. Manfra, "Reduction of Sidegating in GaAs Analog and Digital Circuits Using a New Buffer Layer," *IEEE Trans. Elec. Dev.* **36** (9), 1546–56 (1989)

ENGINEERING RESEARCH AND DEVELOPMENT

Thrust Area Report FY 89

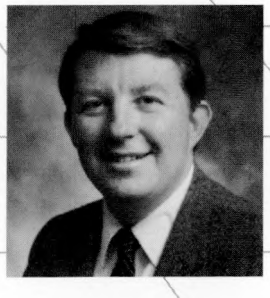
4

Fabrication Technology

The goal of the Fabrication Technology thrust area is to develop high-precision fabrication processes and equipment that are economical. Our work for Laboratory programs focuses on the means to produce components with previously unrealized levels of precision. In FY 89 our research was concentrated in five areas:

- Developing an ultraprecision grinding (ductile grinding) process for the economic preparation of glass and ceramic optical components.
- Initiating new and fundamental work to model, at the molecular level, the deformation processes occurring during the high-speed interaction of either a hard abrasive or a cutting tool with a surface.
- Analytically modeling plasma transport and experimentally verifying a unique cathodic-arc vacuum deposition process for producing high quality thin film coatings.
- Evaluating the economic viability of machining a variety of new single-crystal materials for use as harmonic generators for high-power laser systems.
- Evaluating a high-speed deformation procedure to determine if it can reduce the expense of evaluating full-size shape charges.

Irving F. Stowers
Thrust Area Leader



Ductile Grinding of Glass

Kenneth L. Blaedel and Pete J. Davis
Materials Fabrication Division
Mechanical Engineering

Daniel J. Nikkel, Jr.
Engineering Sciences Division
Mechanical Engineering

Our goal is the realization of a fabrication process that can generate finished precision optical components in an essentially one-step operation. This process, called ductile or shear mode grinding, promises to significantly reduce both the processing time and the overall cost of manufacturing precision optical components. The implementation of this process would significantly benefit the Laboratory's Laser Fusion Program as well as programs involving astronomical telescopes and space defense.

We are making progress in our understanding of the ductile grinding process and the parameters that determine the transition from brittle to ductile material removal. We have evaluated specimens prepared by ductile grinding to determine if the surface is significantly different from conventionally ground and polished optical surfaces, and have found no significant difference. We have modelled the stresses and deformations that occur at the surface and immediately beneath the surface as a single abrasive particle interacts with the surface. The modelling has verified that cracking can be initiated when the particle loading is being either increased or reduced, as has been previously observed.

Introduction

Many Laboratory programs utilize sophisticated and costly optical components in the construction of various analytical instruments. The largest single user is the Laser Fusion Program's Nova laser, which contains nearly $50M of optical components. The construction of an even larger laser fusion driver in the late 1990s to demonstrate break-even energy gains will require an order of magnitude more optical components, but at essentially the same cost as was required to construct Nova. The ductile grinding process is being developed to address this fabrication cost reduction goal, that is, an order-of-magnitude cost reduction for the fabrication of special-purpose precision optics.

The technical challenge of achieving this goal is centered in the understanding of the exact mechanism that allows a brittle material, such as glass, to behave under certain conditions in a ductile fashion, so that it can be ground precisely to shape without sustaining any deleterious subsurface damage or cracking of the surface. To be more specific, when machining a brittle material like glass, there is a fundamental difference between material removal in the brittle regime and in the ductile regime. Brittle removal occurs by propagating cracks and usually accompanies large depths of cut and high grinding forces. It results in a surface with significant roughness and deep subsurface damage. In contrast, ductile removal is promoted by shallow depths of cut and low cutting forces, and leaves a specular surface and low subsurface damage.

In the conventional optical fabrication process, the tens of micrometers of damage left by brittle grinding have to be removed by loose abrasive lapping and then finished by polishing in order to arrive at the correct contour and surface finish. Unfortunately, lapping introduces errors into the contour which must be removed during the polishing process. Because of the potential benefits of minimal subsurface damage and the ability to machine to close tolerances, bound abrasive ductile grinding holds great promise for replacing the abrasive lapping process, and for minimizing or eliminating the amount of polishing required to meet a given component specification.

Progress

We have made progress in three different areas. We have quantified the amount of subsurface damage remaining after ductile grinding, and have found that the ductile-ground surface is difficult to distinguish from a more conventionally polished surface. Secondly, we have modelled the interaction of several indenter shapes with a free surface to better understand what conditions induce, and more importantly prevent, cracking. And thirdly,

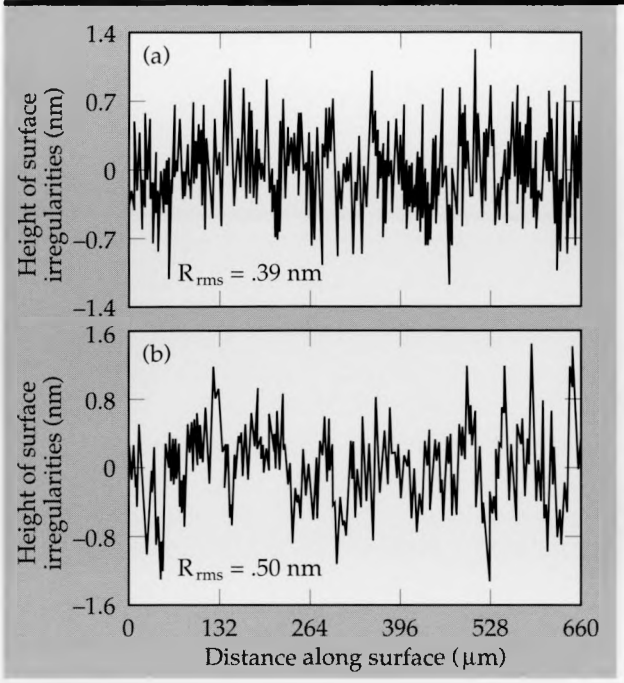


Figure 1. Surface roughness of a polished (a) and ductile ground (b) specimen as measured with a profilometer.

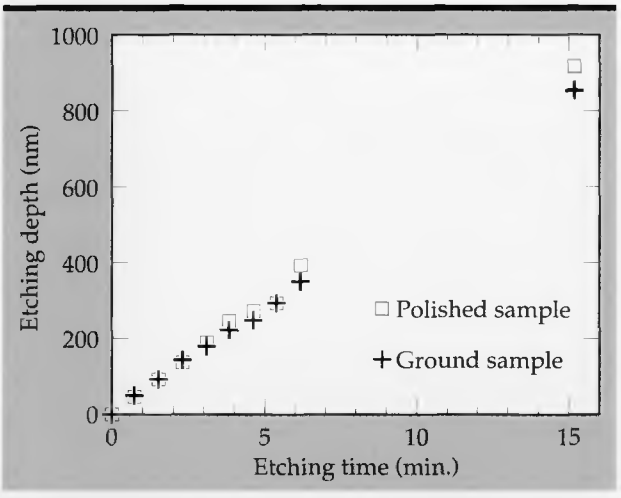


Figure 2. Plot of etch depth versus etching time, showing no significant difference between a ductile-ground and a conventionally polished specimen. The etchant was a solution of 10% HF and 5% HCl by volume in distilled water.

we have developed a highly accurate technique for detecting the occurrence of physical contact between a bound abrasive grinding wheel and the workpiece.

Quantifying Surface and Subsurface Damage

The goal of this portion of the research was to characterize the surface produced by ductile grinding and thereby estimate the depth of damaged material that would have to be removed by polishing such a surface. The strategy was to use a very high quality polished surface as a reference against which to compare a ductile-ground surface. The two surfaces were evaluated by three techniques (chemical etching, response to indentation, and laser damage threshold) and were found to be comparable. We therefore conclude that we can produce a high quality surface by ductile grinding.

The test specimens, made from BK-7 glass, were fabricated by free abrasive grinding with a schedule of progressively smaller grit abrasives. This was done to ensure that no subsurface damage remained after the final polish. One of the specimens was then placed on a rotating spindle, and was precision-ground with a cupped wheel. The diamond grinding wheel (a resin-bonded grinding wheel impregnated with 2-4 μm diamond abrasives and rotating with a surface speed of 13.8 m/sec) was fed into the glass sample rotating off-axis at 100 rpm. A total of 5 μm was removed by ductile grinding to ensure that the surface was characteristic of the ductile grinding process and no artifacts remained from any preceding fabrication processes. **Figure 1** shows the surface profiles of the polished and ground specimens as measured on a WYKO profilometer. The rms roughnesses of the ground and polished specimens are almost equal.

Chemical Etching. Etching is the process that has traditionally been used to reveal damage in glass surfaces. It is known that a variation in the rate of etching is an indication of a difference in subsurface damage. **Figure 2** is a typical test showing depth of material removed versus etching time. The differences in the plotted points for the polished and ductile ground specimens are well within the uncertainty levels typical of etching. There is no indication that the ground and polished surfaces are significantly different from one another.

Response to Indentation. The response of a surface to indentation is an indication of the integrity of that surface and its supporting subsurface. Specifically, if ductile grinding produces a surface which has subsurface flaws, then an indentation test will show signs of reduced "flow pressure," where flow pressure is defined as the ratio of indentation load to projected area of indentation. This is similar to hardness, except that it uses instantaneous load and indentation area, not just the maximum load and residual plastic indentation area. Flow pressure therefore includes elastic, plastic, and inelastic effects. **Figures 3** and **4** show the average flow pressure versus depth of penetration for a 1.6-mm-diam ball indenting the ground and

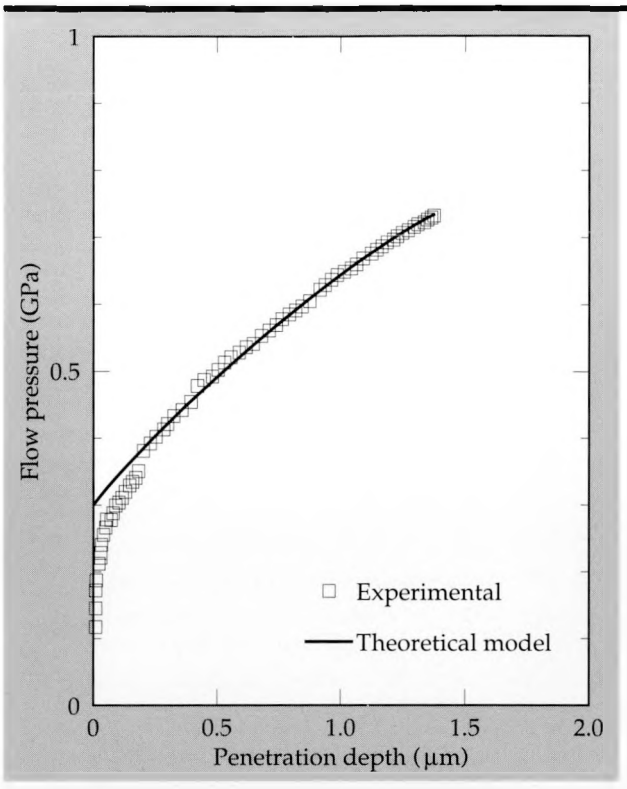


Figure 3. Average flow pressure versus depth of penetration for a 1.6-mm-diam ball indenting a conventionally polished specimen.

polished specimens. The conclusion to be drawn from the indentation tests is that only in the very near surface region are ductile-ground and polished surfaces different; thereafter the surfaces are indistinguishable. There is no evidence of micrometer-size cracks in the subsurface of the ground specimen.

Laser Damage Threshold. The laser fusion program is interested in the laser damage threshold of its optical components, which is also a function of subsurface damage. The laser damage threshold of the ductile-ground and polished specimens was determined by subjecting the surfaces to high-fluence pulses of 1.06-μm laser light. The surfaces were then inspected microscopically for any evidence of damage. The specimens were tested for front-surface as well as rear-surface damage threshold. The results, shown in **Table 1**, demonstrate that the subsurfaces of both the ductile ground and polished specimens are similar in this respect.

Table 1. Laser damage threshold in J/cm² for BK-7 glass tested with 1-nsec pulse of 1.06-μm laser light.

	Front Surface	Rear Surface
Ductile Ground	14.2 ± 2.1 J/cm²	9.6 ± 1.8 J/cm²
Polished	11.9 ± 1.8 J/cm²	11.9 ± 1.8 J/cm²

The objective of these evaluations was to determine whether optical surfaces fabricated by ductile grinding could approach or equal in quality those surfaces produced by conventional high-quality polishing. The evaluations indicate that the ductile-ground surfaces were not significantly different from the conventionally-polished surfaces, and give no indication of any micrometer-size cracks. Our interim conclusion is that ductile grinding can produce a surface well suited to minimize or eliminate any requirement for subsequent polishing.

Computer Modeling of Subsurface Stresses

Ductile grinding involves the action of many abrasive grits bound into a grinding wheel, rapidly indenting and scribing a glass surface. Depending on various parameters such as the depth of cut and the shape of the grit, material may be removed in either a ductile or a brittle mode. With ductile grinding the primary mechanical process consists of plastic deformation, in contrast to brittle grind-

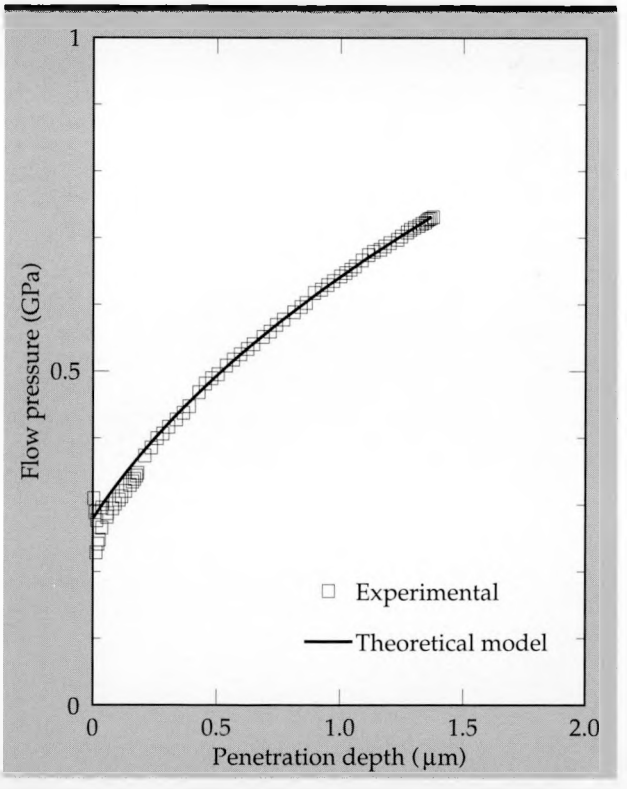


Figure 4. Average flow pressure versus depth of penetration for a 1.6-mm-diam ball indenting a ductile ground specimen. It can be seen that the ductile ground specimen appears "harder" than the polished specimen only for the first 50 nm of penetration, but thereafter is indistinguishable from the polished specimen.

ing where the primary mechanical process is fracture. Knowing what conditions lead to the initiation of fracture helps to provide an understanding of what causes the transition between ductile and brittle behavior in grinding.

Our computer modeling effort is concerned with trying to develop the tools necessary to predict the brittle to ductile transition during glass grinding. The actual physical problem is very complex. The material response of glass is sensitive to such factors as machining rate and temperature. The material properties near the surface are substantially different from the bulk properties, and are significantly affected by an exposure to water or other chemicals. To realistically model all aspects of the problem is extremely difficult, since we would have to include the effects of strain-rate, temperature, and chemistry.

We are using finite element stress analysis to determine the stress state that governs the mode (ductile or brittle) of grinding. We have concentrated on the simpler and more numerically tractable indentation test as a first step to understanding grinding, and have used finite element analysis to model a variety of two-dimensional, axisymmetric indentation problems corresponding to a wide spectrum of grit geometries. We have modelled a spherical, 60° conical, and 120° conical diamond indenter entering elastic-plastic materials (soda-lime glass, fused silica, and BK-7 glass). The calculations show that for large spherical indenters, the highest likelihood of crack initiation will be on the surface of the glass in close proximity to the region of contact with the indenter, which agrees with observations of "cone cracks" formed during such indentation tests.[1] Calculations for sharp conical indenters show that, as seen in **Fig. 5**, the region of plastic deformation is fairly localized near the region of contact. Furthermore, the calculations show a region with a high likelihood of crack initiation beneath the region of plastic deformation, indicating the tendency of the sharp indenter to initiate a crack from within the subsurface region.

Calculations for both blunt and sharp indenters have also shown that the tendency to initiate or propagate cracks continues to increase as the indenter is extracted. This result is also supported by observation.[2] For these calculations the elastic properties for the glass have been assumed to be the same as the bulk properties reported in the literature. While glass scientists suggest that material properties at the surface may be different than in the bulk, no method is available to quantify this. The plastic properties (yield strength and work

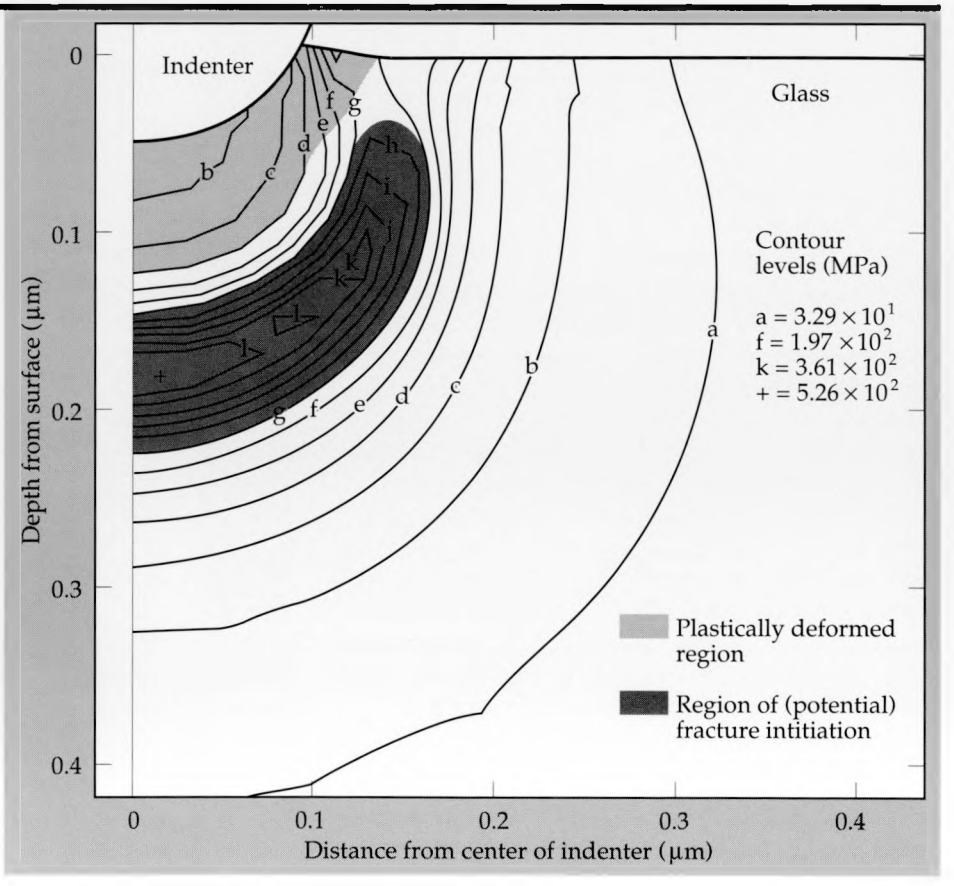


Figure 5. Results of a finite element stress analysis showing the contours of the fracture function for a sharp conical indenter, indicating areas where the material fracture criterion has been exceeded.

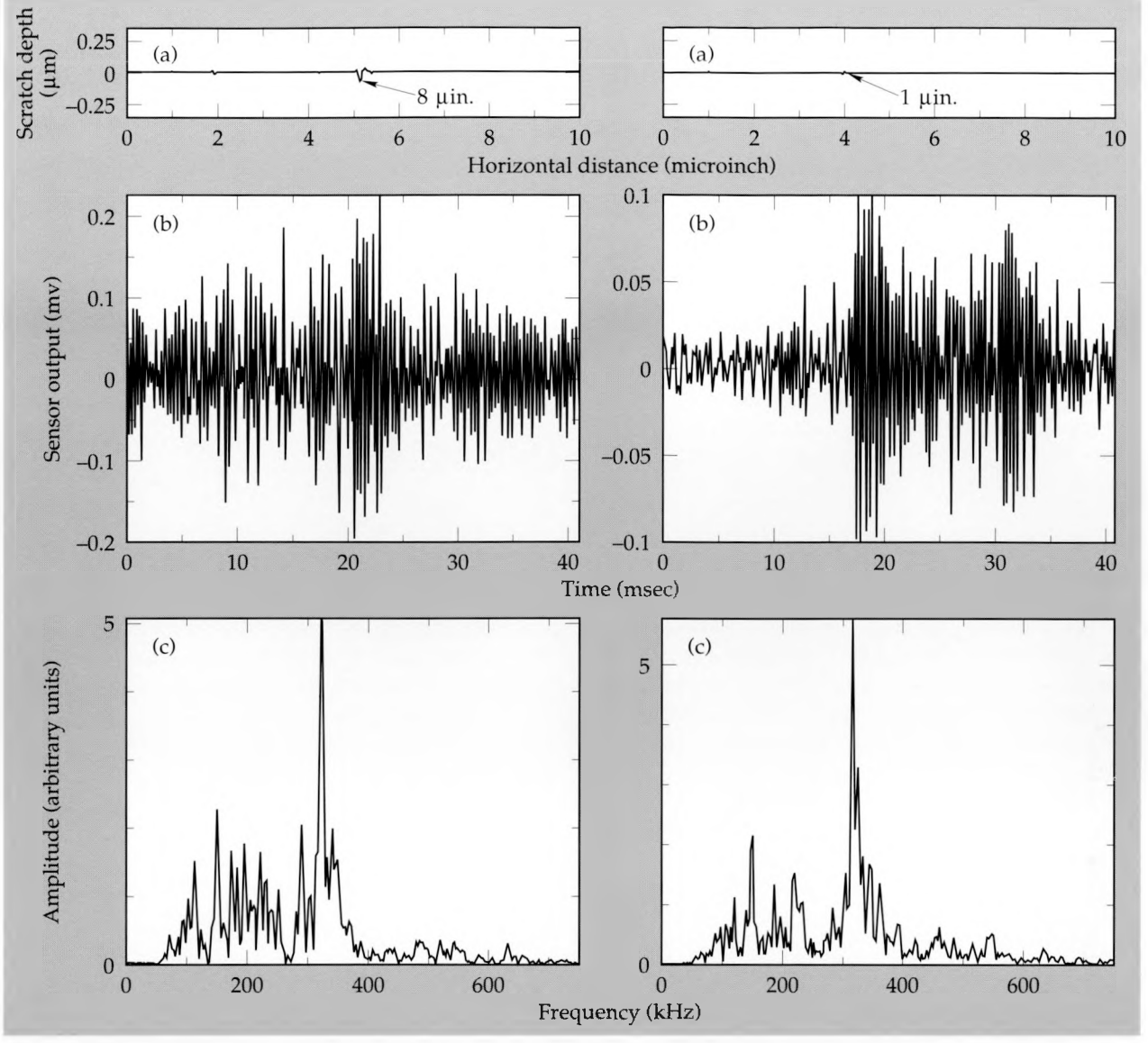


Figure 6. Cross-sectional profile of a scratch (a), acoustic time signal (b), and spectrum (c) made during tests to determine the edge of the acoustic emission grinding wheel.

hardening parameters) have been set to textbook values for our present calculations. The inaccuracy in our characterization of the properties (both elastic and plastic) of the material near the surface should be reduced by experimental work currently being performed at the National Institute for Standards and Technology. These experiments should give us more realistic values for near-surface properties, making the quantitative results of our calculations more representative of the actual physical situation.

Grinding Wheel Edge Detection

In order to make ductile grinding a useable process, we have had to develop several ancillary techniques, one of which was devised to precisely locate the perimeter of the grinding wheel. The depth to which a grinding wheel penetrates a workpiece ("depth of cut") is an extremely important parameter in controlling whether material is removed in the shear or fracture mode. A prerequisite to controlling this depth of cut is the ability to locate the periphery of the grinding wheel. It is for this reason that we have developed a precise measurement technique using acoustic emission (AE). For many of the glasses that we grind, the threshold depth of cut which divides the ductile and brittle grinding modes is about 0.2 μm. Therefore, our objective was to develop a "wheel set station" to detect the location of the periphery of the grinding wheel to about one tenth of the average threshold, or 0.02 μm (20 nm). AE has been used by others to locate a grinding wheel,[3,4] but the repeatabil-

ity of the detection was limited to about 2 μm, or a hundred times less precise than is desired here. Our effort therefore represents a challenging but straightforward application of AE to contact detection.

A second reason for developing this technique is to determine the amount of wheel wear that occurs during one pass of the grinding wheel over the workpiece. Unless compensation is made for wear, any wear of the wheel will cause an error in the contour of the workpiece. The wheel set station allows us to locate the periphery of the wheel, make a grinding pass, and then locate the periphery of the wheel again, thus determining the change in diameter. We can thereafter alter the commanded path that the grinding wheel takes in order to compensate for wheel wear during each subsequent pass.

In the AE edge detection technique, an AE sensor is attached to the back of a glass specimen to listen for the first acoustic signal resulting from contact between the grinding wheel and the specimen. The front surface of the plate represents the reference surface from which the periphery of the wheel can be located. The repeatability by which this technique can locate the edge of the grinding wheel is a measure of success. Our testing arrangement moved the rotating grinding wheel into the workpiece until the onset of an acoustic signal in the range of 100 kHz to 600 kHz indicated contact between the wheel and workpiece. This procedure created scratches whose cross-sections appear in the Talystep profile in **Fig. 6a**. The acoustic emission associated with each scratch is shown in the time domain in **Fig. 6b** and the frequency domain in **Fig. 6c**. This procedure was repeated an average of ten times, with each scratch being made adjacent to the preceding scratch. The scratches in the glass workpiece were then analyzed for repeatability of the depth of cut. The technique has achieved a detection accuracy of 0.12 μm when operating without a fluid coolant. However, the use of a coolant introduces significant additional acoustic noise and masks the first point of contact. Enhanced signal processing has been developed this past year which now allows the wheel edge to be detected to an accuracy of 0.1 μm even with a fluid coolant present.

With enhanced signal processing, the point of contact is detected by integrating the AE signal over a very short time interval. This distinguishes this approach from previous approaches in which the signal was integrated over one or two revolutions of the wheel before a decision was made. Also, contact is defined as the repeating AE signal associated with a given grit striking the reference surface once per revolution. The detection of a single acoustical spike is not considered contact.

We have found that the enhanced signal control detection technique is not sensitive to the wheel speed, grit size, grit concentration, or type of grit on the grinding wheel. The circuit is acoustically quite robust and does not require much tuning for different conditions of grinding. The arrangement has been tested on an experimental bench. However, nonideal circumstances such as poor temperature control of the mechanical components and imperfect electronic components have limited the scratch grinding consistency to about 0.1 μm instead of the goal of 25 nm.

Future Work

Our plan is to make ductile grinding, currently a robust but slow grinding technique, into a much more versatile and productive material removal process. We intend to accomplish this improvement in a number of ways. The first way is to improve traditional machining parameters such as feed rate and surface speed, and to alter various grinding wheel parameters such as average abrasive size and size distribution of the abrasives. The second way is to develop a method to continuously recondition (resharpen or redress) the grinding wheel. We now condition a wheel and then grind until it is worn. The term *worn* refers to a reduction in the volumetric removal rate. Grinding is therefore not a steady-state process. It is our belief that we must continuously recondition the wheel during grinding if ductile grinding is ever to become an economically productive process for moderate-size optical components. The third way to increase productivity is to alter the chemistry under which the grinding takes place. The current method is to use a water-base grinding solution (or coolant) with various additives. Historically, grinding coolant research has been limited to brittle grinding and to polishing. It is not known how well the conclusions from this research will apply to ductile grinding, which removes material under somewhat different conditions.

We will continue to pursue the finite element stress analysis modeling work already under way. The accuracy with which we can characterize the near surface properties of glass (both elastic and plastic) will be improved by precise measurements to be carried out at the National Institute of Standards and Technology. These measurements should give us more accurate values for surface properties, thus making the results of our modeling work quantitatively closer to the actual physical situation.

1. N. P. Suh, *Tribophysics* (Prentice-Hall, Englewood Cliffs, NJ) 1986.
2. B. R. Lawn, T. P. Dabbs, and C. J. Fairbanks, "Kinetics of shear-activated indentation crack initiation in soda-lime glass," *J. Mater. Sci.* 18, 2785 (1983).
3. J. J. Henry, *An Ultrasonic-Acoustic Grinder Setting Station*, Report No. Y-1935, Oak Ridge Y12 Plant, Union Carbide Corp (1974).
4. M. Takahara, "Performance of Vitrified Bonded CBN Wheels and its Applications for the Automotive Industry," Third International Grinding Conference, Soc. Manf. Eng. (1987).

Molecular Dynamics Modeling Applied to Indentation and Metal Cutting Problems

James F. Belak
H Division
Physics Department

Anthony J. DeGroot
Engineering Research Division
Electronics Engineering

William G. Hoover
Applied Science Department
University of California, Davis

Irving F. Stowers
Precision Engineering Program
Electronics Engineering

Carol G. Hoover
National Magnetic Fusion
Energy Computer Center

We have discovered new insights into the behavior of common engineering materials subjected to external forces—surface indentation and orthogonal cutting—as found in high-speed precision machining. To study this behavior, we use a computer simulation method called nonequilibrium molecular dynamics because it is atomistic—it simulates the response of every atom in a material. It can thereby give us more information about material properties, stresses, and deformations than can macroscopic measurements and simulations. Specifically, we used the embedded-atom method to model the atomistic interactions. This new method allows us to accurately simulate the behavior of simple metals.

We simulated a two-dimensional cross section of material containing a few thousand atoms and obtained good qualitative results. Yet, to attain more accurate quantitative analyses, our simulations will require a few million atoms, and hence, more computer processing power. Therefore, we have also developed a strategy for simulating millions of atoms on the new class of supercomputers—the SPRINT and BBN-TC2000 parallel computers.

Introduction

Material removal and build-up of surfaces are important processes in the fabrication of high-precision mechanical parts.[1] Examples of these processes include high-speed precision machining,[2] where spatial tolerances are a few tens of nanometers, and the ductile grinding of glass[3] to produce precision optical components. Macroscopic measurements have yielded new and useful insights into our understanding of surface indentation and cutting.[4] However, little is understood of the basic atomistic mechanisms by which material is removed and deformed, by which the tool-tip and workpiece interact, and by which induced surface and subsurface damage occurs. The two processes mentioned above are particularly affected by atomistic mechanisms.

For example, chip morphology differs dramatically between high- and low-speed machining of steel.[2] **Figure 1** depicts the cutting process. **Figure 1(a)** shows the continuous chip that forms in low-speed machining. The chip's thickness is larger than the depth of cut, and its smoothness indicates a continuous shearing of the workpiece material at the tool tip. The chip formed in high-speed machining [**Fig. 1(b)**] is much different; it contains bands of intense localized shear. Between these bands are large segments of undeformed material.

In the ductile grinding of inherently brittle materials, small abrasive particles (~1 μm) are used to achieve high surface smoothness (~0.001 μm).[3] When the depth of cut is small compared to the abrasive particle size, the mechanism of material removal is similar to low-speed machining of ductile metals. However, at greater removal depths, appreciable cracking and fracture occurs, as

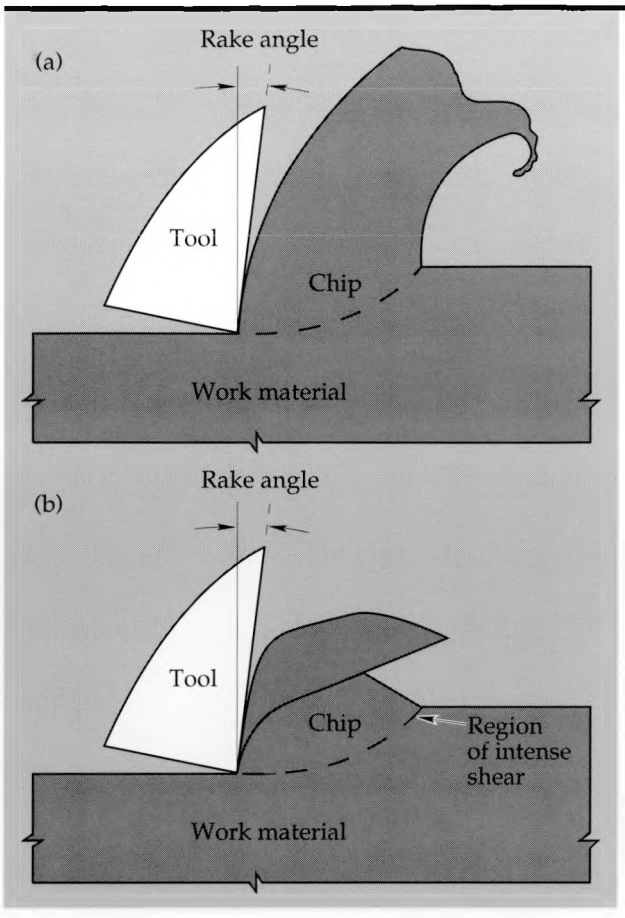


Figure 1. The chip morphology found in the machining of metals. (a) The smooth, continuous chip observed when machining steel at low cutting rates (≤1 m/sec). (b) The segmented chip observed in high-speed machining of steel (≥100 m/sec).

expected for brittle materials.

The non-equilibrium molecular dynamics (NEMD) computer simulation method[5] is ideal for studying these atomistic mechanisms. From our simulations, we evaluate the response of a material subjected to an external force by following the response of every atom in the material. Examples of external forces include: coupling to heat baths, stress fields, and moving boundaries. NEMD simulations also provide new insights into the material stress/strain relationships, which are largely unknown because measurements are exceedingly difficult at the high strain-rates involved in the machining process. These constitutive relationships are necessary input into continuum mechanics* calculations. NEMD thereby augments continuum mechanics, which can neither predict the initiation of fracture nor follow the behavior of a material as it strain-hardens or strain-softens.

To study atomistic mechanisms using the NEMD method, we need to input a description of how the atoms in a material interact. These interactions are affected by the behavior of electrons when the atoms in a material are brought together. We describe three possible electronic behaviors as follows:[6]

- The electrons surrounding each atom remain tightly bound to the atom; their motion is uniquely determined by the motion of the atom. This interaction represents materials that are gaseous at room temperature, such as nitrogen or oxygen
- The electrons are not bound to any atoms; the motion of these "free electrons" is independent of the motion of the atoms. This interaction represents simple metals, such as copper and nickel.
- The electrons are shared by many atoms; their motion depends on the motion of all surrounding atoms. This interaction represents covalently bonded materials, such as glass.

To accurately simulate these different materials, specific models must be used in NEMD simulations. We have two successful models—the Lennard-Jones force model[7] and the recently developed embedded-atom method (EAM).[8] At present, no satisfactory model exists for covalently bonded materials; we will focus on this issue in our future research. The Lennard-Jones force model accurately represents materials that are gaseous at room temperature. It has been used as the predominant method for simulating inter-atomic forces, but it is limited to modeling the first type of electronic interaction. Our FY 89 research focused on the EAM because it successfully models simple metals; never before have free surfaces and cracks in common, practical engineering materials been accurately simulated on an atomistic level.

We also began to work on an algorithm to take advantage of the new class of supercomputers. This new class represents the trend in supercomputing to use a large number of processors all working in parallel on the same problem (hence, the name *parallel computers*). This trend will define supercomputing in the 1990s and well into the 21st century. State-of-the-art parallel computers include the SPRINT (under continuing development at LLNL)[9] and the BBN-TC2000 (recently purchased by LLNL for the Massively Parallel Computing Initiative). Our group is among the first to create an algorithm that exploits the power of these new computers and gives us the data manipulation capabilities we need for our simulation research.

*Continuum mechanics is the study of distributions of energy, matter, and other physical quantities under circumstances where their discrete nature is unimportant. These physical quantities may be regarded as continuous functions of position.

Progress

Our research in FY 89 focused on two activities:

- Developing a strategy for simulating very large systems (several million atoms) on available parallel computers, and
- Simulating common engineering metals using the EAM; specifically, we simulated shallow depths of cut (about ten atomic layers), indentation processes, and the slow collision of disks made up of several hundred atoms.

We present here a description of these activities and the results of our simulations.[10]

Strategy for Using Parallel Supercomputers

A realistic three-dimensional simulation of a shallow depth of cut, ~0.1 μm, requires several billion atoms and generates an unmanageable amount of data.

However, a two-dimensional simulation contains all the relevant physics information for the indentation and orthogonal cutting processes. Thus, we can simulate in two dimensions the same depth of cut using only a few million atoms. Even so, an NEMD simulation of this size requires a supercomputer using an algorithm based on two information management ideas—data storage in limited memory and the division of the simulation problem into many small cells.

First, memory is at a premium. Hence, we require a minimum-storage algorithm. The Störmer algorithm[11] is one of several choices. We use a modified version of this algorithm because it requires only three storage locations per degree of freedom: two spatial coordinates (at different time-steps) and the force. The velocity values are not stored, but we evaluate them by using a central difference when we need to calculate the temperature or stress anywhere in the material.

For comparison, the popular fourth-order Runge-Kutta method[12] requires six storage locations per degree of freedom. This method requires twice as much memory storage as the Störmer method to produce the same level of accuracy. We have found that the accuracy and the central processing time for all commonly used integration algorithms are about the same, as long as the optimum time-step is used.* The Runge-Kutta method can use a time-step four times longer than the Störmer method, but it requires four times as many force evaluations (the most expensive part of any molecular dynamics simulation code).

Our second idea is to divide the molecular dynamics simulation problem into many small cells. By doing so, when we double the number of atoms in a simulation, we double the amount of computer time necessary to process our data. This is the most efficient use of computer time. The size of our cells depends on the interaction range of the inter-atomic potential, which we define as two atomic diameters. In practice, we assign each processor a region containing several hundred cells, which minimizes the amount of information the processors communicate at the end of each time-step. Subdividing requires an additional "linked-cell" list containing one entry per particle. We use this linked-cell storage algorithm in conjunction with the Störmer algorithm to integrate the equations of motion.

Thus, we have developed a minimum-storage algorithm that uses seven storage locations per atom in a two-dimensional simulation (two degrees of freedom per atom). This algorithm has proven to be the most efficient for the large systems we are simulating using the NEMD method, and it most advantageously uses the additional processing power of the SPRINT and the BBN-TC2000. Currently, we can simulate, at most, one million atoms. A SPRINT machine with 1000 times more processors than our current SPRINT will be able to simulate one billion atoms in the same amount of computer time we currently use to simulate one million atoms.

The computer simulation geometry we used for the NEMD simulations described here is much the same as the geometry reported in the FY 88 Thrust Area Report.[13] **Figure 2** represents a two-dimensional, one-atom-thick cross section of material used to model the indentation process. We establish a boundary layer of fixed atoms (not allowed to move), which represents the bulk material at the desired temperature and density. The next two layers (thermostat atoms) are allowed to move and are constrained to the desired temperature.[14] In this way, heat energy from the cutting tool (or the indenter) working at the surface is drawn away as if the cross section were much larger. The remaining 10 to 100 atomic layers are free to move unconstrained. These atoms interact with themselves, with the fixed and thermostat atoms, and with the moving tool. We study the indentation process by moving an indenter (the large ball in **Fig. 2**) into

*All of the algorithms we use are stable up to a certain amount of time in each time-step. This "maximum" time-step is the optimum for running an algorithm. The maximum is a different amount of time for each algorithm. It also depends on the nature of the inter-atomic interactions.

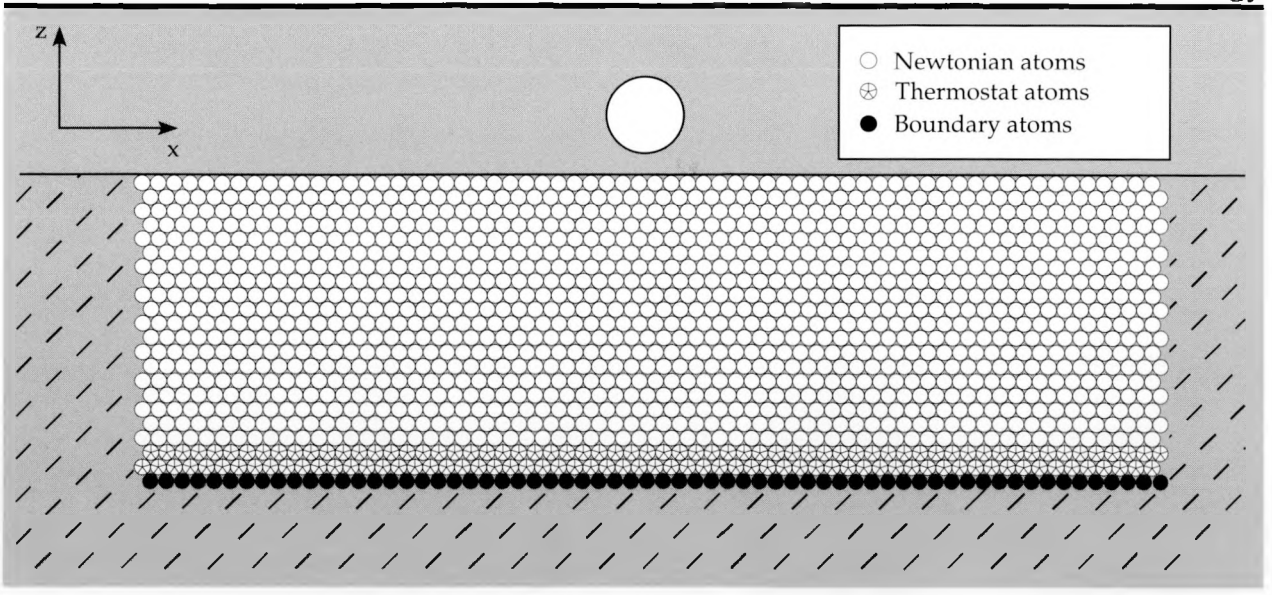


Figure 2. In a typical indenter geometry, the large circular indenter interacts with the atoms according to an extended Lennard-Jones force model. All of the atoms are constrained to move in two dimensions only. The bottom layer of boundary atoms is not allowed to move. The thermostat atoms are maintained at a constant temperature, and the remaining atoms obey Newton's equations of motion.

the top layer of the material. We study the cutting process by scraping a tool across the surface of the material (as in **Fig. 4**). Using the maximum allowable time-step, our cutting rates are limited to greater than 100 m/s for simple metals (high-speed machining).

The Embedded-Atom Method (EAM)

The EAM has revolutionized our ability to represent the inter-atomic forces in simple metals. Using the EAM, we have begun to investigate the response of these simple metals to external forces.

As previously described, metals contain free electrons. The EAM is a simple method that describes the interaction of atoms with these free electrons. **Figure 3** depicts the density of free electrons in a simple metal. Daw and co-workers[8] describe this free-electron interaction energy as the energy required to embed an atom into a region of charge density ρ. Because the free electrons spend most of their time close to the atoms, we can approximate the total charge density at the position of an atom as the linear superposition of the charge clouds onto the atom from its neighboring atoms:

$$\rho_i = \sum_{neighbors} f(r_{ij}) \quad ,$$

where ρ_i is the total charge density at the position of atom i, r_{ij} is the distance from atom i to one of its nearest neighbors (indexed by j), and $f(r_{ij})$ is a function that describes the density of the cloud of electrons around any atom. The energy to embed an atom is written as F(ρ), where F is the embedding function. The total embedding energy for our collection of atoms is the sum

$$F_{total} = \sum_{all\ atoms} F(\rho) \quad .$$

This energy is in addition to the ordinary atom-atom interaction energy described by the Lennard-Jones force model. To attain accurate results using the EAM, we need to use an empirical embedding function. We take $F(\rho) \propto \rho \log \rho$ (following a suggestion of Holian and Voter*). By varying the form of this embedding function and the electron density function, we can adjust our calculated properties to agree with the known properties of the real material.

Molecular Dynamics Simulation Results

We made a movie of our simulation of high-speed machining. **Figure 4** contains two frames from the movie. We shaded the different layers to illustrate how the atoms diffuse. Although the

*Brad Holian and Art Voter (from Los Alamos National Laboratory) have devised a simple embedded-atom potential based on the embedding function $F(\rho) \propto \rho \log \rho$. Their potential closely approximates the three-dimensional cohesive energy and vacancy formation energy of a material resembling copper or nickel within a two-dimensional calculation.

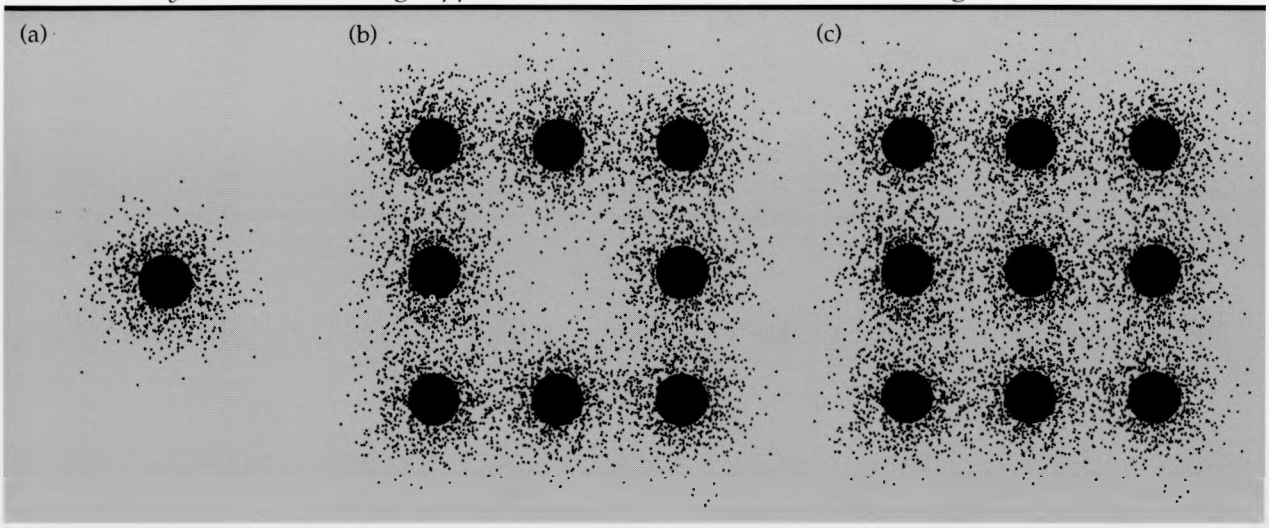


Figure 3. The free electrons in metals give rise to an additional cohesive interaction not described by the Lennard-Jones force model. The "clouds" around the atoms represent the electron density in an idealized two-dimensional crystal. (a) The electron density around a single atom. The solid ball represents the "core" electrons that give rise to the Lennard-Jones interaction. The cloud represents the "valence" electrons that become free when the atom is embedded into a metal. (b) The electron density around several atoms. The space in the middle represents a vacancy in the material. (c) The additional cohesive interaction is the energy required to embed the central atom [from (a)] into the electron density due to its neighbors at the position of the vacancy [in (b)].

depth of cut is only ten atomic layers, the simulation demonstrates the relevant physics of the machining process.

When the tool first impacts the material [**Fig. 4(a)**], the response is elastic, and a compressive shock wave propagates into the underlying material. As the simulation progresses, plastic deformation occurs—the tool performs work that causes permanent changes to the atoms' local environment. We observe the creation of voids where atoms used to be and slip dislocations (planes of atoms sliding across each other). Directly beneath the cutting tool in **Fig. 4(b)**, we observe appreciable subsurface damage, as indicated by the interlayer diffusion. The chip forming directly in front of the tool is qualitatively the same as a chip created in the actual machining process. Inside the chip, we again see voids and interlayer diffusion,

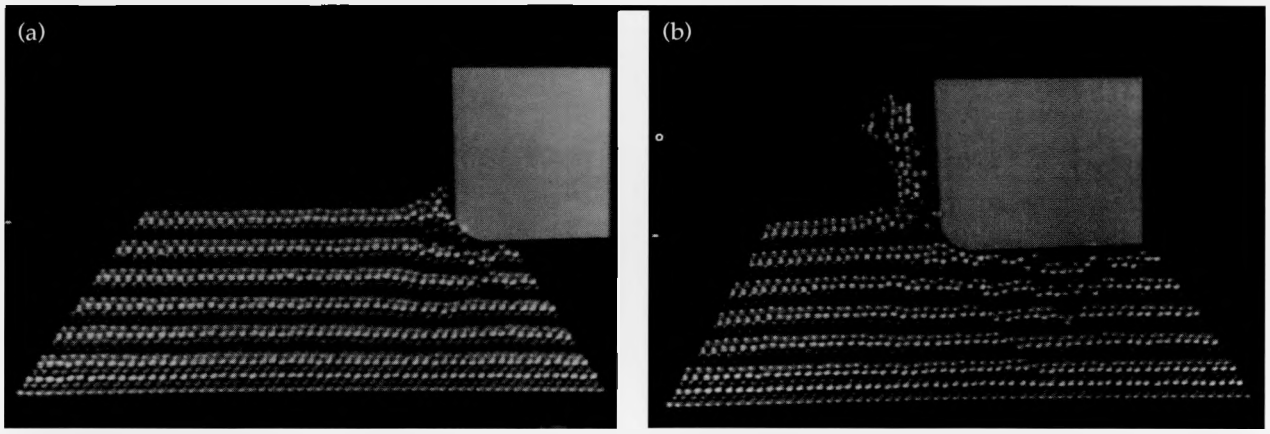


Figure 4. Two frames from our movie of the molecular dynamics simulation of orthogonal cutting of an embedded-atom material. We used 1526 atoms at a depth of cut of ten atomic layers, a temperature of about one-tenth the melting temperature (T_m = 1356 K for copper), and a tool speed of one-tenth the longitudinal speed of sound (V_b = 3700 m/sec for copper) in the embedded-atom material. (a) This frame is near the beginning of the simulation. The shaded bands illustrate atoms initially lying in the same layer, emphasizing interlayer diffusion as the simulation progresses. (b) This frame is near the end of the simulation and shows the same profile shown in Fig. 1(a). The chip is smooth and continuous, appreciable surface and sub-surface damage can be seen behind the cutting tool, and a noticeable slip dislocation propagates down to the lower boundary.

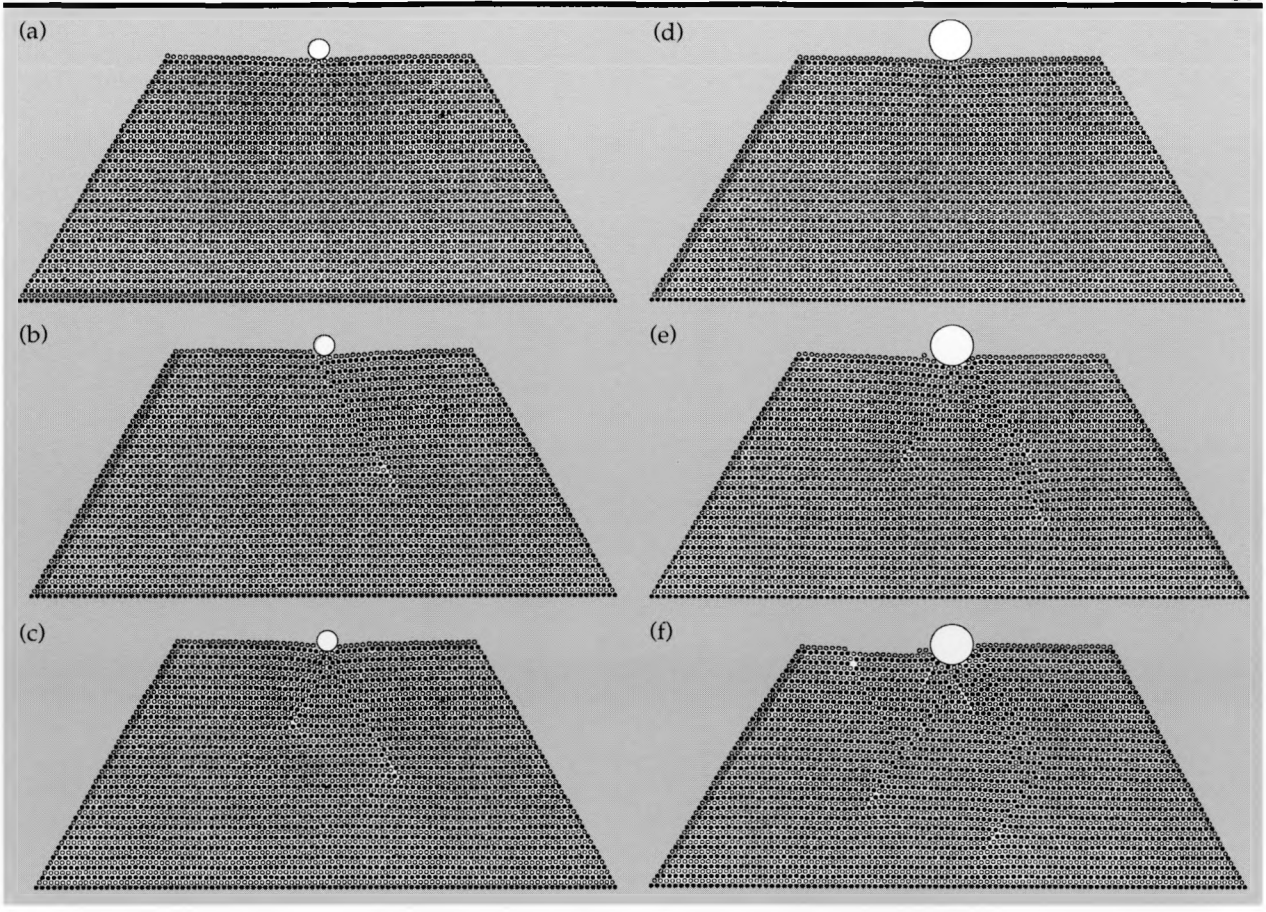


Figure 5. Several frames from our simulation of the indentation process. (a)-(c) Indentation using a 4σ-diam indenter (where σ is the atomic diameter). (d)-(f) Indentation using an 8σ-diam indenter. In both cases, the material contains 4000 atoms at a temperature of one-fourth the melting temperature. The indenter velocity is one-thousandth the longitudinal speed of sound. We shaded alternating layers black and white to illustrate the formation of slip dislocations and plastic deformation. Comparing the two indenter sizes, we observe that the edge of slip dislocations propagates farther into the material under the larger indenter than under the smaller one. Yet, the force needed to initially create a slip dislocation is the same for the two indenters.

which indicate strong shearing at the tool tip.

The morphology of the chip formed in our simulation using the EAM (shown in **Fig. 4**) is qualitatively different from the chip formed using the Lennard-Jones force model.[13] For instance, in the Lennard-Jones model, the chip has a strong tendency to vaporize or disintegrate, whereas in the embedded-atom model, the chip stays together. The EAM more closely models the behavior of metals, which are of greater practical interest.

We also simulated the indentation process using a circular indenter (see **Fig. 5**). We varied the size of the indenter from about 2σ to 20σ (where σ is the atomic diameter). We also varied the number of atoms in the material from a few hundred to a few tens of thousands. Shown in **Fig. 5(a)-(c)** is a 4σ-diam indenter and in **Fig. 5(d)-(f)** is an 8σ-diam indenter. In both cases, the material contains 4000 atoms. As the tool impacts the surface of the material, we observe elastic deformation [**Fig. 5(a) and (d)**], just as in our cutting simulation. As the tool continues to push down, we observe the creation of slip dislocations (plastic deformation), which allow the material to yield [**Fig. 5(b)** and **(d)**]. We also observe material bulging out around the sides of the indenter [**Fig. 5(c)** and **(f)**]. This bulging relieves stress, and the edge of the slip dislocation propagates back toward the indenter. Comparing the two indenter sizes, we observe that this edge propagates farther into the material under the larger indenter than under the smaller one. Yet, the force needed to initially create a slip dislocation is the same for the two indenters.

Because impact is fundamental to material removal processes, we also studied the impact of disks made up of a few hundred atoms. In **Fig. 6**, we show our simulation of this impact. We collided two disks of pure metal traveling at one-tenth

the speed of sound. These disks stick together (cold-weld). However, if surface impurities exist, the disks tend to bounce off of each other.

Future Work

In FY 89, we demonstrated the ability to accurately simulate the behavior of simple metals—copper, steel, etc.—and developed an algorithm that will allow us to run larger and more accurate simulations.

Our work will now focus on accurately modeling the deformation of brittle materials (primarily glass). Glass is difficult to model because of the nature of its electronic motion. A method has been developed to study this problem;[15] however, it can only model, at most, a few independent atoms. The many-body potential gives us another means for modeling glass.[16] It evaluates the interaction of two atoms by accounting for the positions of all neighboring atoms. However, the computer time required to calculate a many-body potential is approximately an order of magnitude more expensive than that required for the EAM. Nevertheless, we plan to incorporate this many-body potential into our computer code for modeling glass.

We also plan to couple molecular dynamics simulations to a finite-element code (a continuum model). Our atomistic simulations give us information unobtainable from continuum methods, but continuum methods are adequate for describing certain conditions. For example, in our indentation studies, the material far from the indenter deforms elastically; this deformation should be adequately described by continuum methods. By coupling a finite-element code with NEMD, we will introduce a moveable boundary layer into our simulation, within which the motion of the atoms is important, and outside of which the finite-element code should sufficiently model the deformation of the material.

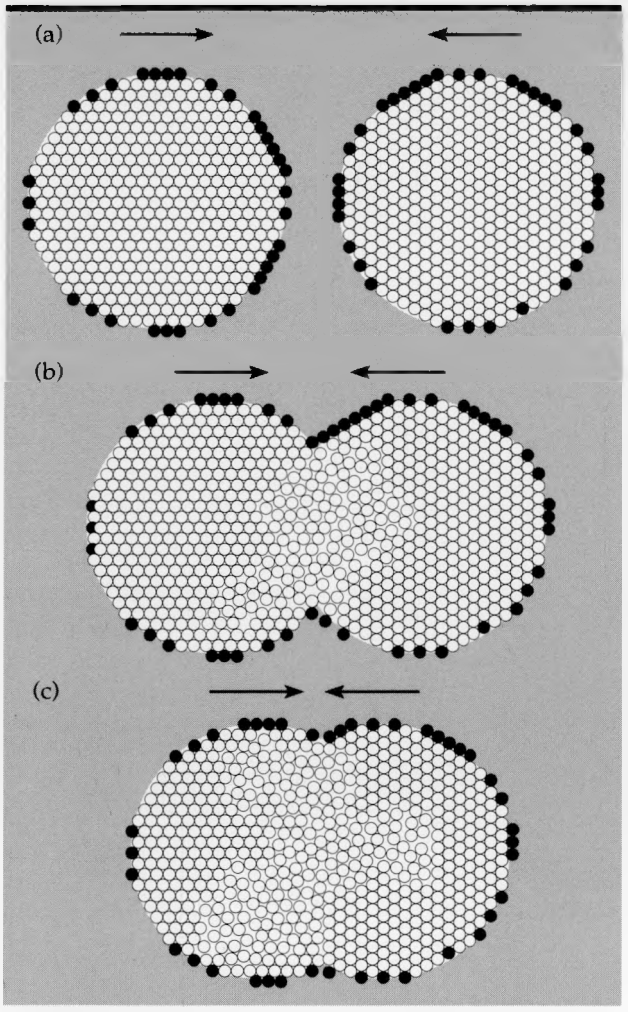


Figure 6. The impact of two disks made up of the embedded-atom material (such as a pure metal). The velocity of the disks is one-tenth the speed of sound. The disks stick together (cold-weld) and do not come apart. If the disks contain surface impurities, they tend to bounce off of each other.

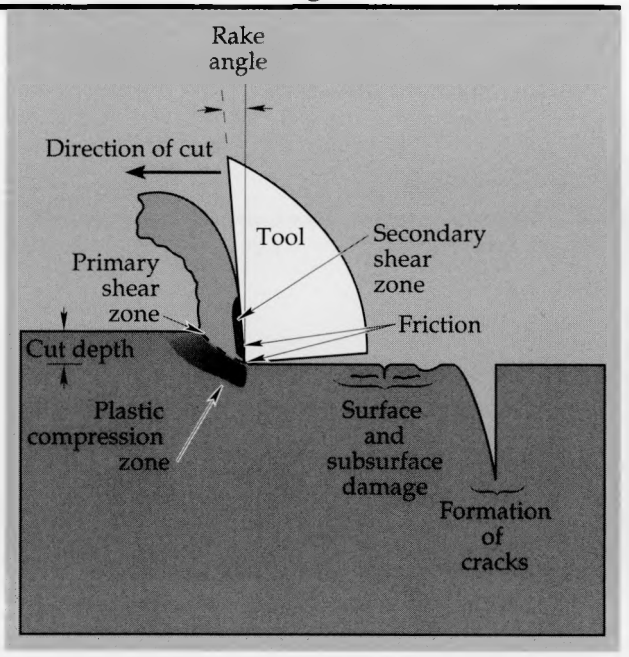


Figure 7. A cartoon of the orthogonal machining process; shown are the intense shearing of material that occurs directly in front of the tool tip, regions of high material compression, and residual surface and subsurface damage. Of particular interest is the formation of a large crack behind the cutting tool. By understanding how cracks form in brittle materials, we can optimize the machining effort—gaining optimum surface quality and lowering end product cost.

Finally, we plan to spend most of our time in FY 90 studying the cutting process. In **Fig. 7**, a cartoon of the cutting process, we illustrate some important features we will focus on. For example, the material near the tip is undergoing strong shear deformations, which depend on material properties, thermodynamic conditions, and cutting rate. We also show some surface and sub-surface damage to illustrate other factors affecting and resulting from the cutting process, such as the material's structure—crystalline or amorphous—and the presence of defects and/or grain boundaries. Behind the tool, a crack is forming. We will study the initiation and propagation of cracks, which are especially important to an understanding of the deformation of brittle materials. We will also study how the tool's work is distributed into heating the material and into plastic and elastic deformations.

Besides the basic physics questions we plan to answer through these activities, we will provide insight into specific engineering problems, such as tool velocities, tool shapes, rake angles, and tool wear. In summary, we hope to understand the underlying physics that determines the hardness of brittle materials, their deformation behavior and yield strength, strain-hardening, and the initiation and propagation of cracks and fracture. With this information, we will perform atomistic engineering, thereby designing the most appropriate materials and tools to use for specific purposes, such as precision optical components.

1. "Fabrication Technology," Section 4 of the *Engineering Research and Development Thrust Area Report FY 88*, UCRL-53868-88, Lawrence Livermore National Laboratory, 4-1–4-31 (1988).
2. R. Komanduri, "High-Speed Machining," *Mechanical Engineering* **107**(12), 64–76 (December, 1985); and M. C. Shaw, *Metal Cutting Principles* (Clarendon Press, Oxford, 1984).
3. K. L. Blaedel, P. J. Davis, and J. Franse, "Ductile Grinding of Brittle Materials," *Engineering Research and Development Thrust Area Report FY 88*, UCRL-53868-88, Lawrence Livermore National Laboratory, 4-26–4-28 (1988).
4. See *Proceedings of the Fifth International Precision Engineering Seminar*, Monterey, Calif. (1989). This seminar was dedicated to research into surface indentation and cutting.
5. For an elementary introduction to computer simulation methods, see H. Gould and J. Tobochnik, *An Introduction to Computer Simulation* (Addison-Wesley, New York, 1988); advanced details of molecular dynamics may be found in W. G. Hoover, "Molecular Dynamics," *Lecture Notes in Physics*, **258** (Springer-Verlag, Berlin, 1986); and M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Clarendon Press, Oxford, 1987).
6. For an excellent account of this view of cohesion in condensed matter, see W. A. Harrison, *Electronic Structure and the Properties of Solids: The Physics of the Chemical Bond* (Dover Publications, Inc., New York, 1989).
7. J. E. Lennard-Jones, "Interatomic Forces," in *Statistical Mechanics: The Theory of Properties of Matter in Equilibrium, 2nd ed.*, by R. H. Fowler, Chapter X 292–337 (Cambridge University Press, Cambridge, 1936).
8. M. S. Daw and M. I. Baskes, "Embedded-Atom Method: Derivation and Application to Impurities, Surfaces, and Other Defects in Metals," *Phys. Rev. B* **29**(12), 6443–6453 (1984); S. M. Foiles, M. I. Baskes, and M. S. Daw, "Embedded-Atom-Method Functions for the fcc Metals Cu, Ag, Au, Ni, Pd, Pt, and Their Alloys," *Phys. Rev. B* **33**(12-I), 7983–7991 (1986); and M. S. Daw, "Model of Metallic Cohesion: The Embedded-Atom Method," *Phys. Rev. B* **39**(11), 7441–7452 (1989).
9. A. J. DeGroot, E. M. Johansson, J. D. Fitch, C. W. Grant, and S. R. Parker, "SPRINT—the Systolic Processor with a Reconfigurable Interconnection Network of Transputers," *IEEE Trans. Nucl. Sci.* **34**(4), 873–877 (1987).
10. For further details, see: W. G. Hoover, C. G. Hoover, I. F. Stowers, and W. J. Siekhaus, "Interface Tribology via Nonequilibrium Molecular Dynamics," *Mat. Res. Soc. Symp. Proc.* **140**, 119–124 (1989); and W. G. Hoover, C. G. Hoover, I. F. Stowers, A. J. DeGroot, and B. Moran, "Simulation of Mechanical Deformation via Nonequilibrium Molecular Dynamics," *CECAM Workshop on Microscopic Simulation of Macroscopic Flows*, to be published (Brussels, August 1989).
11. C. Störmer, "Sur les Trajectoires des Corpuscules Électrisés dans l'Espace sous l'Action du Magnetisme Terrestre avec Application aux Aurores Boréales," *Archives des Sciences Physiques et Naturelles* 24, 221-247 (1907); and L. Verlet, "Computer 'Experiments' on Classical Fluids. Thermodynamic Properties of Lennard-Jones Molecules," *Phys. Rev.* **165**(1), 201–214 (1968).
12. J. B. Scarborough, *Numerical Mathematical Analysis, 3rd ed.* (Johns Hopkins Press, Baltimore, 1955); and W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes: The Art of Scientific Computing* (Cambridge University Press, New York, 1986).
13. W. G. Hoover, C. G. Hoover, and I. F. Stowers, "Interface Tribology by Nonequilibrium Molecular Dynamics," *Engineering Research and Development Thrust Area Report FY 88*, UCRL-53868-88, Lawrence Livermore National Laboratory, 4-11–4-14 (1988).
14. S. Nosé, "A Unified Formulation of the Constant Temperature Molecular Dynamics Methods," *J. Chem. Phys.* **81**(1), 511–519 (1984).

15. R. Car and M. Parrinello, "Unified Approach for Molecular Dynamics and Density-Functional Theory," *Phys. Rev. Lett.* **55**(22), 2471–2474 (1985).
16 F. H. Stillinger and T. A. Weber, "Computer Simulation of Local Order in Condensed Phases of Silicon," *Phys. Rev. B* **31**(8), 5262–5271 (1985); and J. Tersoff, "New Empirical Model for the Structural Properties of Silicon," *Phys. Rev. Lett.* **56**(6), 632–635 (1986).

Cathodic-Arc Process for Thin-Film Coatings

David M. Sanders and Steven Falabella
Materials Fabrication Division
Mechanical Engineering

David B. Boercker
H-Division
Physics Department

The cathodic-arc project is developing an ion-based physical vapor deposition process that promises to produce better thin-film coatings at faster rates than the conventional processes. Improvements desired in thin-film coatings are higher density, greater purity, improved adhesion, and, for ceramic coatings, more complete stoichiometry. The cathodic-arc process has the potential to provide all these improvements, but a problem has been the macroparticle impurities the process produces. We are applying computer models and verifying them with direct measurements of the plasma behavior in an effort to develop efficient means of filtering out the macroparticles. Using the cathodic-arc process in combination with a macroparticle filter we developed, we have demonstrated the ability to produce coatings of excellent thickness uniformity on large parts (up to 30 cm in diameter) at adequate deposition rates, as well as the ability to produce iron coatings significantly lower in oxygen impurity than previously attainable.

Introduction

Many programs at LLNL use thin-film coatings deposited by the physical vapor deposition (PVD) process. Coatings produced by traditional PVD techniques such as electron-beam evaporation and magnetron sputtering are satisfactory for some applications but not for others because of inadequacies in stoichiometry (chemical uniformity), adhesion, density, or resistance to high-energy laser damage. The cathodic-arc process we are developing offers the prospect of overcoming all these difficulties through its ability to provide copious quantities of charged particles (plasma) of virtually any conductive material for deposition as a thin-film coating. Unfortunately, the process also produces small droplets of unwanted material, called macroparticles, which form undesirable structural defects in the coating. Elimination of these macroparticles from the coating is one of our major goals.

The emphasis this year has been on the development of computer codes to model the operation of a magnetic filter designed to remove the macroparticles from the plasma stream emanating from the cathodic arc. We have carried out experiments to evaluate the accuracy of the computer models, and we have combined the cathodic-arc and macroparticle-filter technologies to provide thin-film coatings of superior quality for several Laboratory programs.

Progress

Cathodic-Arc Ion Self-Sputtering

Figure 1 illustrates some of the similarities and differences between three PVD technologies. All three processes take place in a vacuum chamber. The electron-beam (EB) evaporation process in **Fig. 1(a)** relies on thermal evaporation of the source material, which is heated with an electron beam. The atoms from the EB evaporation process are uncharged, and their kinetic energy is limited by the heating process to about 1 eV. Coatings produced by this process are often not fully dense, and they tend to have limited adhesion unless the substrate can be heated to a high temperature. These difficulties become more severe as the coating rate is increased. Moreover, the EB evaporation process normally requires a vertical orientation between source and substrate, with the source beneath the substrate, because of the necessity to contain the molten source material in its water-cooled hearth. However, a distinct advantage of the EB evaporation process is that it can be carried out at extremely high vacuum levels because no process gas is required. This feature can be exploited to produce

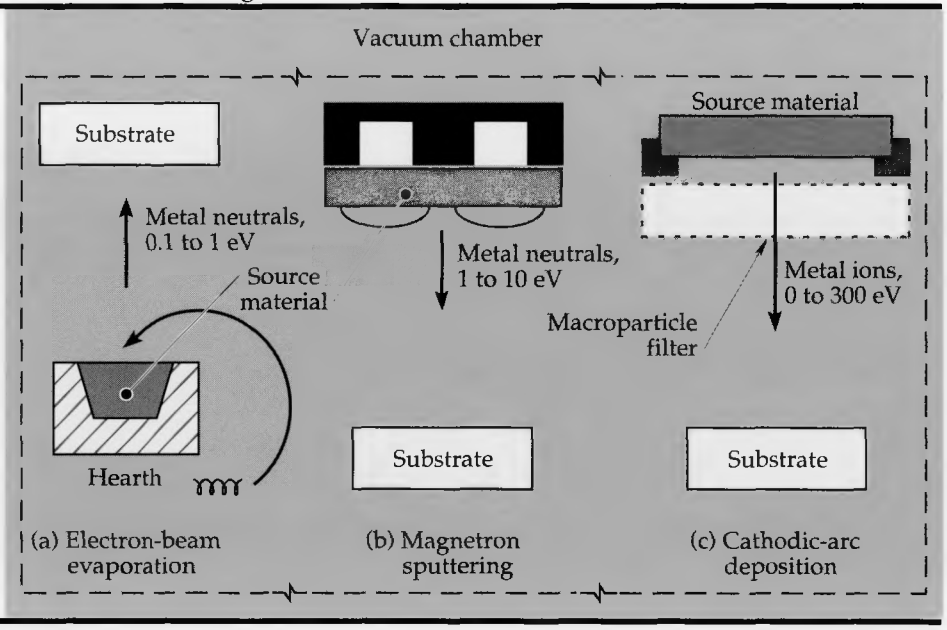


Figure 1. These three physical vapor deposition (PVD) processes are used to produce metallic and ceramic thin-film coatings. Each process has subtle advantages and disadvantages described in the text.

coatings having high purity.

In the magnetron sputtering process, shown in **Fig. 1(b)**, the source material is bombarded with ions of a process gas such as argon. The sputtered particles are in the form of neutral atoms having a somewhat higher kinetic energy than those produced in EB evaporation. This higher energy is thought to be responsible for the improvements in coating density and adhesion sometimes observed. Since magnetron sputtering is normally done using a solid source material (water-cooled to prevent melting), the source-to-substrate orientation of the process is not limited as in EB evaporation. This permits increased latitude in coating parts (substrates) having complex geometries. Although the sputtered particles have energies on the order of 10 eV when they leave the surface of the source material, their energy when they reach the surface being coated is normally lower as a result of collisions with the process gas atoms. In addition, atoms of the process gas are frequently trapped in the coating as impurities.

The cathodic-arc coating process,[1] shown in **Fig. 1(c)**, normally produces large quantities of ionized atoms with high kinetic energy, on the order of 30 to 80 eV; moreover, the kinetic energy can be increased or decreased by applying suitable bias potential to the substrate. Unlike the EB evaporation process, the cathodic-arc process is not limited in its physical orientation relative to the part being coated. The energetic sputtered ions of the cathodic-arc process—as compared with the lower-energy sputtered atoms in the other two processes—can produce coatings having improved density and adhesion, even at high deposition rates. In addition, their higher kinetic energy at the deposition surface makes it possible to produce stoichiometric ceramic coatings via a reactive process over a wide range of process conditions. To fully realize these advantages, however, it is necessary to remove the undesirable macroparticles also produced by the cathodic-arc process.

This year we have demonstrated that the cathodic-arc process for ion self-sputtering can produce nearly pure iron coatings needed for a diagnostic application; these coatings have significantly lower oxygen contamination than had been previously attainable with conventional magnetron sputtering. **Figure 2** shows the spectroscopic signatures of a magnetron-sputtered coating and a coating produced by the cathodic-arc process. The intensities of the two curves are normalized (i.e., equated) in the iron region. Lower on the energy scale, in the oxygen region, the intensity of the magnetron curve is much higher than that of the cathodic-arc curve, indicating a much greater proportion of oxygen in the magnetron-sputtered coating than in the cathodic-arc coating. We attribute the increased purity of the cathodic-arc coating to a higher coating density and to the ability of the process to operate at high vacuum without a sputtering gas in the vacuum chamber. The process produces coatings of very uniform thickness on parts as large as 30 cm in diameter. It has been successfully used to produce coated parts for various Laboratory programs.

Process Modeling

Magnetic Field Modeling. We have used the computer code POISSON, which was developed at Los Alamos National Laboratory, to predict the

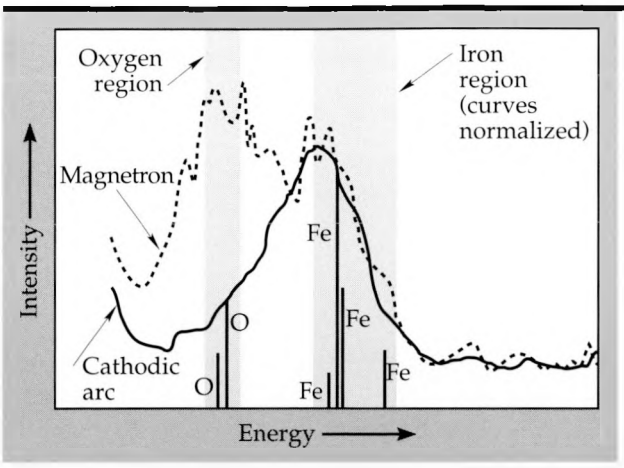


Figure 2. The spectroscopic signature of an iron coating produced by the conventional magnetron-sputtering process compared with that of one produced by the cathodic-arc ion self-sputtering process. The cathodic-arc coating has significantly less oxygen impurity than the magnetron-sputtered coating.

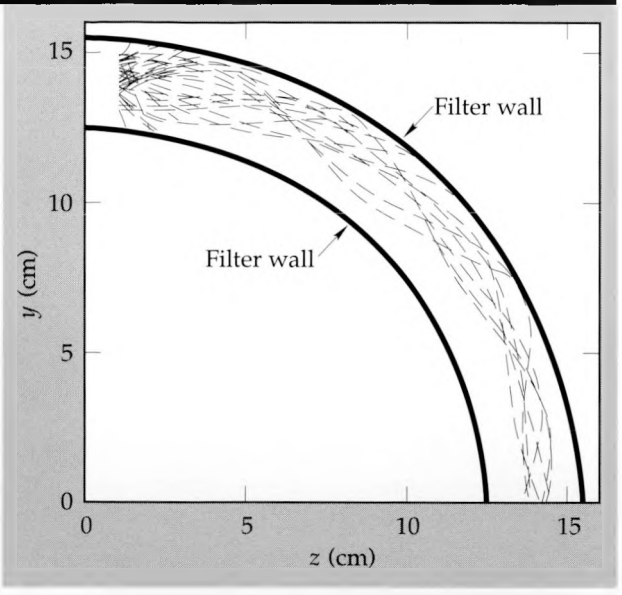


Figure 3. A simplified drawing of the quarter-torus macroparticle filter used to evaluate the QTRTORS computer code. The filter removes the macroparticles from the ion stream while passing most of the ions. Superimposed on the macroparticle filter are the ion trajectories predicted by the code. Ions colliding with the filter wall are interpreted as lost in transit through the filter; thus the ion current diminishes with distance in passing through the filter from upper left to lower right.

magnetic fields inside several macroparticle filters. We verified the code by making magnetic field strength measurements in various filters we have constructed. In addition, we were able to use POISSON to design and build a structure for positioning an electromagnetic arc, under computer control, over different areas of the arc target. This ability is useful for producing multilayer and alloy coatings, and the arc-positioning algorithm makes possible a more complete utilization of the target material with an associated reduction in operating cost.

The plasma transport code, described below, was designed to accept the output from POISSON. This has allowed us to carry out feasibility studies on new designs for macroparticle filters.

Plasma Transport Modeling. Aksenov[2,3] and his co-workers first proposed a macroparticle filter having a solenoid in the shape of a quarter-torus. The magnetic field of the filter constrains the motion of the electrons produced by the arc. It is thought that the ions are forced to follow the electrons through the filter in order to maintain local electrical neutrality of the plasma. This explanation is the so-called "flux tube" model first proposed by Morozov[4] and applied to the quarter-torus geometry by Aksenov.[2]

A major part of our modeling effort this year has been to evaluate the flux tube model's validity and its utility in designing new macroparticle filters having higher efficiency. We have written a code called QTRTORS, based on the flux tube model, to approximate the flow of ions through a filter. The code can handle an arbitrary magnetic-field configuration, and it has been tested using a geometry corresponding to recent experimental data.[5] Because of the simplicity of the model, it is computationally very efficient and has proven useful in making rapid and inexpensive assessments of the qualitative features of alternative filter designs.

Figure 3 shows a simplified configuration of the quarter-torus macroparticle filter used to evaluate the QTRTORS computer code, as well as the ion trajectories predicted by the code. The filter's effect is to remove the macroparticles from the stream of ions, but it also removes some of the ions as well. Ions that collide with the filter wall in **Fig. 3** are interpreted as lost in transit through the filter. **Figure 4** is a plot of the distance required for the ion current to decrease to $1/e$ of its original value as a function of the magnetic field applied to the macroparticle filter. The points represent actual data reported in Ref. 5, while the solid line represents the prediction of our computer code; the agreement is quite satisfactory.

The principal difficulty with the current QTRTORS code is that the electron density distribution is not calculated but rather is obtained from an assumed relation between the electron density and the magnetic field. The motion of ions is then calculated in the fixed electron distribution. While this approach is computationally efficient, it does

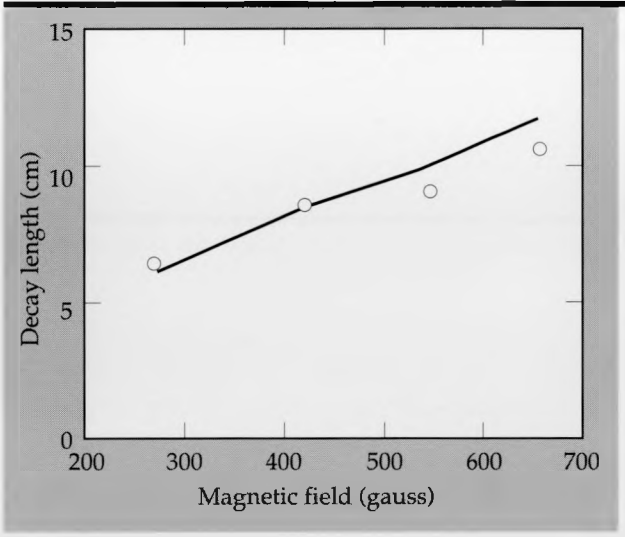


Figure 4. A plot of ion-current decay length (distance required for the ion current to decrease to 1/*e* of its original value) for the ion current in the macroparticle filter in Fig. 3 as a function of the magnetic field applied to the filter. The circles represent data points reported in Ref. 5, while the solid line represents the prediction of our computer code; the agreement is quite satisfactory.

not allow the electrons to respond to the ion motion and thus cannot guarantee that quasi-neutrality of the plasma is maintained. An improved model that will calculate the electron flow pattern self-consistently with ion motion is presently under development.

Cathodic-Arc Ion Source

In parallel with the work outlined above, we constructed and operated a cathodic-arc ion source. This source is currently being optimized to increase its ion transport efficiency. Even when operated in a less than optimal manner, however, the source can produce macroparticle-free ceramic coatings of TiN and Y_2O_3 an order of magnitude faster than has been possible using reactive magnetron sputtering.

Future Work

Next year we will complete the optimization of the cathodic-arc ion source and macroparticle filter to increase the system's transport efficiency. We will continue to use the system to produce coatings of various metals and ceramics, on which we will make property measurements. Our primary focus will be to obtain optimum coating characteristics such as grain size, density, and, in the case of reactive deposition, stoichiometry.

To support these efforts, we will continue to refine our plasma transport model. In addition, we will develop a molecular-dynamics model of the self-sputtering and ion-condensation processes which will enable us to predict the influence of various plasma characteristics such as ion energy on the resulting structure and properties of the coatings. The predictions of both these modeling efforts will be reconciled with experimental observations to verify their accuracy.

1. D. M. Sanders, "Ion Beam Self-Sputtering Using a Cathodic Arc Ion Source," *J. Vac. Sci. Technol.* **A6**(3), 1929-1933 (1988).
2. I. I. Aksenov, V. A. Belous, V. G. Padalka, and V. M. Khoroshikh, "Transport of Plasma Streams in a Curvilinear Plasma-Optics System," *Sov. J. Plasma Phys.* **4**, 425-428 (1978).
3. I. I. Aksenov, A. N. Belokhvostikov, V. G. Padalka, N. S. Repalov, and V. M. Khoroshikh, "Plasma Flux Motion in a Toroidal Plasma Guide," *Plasma Phys. Contr. Fusion* **28**, 761-770 (1986).
4. A. I. Morozov, "Focussing of Cold Quasineutral Beams in Electromagnetic Fields," *Sov. Phys. Dokl.* **10**, 775-777 (1966).
5. J. Storer, J. E. Galvin, and I. G. Brown, "Transport of Vacuum Arc Plasma Through Straight and Curved Magnetic Ducts," preprint of paper submitted to *J. Appl. Phys. Lett.* (1989).

Diamond Turning of Optical Single Crystals

Chol K. Syn, John S. Taylor, and Baruch A. Fuchs
Materials Fabrication Division
Mechanical Engineering

Stephan P. Velsko
Advanced Materials Group
Y-Division

We are studying diamond turning to evaluate whether it is an economically viable method for machining single-crystal materials to the shapes and surface finishes required for various optical components of high-power laser systems. We have identified the range of ductile cutting conditions under which a diamond cutting tool can produce smooth and damage-free surfaces in a number of single-crystal materials of interest to the Laser Program, and we have evaluated the amount of cutting that can be done by a diamond tool on these crystals before it wears to the point where it can no longer maintain ductile cutting conditions. Results are presented here for the single-crystal materials *l*-arginine phosphate (LAP), *l*-arginine acetate (LAAc), potassium dihydrogen phosphate (KDP), calcium fluoride (CaF_2), lithium calcium aluminum fluoride ($LiCaAlF_6$), and lithium niobate ($LiNbO_3$).

Introduction

Single-crystal optical materials are used in frequency converters, laser amplifiers (as host crystals for lasing media), and other optical elements in high-power laser systems. Large diamond-turned KDP crystals used as frequency converters[1] are an indispensable part of the Nova laser and have contributed enormously to the achievement of its many important milestones. Growing of the various single crystals and subsequent fabrication of them into large single-crystal optics of acceptable quality at an affordable cost are important factors to consider in planning for high-power laser systems now and in the future. Single-point diamond turning methods have been shown to provide an economical means (perhaps the only means) of mass fabrication of KDP crystals for the Nova laser. The purpose of the present study is to investigate the diamond turnability of crystals being considered for use in the next generation of high-power laser systems, including the crystals listed in **Table 1.**

Diamond turnability of a given material is determined by two factors: the cutting behavior of the material and the wear behavior of the diamond tool, under a realistically acceptable range of cutting conditions. The work material should be removed in a smooth ductile manner in order to create a smooth surface that is free of surface and subsurface damage. The diamond tool wear should be low enough so that the ductile cutting condition can be maintained until the given cutting operation is completed. The less tool wear, the less degradation of the surface finish and figure accuracy (i.e., how closely the final machined shape of the workpiece matches the intended shape), and hence the more surface area that can be cut before the tool wears out. Tool wear may be the most important parameter determining the quality of the machined surface and the productivity of the machining operation under a given cutting condition.

We have developed a cutting test method we call the "shoulder analysis" technique,[2] which allows us to evaluate the material's cutting behavior and determine the upper limit of the ductile cutting conditions without covering the entire matrix of cutting parameters. We have also developed a procedure called the "plunge-cut replication" method,[3] with which we can measure the diamond tool wear very accurately. Measurement of the surface finish as a function of cutting distance or of material removed provides an indirect measure of the tool wear.

Table 1 lists the six single-crystal types we investigated in this study. The samples were mostly small and few in number. We studied the cutting behavior of these samples and carefully examined the tool wear. The results are promising, showing good surface finish and mostly low or negligible tool wear. Further study seems warranted, including extended tool-wear tests to determine the economics of diamond turning of larger crystals of these materials.

Progress

Experimental Procedure

The crystals we studied were procured from several sources. As shown in **Table 1,** they were generally small and of various shapes. Only the CaF_2, $LiNbO_3$, and KDP crystal samples were studied by the shoulder analysis technique.

Diamond cutting tests were carried out with two precision turning machines, the Pneumo Precision MSE-326 facing machine[4] and the Precision Engineering Research Lathe (PERL-II).[5] A few small crystals were cut only with the Pneumo machine, while other small and large crystals were cut with both the Pneumo and the PERL-II, depending on the cutting requirements. Crystals cut on the Pneumo machine were oriented so that the cutting tool would move in a specified direction relative to the crystal orientation or cleavage planes. Crystals cut on the PERL-II were mounted either on or off the center of the spindle rotation, depending on the crystal size and the cutting condition requirement. As shown in **Fig. 1a,** the center mount produces a spiral cutting pattern centered at the spindle center, and the off-center mount produces a cutting pattern consisting of concentric arcs.

Single-crystal diamond cutting tools were obtained from several commercial sources; they were resharpened after use if required. If the tool wear after a period of use was not detectable under a high-power optical microscope, the tool was reused on a roughing or semifinishing cut without resharpening. Round-nosed tools were used, most of them with a nose radius of 2.54 or 3.18 mm, but a few with a nose radius of 7.62 mm. Each tool was inspected before and after use with a high-power optical microscope, and sometimes with a scanning electron microscope. Only two rake angles—0° and –45°, but mostly –45°—were employed in the cutting tests (see **Fig. 1b**).

Results and Discussion

Table 2 summarizes the experimental results: cutting direction relative to the crystal orientation or cleavage plane, cutting parameters (including depth of cut, feed rate, spindle rotation speed), cutting distance traveled by the cutting tool and amount of material removed by the tool, cutting tool wear, and surface finish achieved. Comments on the cutting behavior, surface finish, and cutting tool wear are given below for individual crystals.

LiNbO3: A crystal of (0001) orientation mounted on the center of the spindle of the PERL-

Table 1. Description of single crystal specimens tested in diamond-turning machining study.

Crystal name (abbreviation), molecular formula	Crystal symmetry, space group	Cleavage plane	Melting point (°C)	Crystal size and shape(mm)	Crystal orientation
Lithium niobate (LN), $LiNbO_3$	Rhombohedral R3C	$\{10\bar{1}\bar{2}\}$	1250	19 × 19 × 6, square	(0001)
L-arginine phosphate (LAP), $^{+}(H_2N)_2 \cdot CNH \cdot (CH_2)_3 \cdot CH(NH_3)^{+} \cdot COO^{-} \cdot H_2PO_4^{-} \cdot H_2O$	Monoclinic, $P2_1$	(100)	~130	44 × 19 × 16, parallelogram	(010)
Potassium dihydrogen phosphate (KDP), KH2PO4	Tetragonal, I42d	None	252.6	50 × 50 × 9, 25 × 25 × 9, square	Type II doubler, surface normal 60° away from z-axis
L-arginine acetate (LAAc), $CH_3 \cdot COOH \cdot H_2N \cdot CNH \cdot NH \cdot (CH_2)_3 \cdot CH \cdot NH_2$	Monoclinic, $P2_1$	None	~260	37 × 12 × 6, parrallelogram	(010)
Calcium fluoride (CaF), CaF_2	Cubic, Fm3m	{111}	1395	50 diam × 25 thick, round disk	Close to (111)
Lithium calcium aluminum fluoride (LiCAF), $(LiCaAl)F_6$	Rhombohedral, $P3_1/m$	None	~810	20 diam × 3 thick, round disk	Close to $(11\bar{2}0)$

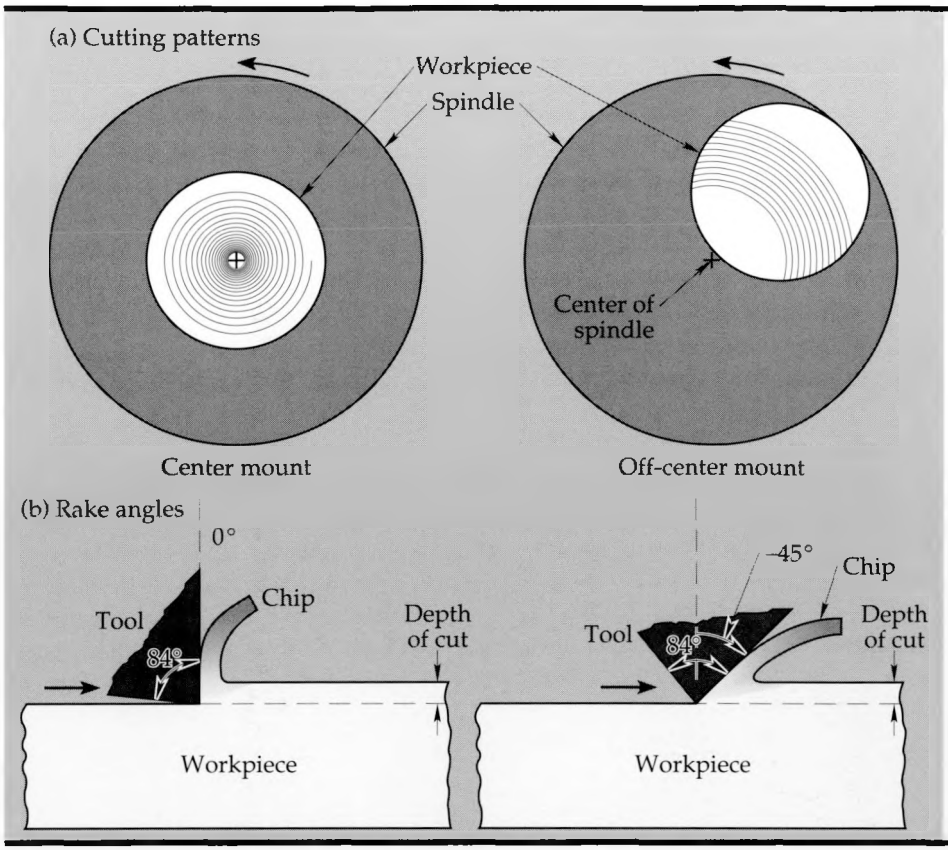


Figure 1. Schematic diagrams of (a) cutting patterns on PERL-II, (b) tool rake angles used in the study.

II for cutting showed a threefold symmetry of the surface finish, indicating the effect of the cleavage planes on the material removal process during the cutting operation, as shown in **Fig. 2.** When the cutting tool moved against the cleavage plane, the material was removed by brittle fracture rather than ductile plastic shear, a phenomenon commonly observed in turning of (111) silicon single crystals[2] as well as single crystals of LAP[6] and CaF_2.[7] Crystals cut along the direction parallel to a cleavage plane on the Pneumo machine showed a rather damage-free surface with a finish of better than 5 nm rms, while those cut on the PERL-II showed residual surface-fracture damage indicating the possibility of tool wear due to insufficient cutting fluid on the workpiece surface during the machining.

The diamond cutting tools showed a fairly serious tool-wear pattern upon repeated cutting operations. The wear pattern shown in **Fig. 3** indicates that the tool wear was caused by chemical reaction with the work material, perhaps by oxidation of the diamond by the niobium oxide in the work material. Also, the cutting edge showed a tendency for transverse thermal cracking as well as adhesion buildup of apparently melted work material, implying that the cutting temperature may have been extremely high. Thus a substantial reduction in the cutting edge temperature should be pursued by selecting a cutting fluid that provides more effective cooling and applying it in a more efficient manner, if the diamond turning of larger crystals is desired. A separate report is being prepared on the diamond turning of this material.

LAP: A detailed separate report on this material is already available.[6] The cutting conditions listed in **Table 2** can produce a surface smoothness better than 0.5 nm rms. Variation of surface smoothness

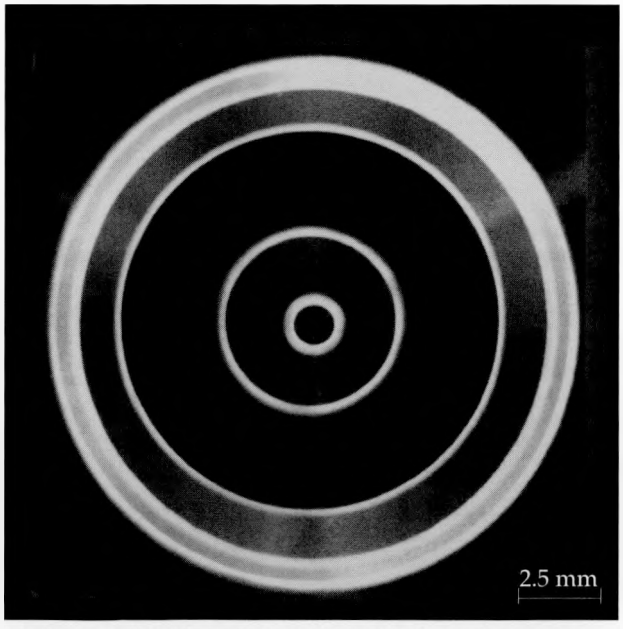


Figure 2. $LiNbO_3$ crystal with multiple cutting zones showing threefold symmetry in the surface finish.

as a function of cutting direction with respect to the crystal orientation was detectable, but it was rather negligible if the cutting conditions were properly controlled.

The cutting tools showed noticeable evidence of edge chipping and adherence buildup of work material on the surface. For the relatively short cutting distances employed in the earlier tests[6] on the Pneumo machine, the chipping tendency was negligible. But in the latest and far more extended cutting tests performed on the PERL-II, especially in roughing and semifinishing cuts, the edge chipping was much more severe, as shown in **Fig. 4.**

KDP: KDP crystals were cut on the PERL-II to provide a comparison with other crystals cut on that machine. One 50-mm-square sample was mounted on the center of the spindle of the PERL-II, and multiple steps were created around the

Table 2. Machining parameters and experimental results for study on diamond turning of optical single crystals.

Crystal	Cutting direction or mode	Cutting conditions DOC (μm)	FR (μm/rev)	Speed (rpm)	Cutting distance (km)	Material removed (mm^3)	Tool wear	Surface finish (nm rms)	Comments
LN	Spiral cut	20 10 20	2.5 6 2.5	1900 1000 1900	0.22	14.7	Rake face cratering and edge thermal cracking	5	• Rainbow finish
	Parallel to cleavage plane	25 10 5	2.5 2.5 2.5	1000 1000 1000	0.85	66		2.6-9.7	
LAP	Parallel to cleavage plane	50 25 12.5 5 2.5	5 5 2.5 2.5 2.5	1000 1000 1000 1000 1000	31.8 5.2 7.1 7.8 1.4	8230 671 252 200 8	Extensive chipping and material buildup	0.65-3.43	• Chipping tendency of tools
KDP	Spiral cut	50 25 12.5	5 5 2.5	1000 1000 1000	3.2 4.7	668 247	No detectable wear	0.75-1.98	• Excellent finish • Very forgiving for different cutting conditions
	Parallel to *a*-axis	50 25 12.5	5 5 2.5	1000 1000 1000	3.0 2.5 5.1	786 327 164			• No detectable tool wear
LAAc	Parallel to (110)	5 2.5	2.5 2.5	1000 1000	7.6 0.8	98 5	Chipping tendency	5.19-11.2	• Chipping tendency of tools • Needs larger and better quality crystals
CaF	Spiral cut	25 5 2.5	5 2.5 2.5	1000 1000 1000	4.0 3.2	144 2.1	Detectable cratering and edge cracking	1.75-3.14	• Threefold symmetry in surface finish
LiCAF	Parallel to a_1-axis	10.0 2.5	5 2.5	1900 1900	1.1	22	No wear detected	1.2-8.2	• Needs larger and better quality crystals

DOC— depth of cut; FR— feed rate; Speed— spindle rotation speed; Cutting distance and material removed are per individual tool.

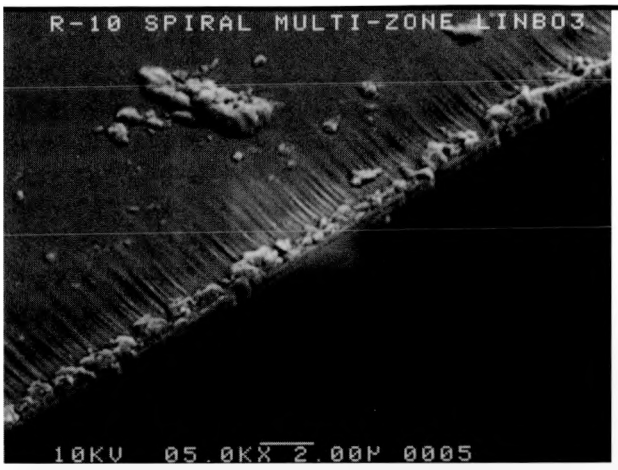


Figure 3. Diamond tool wear in cutting $LiNbO_3$ crystal shown in Fig. 2.

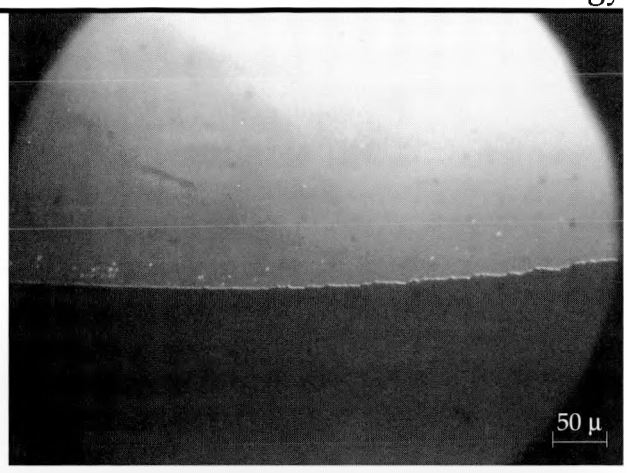


Figure 4. Diamond tool wear in cutting LAP crystal, showing serious edge-chipping tendency.

center during the roughing and semifinishing cuts by stopping the cutting tool slightly behind the termination point of the previous cut, as shown in **Fig. 5.** After the finishing cut on the rest of the surface, three additional steps with a wider interval were created to examine the material cutting behavior on the shoulder of each step.

On the steps of the roughing and semifinishing cuts, one can see almost symmetric lobes of surface fracture area (**Fig. 5**). This surface feature suggests that the prior blanking and handling process of the crystal may have created subsurface damage which, in turn, may have induced surface fracture along a certain crystal orientation during the cutting operation.

No tool wear was detected on the tool used in cutting the 50-mm-square sample, even though a substantial amount of material was removed during the roughing and semifinishing cuts, indicating that KDP may be a much more forgiving material for diamond turning than LAP. Four 25-mm-square samples were also cut on the PERL-II. They were mounted off-center on the spindle, but otherwise the cutting conditions were the same as for the 50-mm-square sample; again no tool wear was detected.

LAAc: These crystals were much smaller than the LAP crystals, but internal cracks and signs of inclusions were visible. The cutting tool used for the roughing cut showed a rather severe edge-chipping tendency, but the tool used for the finishing cut did not reveal any noticeable wear, probably because only a small amount of material was removed. Handling, mounting, and cleaning were rather difficult because of the small crystal size. The machined surface showed signs of surface contamination even after careful and repeated cleaning.

CaF_2: The 50-mm-diameter crystals were mounted for cutting on the center of the spindle of the PERL-II. A single tool was used to finish both sides of a crystal. While the first side was clean and transparent to the naked eyes, the second side showed a visible threefold symmetry pattern on the surface with alternating smooth and rough areas, as also noticed by Decker et al.[7] However,

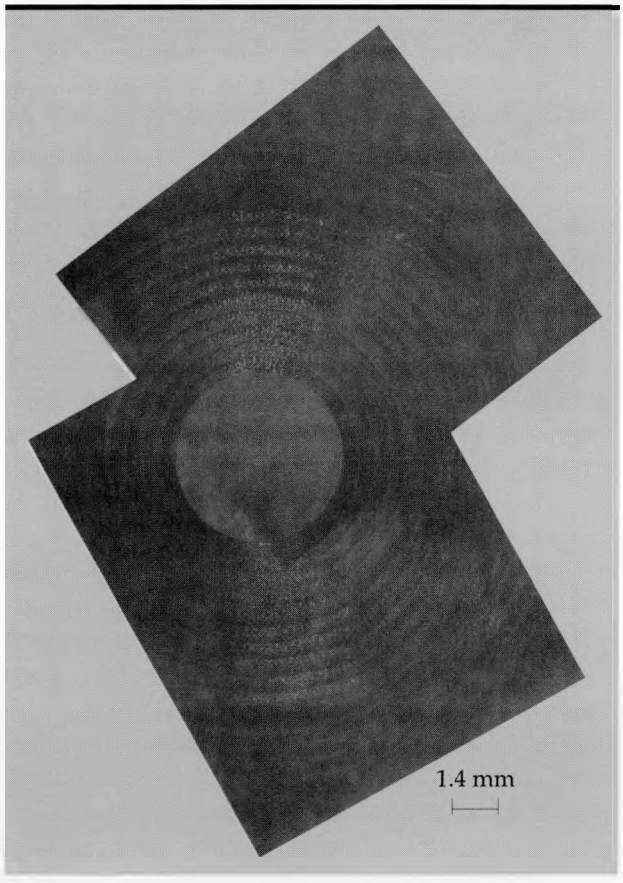


Figure 5. KDP crystal with multiple cutting steps to reveal the subsurface damage pattern.

even the first side showed a threefold symmetry in the surface finish when examined under an optical microscope. This progressive development of the threefold symmetry clearly indicates the effect of increasing tool wear as the cutting distance increases. The cutting tool used in the roughing and semifinishing cuts showed clear evidence of wear similar to that seen in cutting the $LiNbO_3$ crystals.

$LiCaAlF_6$: This round-disk sample with a thin wedge shape was cut from a small, experimentally grown crystal. The sample contained a very large crack, running through its thickness from the edge almost to the center, as well as numerous microscopic cracks and fractures. For cutting, the sample was carefully glued to a glass substrate and mounted off-center on the spindle of the PERL-II so that the large crack was parallel to the cutting direction. It was repeatedly cut many times with the same tool. Intermittent examination of the machined surface showed that the large crack did not grow, while most of the microscopic cracks were gradually eliminated by the machining. The tool used in cutting did not show any detectable wear, but the total cutting distance was rather short (under 1 km).

Other Crystals: Small urea and lithium formate crystals were cut on the Pneumo machine under the same cutting conditions as used for LAAc. Good surface finishes were obtained, comparable to those obtained with the crystals discussed above.

Tool Wear: The edge-chipping tendency of diamond tools used in roughing and semifinishing cuts of LAP and LAAc crystals on the PERL-II is perplexing in light of the softness of these crystals and their low melting points. Since these crystals can disintegrate into water, carbon dioxide, ammonia, and other residues at a temperature higher than their melting points, it may be possible that one or more of the disintegration products react with the diamond to induce chipping at the cutting edge, where the cutting temperature can exceed the melting point of the work material. However, this seems a rather unlikely possibility when one considers the generally recognized chemical inertness of diamond.

A more likely explanation for the edge-chipping observed in diamond cutting tools used on the PERL-II lies in the basic design of the machine as compared with the Pneumo design. Edge chipping may be more prevalent on the PERL-II than on the Pneumo because on the PERL-II the workpiece is mounted on the spindle and the tool on the slide, whereas on the Pneumo machine they are set up in exactly the opposite way. The workpiece on the PERL-II is whirling in a circle at the high spindle rotation speed, which causes the cutting fluid directed onto the workpiece surface to be whirled off. By contrast, on the Pneumo machine the workpiece is moving at the much slower slide speed, and the cutting fluid stays on it much longer. An ample supply of cutting fluid on the workpiece surface is an important factor in controlling the tool wear and the surface finish quality.

Another possible explanation for the edge chipping may have to do with the diamond tools used in our experiments. It is entirely possible that the particular set of diamond tools used in these cutting tests may be more prone to edge chipping than normal diamond tools. In fact, most of our tools were reused and resharpened many times, so that their cutting edges may have been weakened and thus unable to hold up very long in cutting operations. However, some of the same tools were used in cutting KDP crystals without chipping. In any event, diamond turning of LAP or LAAc crystals as large as the KDP crystals used in the Nova laser may require more stringent control of cutting conditions and tool selection.

Shoulder Analysis: We shoulder-analyzed our $LiNbO_3$, KDP, and CaF_2 crystal samples. This analysis technique[2] requires a rapid retraction of the cutting tool during the machining to create a shoulder between the machined and unmachined or previously machined surface, as shown in **Fig. 6.** One looks for the transition in cutting behavior from brittle fracture to ductile shear on the shoulder, and measures the location of the transition point in terms of the separation from the machined surface. If such a transition point is well separated from the machined surface, the brittle fracture occurring on the upper part of the cutting zone of the tool will not leave any damage on the machined surface, since the lower part of the tool is cutting in a ductile shear mode.

The results showed that KDP and CaF_2 crystals could be machined in the ductile-shear cutting mode at a feed rate of 5 μm/rev and a rake angle of –45°, with the transition point separated from the machined surface by 5 and 7.5 μm respectively. $LiNbO_3$ can be shear-cut with the same clearance of the transition point from the machined surface if the feed rate is reduced to less than 0.8 μm/rev, which is a possible though rather unrealistic cutting condition. But such a greatly reduced feed rate will greatly increase the cutting time, which is not desirable since it will in turn increase the possibility of the thermal drift of the machine tool, which may adversely affect the surface finish and figure accuracy of the workpiece.

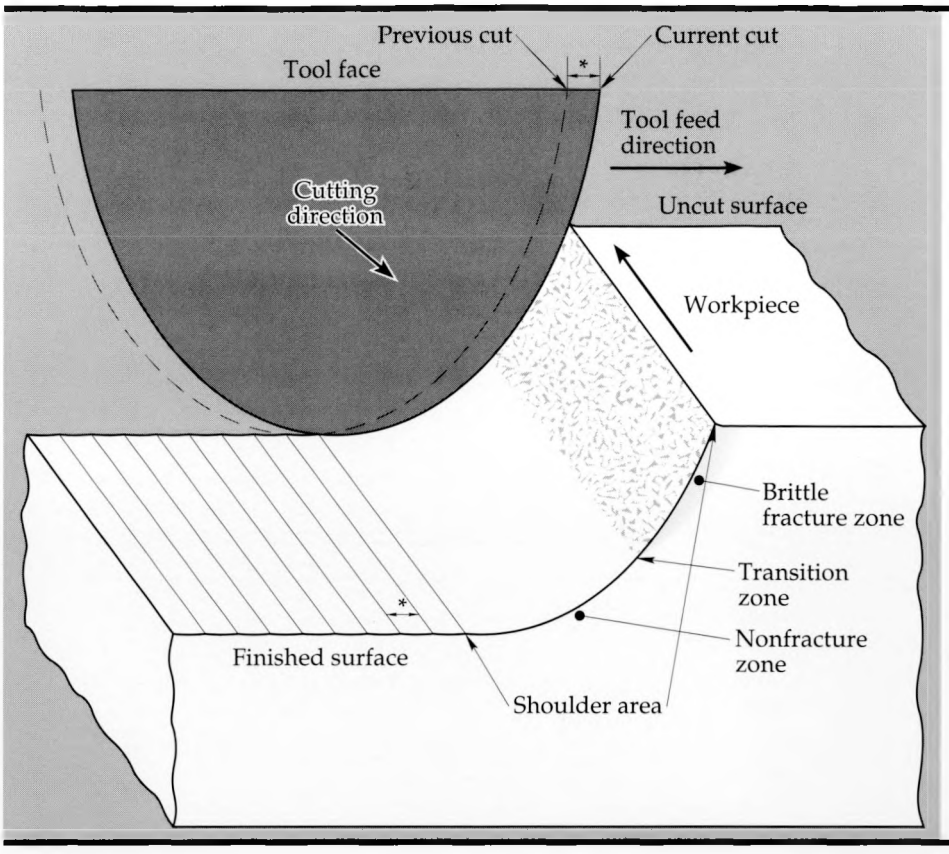


Figure 6. Schematic diagram of the shoulder with a ductile-brittle transition.

Future Work

We plan to continue diamond turning studies on small single-crystal specimens having optical properties that are attractive for applications in high-power lasers. However, our testing of the diamond-turning fabricability of candidate single-crystal materials is only part of the work needed to evaluate whether the materials can be successfully used in optical components for high-power lasers. For example, the candidate materials also need to pass tests for resistance to damage from intense laser beams. In addition, it must be demonstrated that successful diamond turning of the relatively small specimens in our studies can be extended to the considerably larger crystals required in the fabrication of full-size components for high-power lasers.

The LAP and LAAc crystals turned in the present studies were submitted for laser damage tests to measure the damage threshold values, which can be another factor affecting the diamond-turning fabricability of these crystalline optics in comparison with fabrication by other methods such as grinding and polishing. However, the quality of the crystals in terms of purity and inclusions can obscure the results of the laser damage tests, and hence more cutting and damage testing with better crystals seems necessary to establish the damage threshold values and their relationship with the cutting parameters.

Even if an excellent damage threshold value is obtained with small high-quality diamond-turned crystal samples, the overall success of such a crystal material is still dependent on the tool wear characteristics during the cutting of full-size crystals, which may be much larger than the laser damage test samples. Cutting studies done on small and odd-shaped crystals do not show a clear picture of the crystals' cutting behavior. For any single-crystal material seriously considered as a candidate for use in high-power laser optics, a more systematic cutting study using the shoulder analysis technique should be performed on reasonably large crystals (25 to 50 mm in diameter) of good quality.

The present study, like most other diamond-turning studies, bases its judgment as to a material's turnability on whether the machined surface is specular (i.e., mirrorlike) and fracture-free after the final finishing cuts. However, subsurface damage created during the roughing and semifinishing cuts may not be completely removed and could cause problems later. Laser damage tests may be able to detect the existence of such fabrication-induced subsurface damage if an adequate database is accumulated. But the laser damage tests may never be able to ascertain the nature

of such subsurface damage and its structure. Thus it would seem prudent to include a systematic characterization of subsurface damage by suitable techniques such as x-ray microtopography and transmission electron microscopy in our evaluation of the effects of diamond turning on single crystals.

1. M. A. Summers, J. D. Williams, B. C. Johnson, and D. Eimerl, "Harmonic Generation at High Peak Power," SPIE Vol. 622, High Power and Solid State Lasers (1986). See also "Frequency Conversion of the Nova Laser," *Energy and Technology Review,* UCRL-52000-86-4/5, pp. 2–11 (April/May 1986).
2. R. R. Donaldson, C. K. Syn, J. S. Taylor, and R. A. Riddle, "Chip Science: A Basic Study of the Single-Point Cutting Process," *Thrust Area Report FY-87, Engineering Research and Development,* UCRL-53868-87, pp. 6-12–6-15. See also C. K. Syn, J. S. Taylor, R. R. Donaldson, and S. Shimada, "Ductile-Brittle Transition of Cutting Behavior in Diamond Turning of Single Crystal Silicon," Proc. 1988 Japan Society of Precision Engineering Annual Spring Meeting, Meiji University, Kawasaki, Japan, April 16–18, 1988, pp. 735–36.
3. C. K. Syn, J. S. Taylor, and R. R. Donaldson, "Diamond Tool Wear vs. Cutting Distance on Electroless Nickel Mirrors," Proc. SPIE International Society for Optical Engineering, San Diego, California, August 18–22, 1986. See also UCRL-95513.
4. B. A. Fuchs, P. P. Hed, and P. C. Baker, *Fine Diamond Turning of KDP Crystals,* UCRL-93974 (February 1986).
5. R. R. Donaldson and D. C. Thompson, "Design and Performance of a Small Precision CNC Turning Machine," *CIRP Annals,* Vol. 35, No. 1 (1986). See also D. C. Thompson, "Precision Engineering Research Lathe," *LLNL Laser Program FY 83 Annual Report,* UCRL-50021-83.
6. B. A. Fuchs, C. K. Syn, and S. P. Velsko, *Diamond Turning of L-Arginine Phosphate, a New Organic Nonlinear Crystal,* UCRL-100404 (December 1988).
7. D. L. Decker, D. J. Grandjean, and J. M. Bennet, "Optical and Surface Physical Characteristics of Diamond-Turned Infrared Window Materials," *Proc. 1979 Boulder Damage Symposium on Laser Induced Damage in Optical Materials,* NBS Special Publication 568, p. 199.

Evaluation of Electroformed Copper for Shaped-Charge Applications

Jack W. Dini
Materials Fabrication Division
Mechanical Engineering

William H. Gourdin
Chemistry and Materials Science Department

We observed the behavior, in explosive tests, of conical shaped-charge liners electroformed in a proprietary acid copper sulfate solution. We then compared the test results to the dynamic failure observed in rapidly expanding ring experiments performed on the same material and to previous ring test results from benchmark wrought copper rings. In their respective tests, the electroforms—both liners and rings—exhibited poor ductility at the high strain rates encountered, suggesting that the tests may be comparable in evaluating high strain rate performance of materials. This comparability of tests, in turn, suggests that the electromagnetic ring test, which is considerably less expensive to perform than explosive tests on full-size liners, can be used as a screening test in evaluating materials for shaped-charge liner applications.

Introduction

In this project, we observed the behavior, in explosive tests, of conical shaped-charge liners electroformed in a proprietary acid copper sulfate solution (UBAC No. 1, OMI International Corp., Warren, MI) and compared the test results to the dynamic failure observed in rapidly expanding ring experiments performed with ring specimens electroformed from the same material. Results of previous tests on wrought rings,[1,2] including several rings formed from liners known to perform well, served as benchmarks for the current expanding ring tests. This work is important to those LLNL programs that are studying materials or processes which will enhance the performance of shaped charges.

A typical conical charge is shown in cross section before detonation in **Fig. 1** and during detonation in **Fig. 2**. As illustrated, the collapse of the conical liner material on the centerline forces a portion of the liner to flow in the form of a jet (**Fig. 2**) in which the jet tip velocity can be in excess of 10 km/sec. When this extremely high-energy jet strikes a metal plate, it causes a deep cavity to be formed. Because of the presence of a velocity gradient, the jet will stretch until it fractures into a column of jagged particles. Once the jet breaks into fragments, its ability to penetrate is dramatically reduced.[3] **Figure 3** shows a degradation in jet quality for wrought copper liners as grain size increases to 50 μm or larger. At 10-μm, 20-μm, and 30-μm grain size, the jet particles appear to be coherent, ductile, and well aligned, but at 50-μm grain size, the jet particles begin to fragment in an apparently brittle fashion and the jet quality becomes progressively poorer as the grain size increases.[4]

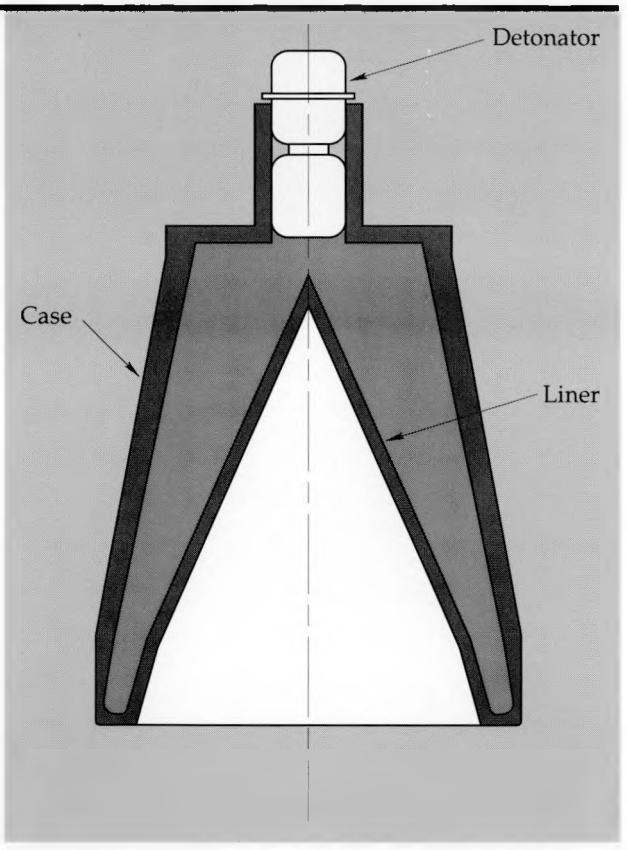


Figure 1. Cross section of a conical shaped-charge before detonation.

The fine, uniform grain size associated with the greatest penetration of a conical shaped-charge liner can be difficult to produce in a pure metal

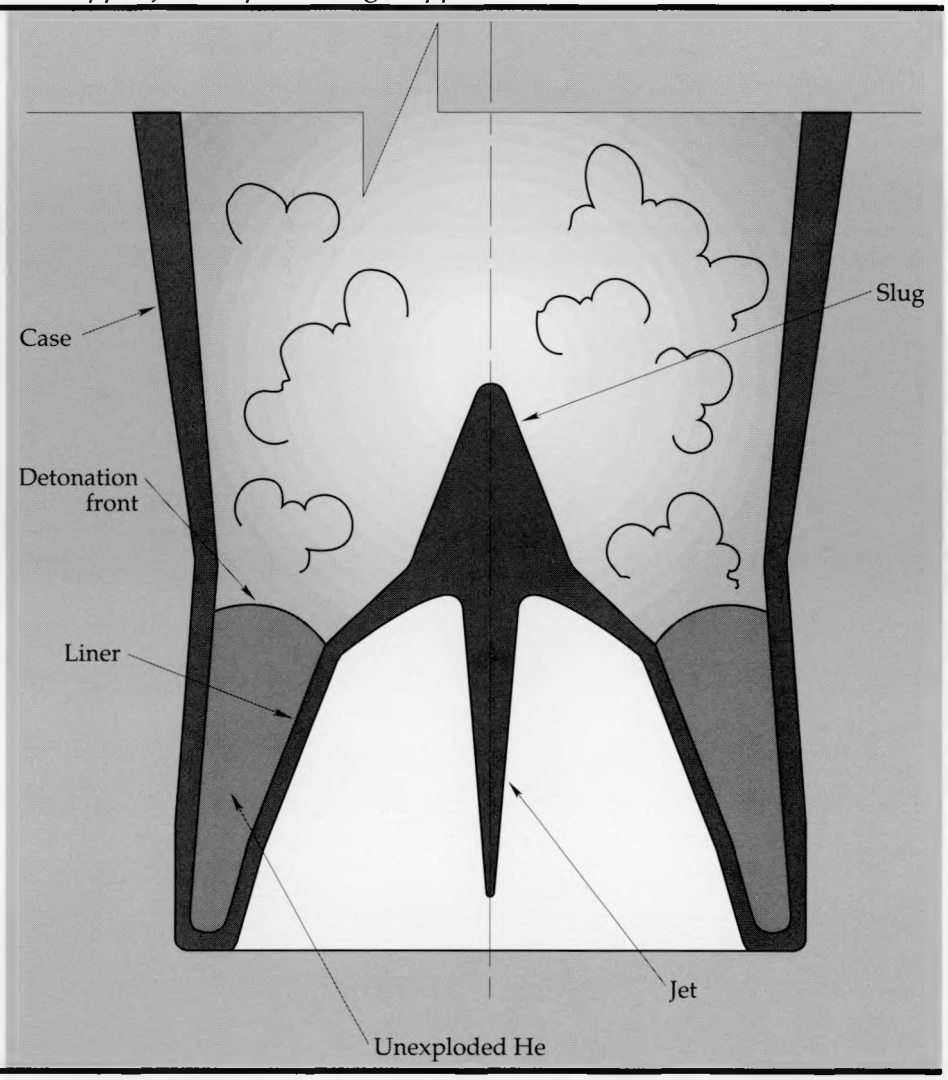


Figure 2. Cross section of a conical shaped-charge during detonation, illustrating the simultaneous formation of a low-velocity slug and a high-velocity jet.

when traditional deep drawing and coining or spin forming operations are used. Alternative methods that permit better process control or produce a refined structure are, therefore, of interest. Electroforming—the production or reproduction of articles by electrodeposition of material onto a mandrel or mold that is subsequently separated from the deposit—is one alternative offering a rapid and economical means of fabricating liners in a variety of shapes.

The two shaped-charge liner electroforms fabricated for our tests were 42° cones with a wall thickness of 2 mm, a length of 120 mm and an open end diameter of 94 mm. The electroformed ring specimens for the expanding ring tests had an inner diameter of 3.1 cm, a wall thickness of 1 mm, and a height of 1 mm. These were fabricated by thickly plating copper on aluminum mandrels, machining the outer diameter to final dimension, cutting the plated mandrels into individual 1-mm-thick slices, and then chemically dissolving the aluminum mandrel material.

For testing, the shaped-charge liners were assembled into test bodies and loaded with explosive. They were then fired at long stand-off distances and photographed with triple-flash radiography. The ring specimens were tested by the electromagnetic ring expansion test, which has been used at LLNL for several years to obtain dynamic tensile and failure data for metals at peak strain rates of approximately 10^4/sec and higher.[5,6] The principle is straightforward (**Fig. 4**). A capacitor charged to several thousand volts rapidly discharges through a solenoid surrounded by the specimen ring. As in a common electrical transformer, the current in the solenoid induces a large current in the ring that interacts with the magnetic field of the solenoid and its own magnetic field to produce a large outward force, causing the ring to expand rapidly. Two small explosive detonators act as fast (<50-nsec) switches: the first initiates the experiment, and the second removes the capacitor from the circuit at a time chosen to minimize residual magnetic forces and subsequent specimen heating.

Progress

Explosive tests of the two electroformed conical shaped-charge liners showed that the jet was quite unstable and exhibited poor ductility similar to that shown for wrought copper with 50-μm grain size (**Fig. 3**), this in spite of the fact that the grain size of the electroformed copper was extremely small (unresolvable at a magnification of 200 times).

Electromagnetic ring expansion tests were performed on the electroformed copper rings in the as-deposited condition and as heat treated at 250°C and 350°C for 1 hour to stabilize the deposit. The overall strains at failure from these tests are compared in **Table 1** with those for wrought oxygen-free electronic (OFE) copper with 10-μm grain size[1,2] and wrought OFE copper with 25-μm grain size, the latter taken from 81-mm shaped-charge liner. As the table shows, the electroformed specimens showed a significantly lower overall strain to failure than either of the wrought materials.

Both the explosive tests with the electroformed shaped-charge liners and the ring expansion tests with the electroformed rings revealed poor ductility in the electroformed specimens. Such test results suggest that the two tests are comparable in evaluating high strain rate performance of materials. This, in turn, suggests that the electromagnetic ring test, which is considerably less expensive to perform than explosive tests on full-size liners, can be used as a screening test in evaluating materials for shaped-charge liner applications.

Table 1. Overall strains achieved for five ring materials (electroformed UBAC copper and OFE [Oxygen-free Electronic Grade] copper) tested in the electromagnetic expanding ring tester.

Material	In (r/r_0)
UBAC—as deposited	0.38
UBAC—heated at 250°C for 1 hour	0.42
UBAC—heated at 350°C for 1 hour	0.44
OFE with 10-μm grain size	0.49
OFE liner material, 25-μm grain	0.52–0.56

Table 2. Vickers hardness of electroformed UBAC and OFE (Oxygen-free Electronic Grade) copper.

Materials	Vickers hardness (kg/mm^2)
UBAC—as deposited	114 ±3
UBAC—heated at 250°C for 1 hour	101 ±4
UBAC—heated at 350°C for 1 hour	93 ±4
OFE with 10-μm grain size	60

OFE ring specimens with 10-μm grain size, prepared from heavily worked starting material (Vickers hardness 130 kg/mm^2), were heat treated for 10 minutes at 300°C in a sand bath fluidized with flowing argon to produce the final grain size and a hardness of 60 kg/mm^2. The hardness of the electroformed copper rings, in contrast, drops relatively little with much longer heat treatments at comparable temperatures, as shown in **Table 2**.

The dramatic drop in the hardness of the OFE material after such a brief heat treatment compared with the persistently high hardness of the electro-

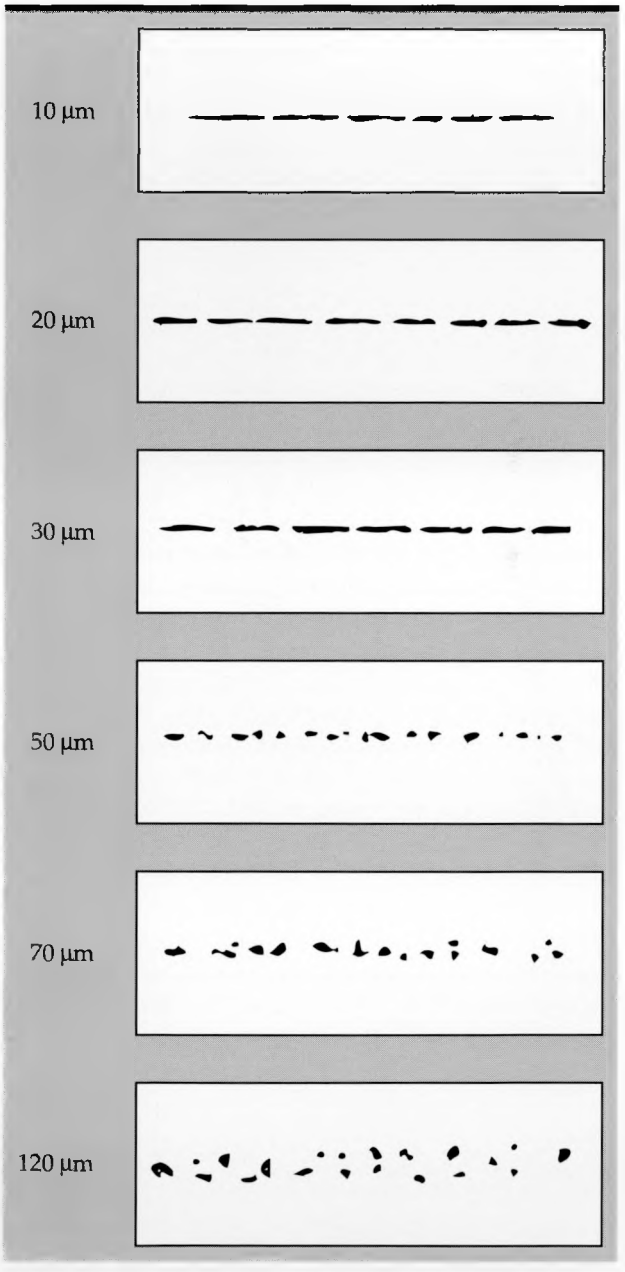


Figure 3. Flash radiographs of jets from a Ballistic Research Laboratory 8-mm shaped-charge liner showing jet breakup as a function of grain size (Ref. 4). Jet quality degrades as grain size increases to 50 μm or larger.

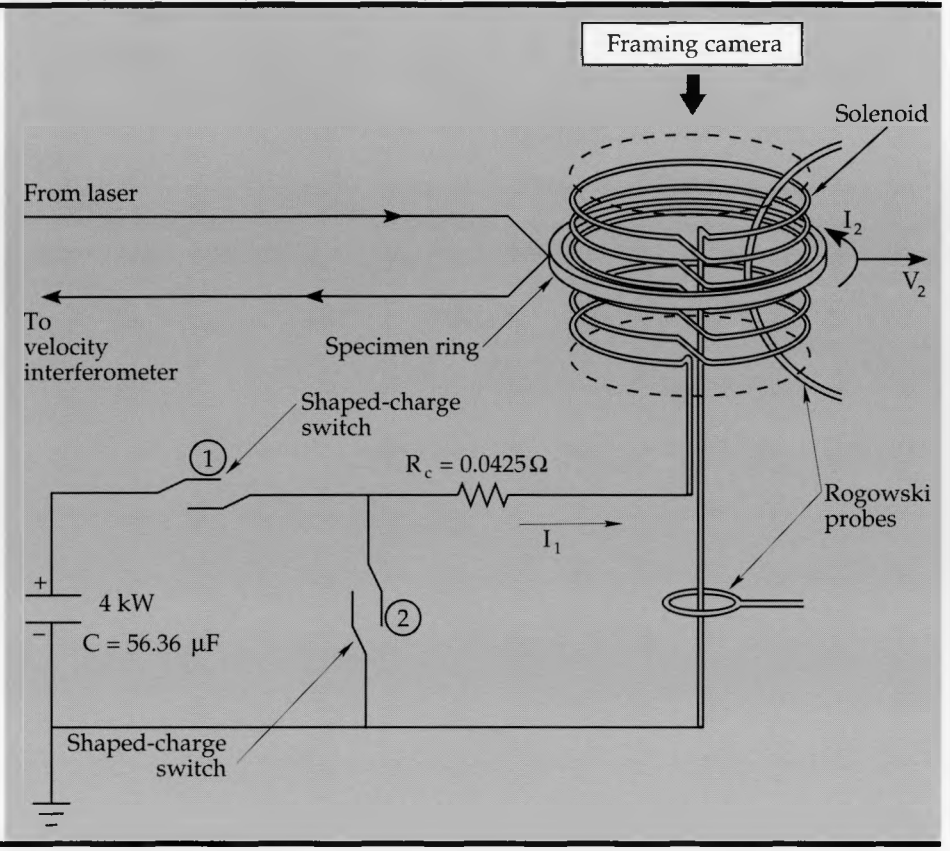


Figure 4. Apparatus used to electromagnetically expand ring specimens. Rogowski probes measure currents from which are calculated the temperature rise from Joule (resistive) heating. The ring expansion speed is calculated from interference fringes obtained with a velocity interferometer.

formed copper even after 1 hour at 350°C suggests that impurities in the electroformed copper may inhibit dislocation and grain boundary motion and arrangement, effectively preventing recovery and recrystallization. While such behavior appears generally consistent with the high impurity levels in the electroformed copper (**Table 3**) and the observation of void formation at grain boundaries for specimens heated at 450°C and higher,[7] we have not yet attempted a detailed characterization of either the role or distribution of solutes in our specimens. We speculate, however, that impurities may modify the constitutive behavior or produce grain boundary embrittlement that leads to plastic instability and failure at small overall strains as compared with purer materials of comparable grain size. Indeed, other researchers have shown that a few parts per million of sulfur embrittles cast, high-purity (99.999%) copper.[8]

Table 3. Impurities in parts per million (ppm) for two forms of copper utilized in the electromagnetic expanding ring tester.

Copper	C	H	O	N
Proprietary acid copper sulfate, UBAC No. 1, OMI International Corp., Warren, MI	50	7	70	10
OFE (Oxygen-free Electronic Grade)	<10	1	50	<10

Future Work

Our future work will evaluate copper deposits with lower impurity contents than were present in the copper used in this study. For example, deposits produced in a cyanide solution have markedly fewer impurities than do the acid sulfate deposits used in this present study (**Table 3**), and specimens thus deposited will be evaluated. We also plan to evaluate an electrodeposited alloy of 50% nickel-50% cobalt, which is known to exhibit superplastic behavior.

References

1. W.H. Gourdin, "Constitutive Properties of Copper and Tantalum at High Rates of Tensile Strain: Expanding Ring Results," *Inst. Phys. Conf. Ser. 102: Session 5*, 221–226 (1989).
2. W.H. Gourdin, "Metallurgical Effects on the Constitutive and Fragmentation Behavior of OFHC Copper Rings," *Proceedings of the American Physical Society Topical Conference*, Monterey, CA, July 20–23, 1987, p. 351.

3. W.P. Walters and A. Zukas, *Fundamentals of Shaped Charges* (John Wiley & Sons, 1989).
4. M.L. Duffy and S.K. Golaski, "Effect of Liner Grain Size on Shaped Charge Jet Performance and Characteristics," BRL-TR-2800 (April 7, 1987).
5. W.H. Gourdin, "Analysis and Assessment of Electromagnetic Ring Expansion as a High Strain Rate Test," *J. Appl. Phys.* **65**, 411 (1989).
6. W.H. Gourdin, S.L. Weinland, and R.M. Goling, "Development of the Electromagnetically Launched Expanding Ring as a High Strain Rate Test Technique," *Rev. Sci. Instrum.* **60**, 427 (1989).
7. J.C. Farmer, H.R. Johnson, H.A. Johnsen, J.W. Dini, D. Hopkins, and C.P. Steffani, "Electroforming Process Development for the Two-Beam Accelerator," *Plating & Surface Finishing* **75**, 48 (March 1988).
8. M. Myers and E.A. Blythe, "Effects of Oxygen, Sulphur, and Porosity on Mechanical Properties of Cast High-Purity Copper at 950°C," *Metals Technology* **8**, 165 (May 1981).

ENGINEERING RESEARCH AND DEVELOPMENT

Thrust Area Report FY 89

Materials Engineering and Science

The objective of the Materials Engineering and Science thrust area is to enhance our understanding of the physical and mechanical behavior of structural materials in areas of interest to Laboratory programs. Thrust area activities are currently focused on deformation and failure modeling and advanced engineering materials.

In the area of deformation and failure modeling we are expanding our knowledge of material deformation and failure behavior. The common objective is to provide Engineering and Laboratory programs with an enhanced ability to do finite element modeling and engineering analysis. Many Laboratory programs are making extensive use of modeling (much of it using finite element codes) to perform design, feasibility, and general engineering analysis studies. Programmatic requirements in this area inspired two research and development projects during FY 89:

- Numerical modeling of crack growth at interfaces
- Modeling and experimental measurement of residual stress.

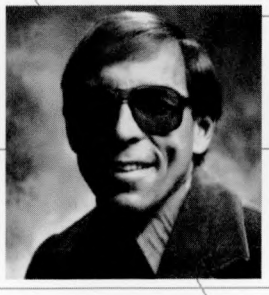
Thrust area research in advanced engineering materials is directed toward fiber reinforced composites and superplastic materials. Fiber reinforced composites have received considerable research attention and programmatic application at LLNL because of their unique properties, including high specific strength, high specific stiffness, composition of low Z atoms, corrosion resistance, and the possibility for a low coefficient of thermal expansion. These properties can be tailored to specific applications. Programmatic requirements have inspired three research and development projects during FY 89 and FY 90:

- Failure characterization of composite materials
- Studies of residual stress in composite materials
- Creep modeling for fiber reinforced epoxy resin composites.

Superplastic materials are crystalline solids that can be deformed in tension to such an extent that large strains will be attained at very low flow stresses. These materials, which deform like hot glass, permit components to be formed into shapes whose dimensions are very close to those desired in the final product (net shape processing). Thus, machining and machining-related operations can be reduced or eliminated. Our work in this technology has been stimulated by U.S. industry, which has demonstrated a strong interest in superplasticity for net shape processing. Currently the Laboratory is engaged in a collaborative research and development project with three industrial partners to develop the technology for commercial production of superplastic, ultra-high-carbon steels. Superplastic forming can also reduce environmental, safety, and health problems in the DOE nuclear weapons complex through the reduction of toxic and radioactive scrap produced during the fabrication of components. Two projects are currently being conducted on superplastic materials:

- Modeling the influence of structural change on superplasticity
- Superplastic ceramics.

Donald R. Lesuer
Thrust Area Leader



Creep Model for Fiber Reinforced Epoxy-Resin Composites

William W. Feng
Engineering Sciences Division
Mechanical Engineering

We have developed a mathematical model for creep in a fiber-reinforced epoxy-resin laminated composite, based on both micromechanics and macromechanics theories. Creep for fiber reinforced epoxy-resin composites is solved by the macromechanics model, via the finite element method. We have developed a simple experimental finite element computer program that incorporates a recurrence formula to evaluate the hereditary integrals for solving two-dimensional anisotropic viscoelasticity problems. With the recurrence formula, the value of a hereditary integral at any given time step depends only upon the value at the previous time step. Thus, no data storage is required and computing time is reduced. The experimental computer program will be extended to the NIKE3D structural analysis program, and the results will be checked against data to be obtained from sample composite laminates applicable to nuclear directed energy weapons development.

Introduction

Creep is the time-dependent deformation that distinguishes viscoelastic solids from purely Hookean elastic solids. A Hookean solid will elastically deflect instantaneously, and the stress generated by strain will not exhibit any relaxation whatever. A purely viscous fluid with a strain imposed upon it will generate a stress only if the amount of strain is changing, and the stress generated by strain will relax instantaneously. A viscoelastic material exhibits behavior characteristics of both purely viscous fluids and purely elastic solids, and the stress produced by strain relaxes at a finite rate.[1] Predicting creep in a composite structure requires the modeling of the structure under a variety of environmental conditions and applied loads. This calls for both a mathematical model of the creep phenomenon and a structural analysis program that includes the creep model. In this project, we have developed both.

The viscoelastic behavior of epoxy resin is the cause of creep effect in composites. Creep may result from externally applied loads, internal postcure residual stress, environmental conditions, or a combination of these factors. The residual stresses in a composite that result from curing can cause creep with no external load. Unpublished data from the aerospace industry indicate that creep can occur at a number of levels in dry samples of carbon-fiber-reinforced plastic with no external loading. At the uniaxial ply level, the stresses between the fibers and the resin are a source of creep. At the laminate level, the stresses between the independently oriented orthotropic plies are a source of creep at the resin-rich ply interfaces. An epoxy adhesive joint will also creep.

Composite materials are complex in nature, and their properties are anisotropic. The creep rate is usually related to environmental conditions. At the glass transition temperature a resin goes from a glassy state to a rubbery state and creep becomes rapid. An increased moisture content lowers the glass transition temperature of a resin.[2] Exposure of composite structures to space environments may also change the creep rate. Radiation interacts with the epoxy matrix to embrittle the composite. Atomic oxygen flux in a low earth orbit will erode a composite and change its properties. A simple creep model that is applicable to a variety of laminates, and a finite element structural analysis program, are logical first steps in performing a parametric analysis of these and other factors. In this project we are developing a finite element program that will consider these complex physical problems in composites.

Long term test data on creep in composite structures is scarce and the proposed directed energy weapons applications at LLNL are pushing the state of the art in dimensionally precise composite design. To address this need for an increased data base and improved theoretical modeling of the creep phenomenon, we have scheduled a number of experimental tests during the upcoming year. The experimental results will be checked with the numerical results obtained from the finite element code calculations.

Progress

Micromechanics Model

We are using a micromechanics model developed earlier for studying the residual stress of epoxy-resin composites due to curing. The viscoelastic properties of a unidirectional composite lamina have been theoretically obtained. The micromechanics of composite materials can describe the relationship between the properties of the constituents and those of the composite. We thereby obtain a relation describing the creep compliances of an orthotropic composite as a function of the properties of the constituent matrix and fibre. In this model the fiber is assumed to be elastic, the matrix is assumed to be viscoelastic, and the fiber volume content is assumed to be constant.

Based upon the micromechanics mathematical model, we have derived the unidirectional material properties for composite laminas. The constitutive equations are written in the incremental form and a recurrence formula has been obtained. With the recurrence formula, the value of a hereditary integral at any given time step depends only on the value at the previous time step. Thus, no data storage is required and computing time is reduced. In addition to these advantages in computation, the recurrence formula reduces the effort required to implement viscoelasticity in the existing finite element codes, as discussed below.

We now use only the linear viscoelastic constitutive relationship to calculate the matrix properties. Next year the micromechanics model will be extended to include nonlinear material effects, using Schapery's constitutive equation for nonlinear viscoelastic material.

Finite Element Program

A simple but functional finite element program has been written, implementing both linear and quadratic quadrilateral elements for the orthotropic plane-stress and plane-strain viscoelastic problems.

The constitutive equation for linear viscoelastic materials can be described by the Boltzmann superposition integral equation, which has the form

$$\sigma = \int_0^t G(t-\tau)\frac{\partial \varepsilon(\tau)}{\partial \tau}\, d\tau \quad . \tag{1}$$

This constitutive **Eq. (1)** for viscoelastic materials is a special case of the Volterra integral equation of the first kind, or

$$f(x) = \phi(x) - \lambda \int_0^b k(x,y)\,\phi(y)\, dy \quad . \tag{2}$$

A recurrence formula for the Volterra integral equation has been obtained by Feng.[3] Applying the same formulation to the Boltzmann superposition integral, the recurrence formula for the linear viscoelastic constitutive equation becomes

$$\sigma_{n+1} = \sigma_n + \Delta t\,\sigma_n^1 + \cdots + \frac{(\Delta t)^{m-1}}{(m-1)!}\,\sigma_n^{m-1} + \cdots + G\left(\frac{\Delta t}{2}\right)(\varepsilon_{n+1} - \varepsilon_n) \quad . \tag{3}$$

The σ_n^r in the above equation denotes the Boltzmann superposition integral **[Eq. (1)]** with the r^{th} order differentiation of $G(t)$ with respect to t. The value of Dt is a small increment of time from the time step n to the time step $n+1$.

When the relaxation function is

$$G(t) + \sum_{i=1}^{n} C_i\, e^{-\gamma_i t} \tag{4}$$

the recurrence formula is further reduced to

$$\sigma_{n+1} = \sum_{i=1}^{n} \varepsilon^{-\gamma_i \Delta t}\sigma_n + \sum_{i=1}^{n} C_i e^{-\gamma_i \Delta t/2}(\varepsilon_{n+1} - \varepsilon_n) \quad . \tag{5}$$

In order to solve viscoelastic problems by the finite element method, we need a set of field equations. The equilibrium equations in terms of stress are:

$$\frac{\partial \sigma_x}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} = 0 \quad ,$$

$$\frac{\partial \tau_{xy}}{\partial x} + \frac{\partial \sigma_y}{\partial y} = 0 \quad . \tag{6}$$

The incremental strain-displacement relations are

$$\widehat{\varepsilon_{xy}} = \frac{\partial \hat{u}}{\partial x},\ \widehat{\varepsilon_y} = \frac{\partial \hat{v}}{\partial y},\ \widehat{\tau_{xy}} = \frac{\partial \hat{u}}{\partial y} + \frac{\partial \hat{v}}{\partial x} \tag{7}$$

where $\hat{\varepsilon}_x$ is the incremental strain of ε_x from t_n to t_{n+1}, and $\hat{u}$ and $\hat{v}$ are the incremental displacements from t_n to t_{n+1}.

The equilibrium equations shown as **Eq. (6)** can be expressed in terms of the incremental displacements $\hat{u}$ and $\hat{v}$. Following the standard variational formulation for finite element problems, the governing equations reduce to

$$\left[K^{11}\right]\left\{\hat{u}\right\} + \left[K^{12}\right]\left\{\hat{v}\right\} = \left\{F^1\right\} \quad ,$$

$$\left[K^{21}\right]\left\{\hat{u}\right\} + \left[K^{22}\right]\left\{\hat{v}\right\} = \left\{F^2\right\} \quad , \tag{8}$$

where K_{ij}^{11}, K_{ij}^{12}, K_{ij}^{22}, F_i^1, and F_i^2

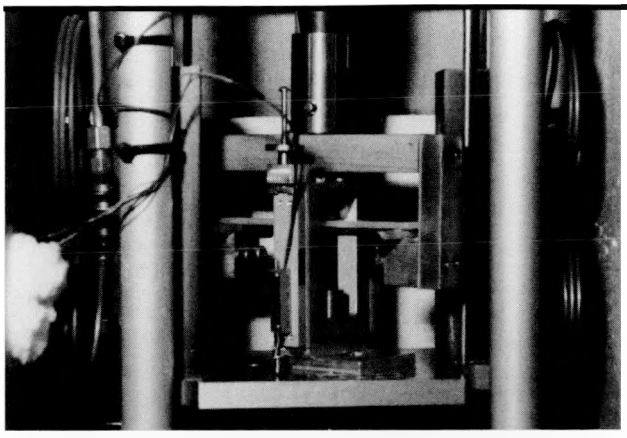


Figure 1. The test apparatus for three-point-bend creep test at temperature.

are functions of the viscoelastic material properties and the interpolation function used in the finite element formulation.

The code is working now on an IBM PC and can readily be transferred to a Cray machine where speed and storage space (i.e., the number of degrees of freedom and number of elements) can be increased drastically.

The purpose for developing the finite element program is to study the numerical procedures that will be used later for incorporating the viscoelastic constitutive equations into the general purpose codes, such as NIKE2D, NIKE3D, DYNA2D, and DYNA3D. As one can see from this paragraph, the incorporation of the viscoelastic constitutive equation into the finite element codes is not as difficult as many finite element code developers anticipated. This is primarily a result of the introduction of recurrence formula.

Currently, a new recurrence formula is being incorporated into the finite element code. The new formula does not restrict the form of creep function. The creep function may be of any differentiable functions.

Experimental Tests

The test fixtures for creep tests have been designed and fabricated. There are two kind of tests: the uniaxial tests and three-point-bend tests. **Figure 1** shows the fixture for three-point-bend creep tests and **Fig. 2** shows the fixture for uniaxial creep tests. The tensile creep testing and the three-point-bend creep testing facilities consist of six stations each. Tests can be run in all stations concurrently under microcomputer control. The constant load is generated by weights acting on the end of the lever for tensile tests and by weights acting at the middle of the span for three-point-bend tests. The strain is measured using Bourns Pots extensometers for the three-point-bend test and by strain gauges for the tensile test. Signals are then fed to a signal conditioning unit, which sends the transformed digital data through an interface to a microcomputer. The experimental apparatus has been designed and preliminary tests have been performed. The creep tests for an epoxy and a composite material will be performed next year.

Future Work

Next year, we will perform the experimental tests and continue with the theoretical development. Each resin system will be tested at six different temperature levels. The time-temperature correspondence principle will be used to evaluate the creep compliance for the epoxy resin. The loading duration will be 100 hr for both three-point-bend and uniaxial tensile creep tests.

For composites, the laminate [452]s will be used to evaluate the interlaminar shear strain, and the [904]s will be used to evaluate the transverse strain response. The test similar to the pure epoxy resin will be repeated and the viscoelastic constitutive relation for a unidirectional composite will be obtained.

Tests at different stress levels will also be carried out. These test will be used to evaluate the validity of the linearity assumption. Should the linearity assumption be impossible to satisfy, five different stress levels ranging from 10 to 80% of the short-term ultimate tensile strength will be tested. The Scherpy's nonlinear viscoelastic constitutive equation will be used.

The experimental finite element code will continue to be used to develop the numerical procedures for studying the residual stress due to curing

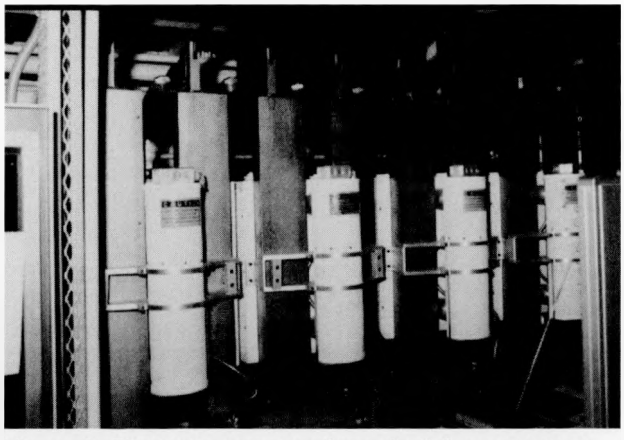


Figure 2. The test apparatus for uniaxial creep test at temperature.

and for studying the nonlinear material effect.

The theoretical formulations and the numerical logics developed with the experimental finite element code will be incorporated into NIKE3D next year by R. Logan.

1. E. W. Billington and A. Tate, *The Physics of Deformation and Flow*, (McGraw-Hill, 1981).
2. S. W. Tsai, *Composites Design*, Think Composite, Dayton, OH, 1987.
3. W. W. Feng, *A Recurrence Formula for Viscoelastic Constitutive Equation*, Lawrence Livermore National Laboratory, Livermore, CA, UCRL–94175 (1986).

Experimental Solutions to Residual Stresses in Carbon/Epoxy Composites

Jay Brentjes
Engineering Sciences Division
Mechanical Engineering

When a fiber-reinforced composite laminate is cured at elevated temperatures, the differences in fiber and matrix properties will result in highly significant residual stresses. These stresses, if not properly understood and minimized, can cause severe distortions in the final part. In addition, these residual stresses and distortions will change with time due to relaxation and the viscoelastic behavior of the matrix. This short project has provided a significant degree of insight concerning the manufacturing process parameters which affect the amount of residual stress. The most influential variable for any one resin/fiber combination is the cool-down rate of the cure cycle. Other objectives of this project included the study of fiber orientation laminate thickness, post cure and thermal cycling, as well as the effects of residual stress on the stability of certain part geometries, e.g., right angle sections.

Introduction

Despite the fact that fiber-reinforced composites are more difficult to fabricate into simple structures than conventional materials, the last two decades have seen a tremendous surge in the use of this materials approach. The advantage of being able to tailor high strengths and stiffnesses at substantial weight savings is paramount in all space and aircraft applications, and has outweighed some of the unpredictable or poorly understood behaviors of these materials during and after their manufacture.

Residual stresses between fibers and matrix materials can cause distortions, warps, and twists in otherwise flat or straight parts. Analytical studies of these residual stresses have been undertaken in the past. Work by Bill Feng, for example, has shown that these thermally induced stresses can be reduced with time due to viscoelastic relaxation of the matrix material.

This project was initiated to study composite residual stresses of composite laminates experimentally and to determine ways of reducing their magnitude and relaxation time. In particular, the induced stresses of carbon/epoxy laminates were selected because of the wide usage of this material combination. However, the research results should apply equally to other fiber/matrix types. For the same reason it was decided to concentrate on commercially available unidirectional preimpregnated tapes, commonly used at LLNL for a variety of weapon and laser programs. This material is fabricated into components by stacking a predetermined number of layers of this tape on or around a tool surface and curing the resulting laminate under heat and pressure. In order to best utilize the strength and stiffness features of the fibers, the various layers are rotated with the fibers oriented in different directions. And this is where the problem of residual stresses originates. The coefficient of thermal expansion of carbon fibers is nearly zero, whereas that of epoxy resin is quite large (approximately 40×10^{-6} in./in. °F). Hence, when the epoxy in the layers is cured at 350°F, the laminate at any other temperature tends to curl, twist, or distort. The strong, stiff fibers, which change very little in length with temperature, dominate in the 0° or fiber direction. However, at 90° (the cross-direction) the resin preponderates and causes that layer to shrink. The result is that a laminate made with fibers oriented at 0° and 90° comes out curled as shown in **Fig. 1.**

Determining the amount of curl and attempting to reduce it by manipulating geometry, material properties, cure parameters and relaxation time were the objects of this study.

Procedure

The first part of this empirical study was directed toward measuring and obtaining a better understanding of a variety of parameters affecting laminate distortion and relaxation behavior. A simple four-ply laminate as shown in **Fig. 1** was selected. A series of specimens 2 in. wide × 20 in.

long was prepared, cured, and clamped in a fixture as shown in **Fig. 2.** Measurements of heights h_1 and h_2 were then made as a function of time. The specimens were all made using the same prepreg, namely Hexcel T300 tape 145 g/m^2 areal weight with F263 epoxy resin. The parameters and variables investigated included:

- Cure cycle—effect of heat-up rate and effect of cool-down rate
- Post cure at elevated temperature
- Heat cycling—hot and cold
- Relaxation time and rate for all of the above
- Relaxation rate of a restrained vs unrestrained laminate.

In addition, four unbalanced 12-in.-square laminates were made with 2, 6, 8, and 16 plies to investigate the effects of panel thickness.

The second series of experiments was started to determine the behavior of right-angle parts. Previous experience has shown that laminated components with right angles cured at elevated temperatures would come out with some spring back, or with the right angles changed to approximately 87° to 88°. This will occur with balanced or symmetric laminates such as $0/90_2/0$. (Note that this designation indicates a laminate with one ply in the 0° direction, two 90° plies, plus one more 0° ply on the outside.) **Figure 3** shows how this behavior is attributed to the change in thickness t, again caused by differences in the coefficient of thermal expansion (CTE). In effect, it can be shown that this change in thickness causes an increase, $\Delta\theta$, entirely because of large changes in resin thickness and very small changes in the arc length of the inner and outer plies, i.e., small changes in fiber length.

Aluminum tooling was made to fabricate 90° pieces 2 in. wide with 8-in. leg lengths. Brackets were also made to mount these parts on a flat bench and to measure changes in angle with time. A few parts were made which indeed demonstrated the angular deviation. However, the tooling corners were too sharp and actually cut the fibers in the outer plies. This caused erroneous results. Before changes in the tooling could be made, the project was stopped because of higher priority commitments.

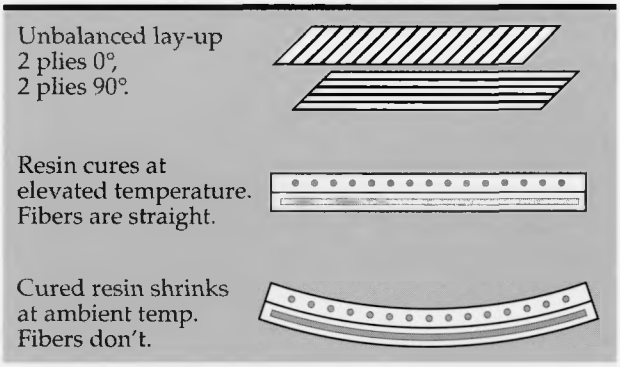


Figure 1. Residual stress mechanism in unbalanced composite lay-up. As composite parts cool after resin is cured, residual stress causes distortion, or curl.

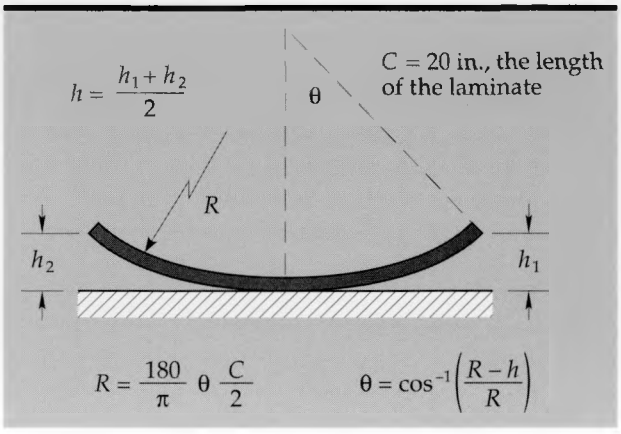


Figure 2. Method for measurements of curl in unsymmetric laminates. These relationships can be solved iteratively for R using experimentally-derived values for h.

Results

The study of cure cycle variations using thin, narrow strips provided the most interesting insight into the residual stress problem. Samples which were cured in the autoclave using a very slow cool-down rate clearly had less curl or residual strain than those which were cooled as rapidly as pos-

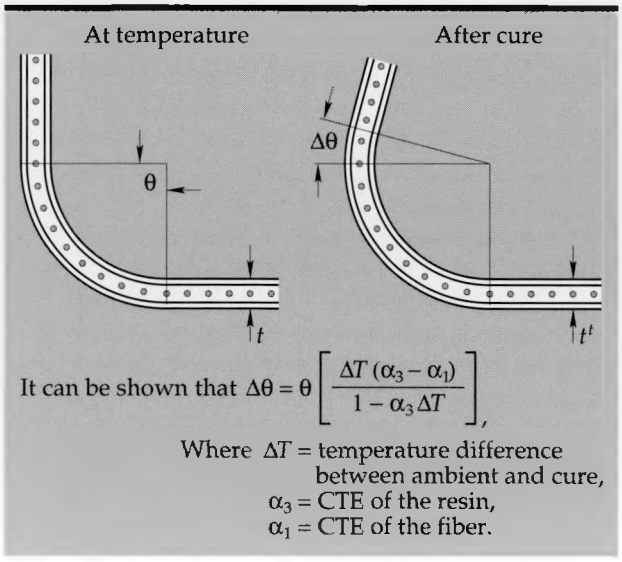


Figure 3. Variations in composite geometry due to differences in thermal properties. The equation shown here is derived from geometric considerations and dimensional changes due to thermal expansion.

sible. In addition, it was found that the relaxation of all these samples followed a pattern similar to that shown in **Figs. 4 and 5.** The values for height h, which is the average height of the two ends above the base, were measured along with the ambient temperature. The plotted values were not adjusted for temperature, but it should be noted that values above the lines were consistently observed when the ambient temperature was 2°–3°C below normal.

This temperature effect was clearly demonstrated with samples exposed to elevated temperatures. At about 320°F the unrestrained samples were essentially flat. This indicates that the samples were actually gelled, set and cured at that temperature and that this is the zero stress condition. Height measurements of several samples at varying temperatures were started, but again this work was not completed.

Figure 6 shows the effect of cool-down rate on the initial curl or leg height. Two lines are shown because the cool-down rate was not completely

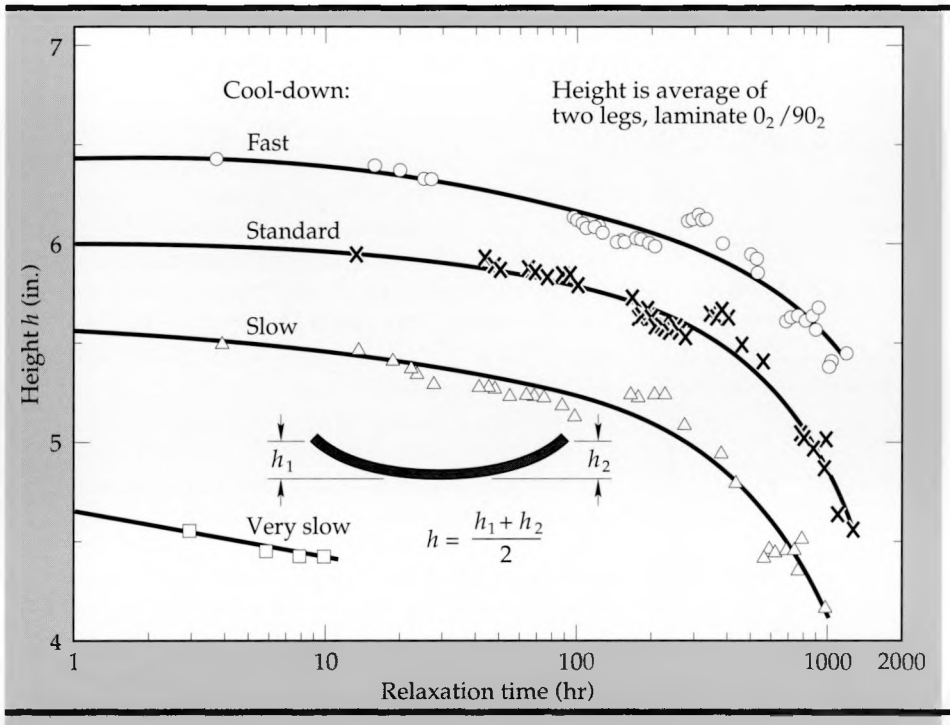


Figure 4. Relaxation of residual stress for several nonsymmetric laminates. Residual strain appears to be inversely proportional to the cool-down rate, so a slow cool-down can reduce residual stress and thus reduce the resulting distortion of composite parts.

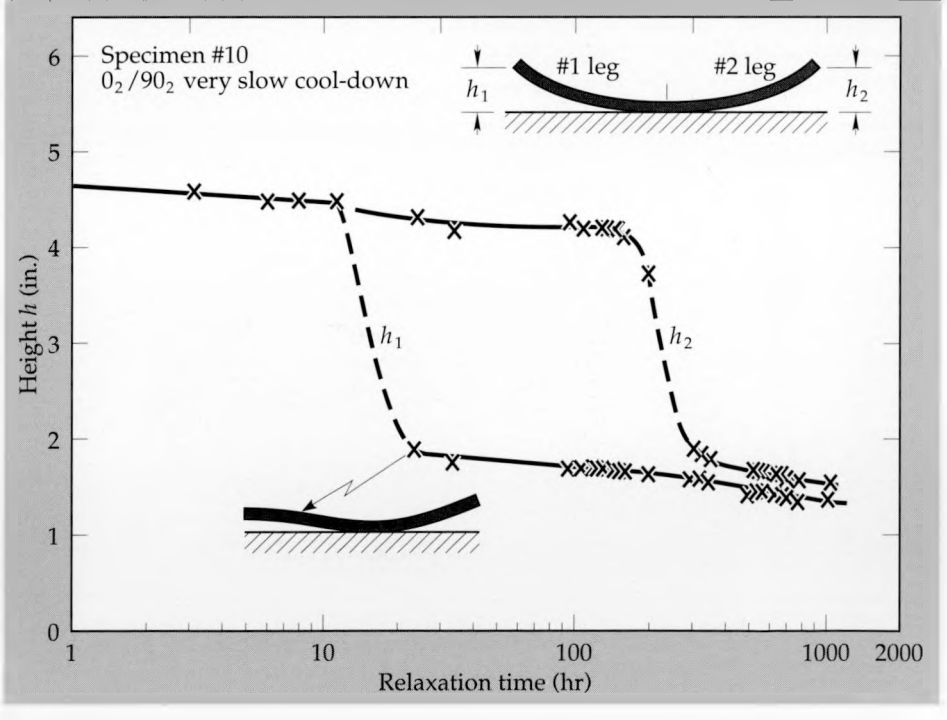


Figure 5. Relaxation behavior of a laminate fabricated with a very slow cool-down rate. Note the anticlastic or reverse curl in legs h_1, and then h_2 after 10 and 200 hr of stationary relaxation.

linear. The top line was the cool-down rate from 365 to 300°F; the lower line is for 300 to 120°F. In both cases it is clear that slowing the cool cycle makes a significant impact. This work could have been extended to extremely long cool cycles; however, this may become rather impractical in production. The slowest cure cycle took a total of 13 hr **(Fig. 7).** As a result of this part of the study, new cure cycles were initiated at LLNL with longer cool-down times, which were increased from 70 min. to 3.5 hr.

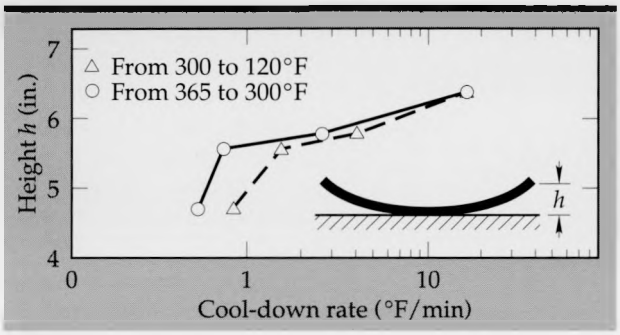


Figure 6. Effect of cool-down rate on average leg height.

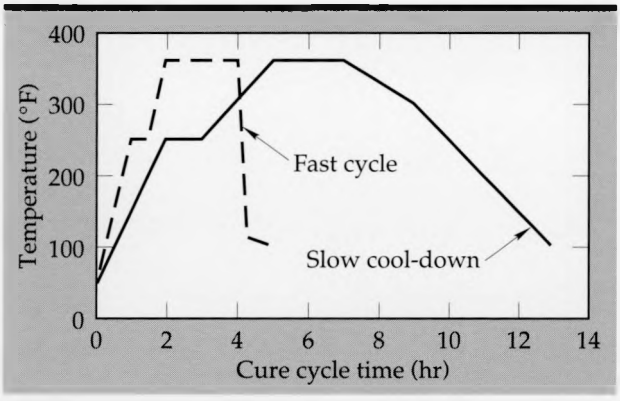


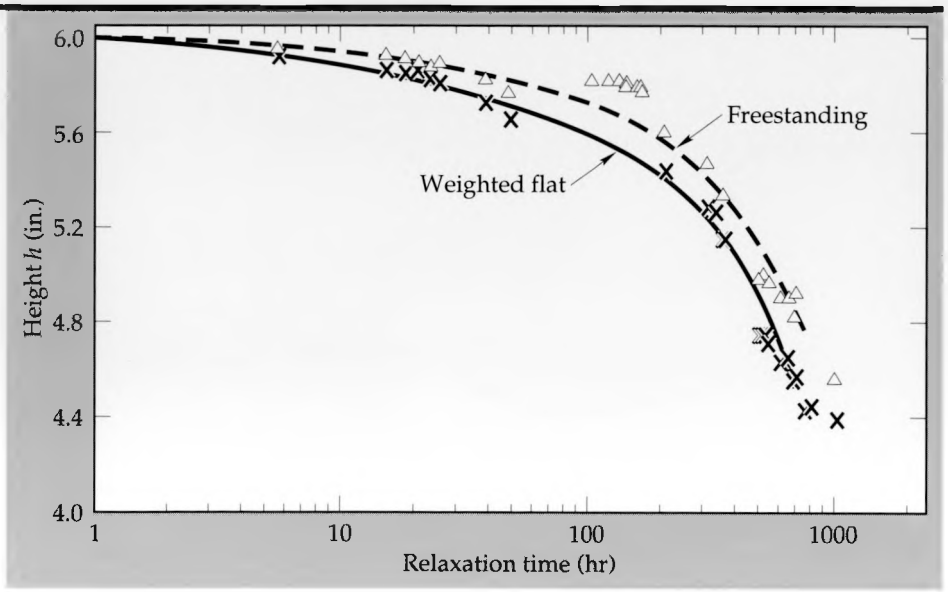
Figure 7. Typical temperature time profiles for LLNL autoclave epoxy curing process.

In order to measure the effect of induced stress on relaxation rate, two identical specimens were compared under different conditions. One was left free standing, while the second was weighted flat with lead weights. Measurements were made periodically and the weights were immediately replaced. **Figure 8** shows the difference with time. The significance of this is that residual stress in most fabricated parts will relax somewhat faster because of the higher level of initial stress that exists. In other words, the higher the induced stress, the faster the viscoelastic effects tend to relax this stress. Further detailed experiments are needed in this area, and the results of this study should be compared with the previous analytical work by Bill Feng.

The four panels made at different thicknesses again showed that the residual of induced stress affects the amount of curl in a nonsymmetric layup. The thicker the panel (0/90 vs $0_8/90_8$), the less curl. That is not to say that thicker laminates have less stress. To the contrary, less curl could mean that the stresses are not as much relieved. In fact, the thicker panels clearly showed resin cracks at the surface and edges as a result of the severe stresses between the fibers and resin.

Another interesting and as yet unexplained phenomenon was the reversal of relaxation when samples were once again exposed to elevated temperatures. When a sample which had relaxed for several hundred hours was cycled from 350°F back to ambient, the amount of curl increased and all the relaxation disappeared. It was as if the lost stresses were reintroduced because of the elevated temperature exposure. This is contrary to what other

Figure 8. Effect of induced stress on composite relaxation behavior. Relaxation is faster when induced stress levels in the composite are higher.



researchers have claimed to be a means of reducing residual stress, namely thermal cycling at elevated temperatures. The reheating and post curing was repeated several times with the same result. Parts went back to their original condition immediately after the original cure cycle. Two cured and relaxed strips were even reheated with lead weights holding them flat and a programmed cool rate lasting sixteen hours. The amount of curl of the cooled samples was more than when they were reheated.

Despite some of the inexplicable behavior we observed, the work performed under this LLNL-funded development project certainly has provided better insight into the effects of certain cure parameters on residual stress. Likewise, the relaxation phenomenon is an important consideration at first glance, and has correlated well with analytical studies. However, several additional questions need to be more fully addressed. Parts of this project which could not be started or completed include:

- Effect of cure temperature and use of different resins
- Variations in fiber orientations and stacking sequence
- Changes in material CTE and the use of fibers such as Kevlar and fiberglass
- Correlation of the experimental data with the analytical work of Bill Feng.

Acknowledgements

The author extends his gratitude for valuable suggestions provided by Bill Feng; thanks also to Steve Winchester, Derrick Mathews and Paul Milligan for fabricating specimens and making numerous measurements.

Failure Modeling in Finite Element Analysis

Daniel J. Nikkel, Jr.
Engineering Sciences Division
Mechanical Engineering

Robert A. Riddle and Ronald D. Streit
Weapons Engineering Division
Mechanical Engineering

The ability to model failure in structural materials is important in the design and analysis of structural components. We have developed analytical tools to predict structural component failure caused by crack growth and plastic deformation. As a part of the ongoing effort to improve J-integral numerical analysis of fracture problems using the Lab's finite element codes, we have implemented an elastic-plastic material model into NIKE2D which accounts for the porosity changes observed in materials near the site of a ductile failure. We have also investigated the possibility of incorporating new techniques into the codes for adaptively modifying the finite element mesh, thereby improving our ability to model problems in which the dynamic evolution of cracks is taken into account.

Introduction

Numerical modeling using finite element analysis has become an indispensable tool for the development and evaluation of engineering designs and the solution of general mechanics problems in a wide variety of Laboratory programs. Through the use of the Lab's general purpose codes such as NIKE[1,2] and DYNA[3,4] an analyst can predict stresses and strains in a structure composed of a number of different parts, with the response of each part being modeled with a variety of continuum material models. A knowledge of the stress and deformation in a structure is certainly important in the process of developing a design or understanding a mechanics problem. However, another equally critical aspect of engineering design is the ability to predict failure in structural components.

The notion of failure is rather broad. It ranges from a reduction in stiffness of a material and its inability to support a required load due to plastic flow or other accumulated damage to the actual breaking of a component due to the growth of defect cracks in the material. While theoretical plasticity is a relatively mature subject and the current analysis codes provide material models which account for plastic flow, our ability to understand and model other mechanisms of failure such as accumulated damage and crack propagation is less well developed, and it is with these failure mechanisms that we are concerned here. Specifically, we will discuss failure due to cracking from ductile fracture.

One difficulty with the modeling of failure is that it is often governed by competing mechanisms. Trying to model them all with the appropriate interaction coupling between them would be a formidable task. In practice, the modeling of failure can be approached from two different perspectives. In one approach, a macroscopic viewpoint is taken, and a critical damage criterion is developed in terms of continuum variables such as stress and strain, and their history.[5] Substantial experimental testing is required to validate criteria proposed in this way. Sometimes additional material parameters (or internal state variables) are introduced into the formulation which affect the damage criteria for which constitutive relations and evolution equations are proposed. However, one must be cautious not to introduce such parameters in an *ad hoc* manner and thus to render the theoretical formulation merely a fancy method of curve fitting. An alternative approach is to attack the problem of failure from a more microscopic point of view, as in the case of traditional fracture mechanics. In this approach the material is modeled with appropriate flaws, and the stability of the pre-existing flaws is analyzed by solving the corresponding local boundary value problem which is on the scale of the flaw size. It is this approach, within a numerical context, which we will discuss here.

This thrust area has had a continuing interest in developing a better understanding of failure processes, and has been endeavoring to develop material test methods and analytical tools which can be used in conjunction with Laboratory codes to

model material failure. Numerical analysis of materials with geometric discontinuities such as cracks is difficult when using finite elements. The discontinuities can induce numerical oscillations in the stress, stress singularities, and dependence of the solution on the mesh size. In order to get around the difficulties found in the regions near defects or crack tips, a path-independent contour integral, known as the J-integral, is commonly used in fracture mechanics to uniquely characterize the singularity. It can be calculated from the value of the field variables along a contour which lies in a region where the solution is better behaved. The value of the J-integral can be interpreted as the energy available for crack growth, and for elastic materials it can be related to stress intensity factors arising in linear elastic fracture mechanics. Crack propagation is assumed to occur when the J-integral reaches a critical value, which in the case of Mode I loading (opening the crack by deformation normal to the crack surface) is denoted by J_{Ic}, and in the case of Mode II loading (relative in-plane shearing of the crack surfaces) by J_{IIc}. The values of J_{Ic} and J_{IIc} are taken to be material constants to be determined from experimental testing.

The J-integral, which is based upon the work of Eshelby,[6,7] was introduced in the field of fracture mechanics for use with both elastic and elastic-plastic materials, but the formulation was limited in that it did not account for thermal and dynamic effects.[8,9] Subsequent formulations have sought to overcome these limitations. One such formulation for the value of the J-integral over an area A which is bounded by a curve Γ, which accounts for thermal effects is given by[10]

$$J = \int_{\Gamma} W^e \, d x_2 - \int_{\Gamma} \sigma_{ij} \, \nu_j \, \frac{\partial u_i}{\partial x_1} \, d\Gamma + \alpha \int \int_A \sigma_{kk} \, \frac{\partial \theta}{\partial x_1} \, dA \quad , \qquad (1)$$

where

$$W^e = \frac{\sigma_{ij}\left(\varepsilon^t_{ij} - \varepsilon^t_{ij}\right)}{2} \, ,$$

and

x_1, x_2 = coordinate directions,
u_i = displacements,
ν_i = surface (contour) normal,
α = cofficient of thermal expansion,
σ_{ij} = stress tensor,
ε_{ij} = strain tensors, $\varepsilon^t_{ij} = \alpha\theta\delta_{ij}$,
θ = temperature change from reference state.

One of the results of our ongoing work in the area of failure modeling has been the development of a J-integral post-processor which operates on the numerical output of the NIKE2D finite element code.[11] This post-processor has been continuously updated and improved and is now capable of dealing with fracture problems involving thermal stresses.[12] The use of numerical analysis using this J-integral post-processor has been extensively validated with material testing using a variety of materials such as depleted uranium (D38), HY130 steel, 7075-T6 aluminum, and soda-lime glass.

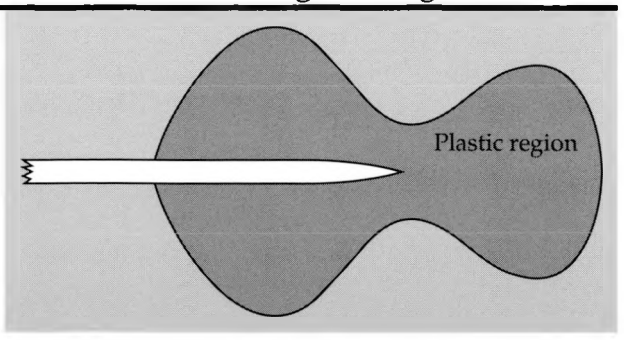


Figure 1. Schematic view of cracking in a ductile failure showing the plastic zone whose size is of the same order of magnitude as the flaw size.

In real materials, the stress singularities at crack tips correspond to regions of intense plastic deformation. In materials which are relatively brittle, the region of plastic deformation is small in comparison with the crack size and is typically ignored. In ductile materials, the size of the plastic deformation region can be of the same order of magnitude as the crack size (**Fig. 1**). This region of intense plastic deformation may be termed a "damage process zone," a concept which may be used to relate microstructural critical damage events and local constitutive properties to the macroscopic fracture toughness of a material.[13] Ductile fracture in metals can involve the generation of considerable porosity via the nucleation, growth, and coalescence of voids under the action of high triaxial tensile stresses near the crack tip.[14] There is a considerable amount of literature on this subject, which is discussed in a recent review article.[15] Traditional models of plasticity in metals hold that yield is independent of the mean-normal component of the stress (i.e., pressure or triaxial tension) and that the material is plastically incompressible (i.e., no plastic volume change). In order to accurately model the elastic-plastic behavior of ductile materials near a crack tip and to account for the observed porosity changes, it is necessary to adopt a more general set of constitutive functions for characterizing elastic-plastic materials.

Progress

Material Model for Failure Process Zone

In problems involving inhomogeneous materials, such as those containing inclusions or voids, it can be advantageous to consider an idealized model on the microscale which after appropriate volume averaging simulates the actual material behavior on the macroscale. This approach is referred to as micromechanics. An approximate plastic constitutive theory which accounts for void nucleation and growth has been developed by Gurson.[16] This model is based upon a micromechanical analysis of a unit cell containing a single void in an incompressible rigid—perfectly plastic matrix material. This unit cell is used as a simple analog from which approximate macroscopic constitutive relations (yield function and flow rule) are deduced by exploiting numerical approximations and bounding theorems derived from the micromechanical analysis. It should be noted that while the material in the micromodel is very idealized (i.e., incompressible rigid—perfectly plastic) the macroscopic constitutive equations derived from it describe a material which is not so constrained. Part of our effort this year has been to implement a new material model into NIKE2D, based on Gurson's model, in order to better represent the elastic-plastic behavior of material in the failure process zone. In the case of axisymmetric deformation the macroscopic yield function is of the form

$$f = \frac{\tau_{ij}\tau_{ij}}{\sigma_y^2} + 2\phi \cosh\left(\frac{\sqrt{3}}{2}\frac{\sigma_{\beta\beta}}{\sigma_y}\right) - 1 - \phi^2 \ , \qquad (2)$$

where

τ_{ij} = deviatoric stress,
σ_y = yield stress of matrix material,
ϕ = void volume fraction,

and β is an index taking the values 1, 2.

Plastic flow is governed by an associated normality flow rule. The change in porosity is related to the plastic volume change through

$$\dot{\phi} = (1 - \phi)\, \dot{e}^{p}_{kk} \ . \qquad (3)$$

It is worth noting that in the absence of voids; i.e. $\phi = 0$, this reduces to the usual von Mises plasticity model.

Assessment of Adaptive Mesh Techniques

We have investigated the possibility of developing other techniques which would allow us to more accurately analyze material failure. Fracture by its very nature is a dynamic process. When using finite elements to analyze a flaw, a finer mesh is used in the region near the crack tip due to the high gradients of field variables there (**Fig. 2**). However, as the crack grows the position of the crack tip changes while the position of the mesh remains fixed. Currently the codes allow us to statically analyze a flaw and use the tools indicated above to determine whether it is stable or whether it will grow. If it is unstable, one can manually rezone the mesh and separate nodes to allow for analysis of the crack in a new configuration. This is a tedious and inefficient procedure. The codes also provide a capability to model material separating along a "slideline" when a criterion based on accumulated plastic strain is satisfied. This is not wholly satisfactory either, since one must position the slideline (and hence determine the direction of crack propagation) *a priori* rather than having it be determined as the calculation proceeds.

Ideally one would like to have the crack stability analyzed as a calculation proceeds. Then, if it is determined that the flaw will grow, the mesh should automatically be modified, both to increase the flaw size and to refine the mesh in the new crack tip position. This requires so-called adaptive finite element procedures. Techniques of this kind have been incorporated into the Lab's geomechanics code FEFLAP,[17] and are the subject of considerable research.[18] Due to the complexity of the NIKE and DYNA codes it is unlikely that such features will be introduced in the near future. However, eventually NIKE and DYNA will have this capability. Near-term improvements in this area will have to come from a smoother coupling between the J-integral analysis (or other failure criteria) and the existing manual rezoning capability of the codes.

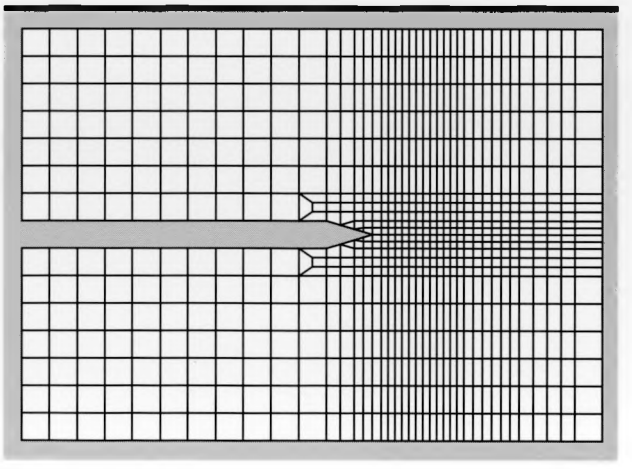


Figure 2. Finite element mesh of material in the region near the tip of a crack. The mesh is fine near the crack tip so that the high gradients in field variables can be more accurately represented.

Future Work

It remains to evaluate the effectiveness of using the new material model for analyzing cracks in ductile materials. Prior computational problems involving flaws in ductile materials will be run again to provide a basis for comparison. We intend to continue our efforts to improve our understanding of material failure, both through the development of new testing techniques and by improving our computational tools. On the computational side, the implementation of techniques to better analyze the dynamic aspect of the failure process will be a priority.

1. J. O. Hallquist, *NIKE2D—A Vectorized Implicit, Finite Deformation Finite Element Code for Analyzing the Static and Dynamic Response of 2D Solids with Interactive Rezoning and Graphics*, Lawrence Livermore National Laboratory, Livermore, CA, UCID-19677, Rev. 1 (December 1986).
2. J. O. Hallquist, *NIKE3D: An implicit, finite-deformation, finite element code for analyzing the static and dynamic response of three-dimensional solids*, Lawrence Livermore National Laboratory, Livermore, CA, UCID-18822, Rev. 1 (July 1984).
3. J. O. Hallquist, *DYNA2D—An Explicit Two-dimensional Hydrodynamic Finite Element Code with Interactive Rezoning and Graphical Display*, Lawrence Livermore National Laboratory, Livermore, CA, UCID-18756, Rev. 8 (March 1988).
4. J. O. Hallquist and R. G. Whirley, *DYNA3D User's Manual (Nonlinear Dynamic Analysis of Structures in Three Dimensions)*, Lawrence Livermore National Laboratory, Livermore, CA, UCID-19592, Rev. 5 (May 1989).
5. M. L. Wilkins, R. D. Streit, and J. E. Reaugh, *Cumulative-Strain-Damage Model of Ductile Fracture: Simulation and Prediction of Engineering Fracture Tests*, Lawrence Livermore National Laboratory, Livermore, CA, UCRL-53058 (1981).
6. J. D. Eshelby, "The Continuum Theory of Lattice Defects," in *Progress in Solid State Physics*, **3**, F. Seitz and D. Turnbull, eds. (Academic Press, New York, 1956) p. 79.
7. J. D. Eshelby, "Calculation of Energy Release Rate," in *Prospects of Fracture Mechanics*, Sih, Van Elst, and Broek, eds. (Noordhoff, 1974) pp. 69–84.
8. G. P. Cherepanov, "Crack Propagation in Continuous Media," *USSR Journal of Applied Math. and Mech.* Translation, **31,** 504 (1967).
9. J. R. Rice, "A Path Independent Integral and the Approximate Analysis of Strain Concentrations by Notches and Cracks," *Journal of Applied Mechanics*, **35,** 379–386 (1968).
10. S. Aoki, K. Kishimoto, and M. Sakata, "Elastic-Plastic Analysis of Cracks in Thermally Loaded Structures," *Engineering Fracture Mechanics*, **16**(3), 405–413 (1982).
11. R. A. Riddle, *The Application of the J Integral to Fracture under Mixed Mode Loading*, Lawrence Livermore National Laboratory, Livermore, CA, UCRL-53182 (June 1981).
12. R. A. Riddle and R. D. Streit, *The Effects of Thermal Stresses in the Numerical Modeling of Crack Growth*, Lawrence Livermore National Laboratory, Livermore, CA, UCRL-53868-88 (1988).
13. R. O. Ritchie and A. W. Thomson, "On Macroscopic and Microscopic Analyses for Crack Initiation and Crack Growth Toughness in Ductile Alloys," *Metallurgical Transactions A*, **16A,** 233–248 (1985).
14. H. C. Rogers, "The Tensile Fracture of Ductile Metals," *Trans. Metall. Soc. AIME*, **218,** 498 (1960).
15. D. R. Curran, L. Seaman, and D. A. Shockey, "Dynamic Failure of Solids," *Physics Reports*, **147,** 253–388 (March 1987).
16. A. L. Gurson, "Continuum Theory of Ductile Rupture by Void Nucleation and Growth: Part 1—Yield Criteria and Flow Rules for Porous Ductile Media," *J. Eng. Mater. and Tech.*, **99,** 2–15 (1977).
17. A. R. Ingraffea, R. J. Shaffer, and F. E Heuze, *FEFLAP: A Finite Element Program for Analysis of Fluid-Driven Fracture Propagation in Jointed Rock, Vol. 1: Theory and Programmer's Manual*, Lawrence Livermore National Laboratory, Livermore, CA, UCID-20368 (March 1985).
18. J. T. Oden and L. Demkowicz, "Advances in Adaptive Improvements, A Survey of Adaptive Finite Element Methods in Computational Mechanics," in *State-of-the-Art Surveys on Computational Mechanics*, A. K Noor and J. T. Oden, eds. (ASME, New York, 1989) Ch. 13.

Characterizing the Failure of Composite Structures

Scott E. Groves
Engineering Sciences Division
Mechanical Engineering

We have succeeded in developing a multiaxial test system that has generated increasingly accurate information on the three-dimensional behavior of continuous fiber composites. The resulting data has given us a more accurate picture of the strengths and failure modes of these complex materials. Using this data, composite materials engineers can more accurately predict whether a specific design will sustain a given load; alternatively, they can use the information to design a stronger, lighter composite. In this fiscal year our goal has been to develop a more accurate three-dimensional failure criterion based on the data generated by our test system. We are also investigating the dynamic behavior of composites at strain rates up to 1000 in./in./sec.

Introduction

A composite material is made of two or more materials combined to create a better material system. In a broad sense, metal alloys and plastics are composites, but their constituent materials are combined at the microscale. In a narrower sense, a composite material is one in which the constituent materials are combined at a macroscopic level; thus, the individual behavior of each material is retained and directly affects the composite's properties.

There are two main types of these materials - particulate composites and fiber composites. In particulate composites, small particles are uniformly distributed and embedded in a matrix system. In fiber composites, the reinforcing fibers — either continuous or chopped fibers — are embedded within a matrix. At LLNL, our major interest is in continuous-fiber composite materials, and these are the composites we discuss in this article.

Composite materials have been widely used for decades. However, the demands placed on composite materials by LLNL applications are very different. For example, composites are attractive to Laboratory weapons applications for their light weight and high strength. Such composites are currently being considered as candidates for outer weapon cases because they could provide structural strength and yet save weight. They are also being considered for certain projects for the Strategic Defense Initiative, where very stiff, very lightweight structual components are needed that undergo only minimal dimensional changes with changes in temperature.

Most of these applications are concerned with ultimate strength and one-shot life. However, even though composites are not new, for our purposes much information on their mechanical characteristics has yet to be obtained. Extensive data bases on material properties and accurate models of their behavior must be developed if structural designers are to avoid the costs and pitfalls of a cut-and-try approach and efficiently tailor composites to specific Laboratory applications.

Advantages and Disadvantages of Composites

Fiber-composite materials offer two major advantages over metals: high specific strength and high specific stiffness. In applications where weight is a limiting factor, fiber composites are very attractive because they provide good strength and stiffness with very little weight. At the Laboratory we have examined a number of composites that have greater tensile strength by weight than metals (**Fig. 1**). One of them, Toray 1000, is currently being considered for a number of Laboratory applications.

Fiber composites offer the possibility of selecting a reinforcing fiber and an appropriate matrix and then orienting the fibers within the laminate to create a composite material with the desired strength, stiffness, weight, wear resistance, and/or thermal expansion. Our goal is to understand the behavior and properties of the fibers, matrices, the composite material, and various laminates so that we can tailor composites to specific applications. Diagrams of unidirectional and quasi-isotropic composite laminates are shown in **Fig. 2(a)**, and a

photomicrograph of a typical fiber composite is shown in **Fig. 2(b)**.

A major disadvantage of these materials is that the improved properties are exhibited only in the plane of the fibers—that is, only in two dimensions. This generally restricts the use of composites to two-dimensional structures since these materials tend to be very weak in the third (out-of-plane) direction. The structural designer must assure himself that the out-of-plane loading conditions for the intended application are acceptable (i.e., within the strength limits of the desired material). Another problem with fiber composites is that it is difficult to fabricate a composite part without flaws (e.g., voids or microscopic cracks) that can lead to the premature failure of the composite part.

Fabricating Fiber Composites

Two methods— laminating and filament winding—are used to fabricate fiber-composite parts, depending on the nature of the part. To fabricate a laminated part, we use pre-impregnated tape of continuous fibers embedded in a thin film of resin that is layered onto a nonstick backing paper. The layers (plies) of pre-preg tape are stacked in the desired orientation and shaped over a mold or mandrel. The polymeric resin is usually dormant at room temperature and is cured at high temperature, usually in a laminating press or autoclave and usually under vacuum. **Figure 3** shows some of the steps involved in fabricating a nose cone using pre-preg tape.

In filament winding, a continuous fiber, coated in resin, is wound around a mandrel (**Fig. 4**). A numerically controlled winding machine is used to carefully control the winding pattern and the tension at which the fiber is laid on. After the desired thickness is obtained, the part is then cured.

It is possible to create very complex shapes using both fabrication methods (**Fig. 5**), with filament winding being more efficient for spherical or cylindrical parts (e.g., pressure vessels), and lamination more efficient for basically flat parts (e.g., turbine blades, wing tips, structural body parts).

Analyzing the Performance of Composite Materials

Composite materials are difficult to analyze and model because they are orthotropic. That is, they have three mutually perpendicular planes of symmetry (**Fig. 6**) and thus their material properties are markedly different in these different directions. Nine individual material constants are needed to model the elastic constitutive behavior of orthotropic materials, compared to only two constants needed for isotropic materials (e.g., most metals). The experimental effort required to characterize the material constants is very complex and repre-

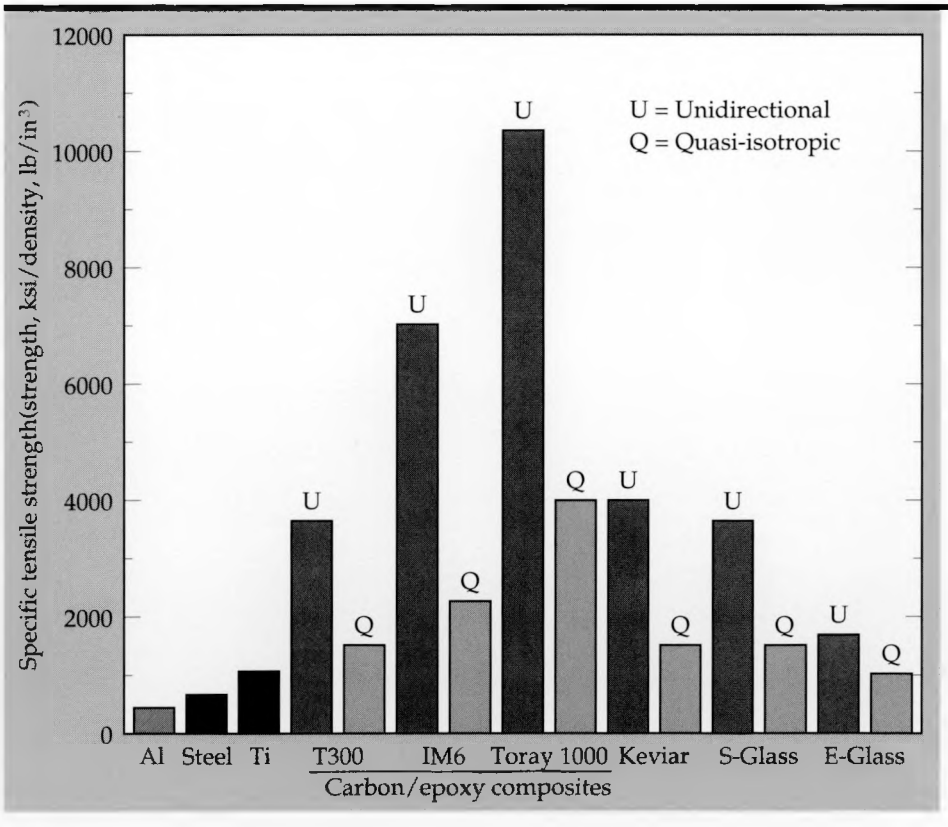


Figure 1. Comparison of specific strength of various metals and composites. The carbon/epoxy composites shown are in use at the Laboratory.

sents a significant effort of LLNL's composite research.

One of the most difficult analytical problems encountered with composite materials is that of predicting failure response. Composite materials, especially those with multiple plies at different orientations or multiple layers of filament-wound fibers, often contain myriad internal failure surfaces (voids, bubbles, fiber flaws, etc.). With so many possible sites at which a flaw or crack can begin to grow and lead to the failure of the composite part, it is very difficult to establish an accurate or appropriate failure load. Many failure theories have been developed, but they have generally proven to be inadequate. Work is underway at LLNL to improve our understanding of failure and to develop a more adequate set of failure criteria.

Strength of Composite Materials

We are currently evaluating the strength of composite materials and predicting their failure under

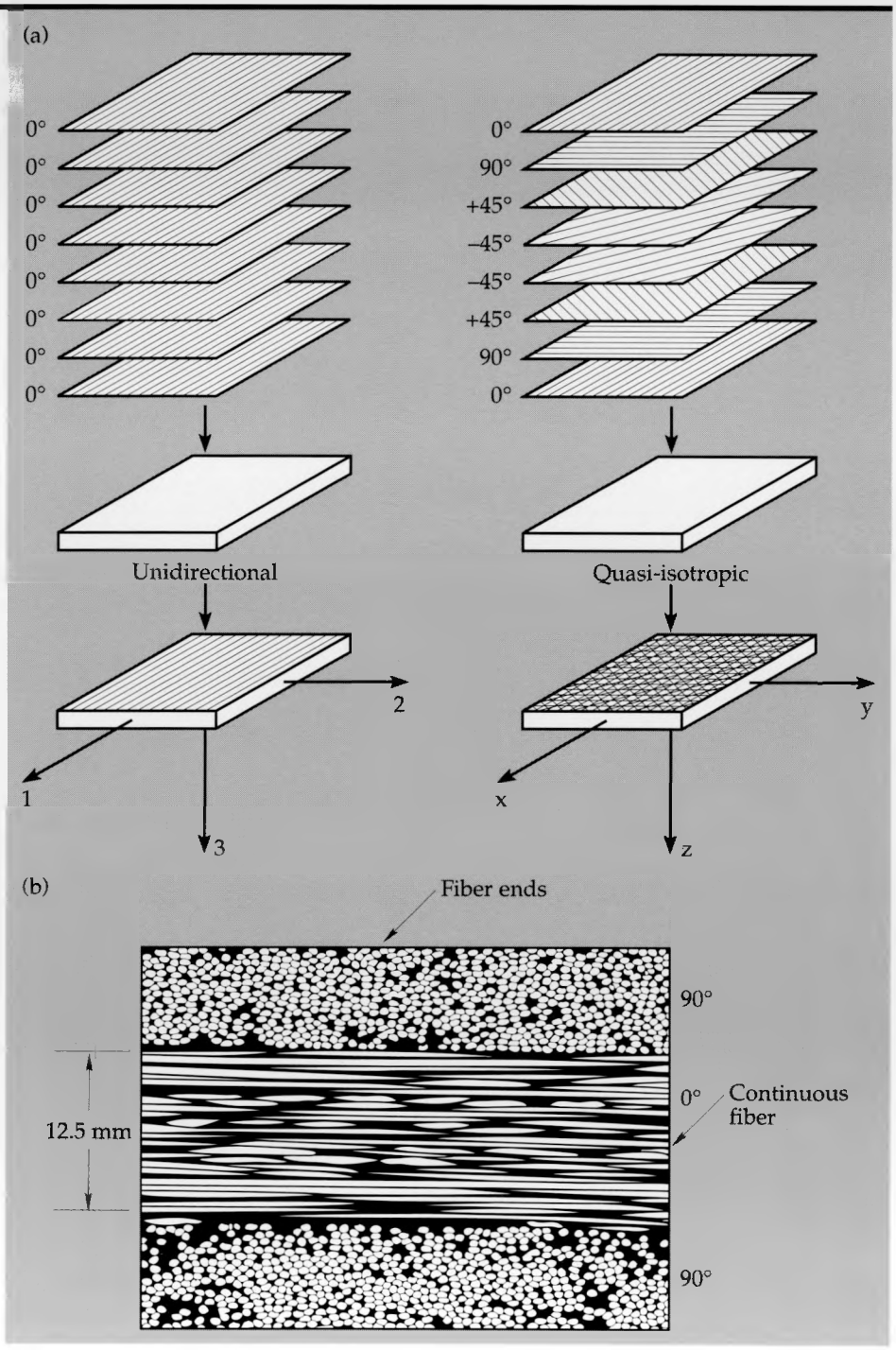


Figure 2. (a) Diagram of unidirectional and quasi-isotropic laminates. (b) Photomicrograph of a T300/F584 graphite/epoxy composite laminate.

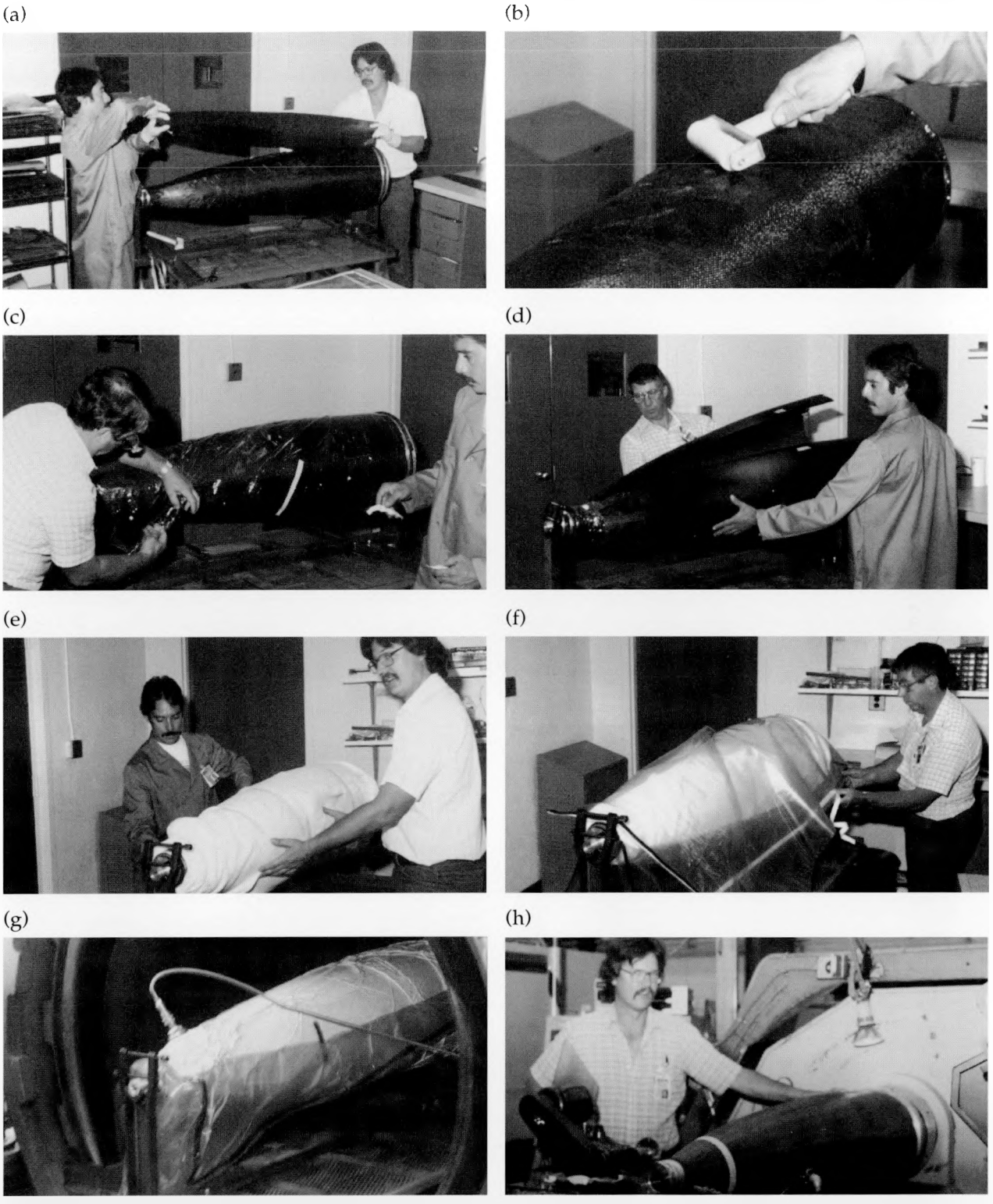


Figure 3. The steps involved in fabricating a nose cone using pre-preg tape. (a) Pre-preg tape is cut into four wedges, which are draped around the aluminum mandrel. A total of 10 plies (40 wedge sections) are built up; seams are staggered between plies. (b) Wrinkles are rolled out as must as possible. (c) Perforated release film is draped around the pre-preg lay-up. (d) A three-piece "clam shell" is fitted over the lay-up; pieces are held in place with nylon tape. (e) Breather cloth is placed over the clam shell. (f) The assembly is enclosed in a vacuum bag; edges are sealed with heat-resistant vacuum tape. (g) A vacuum fitting is hooked up inside the autoclave and the part is cured at 175°C and 565 kPa. (h) After the cure cycle, the part is trimmed to length on a slowly rotating lathe using a diamond-coated saw.

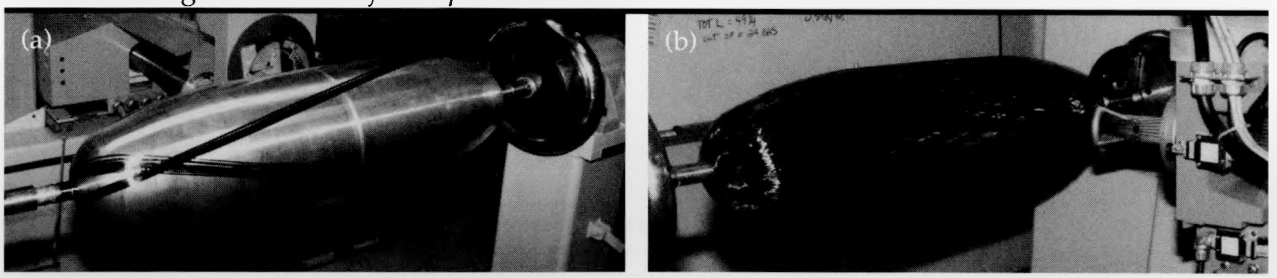


Figure 4. A spherical part being fabricated by filament winding. The continuous fiber, coated in resin, is wound around a mandrel; the fiber tension and winding pattern are controlled by a numerically controlled winding machine.

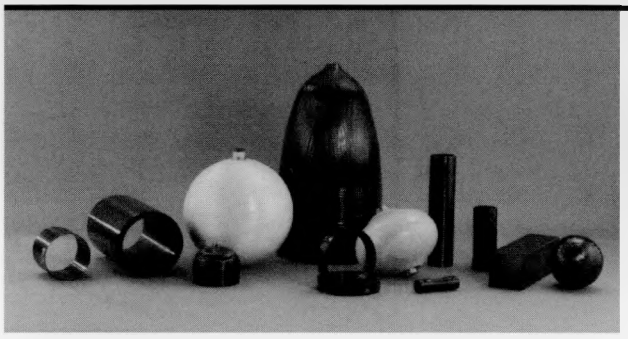


Figure 5. Variety of complex shapes that we can now wind.

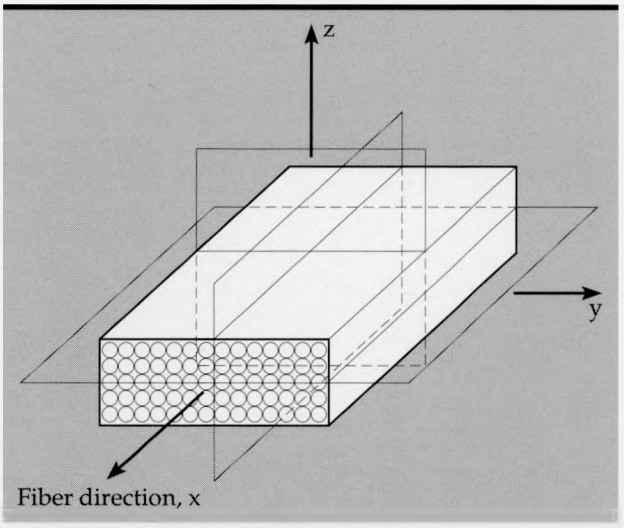


Figure 6. Planes of symmetry in a unidirectional composite lamina. Composite materials are orthotropic; that is, they have three mutually perpendicular planes of symmetry.

thermal-mechanical loading. In particular, we are attempting to develop a new, more appropriate failure criterion. The key to developing such a criterion is the ability to determine experimentally the necessary failure data. This is not an easy task. Beyond this, in order to evaluate the failure criterion in a multiaxial stress field, we must first be able to generate controlled multiaxial failure surfaces. The most common way to do this for metals (isotropic materials) is combined axial tension and torsion of tube specimens. Although a similar procedure can be used for composites, it is not without problems. For instance, it is almost impossible to grip the composite tube, in order to apply the loads, without inducing failure in the grip region.

To eliminate this problem, we designed a composite tube that flared at both ends and then designed grips to hold the flared tube stationary **[Fig. 7(a)]**. This unique combination of tubes and grips worked: it minimized the stress that was previously concentrated in the grip region and transferred it uniformly along the length of the composite. We can now subject these tubes to combinations of tension, compression, torsion, and internal pressurization to generate the failure surfaces. Moreover, we can be confident that the resulting data reflect material failure as opposed to grip failure. The development of these tools has been a critical step toward our goal of successfully predicting the strength and failure modes of fiber composites.

Figure 7(b) shows the tube and grips after they have been placed within a numerically controlled, serval-hydraulic test instrument. Using this instrument, we have successfully applied axial load, torsion, and internal pressure simultaneously to a T300/F263 pre-preg tape. The results have been used to graph the three-dimensional failure surface shown in **Fig. 8**. We did encounter a problem, however, when we pressurized the tube: it leaked. It appears that once a small crack forms, the internal pressure is lost, even though the composite may be structurally sound. We are currently investigating different types of nonstructural internal bladders to help increase the internal pressure.

Our success in loading composites to failure has led us to a point where we can establish realistic criteria for strength and failure in composites. This, in turn, has allowed us to begin altering fabrication parameters so as to increase the strength of a given composite. Currently, we are using the test apparatus in an experiment to increase the compressive strength of a composite (Toray 1000/DER 332-T403) of interest to the Laboratory's Earth

Penetrator project. Our goal is to obtain the highest specific compressive strength possible.

We have found several ways to increase compressive strength, one of which is to increase the percentage of fiber in the composite. (Theoretically, 60% is the optimum fiber volume,[1] but we have not yet verified that figure.) One way to achieve as high a fiber volume as possible is to make the composite as compact as possible. We have found that by manipulating the parameters of two of the fabrication steps, filament winding and curing, we can increase the compaction of the composite. In one of the steps, filament winding, we wound fiber around a mandrel under 8 lb of tension, a much higher tension than previously tried (any greater tension could damage the fiber). We achieved a fiber volume of 56%, close to the optimum figure quoted above. In the second method, we cured the composite in an autoclave, which allowed us to use pressure to squeeze resin out of

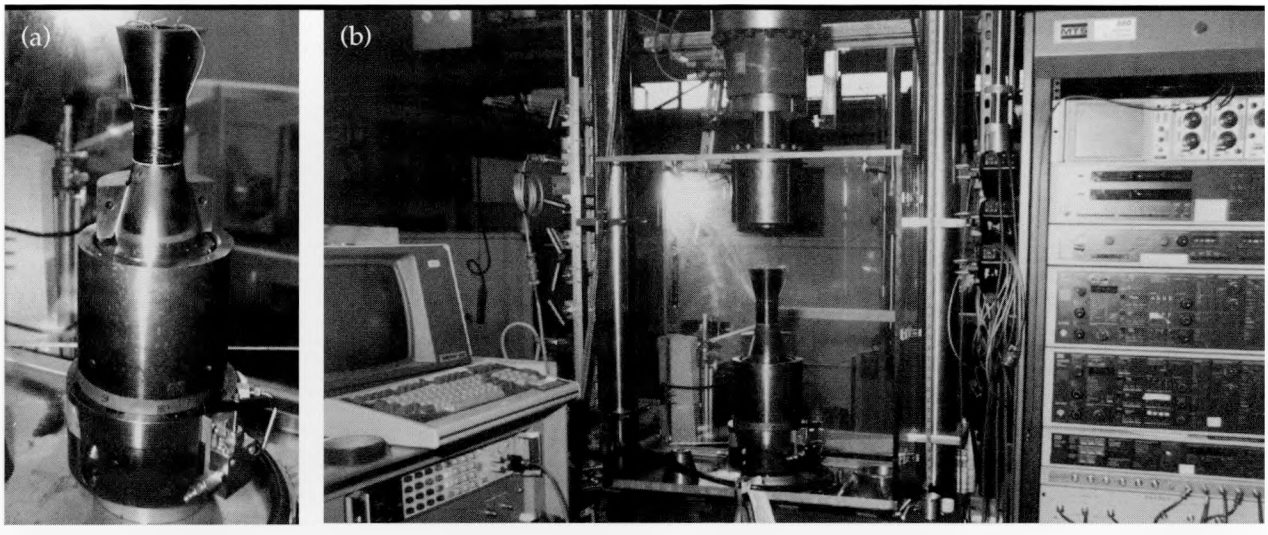


Figure 7. The 2-in.-diam flared composite tube we designed to minimize the stress that is concentrated in the grip region (a) and a photo of the multiaxial test system (b).

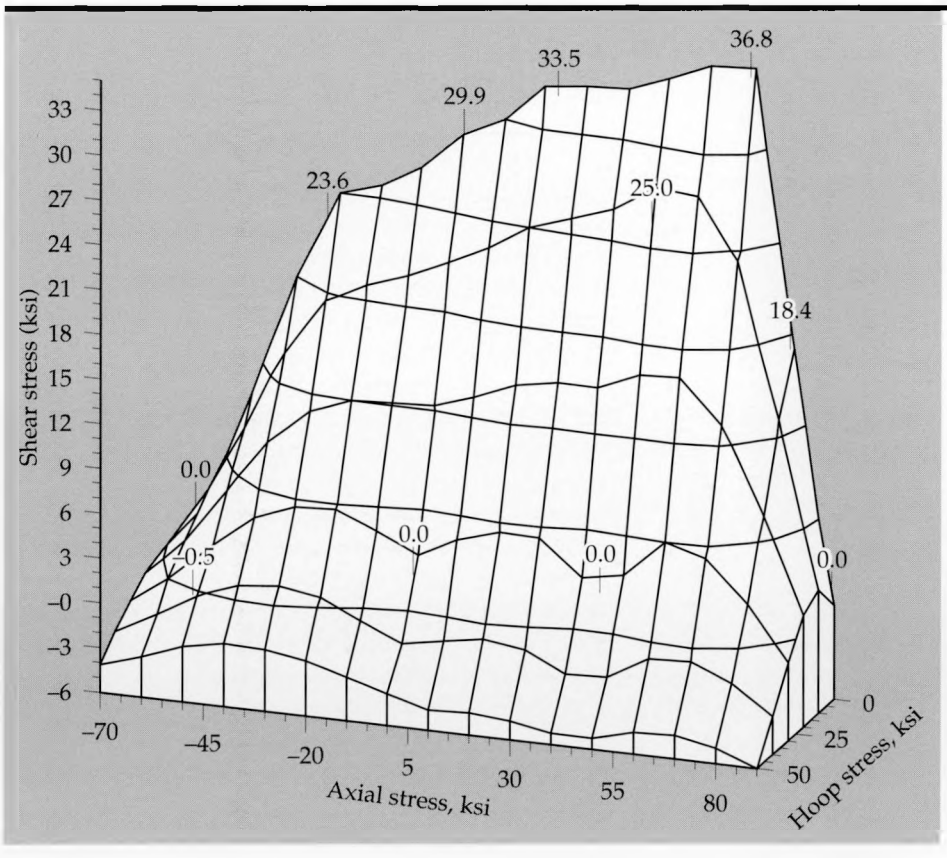


Figure 8. Multiaxial failure surface generated using the test system.

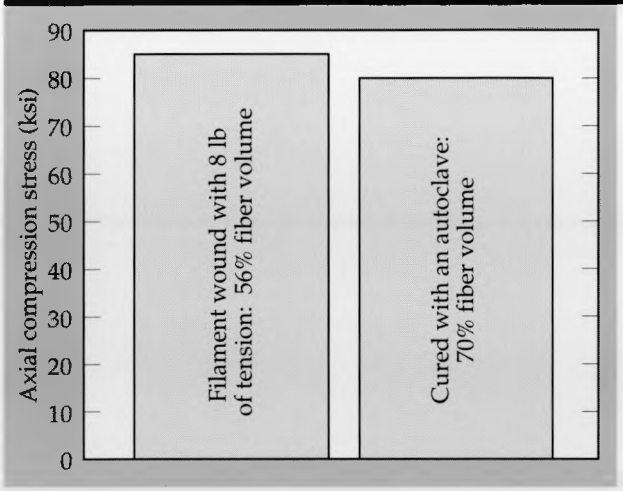


Figure 9. Comparison of the compression strength of a composite wherein two different fabrication techniques were used to increase the fiber volume. The bar on the left represents a composite wound with 8 lb of tension, resulting in a fiber volume of 56%. The bar on the right represents a composite that was cured in an autoclave, resulting in a fiber volume of 70%. The right hand composite failed earlier, even though it had a higher fiber volume, because it had been wrinkled during the curing process.

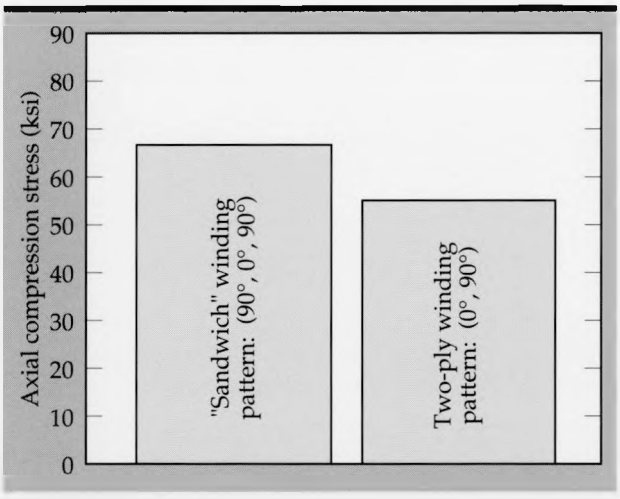


Figure 10. Comparison of the compression strength of a composite wound with two different winding patterns. The pattern shown at the left represents a "sandwich" approach, where the axial fibers (0°) have been sandwiched between two sets of hoop fibers (90°). This winding pattern results in a composite with greater compressive strength than a composite wound with just one set of hoop fibers.

the composite, thereby making the tube as compact as possible. The second technique resulted in a fiber volume of 70%. Next we tested both tubes to failure and found that the first technique was more

Figure 11. Pinned cylindrical shear joint for a graphite/epoxy tube.

successful: the tube withstood greater stress before failing (**Fig. 9**), even though it had a lower fiber volume. The reason that the tube with the higher fiber volume failed earlier was that the pressure applied during the curing process severely wrinkled the fibers, reducing the tube's compressive strength. Defects such as wrinkling, fiber misalignment, fiber waviness, and voids can all be easily introduced during the fabrication process and need to be guarded against. They can result in fiber microbuckling (local failure of a fiber), which in turn can lead to the premature failure of the composite, rather than to compressive fiber failure wherein the composite sustains the predicted load. In our experiment, we have abandoned using an autoclave to increase fiber volume, but have continued to wind fiber at 8 lb of tension.

Another way to increase compressive strength is to vary the ply layup parameters; that is, the angle at which the fiber is wound around the mandrel,

the number of layers wound at a given angle, and the order of the various layers. We are currently investigating a composite whose fibers are wound at 0°, 45°, and 90° angles (actually, the angles are 1.5°, 45°, and 89°; the winder has limitations that prevent it from winding at exactly 0° and 90° angles, but the convention is to refer to the angles as 0°, 45°, and 90°, so we will continue to do so). The reason that we wind at a variety of angles is that if we just wound at one angle (say 0°) the tube would collapse when subjected to axial compression: it would split down its length as though it were a barrel whose staves had caved in. To prevent such splitting, we wrap a layer of 90° hoop fibers around the 0° fibers. Now when we compress the structure, we find that the failure mode is one in which the axial fibers buckle inwardly (the hoop fibers stop the axial fibers from bowing outwardly). That is, when we compress the composite, the compressive force on the axial fibers (0°) is transferred to the hoop fibers (90°) as tensile stress. Because the tensile stress of the hoop fibers is much greater than the compressive strength of the axial fibers, the latter will buckle inwardly. To prevent inward buckling, we next put together a fiber sandwich: we wound a set of 0° axial fibers between two sets of 90° hoop fibers. **Figure 10** shows the success of the sandwich approach: a (90°, 0°, 90°) pattern exhibited greater compressive strength than the (0°, 90°) pattern. We are currently investigating whether fibers at a 45° angle will also increase compressive strength.

In FY 90 we will investigate several other questions concerning compressive performance:

- How sensitive is the compression performance to the angle of the axial fiber?
- Does increasing the stiffness of the hoop fibers increase compression strength?

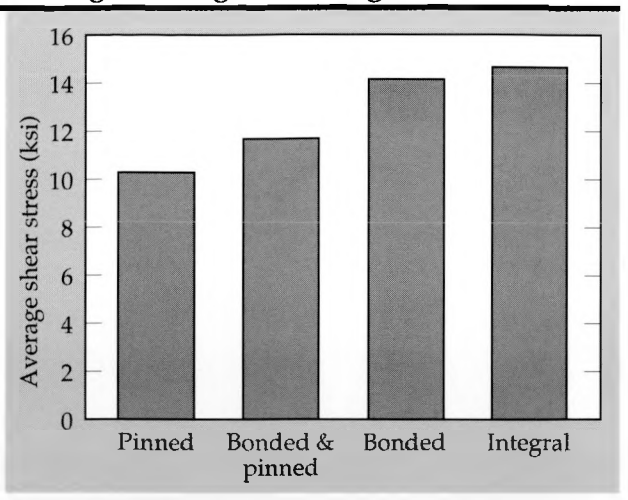


Figure 12. Strength of cylindrical shear joints for 2-in.-diam graphite/epoxy tubes.

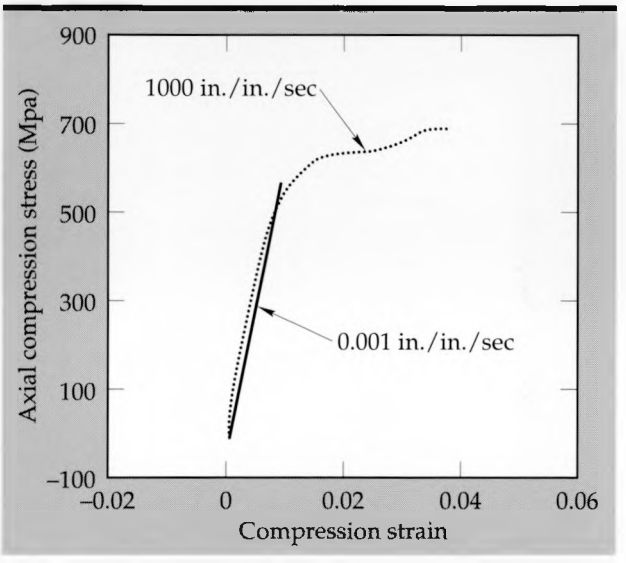


Figure 13. Comparison of the effects of high and low strain rates on a graphite/epoxy composite. We used a Hopkinson bar to induce axial compression.

Tests of Cylindrical Shear Joints

In addition to using the multiaxial test apparatus to determine the compression strength of fiber composites, we have also used it to evaluate the shear strength of four different cylindrical joints made of graphite and epoxy. The four designs consist of a bonded wedge cone, pinned wedge cone (**Fig. 11**), bonded and pinned wedge cone, and an integrally wound wedge cone. All four designs were similar to our multiaxial test part: they all had an axial length of 1 in. and a 15° tapered cone. Our goal is to achieve an average shear strength of 8000 psi along the joint. This value was derived by calculating the maximum shear strength of the epoxy system used in the bonding. As can be seen in **Fig. 12**, all our results were above this value. We believe the increase in available shear strength can be attributed to the hydrostatic compression effects from the gripping mechanism (we discuss this work in more detail in **Ref. 2**).

Out of Plane Failure

To help us further understand the complex failure performance of composite materials, we have developed out-of-plane shear loading fixtures, a three-dimensional compression fixture, and an external ring compression fixture. To graph a com-

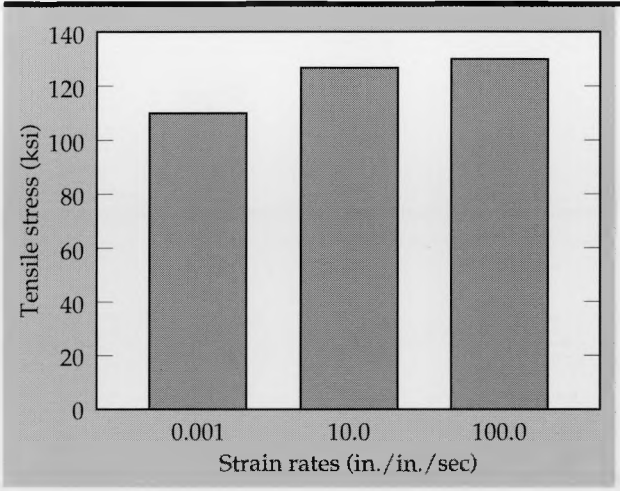


Figure 14. Comparison of the effects of high and low strain rates on a 0.5-in.-diam composite. We used a high rate servo hydraulic MTS machine shown in Fig. 7(b) to induce axial tension.

plete failure surface we must merge the data on out-of-plane stress fields obtained with these tools with the data on in-plane stress fields generated with the multiaxial test apparatus. We do not yet have the ability to load a part in three dimensions.

Dynamic Properties

So far in this paper we have been discussing loading composites at very low strain rates: approximately 0.001 units of displacement (in./in.) per second. We have also developed techniques for loading composites at much greater stress/strain rates because some of our composites may have to withstand such rates. (A composite could be used, for example, to protect the innards of a missile from an explosion close by.) Our purpose is to develop a strain-rate-sensitive material model for inclusion in the Laboratory's DYNA code.

In our first tests we used a Hopkinson bar (a standard device for measuring stress/strain behavior at high rates)[3] to test a composite at a strain rate of 1000 in./in./sec (**Fig. 13**). As can be seen, the toughness (the area under the curve) is approximately eight times greater for a composite subjected to a higher stress/strain rate. Although we do not yet understand the mechanisms behind this result, a working hypothesis is that the increased toughness can be attributed to several factors, including the viscoelastic behavior of the epoxy and to the formation of noncoalescing internal damage mechanisms such as microcracking. Also at these higher strain rates, it is possible that the loading rate is fast enough to prevent the usual delamination of the composite before failure.

All the above tests have been conducted in compression; however, we have also developed a 0.5-in.-diam version of the composite tube (as opposed to the 2-in.-diam one) and have measured its stress/strain rate in tension at rates up to 100 units of displacement per second. As **Fig. 14** shows, we successfully stressed the composite in tension and again found that, as in compression, the greater the stress/strain rate, the higher the tensile strength.

Future Work

In FY 90 we will generate multiaxial failure data for the composite T1000/DER-T403. Our plan is to vary the winding patterns of this composite and see what changes in behavior occur. We will also evaluate several published failure criteria[4–6] and begin to formulate our own more general criterion. As part of that formulation, we will continue to generate high strain rate data using the drop tower and the Hopkinson bar.

1. Joint work between Chemistrys' *"Polymer Composites Group"*, and Engineerings' *"Materials Test and Evaluation Section"*.
2. S. E. Groves, B. Sanchez, and F. Magness, *"Evaluation of Cylindrical Shear Joints for Composite Materials,"* to be published.
3. S. E. Groves and D. H. Lassila, *"High Strain Rate Testing of Composite Materials Using the Hopkinson Bar,"* LLNL work in progress, 1989.
4. R. M. Jones, *Mechanics of Composite Materials*, McGraw-Hill, 1975.
5. R. M. Christensen and S. R. Swanson, *"Evaluation of a New Failure Criterion for Fibrous Composite Materials,"* UCRL-98725, May 1988.
6. W. Feng, *"A Finite Strain Elasticity Approach to Failure Modelling of Composite Materials,"* LLNL work in progress, 1989.

Modeling and Experimental Measurement of Residual Stress

Elane C. Flower
Engineering Sciences Division
Mechanical Engineering

Our objective is to determine residual stress through finite-element modeling and with reliable experimental methods. We have previously demonstrated our ability to calculate with the finite-element method and experimentally measure deformation-induced residual stress with neutrons and x rays and by mechanical and ultrasonic techniques.[1] From our studies, we have concluded that diffraction is the preferred experimental method and that neutron or high-energy radiation can be used to determine residual stress in the bulk or volume of the specimen. Measurements in the bulk or volume compare favorably to our finite-element analysis results. With researchers at the University of California, Davis, we have designed and analyzed an advanced, high-energy x-ray system that could be built at LLNL and potentially obtain both surface and bulk strain measurements. In March, 1989, we carried out experiments at the Stanford Synchrotron Radiation Laboratory, using synchrotron radiation, an intense, highly focused, x-ray tunable radiation source with a wide range of available energies, to determine residual stress. These were the first experiments nationally to try to measure macro residual stress in metal components. With researchers at the University of California, Berkeley, we have successfully demonstrated a method that uses longitudinal acoustic waves to determine residual stress in planar aluminum specimens. We have also modeled and carried out additional neutron diffraction measurements of residual stress in a textured, hexagonal, close-packed alloy that exhibits a high degree of plastic anisotropy (directionally dependent flow properties). To model the anisotropic flow behavior, we used a Hill's criterion developed and implemented into our NIKE2D finite-element code. Although we could model anisotropic flow behavior, our calculations showed that we could not account for the effect of grain-to-grain interaction strains that develop in this class of metals.

Introduction

LLNL designs and manufactures complex components that must perform reliably for many years. Residual stress has been associated with the premature failure of components or, conversely, can be introduced to improve component life. As our first step in managing residual stress, we are developing methods and expertise that can be applied in the early stages of the design process.

Residual stresses are self-equilibrating stresses that result from non-uniform plastic deformation and remain in a body under zero external forces. The three main types of residual stresses are widely designated as being of the first, second, and third kind. Residual stresses of the first kind (macro residual stresses) are homogeneous; that is, they are constant in magnitude and direction over a large area—such as several grains (millimeters to centimeters)—and are equilibrated within the whole body. Residual stresses of the second kind (micro residual stresses) are nearly homogeneous across a grain or parts of a grain (up to millimeters) and are equilibrated across a number of grains. Residual stresses of the third kind are inhomogeneous across submicroscopic areas (100 Å to 1000 Å) and are equilibrated across small parts of a grain. The total residual stress at a point in a material results from the superposition of residual stresses of the first, second, and third kind. This is represented schematically in **Fig. 1**, which shows residual stresses forming from the quenching of a two-phase metal cylinder.[2]

Determining residual stress is a formidable task that we have found best accomplished by applying a combination of methods. This combination includes finite-element modeling, diffraction, ultrasonics, and mechanical methods. Experimental methods do not measure residual stress directly, but rather a unique property that can be related to a stress value. Some methods will yield information on residual stress near the surface while other methods determine residual stress in the volume of

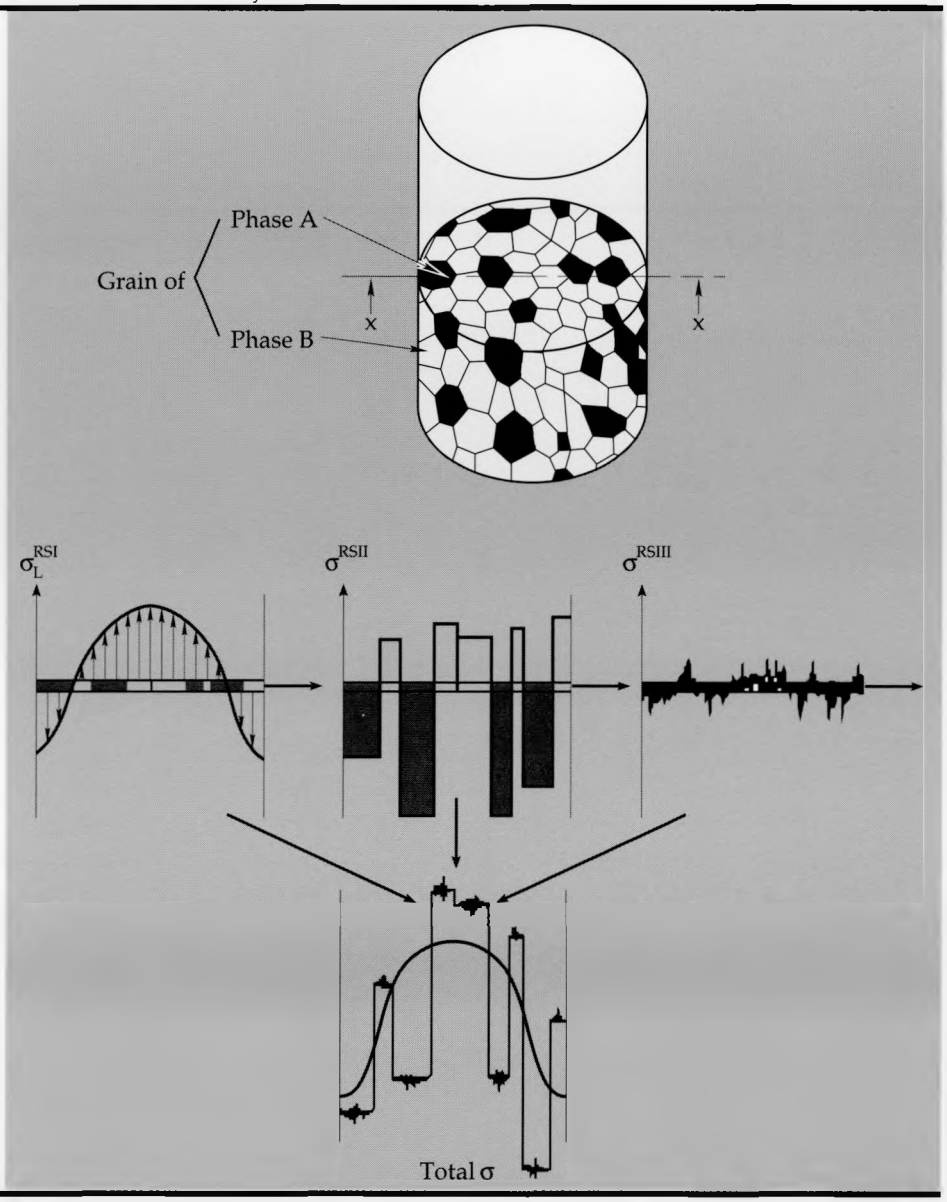


Figure 1. Schematic representation of the three kinds of residual stresses (shown as RSI, II, and III) that can develop in a quenched two-phase metallic cylinder. Macro residual stresses, graphically depicted in RSI, develop from the temperature gradient. Micro residual stresses, graphically depicted in RSII, develop from the difference in coefficients of thermal expansion between the two phases. Micro residual stresses, graphically depicted in RSIII, develop from internal stresses within the grains. The total residual stress at the position x to x is graphically depicted in the bottom figure.[2]

the body. **Table 1** summarizes the methods we have used to determine residual stress. To date, our experiments have concentrated on deformation-induced residual stress in metal components, but the knowledge and capabilities we have gained are applicable to residual stresses from sources other than deformation (i.e., welding and heat treatment) and for additional materials such as ceramics and metal matrix composites.

Our first experiment focused on modeling and experimental measurement of deformation-induced residual stress in an austenitic, stainless steel ring. This ring was compressed approximately 3% diametrically in a mechanical testing device, and the load was removed. To model the compression and unloading of the ring, we used two- and three-dimensional finite-element analysis. From our analysis, we determined that non-uniform plastic deformation produced measurable residual stress at the poles and waist of the ring. Because of the symmetry of the ring, however, we only needed to make experimental measurements of residual strains in two locations: 0 and 90 degrees.

Measurements of residual stress on the ring specimen were made with four separate techniques: neutron diffraction, x-ray diffraction, hole-drilling, and acoustoelasticity. Our goal in making these measurements was to compare our ability to model residual stress with the results obtained experimentally. Details of the results were presented previously.[1]

Non-destructive evaluation with diffraction was shown to be the preferred method for determining residual stress.[3] The use of neutrons, which penetrate more deeply than x-rays, provide a measure of residual strain in the volume of the specimen and are not as affected by surface treatments such as machining. But, in both neutron and x-ray dif-

Table 1. Characteristics of methods used to determine residual stress.

Method	Measured quantity	Kind(s) of residual stress determined	Character
Mechanical	Macroscopic surface strains via strain gage	First	Destructive
Conventional x-ray diffraction	Homogeneous lattice strains	First and second	Nondestructive (surface)
Synchrotron x-ray diffraction	Homogeneous lattice strains	First and second	Nondestructive (surface and bulk)
Neutron diffraction	Homogeneous lattice strains	First and second	Non-destructive (bulk)
Ultrasonic	Time-of-flight longitudinal waves	First, second, and third	Nondestructive (bulk, whole-field)
Finite-element method	Residual elastic strains/ residual stress	First	Numerical predictions (whole-field bulk)

fraction methods, strains are measured on a particular set of crystallographic (hkl) planes. For practical engineering problems, it is necessary to relate strains measured by diffraction to values of engineering stress. We have evaluated algorithms and procedures to convert the measured properties to the more commonly used values of engineering stress. If the material is homogeneous, isotropic, and has crystallographic elastic constants similar to the polycrystalline, this conversion is accomplished through the simple equation using Hooke's Law. If the crystallographic elastic constants are anisotropic, such as they are in stainless steel, strains are related to stress through Hooke's Law, but by using single-crystal approximations such as those of Reuss (assuming a constant stress in all grains), Voight (assuming a constant strain in all grains), or an averaging method. In our experiments, the Reuss approximation and the averaging method using polycrystalline elastic constants gave good correlation between residual stress measured by neutron diffraction and that calculated with the finite-element method.[4]

Progress

Modeling and measuring residual stress in an anisotropic metal.

Some important engineering metals and alloys are non-cubic, texture readily, and exhibit plastic anisotropy, such as beryllium, titanium, and uranium. Residual stresses of the second kind—grain-to-grain interaction stresses—are prone to develop during deformation or heat treatment in these metals and will be superimposed on the macro residual stresses. To study residual stresses in this class of materials, we chose Zircaloy-2, which has a hexagonal close-packed crystal structure as a model material. We were able to cost share a portion of our experiments with Chalk River Nuclear Laboratory.

Our test specimen was a beam 12.5 mm thick, 25.4 mm wide, and 127 mm long that was machined from hot-rolled, annealed bar stock and subjected to four-point bending. To determine the mechanical properties of the bar stock, we conducted tensile and compression tests. The flow properties of specimens taken from the bar depended on the direction from which they had been taken, i.e., transverse vs axial. The test specimens also exhibited a strength differential, both in yield strength and hardening, that depended on whether they were tested in tension or compression. This highly anisotropic behavior was a result of the bar stock's crystalline structure and a high degree of preferred orientation of the crystallites, more commonly called texture.

So that we might determine the texture of the bar stock, texture measurements normal to the transverse direction were made at Chalk River Nuclear Laboratory. In this orientation, the poles of the (10-10) and (10-11) planes were predominant and nearly equally represented. No poles of the basal (0002) planes were found normal to the trans-

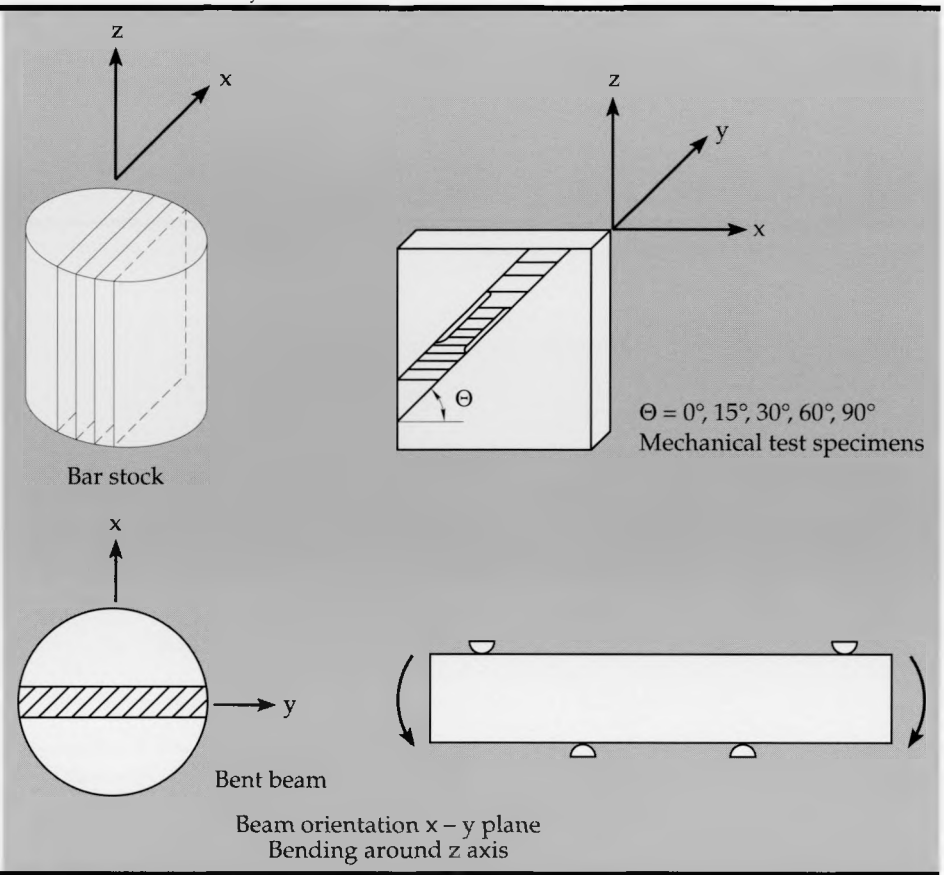


Figure 2. Orientation of mechanical test specimens and bent beam from the bar stock.

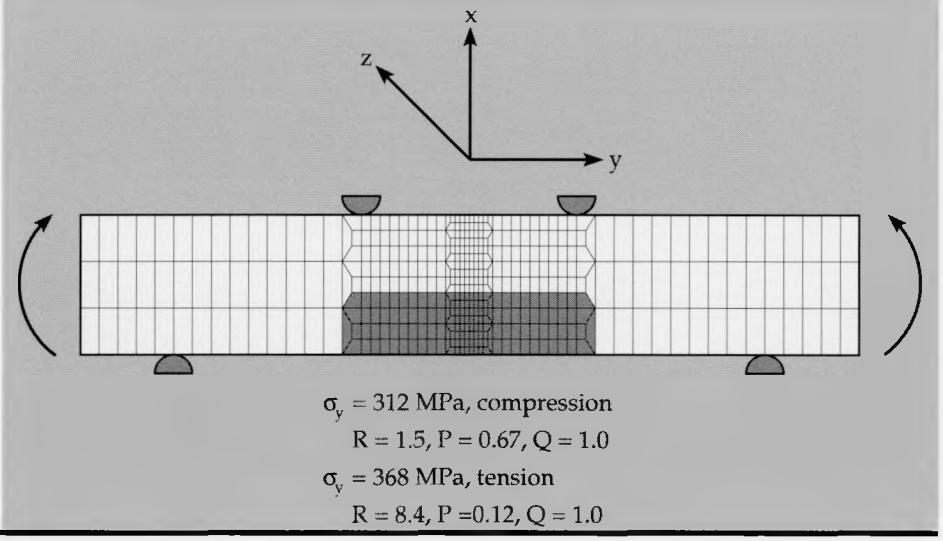


Figure 3. Finite-element model for the Zircaloy-2 bent beam. A composite mesh was used to account for the strength differential, which was predominant over the plastic anisotropy.

verse direction; instead, they were randomly oriented primarily perpendicular to the bar axis.

To model the beam specimen with finite-element modeling, we used Hill's anisotropic flow criterion that had been implemented into NIKE2D.[5] This criterion is commonly used to account for normal anisotropy, as found in sheet products; that is, where the mechanical properties in the plane of the sheet are nearly isotropic. From our mechanical tests, we were able to obtain a yield surface and determine the constants for the Hill model. We assumed that the bar was plane stress and isotropic in the x–y plane; bending of the bar was around the z axis (**Fig. 2**). Our constitutive law cannot account for the difference in strength between tension and compression, so we used a composite mesh with appropriate material properties for the area of the model dominated by tension and that dominated by compression (**Fig. 3**). Strain gage measurements were made on the actual test specimen, and the model was deformed to duplicate the same amount of strain before unloading.

Neutron diffraction measurements were carried out on the L3 triple-axis spectrometer at the Chalk River reactor, which was operated in the diffraction mode. Strain on several hkl planes was measured in the transverse, longitudinal, and normal directions (e_x, e_y, and e_z) at the bar midsection, from the bent outer surface to the inner surface. As noted earlier, strains measured by diffraction will include the effects of residual stresses of the first and second kind. Residual stresses of the second kind develop as a result of compatibility being maintained between grains. **Figure 4**, which shows neutron diffraction measurements of the longitudinal, transverse, and normal strains, illustrates that the NIKE2D calculations follow closely those of measurements made on the (10-12) planes. Measurements of longitudinal strain made on the (0002) planes were of the opposite sign (in compression rather than tension) on the side that was in compression during bending. This can be explained by the simple two-grain model in **Fig. 5**. The grain-interaction strains (from residual stress of the second kind) can account for the sign difference. Slip is not favorable when applied stress is normal to the (0002) planes and the (0002) planes remain strained from the applied stress. Twinning occurs at lower strain levels in tension, thus eliminating strain on the (0002) planes on the side of the beam that was in tension. As expected, strains measured in the transverse direction (**Fig. 4b**) show the same effect on the tension side of the beam as a result of the Poisson effect.

Our research shows that the correlations between strains calculated with NIKE2D and those measured with neutron diffraction are in good agreement. In cubic, non-textured metals and alloys, we can relate measured strains to engineering stress with Hooke's Law and various single-crystal

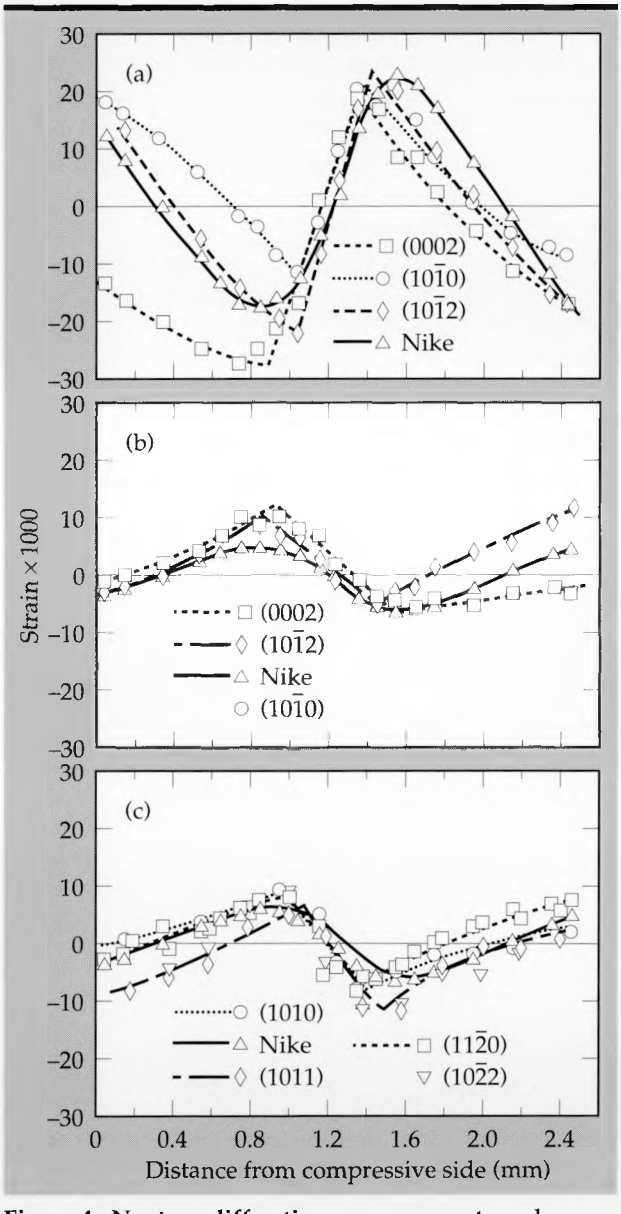


Figure 4. Neutron diffraction measurements and NIKE2D results of (a) longitudinal, (b) transverse, and (c) normal strains in bent beam.

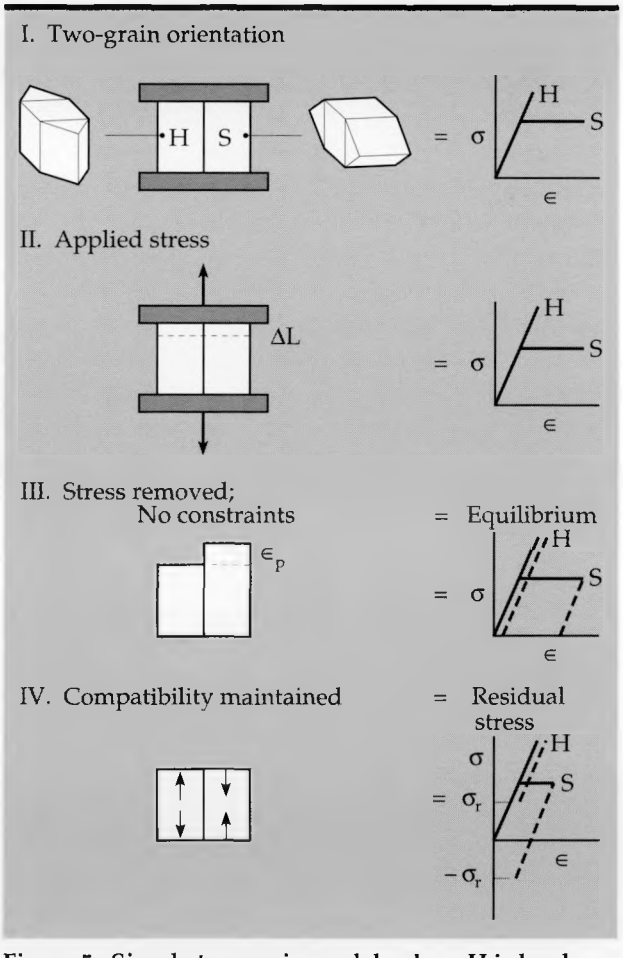


Figure 5. Simple two-grain model, where H is hard direction (no slip) and S is soft direction (easy slip).

approximations. In highly textured, non-cubic materials, correlations between calculated and experimental results depend on both the choice of hkl and the deformation mode. The NIKE2D calculations can account for the plastic anisotropy, and we were able to adjust our model to account for the strength differential. We could not account for grain-to-grain interaction strains, however, but were able to explain their effect with a simple qualitative model.

Work at Chalk River Nuclear Laboratory over the past five years has resulted in a computer program called QUEST[6] that, although it cannot predict residual macro stresses, can predict residual grain-interaction stresses in samples that have been deformed in uniform tension or compression. The analysis assumes that the total strain is the same in every orientation of grain (Taylor, upper-bound model), but allows for the amounts of elastic, plastic, and thermal strain to vary according to the orientation of the grain. Future goals would be to develop a hybrid code that couples QUEST with the NIKE2D finite-element code that would be able to calculate both macro and micro stresses (RS1 and RS2).

Ultrasonic determination of residual stress, UCB

Our colleagues Professor George G. Johnson and Dr. Jay J. Dike at U. C. Berkeley have completed a study of the acoustoelastic technique, using longitudinal waves to evaluate residual stress.[7] The acoustoelasticity theory describes how the velocity at which elastic waves propagate in a solid is affected by the state of stress in that solid. The theory has been applied to planar geometries, using longitudinal waves propagating normal to the plane of stress. The spatial variations in longitudinal wave velocity are used to determine stress fields for each component of the two-dimensional stress tensor for a body in a state of plane residual stress.

In isotropic materials, changes in longitudinal wave speed may be related to the sum of the in-plane normal stresses. This work has shown that the individual components of stress may be extracted from this sum by manipulating the equations of equilibrium to obtain Poisson's equation for shear stress. Solution of Poisson's equation yields the shear stress field, which may then be used in the equilibrium equations to compute the individual normal stress components. Part of our research project was to extend the algorithm to obtain the values of the individual stress components in anisotropic materials. This was achieved using an iterative scheme.

A phase-measuring system was developed to obtain the velocity information required to compute residual stresses. Two different phase-measuring systems are required to measure thickness and the time of flight of the acoustic wave. The measurements provide an estimate of the spatial variation of the wave velocity.

The algorithms and the measuring system developed at UC Berkeley were used to determine residual stress in a 6061-T6 aluminum ring that had

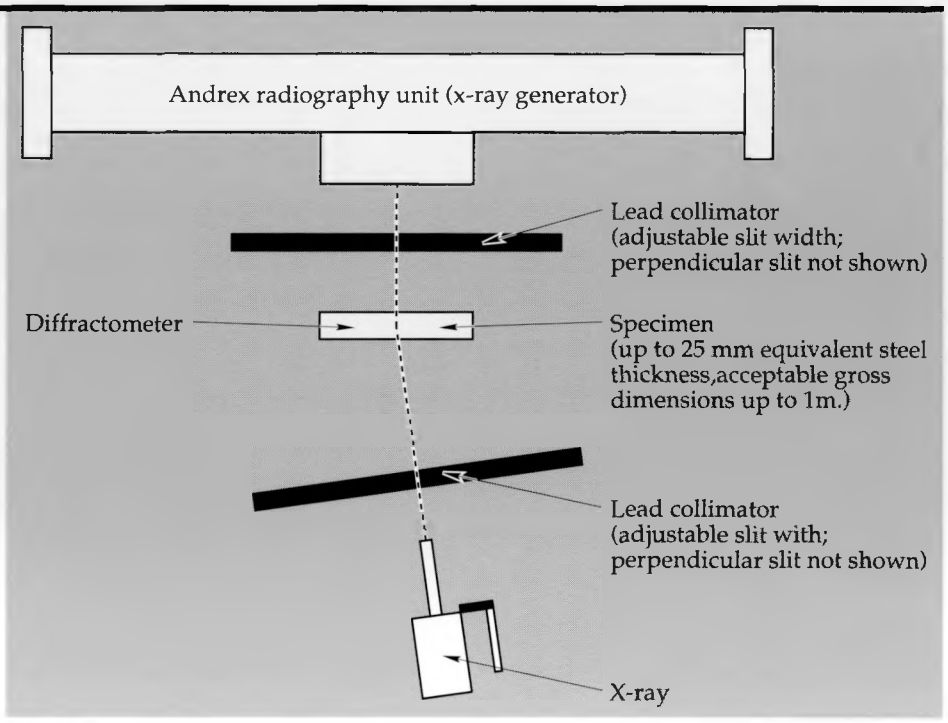


Figure 6. Schematic overview of the pre-prototype experimental apparatus, showing the redesigned diffractometer.

been plastically deformed and unloaded. Then the results were compared to finite-element calculations performed at LLNL. Comparisons of the stress fields with those from finite-element modeling solutions show that the experimental technique does indeed provide spatially detailed information about residual stress. Subsequent similar experimental measurements on the stainless steel ring did not have adequate resolution because of the low value of the acoustoelastic constant of the stainless steel, which is an order of magnitude lower than that of the aluminum. The same was true of the Zircaloy specimen.

In conclusion, residual stress can be measured successfully using longitudinal acoustic waves in planar aluminum specimens. Measurements in metals with lower acoustoelastic constants do not provide a high enough degree of resolution.

Design of triaxial residual stress measurement system using high-energy x-ray diffraction, UC Davis.

With our colleagues Professor James Shackelford, Dr. Brian Brown, and Jun Park at the University of California, Davis, we have been designing a triaxial, residual-stress measurement system that uses high-energy x-ray diffraction. From our previous work and a comprehensive literature review, we have determined that diffraction is the preferred experimental method for measuring residual stress. Conventional x-ray sources can only determine planar near surface stresses (e.g., 1 to 3 μm in iron), while high-energy sources of 10 to 100 keV could penetrate 2 to 3 cm in iron. Measurements within the bulk can provide the whole strain tensor and avoid surface strains associated with machining or other surface treatments.

The prototype system that was built for bulk residual stress measurements has five subsystems:

1. X-ray generator
2. Diffractometer
3. X-ray detection
4. Data acquisition
5. Data analysis

Following the initial experiments, the diffractometer was redesigned. The key features of the redesign include (1) thicker collimator plates, (2) variable collimator slit widths, and (3) accommodation of much larger samples. A schematic of the x-ray generator, redesigned diffractometer, and x-ray detection unit is illustrated in **Fig. 6**.

The data-acquisition system was being developed around the LabVIEW program. However, the detector is capable of producing data at a speed of about 100 times that of the data-acquisition board, so the system had to be modified.

We have also reviewed the traditional and recent advances of analysis of x-ray diffraction data for residual stress information and have extended the capability for the purpose of determining the x-ray elastic constants of engineering materials.[8]

Residual stress measurement with synchrotron radiation.

Another source of high-energy x rays is synchrotron radiation, where a packet of high-energy electrons is accelerated, in a "circular" ring, to speeds close to the speed of light. The centripetal acceleration due to the curvature of the path causes the electrons to emit electromagnetic radiation that will, depending on the energy of the electrons, have a range of wavelengths. The energy will also determine the depth of penetration. With our colleagues at UC Davis, we have conducted the first macro residual stresses nationally, using synchrotron radiation at the Stanford Synchrotron Radiation Laboratory. Benefits of a synchrotron source to our studies include a highly intense and monochromatic beam with variable energies that allow significant sample penetration and very low beam divergence. Details of this work are reported elsewhere.[9]

Conclusion

Our modeling studies and experimental measurements are complete, and we can now model successfully with finite-element method, deformation-induced residual stress. The models we have developed aid in understanding residual stress in materials that are highly textured, non-cubic, and exhibit flow anisotropy, and our evaluation of several experimental methods for determining residual stress has convinced us that diffraction with penetrating radiation (neutron and high-energy x rays) is the preferred method for determining bulk residual stress. We have analyzed and developed a prototype high-energy x-ray diffraction measurement system that could be built at LLNL. We have also aided in the development of an ultrasonic method that uses longitudinal acoustic waves to determine whole-field, bulk residual stress in planar aluminium specimens. In conjunction with our project, we successfully applied (for the first time nationally) synchrotron radiation at Stanford Synchrotron Radiation Laboratory to measure macro residual stresses. We are now collaborating with Chalk River Nuclear Laboratory and the Los Alamos National Laboratory, using a pulsed neu-

tron source to measure macro residual stress in our experimental ring at the Los Alamos Neutron Scattering Center.

1. E. C. Flower, "Finite-Element Modeling and Experimental Measurement of Residual Stress," *Engineering Research and Development Thrust Area Report 88*, Lawrence Livermore National Laboratory, Livermore, CA, UCRL-53868-88, pp. 5-23–5-26
2. E. Maucherauch and K.H. Kloos, "Origin, Measurement, and Evaluation of Residual Stresses in Science and Technology, " *Proceedings of International Conference on Residual Stress*, Vol. 1, Informationsgesellschaft, Verlag (1986).
3. J. F. Shackelford and B. D. Brown, *A Critical Review of Residual Stress Technology, International Advances in Nondestructive Testing*, Lawrence Livermore National Laboratory, Livermore, CA, to be published.
4. E. C. Flower, S. R. MacEwen, T. M. Holden, and R. R. Hobsons, "A Comparison of Finite-Element Analysis with Neutron Diffraction and X-Ray Measurements of Residual Strains in Stainless Steel," *Proceedings of International Conference on Residual Stress*, UCRL-99761, Lawrence Livermore National Laboratory, Livermore, CA (1988).
5. R.W. Logan, *Incorporating Non-Quadratic and Crystallographic Yield Surfaces into Finite-Element Codes*, Los Alamos National Laboratory, Los Alamos, NM, LA-CP-88-186 (1988).
6. S. R. MacEwen, C. Tome, and J. Faber, "Residual Stresses in Annealed Zircaloy," *Acta Metall.*, **37**(3), 979-989.
7. J. J. Dike, *An Acoustoelastic Technique Using Longitudinal Waves for the Evaluation of Residual Stresses*, Ph.D. thesis, U.C. Berkeley, Berkeley, CA (1989).
8. J. S. Park, and J. F. Shackelford, *Analysis of X-Ray Diffraction Data for the Characterization of Residual Stress*, UC Davis, internal report, 1989.
9. J. F. Shackelford, B.D. Brown, and J. S. Park, *Modeling and Experimental Measurements of Residual Stress Using Synchrotron Radiation*, Lawrence Livermore National Laboratory, Livermore, CA, UCRL-21197 (1989).

ENGINEERING RESEARCH AND DEVELOPMENT

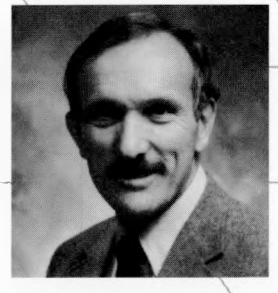
Thrust Area Report FY 89

Microwave and Pulsed Power

Our thrust area conducts research into the technologies of high-power microwaves and pulsed power. The results benefit a number of high-energy programs at the Laboratory and DOD initiatives outside the Laboratory. In FY 89 our research projects included:

- *Investigating high-power, optically triggered switches.* Our research has produced an optically switched, solid-state microwave source with a narrow pulse (<1 nsec) that generates several megawatts of power. This pulser can be used in high-power microwave weapons, ultra-wideband radar systems, and can also be used commercially.
- *Studying microwave sources powered by electron beams.* As part of this study, we are developing a backward wave oscillator (BWO). This device is an efficient source of high-power microwaves and can be used in HPM weapons and to power high-gradient, radio-frequency accelerators.
- *Developing a more accurate model of magnetic switches.* The model is being used to predict the performance of high-power modulators and linear induction accelerators.
- *Completing the development of a high-current ignitron switch for the Inertial Confinement Fusion Program.* Using the new ignitron we successfully switched 900 kA of current and 250 C of charge.

Wayne Hofer
Thrust Area Leader



A Long-Pulse, 50-kV Electron-Gun Modulator

Ron Kihara
Nuclear Energy Systems Division
Electronics Engineering

We designed, analyzed, and built a 50-kV electron-gun modulator in response to a request from one of the Laboratory programs. An optimized Rayleigh pulse-forming network (PFN) with mutually coupled interstage inductors generates the output of the modulator. The output is a rectangular pulse that has a 100-μsec flat-top, a full-width-half-maximum (FWHM) of 135 μsec, and a 10 to 90% rise time of 11 μsec. The flat-top ripple varies from ±0.35% to ±0.50% of peak voltage over the operating range of the PFN. We designed the modulator to drive an electron gun having an impedance of 5 kΩ, in shunt with a ballast resistor of 155 Ω. The completed modulator achieved the requested design goals more simply and cost effectively than anticipated.

Introduction

In FY 88, one of the Laboratory programs asked the Microwave and Pulsed Power thrust area to design and develop a 50-kV electron-gun (e-gun) modulator, which would be used in the development of improved e-gun designs. The guns used now operate continuously at average powers greater than 300 kW. However, the experimental facility does not require a continuous or dc electron beam to characterize experimental designs because the beam reaches dc stabilization in 30 to 40 μsec. Therefore, using a modulator with a short pulse (<100 μsec) and a low-duty cycle (<10^{-4}) is advantageous because new designs can be characterized at a much lower level of power dissipation (<100 W). The initial request was for a design based on a beam-power tetrode tube, type 4CW100,000, and the parameters listed in **Table 1.**

Upon examining the listed parameters and performing some preliminary analysis and design, we proposed replacing the tetrode with a spark-gap-switched PFN because we felt that it would achieve the same goals at substantially lower cost and effort. Our counter proposal was accepted and resulted in the modulator delivered.

Table 1. Minimum required electron-gun modulator specifications

Maximum output voltage: 45 kV
Pulse width: ≥120 μsec FWHM
Flat-top: 100 μsec for ±1% ripple
Max. repetition rate: 1 pps
Ballast matched into a 5-kΩ e-beam diode

Design and Analysis

Design Considerations

The originally requested design would have generated a rectangular pulse across the e-gun by commutating a series capacitor with a beam-power tetrode. The tetrode would have been required to both open and close the circuit between the gun and the capacitor.

Our alternative approach is less complex and more cost effective. It uses the properties of a PFN to generate a rectangular pulse and to turn off the e-gun. Unlike a capacitor, a PFN generates a rectangular discharge waveform when switched into a matched load. Since the output voltage of the PFN drops to zero when the line is discharged, it turns off the e-gun, thereby eliminating the need for the output switch to shut off the e-gun. As a result, we could replace the originally requested beam-power tetrode design with a new design based on a PFN using a much cheaper and simpler closing switch, such as a spark gap. By not using the beam-power tetrode, we eliminated the need for the filament, bias supplies, and amplifier chain required to drive a hard-tube modulator; thus, we considerably reduced circuit complexity and cost.

Network Modeling and Simulation

We analyzed two types of voltage-fed pulse-forming networks for properties that met the re-

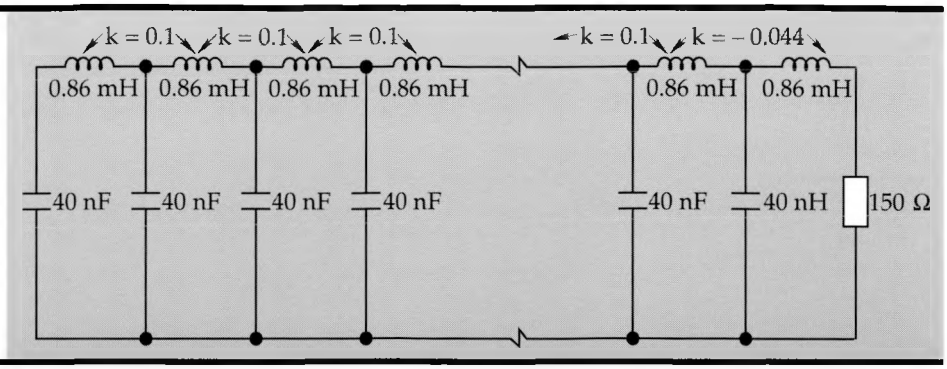


Figure 1. Schematic diagram of final PFN design. Positive mutual couplings between inductors are flux aiding, and negative couplings are flux opposing.

quested specifications: a Guillemin type "B" network of five to seven sections and a 10-section Rayleigh line. We tuned each network in computer simulation using the Microcap II circuit analysis program. Standard commercially available values of capacitors were used in the simulation to generate the initial design, with the additional constraint that all capacitors be identical, or even multiples of each other. The basic designs were initially optimized with the inductors left uncoupled. Ballast loading on both networks allowed us to use standard capacitor values and reasonably sized inductor values in the design. In the application, the ballast load also stabilizes the output voltage against fluctuations in gun impedance.

We rejected the type "B" network early in the design process because it could not be tuned for less than 2% peak-to-peak ripple in simulation and still meet the desired rise time parameter. We unsuccessfully attempted to smooth the flat-top further by adding more sections to the PFN. Increasing the number of sections from five to seven had very little effect on the ripple amplitudes, although the ripple frequency increased. When inductive smoothing was applied to the seven-section lines, the output pulse was identical to the smoothed five-section line, negating the advantage of the extra sections. A requirement for five different values of inductance and two values of capacitance also complicated the design.

To simplify the final design and to obtain optimum performance, we used the Rayleigh line. It easily tuned to less than 1% flat-top ripple while maintaining a rise time parameter of less than 10%, which is well within stipulated specifications. Also, capacitors in this design were all identical, as were all inductors except one, which greatly simplified the task of construction.

The next stage in the design took advantage of mutual coupling between the inductors to smooth the flat-top even more. The final design **(Fig. 1)** had a mutual coupling of 10% between all inductors except the last, which was –4.4%. The flat-top ripple in simulation was 0.1% peak-to-peak.

Construction of Prototype and Final Design

We constructed a small, low-voltage, five-section prototype to verify the simulation and other calculations used in the design. The prototype behaved almost exactly as simulated. We then finalized the design and started construction of the modulator.

The completed PFN **(Fig. 2)** did not match the simulations as well as the prototype did, and we had to tune it by adjusting the mutual couplings

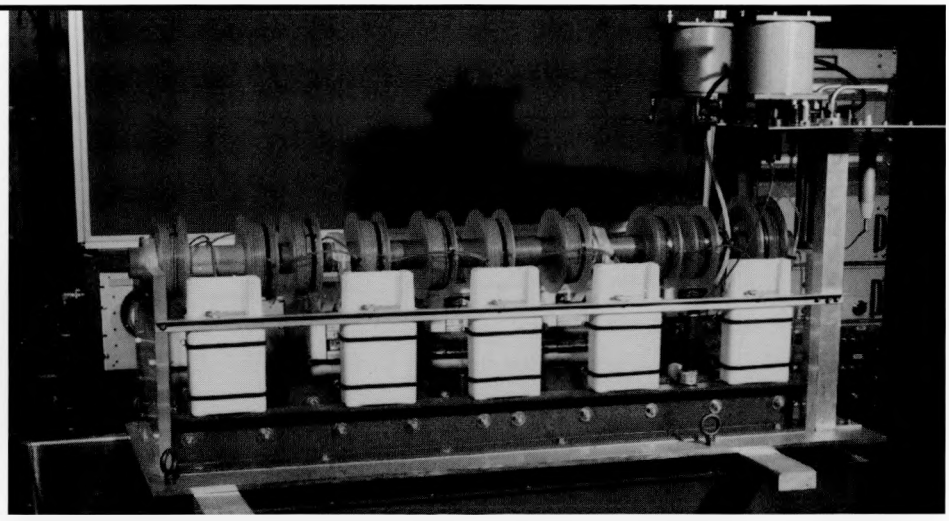


Figure 2. The completed PFN waiting to be installed in the system.

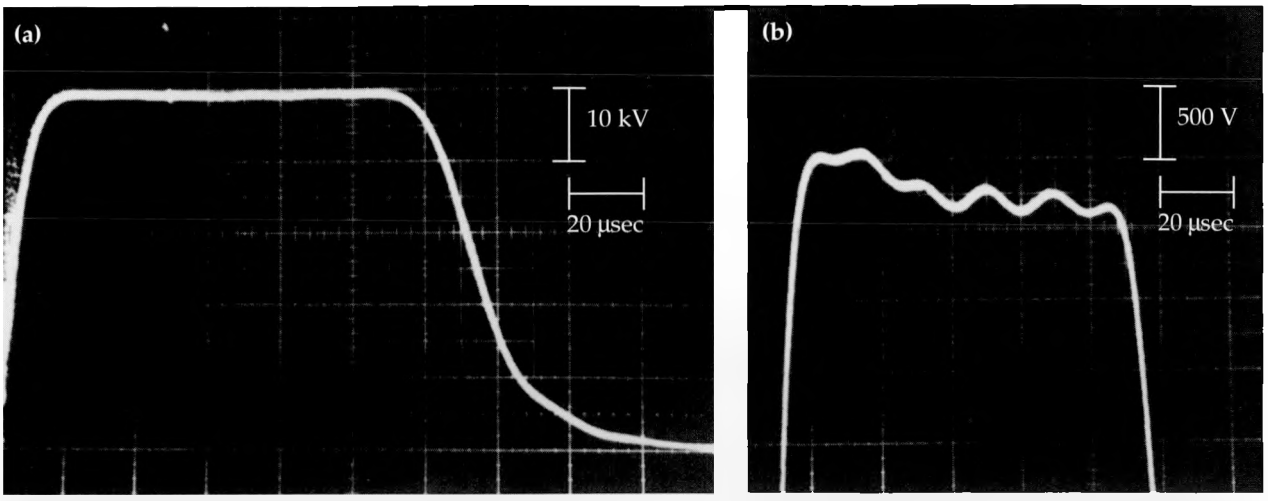


Figure 3. (a) Output pulse from PFN, 100-kV charge, 50 kV into a matched 150-Ω load. The maximum flat-top ripple (±0.5%) over the operating range of the PFN occurs at this voltage (the trace is inverted). (b) Same pulse magnified vertically twenty times (1% of full output voltage per division). The flat-top ripple amplitude is ±0.5% (the trace is inverted).

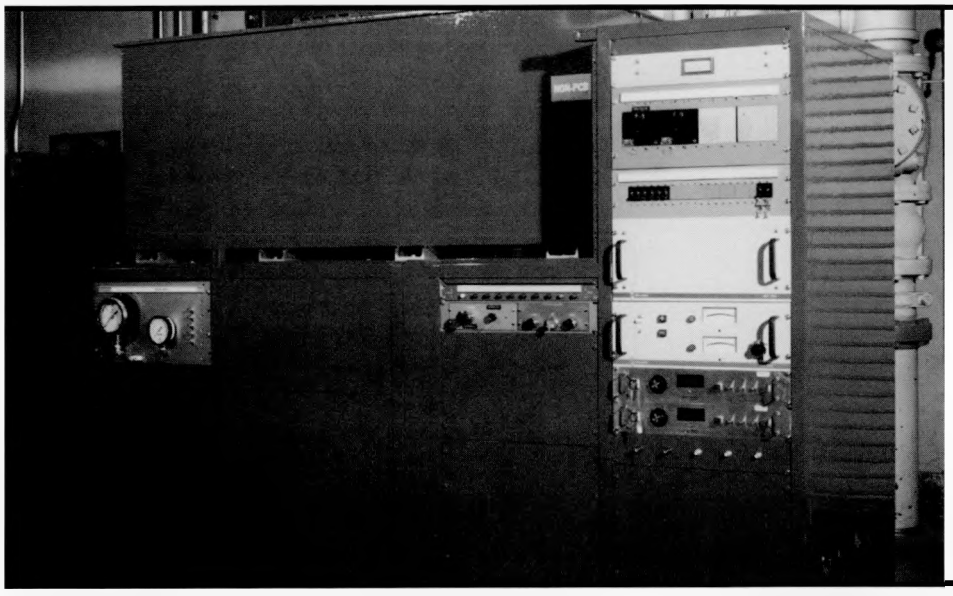


Figure 4. Completed modulator system. It includes interlocks, a triggering system, a recirculating copper-sulphate system for the ballast resistor, and heat exchangers for same. This modulator was designed and built in FY 89 to fill a need for an electron-gun driver.

between interstage inductors. Voltage-dependent effects in the capacitors complicated tuning and caused the amplitude of the flat-top ripple to shift with changes in the operating voltage. To minimize the ripple over the operating range, we tuned the PFN to 40 kV. The maximum flat-top ripple over this range occurs at 50-kV output **[Fig. 3(a) and (b)]**. Even this amount of ripple does not exceed ±0.5%, or one-half of the maximum ripple specified; thus, it is well within design guidelines.

The delivered system **(Fig. 4)** is a completely self-contained turnkey system. It includes interlocks, a triggering system, and a recirculating copper-sulphate system and heat exchangers for the ballast resistor.

High-Power Backward-Wave Oscillator

**J. Frank Camacho, Brian R. Poole,
E. Tom Rosenbury, and Mark Rhodes**
Nuclear Energy Systems Division
Electronics Engineering

We are identifying and developing efficient sources of high-power microwave (HPM) radiation that are driven by a relativistic electron beam (REB). In FY 89, we began to investigate backward-wave oscillators (BWOs), with two immediate purposes: to establish a backward-wave oscillator program at LLNL and to conduct an experiment based on a Soviet BWO design to evaluate Soviet capabilities in this technology. In FY 90, we will expand our efforts to identify and overcome some of the factors that presently limit the capabilities of BWOs. Our ultimate goal is to develop a high-power microwave source capable of efficiently generating on the order of 1 kJ of energy per pulse in the cm-wave regime. Such a microwave generator has possible applications in the areas of HPM directed-energy weapons; HPM lethality and vulnerability studies; advanced particle accelerator concepts; plasma heating for thermonuclear fusion; ultra-high-power radar, communications, and jamming; and electromagnetically pumped free-electron lasers.

Introduction

Our previous work in high-power microwave generation focused on a beam-plasma interaction experiment.[1] However, as our research and consultations with knowledgeable individuals in the field revealed that backward-wave oscillators appeared to be the most promising candidates for long-pulse HPM sources, we redirected our efforts and established a BWO research program.

Backward-wave oscillators belong to a class of microwave devices that use a slow-wave structure, so named because the phase velocity of the electromagnetic waves that propagate in the structure is slower than the speed of light. A typical BWO consists of a waveguide that has walls with periodic corrugations (e.g., sinusoidal ripples). An annular REB is injected into it, guided by an axial magnetic field. **Figure 1** depicts this configuration. Microwaves are generated via an instability in which the slow space-charge wave on the beam couples resonantly to the electromagnetic waveguide modes of the structure. The device is an oscillator because the amplitude of the microwaves that arise out of noise in the system grows in time due to the instability. (This amplitude does not grow indefinitely; eventually, the wave intensity does reach some saturation level.) The energy produced by a BWO propagates in the direction opposite to that of the beam's propagation (hence the name "backward-wave" oscillator). This intrinsic feature of the device is crucial, as the negative group velocity of the resonant structure modes is the mechanism by which energy is transmitted from the downstream end of the device back to the beam inlet, thereby providing the feedback necessary to create an oscillator. Provision is also made in the structure for reflecting the microwave energy at the point where the REB is injected so that the waves are emitted from the output end of the BWO. In addition to feedback, a threshold value on the electron beam current must be exceeded in a finite-length device in order to provide the gain required to make the system break into spontaneous oscillations. This current—referred to as the "start current"—is a function of the device and beam parameters.

Backward-wave oscillators have been proven by both U.S. and Soviet researchers to be efficient sources of high-power microwaves. (In addition, BWOs offer potential advantages over competing HPM generators: the REB required can be produced relatively easily with standard pulsed-power technology; microwave extraction from BWOs is fairly simple; and BWOs can be designed as compact and robust devices.) A comparison of microwave output power and frequency from BWOs with the results from other types of microwave sources in the U.S. and U.S.S.R. is plotted in **Fig. 2.** Perhaps the most impressive data is that from a Soviet experiment[2] in which 15 GW of output power at 9.5 GHz with 50% efficiency was attained by using a two-stage BWO, yielding $P/\lambda^2 = 1.5\ \mathrm{GW/cm^2}$. ($P/\lambda^2$ is a figure of merit for HPM generators obtained by dividing the peak

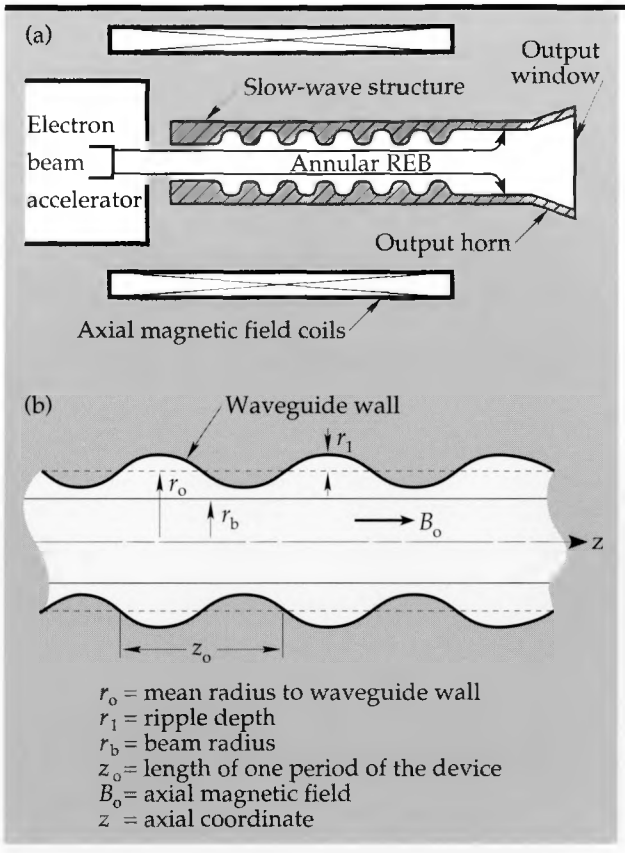


Figure 1. Conceptual diagram of a backward-wave oscillator (BWO). (a) An annular relativistic electron beam (REB) is injected into a corrugated-walled waveguide (a "slow-wave" structure). The structure modes couple resonantly with the slow space-charge wave on the beam. The resulting instability leads to the growth in time of the electromagnetic waves initially created when the REB is injected. This instability is fed by the kinetic energy of the electrons in the REB, which is transferred into the electromagnetic field energy of the generated waves. The coils generate an axial magnetic field that guides the REB. (b) Expanded view of the slow-wave structure showing the important dimensions.

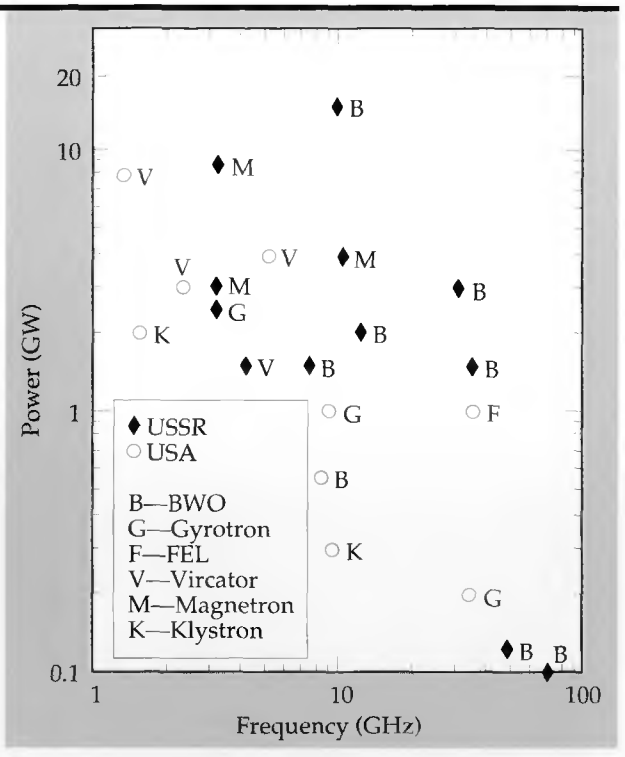


Figure 2. Plot comparing results of various American and Soviet HPM sources. The 15-GW, 9.5-GHz result was achieved by Soviet experimenters using a two-stage BWO. The higher frequency data points are significant, as they demonstrate that BWOs can generate considerable power beyond 10 GHz.

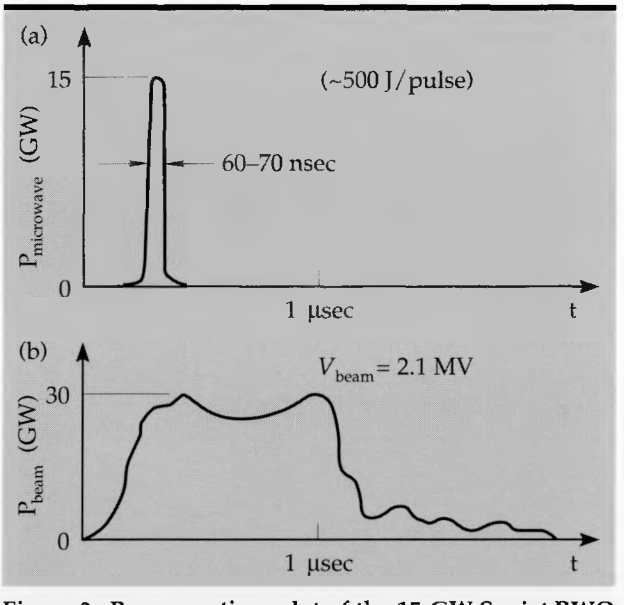


Figure 3. Power-vs-time plot of the 15-GW Soviet BWO result showing the premature termination of the HPM pulse (a) relative to the REB pulse (b). At peak power, the microwaves were generated quite efficiently (50%), but their pulse length is much shorter than the 1-μsec REB pulse, thus limiting the total amount of microwave energy in the HPM pulse to about 500 J. The reasons for the short duration of the microwaves are not known.

microwave power by the wavelength squared; for a fixed size of antenna, neglecting atmospheric absorption, the power density that can be delivered to a target is proportional to P/λ^2.) In the U.S., a plasma-filled BWO at the University of Maryland[3] has generated about 400 MW of power at 8.4 GHz with 40% efficiency, yielding $P/\lambda^2 = 31.4$ MW/cm^2. In these and other BWO experiments, however, the total duration of the microwave pulse has been limited to less than 100 nsec (with total microwave energy produced being at most 500 J), whereas the duration of the REB pulse is usually longer (as much as 1 μsec in some Soviet experiments; see **Fig. 3**). This premature termination of the microwave pulse is characteristic of BWOs and other HPM sources, but the reasons for it are not at all

clear. We are thus primarily researching the question of why the microwave pulse terminates before the REB pulse. An answer to this question may lead to the development of BWO/HPM sources producing considerably longer microwave pulses that can deliver more total microwave energy (~ 1 kJ or more) than has yet been attained in HPM devices.

Progress

As much of the most impressive high-power BWO experimental work has been conducted in and reported by the Soviet Union, there is substantial interest in duplicating some of that nation's work to check their reported results. Therefore, having obtained a slow-wave structure which was constructed at Sandia National Laboratory at Albuquerque, we designed a BWO experiment closely resembling a Soviet experiment[4] for which about 1 GW of microwaves at about 9.4 GHz was reported. The immediate goal is to obtain a set of experimental results to use as a basis for understanding the Soviet data and assessing their BWO capabilities. A drawing of the proposed experimental design is shown in **Fig. 4**.

An important aim in our experiment is to have sufficient variability in its design to allow an adequate scan of all the important parameters. With such flexibility, we can both assess the Soviet results with more reliability and expand our understanding of BWO operation in general. To this end, we plan to conduct experiments in which the axial magnetic field, the beam current, and the beam radius will all be varied over a suitable range of values (0.6 to 3.0 Tesla, 1 to 7 kA at 1 MeV for about 60 nsec, and 7, 8, and 9 mm, respectively). We estimate that most of the energy will be in the TM_{01} mode of a circular waveguide at about 8.5 GHz, with many megawatts of output power (possibly as much as 1 GW). A pulse spectrometer (filter bank with diodes) and a dispersive delay line will be used to measure the power level and frequency bandwidth of the output signal, and various other diagnostics will be employed to measure several beam parameters. We may also use other diagnostics that can yield mode structure information. All these collected data will be analyzed as a function of the variable experimental parameters mentioned above, with the results serving as the basis for the comparison of our results with the Soviet results and for our evaluation of their capabilities in BWO technology. This first experiment

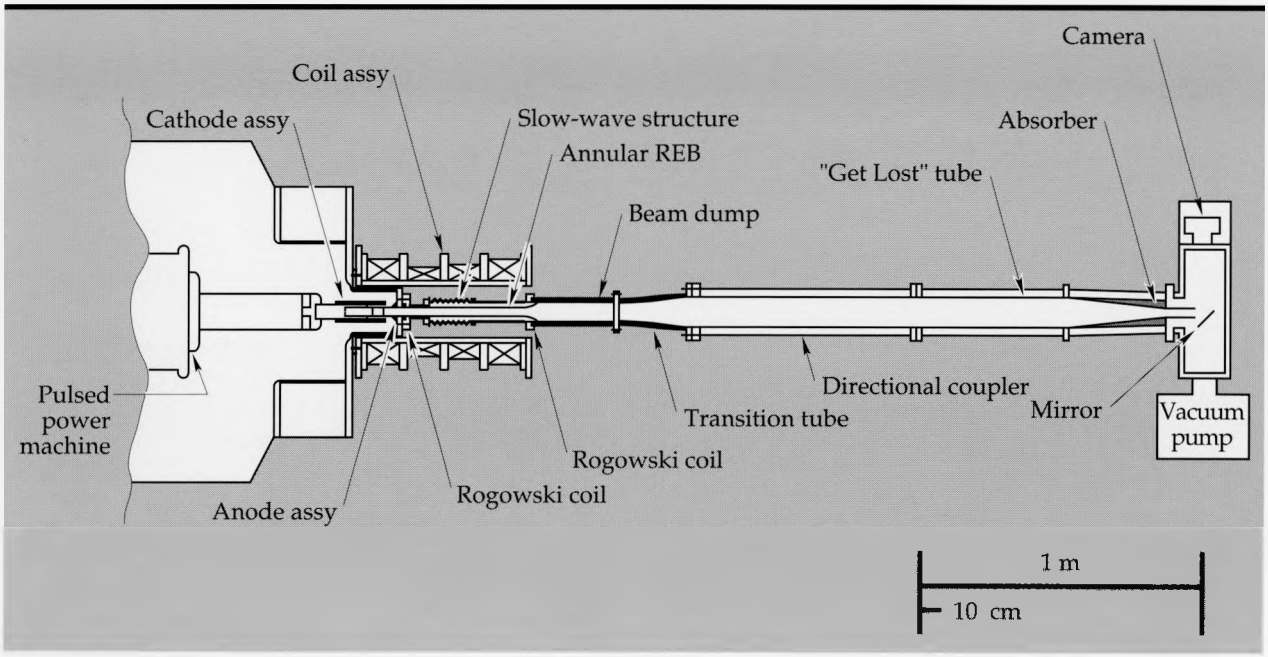


Figure 4. The BWO experimental design. A 1-MeV electron beam will be injected into the slow-wave structure, where the interaction that gives rise to the HPM occurs. These microwaves will then propagate down a circular waveguide and through a beam dump section consisting of a graphite-coated wall, at which point the trajectories of the beam electrons terminate. The microwaves continue through a transition region and into a directional coupler section. From there we will send an attenuated signal to a microwave spectrometer and dispersive delay line (not shown) to obtain power and bandwidth information. A "get lost" tube and absorbing section will then serve as a termination for the microwaves. At this end of the apparatus, we may also employ a camera or other type of diagnostic to obtain further data, such as mode pattern information and microwave power.

will lay a strong foundation for our future work with BWOs.

Along with this experimental design effort, we have also done extensive theoretical and computational work. We have, for example, surveyed the literature and studied linear BWO theory. Using this theory, we can predict basic quantities such as operating frequency, growth rate of the microwave energy, mode structure of the microwaves, and start currents for a given BWO device and REB. We have also written a number of computer codes to work on various BWO-related problems, such as obtaining dispersion relations, solving for the modal structure of a periodic waveguide, and computing start currents. An example of the result from a dispersion relation calculation is given in **Fig. 5**.

We have also used the particle-in-cell (PIC) code CONDOR to model the annular diode behavior in a finite magnetic field and to verify the design criteria for the diode. In addition, the electron trajectories in the beam dump have been computed, yielding information used to establish design criteria for the beam dump section. Furthermore, CONDOR will be used to model the composite BWO and beam dump system in an effort to obtain an estimate for the microwave power that will be generated and to examine nonlinear effects that ultimately determine the efficiency of the BWO.

Through these and other calculations, we have established a sound basis for our experiment. Our theoretical and modeling work has given us a good understanding of basic BWO physics, and the computational tools that we have developed are useful both for the design phase of the experiment and for subsequent analysis of the results.

Future Work

After conducting the above-described experiment and the analysis of its data, we plan to design another experiment to identify those factors limiting the total energy that can be delivered in a single BWO pulse. Such work may enable us to devise techniques and designs that can defeat this limitation, thus increasing the prospects of building a source that is not only able to produce HPM (i.e., several gigawatts) radiation but also capable of delivering a considerable amount of energy (~ 1 kJ or more) in a given pulse.

We plan to continue our focus on such basic BWO physics issues as the dispersion relation, start currents (in both the low- and high-current regimes), and finite magnetic field effects (both in terms of macroscopic beam stability and the effects of cyclotron waves). Saturation of the microwave output (which determines the ultimate efficiency of the device) and other nonlinear phenomena will also be investigated. The most important questions, of course, concern the prolongation of the microwave pulse for the duration of the REB pulse. Intense microwave fields are generated within the BWO, and their influence on the space-charge limiting current of the beam may be a limiting factor in the duration of the microwave pulse. Other mechanisms that might terminate the microwave pulse prematurely are beam filamentation, electric breakdown, field emission, and beam energy deposition to the waveguide walls (which could create a plasma that cuts off the microwaves). We plan to conduct a systematic study to determine which, if any, of these effects is responsible for the early shutoff of the microwaves.

We will also pursue the selection of a suitable pulsed-power machine and the design of a diode that will enable us to investigate the pulse length issues thoroughly. In attempting to achieve higher microwave energy per pulse, we may consider extending our REB pulses to at least several hundred nanoseconds.

We will continue to make extensive use of computer simulations to confirm our analytical predictions and to help us understand the various phys-

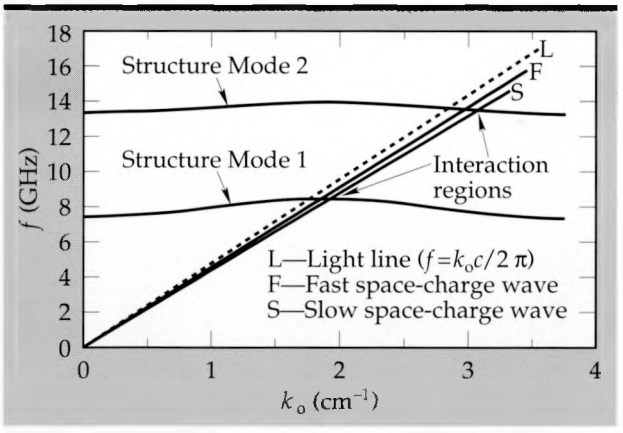


Figure 5. The uncoupled (i.e., beam and structure modes are treated independently) dispersion relation [frequency (f) vs axial wave number (k_0)] for a slow-wave structure with mean radius r_0 = 1.5 cm and ripple depth r_1 = 0.4 cm. The lines intersecting the structure modes are the light line (L), the fast space-charge wave (F), and the slow space-charge wave (S). (These latter two are modes on the electron beam.) The beam parameters are as follows: beam energy V_b = 1 MeV, beam current I_b = 1 kA, and beam radius r_b = 8 mm. The BWO interaction occurs in the vicinity of the intersection of the slow space-charge wave of the beam and the structure modes. We are particularly interested in the intersection of the slow space-charge wave and the first structure mode, as that should be the dominant one.

ics and design issues. We will also consider possible variations of the basic BWO by examining the theory and simulation of a two-stage device and/or a plasma-filled device, as these two approaches have yielded impressive results.

Our research in BWO technology will advance LLNL's technological capabilities in the area of HPM sources and help to establish a national laboratory effort in these devices. Our overall purpose is to complement and expand upon current university-based research and to verify and understand the results that the U.S.S.R. has reportedly achieved. This work will enable us to assess the prospects for developing the technology into an HPM source that is capable of delivering large amounts of energy and that can be utilized in defense and energy applications.

1. J. F. Camacho, B. R. Poole, D. J. Meeker, and B. Chang, "High-Power Microwave Generation Research," *Thrust Area Report FY 88*, Lawrence Livermore National Laboratory, Livermore, CA, UCRL-53868-88 (1988), pp. 6–7 to 6–16.
2. S. P. Bugayev, V. I. Kanavets, A. I. Klimov, V. I. Koshelev, A. I. Slepkov, and V. A. Cherepenin, "Interaction of an Electron Beam and Electromagnetic Field in a 10^{10}-W Multi-Wave Cherenkov Generator," *Sov. J. Comm. Tech. Electr.* **32**, 79–87 (1987).
3. Y. Carmel, K. Minami, R. A. Kehs, W. W. Destler, V. L. Granatstein, D. Abe, and W. L. Lou, "Demonstration of Efficiency Enhancement in a High-Power Backward-Wave Oscillator by Plasma Injection," *Phys. Rev. Lett.* **62**, 2389–2392 (1989).
4. Yu. F. Bondar, S. I. Zavorotnyi, A. L. Ipatov, N. I. Karbushev, N. F. Kovalev, O. T. Losa, G. P. Mkheidze, A. A. Ovchinnikov, A. A. Rukhadze, and L. E. Tsopp, "Generation of High-Frequency Radiation in a Carcinotron with a Relativistic Beam," *Sov. J. Plasma Phys.* **9**, 223–226 (1983).

The Development of Mega-ampere, Kilocoulomb Commercial Ignitrons

Ron Kihara
Nuclear Energy Systems Division
Electronics Engineering

Our overall goal is to develop and qualify devices capable of repetitively switching high peak currents (>1 MA) and large quantities of charge (>500 C). In FY 89, we transferred our hollow-anode ignitron technology to industry and developed a commercial tube using this technology. The new tube, designated an NL-9000, was tested up to levels of 770 kA and 270 C per shot. A life test was also completed at levels of 500 kA and 200 C per shot; the tube was able to survive 728 shots before failing. At this level, standard commercially available tubes fail after 1–5 shots. We are working on modifications to extend the lifetime of these tubes by an order of magnitude. We are also gathering lifetime statistics on a sample of tubes large enough to enable us to make reliable mean time before failure (MTBF) predictions.

Introduction

The requirements imposed by rapidly evolving, advanced technologies have greatly surpassed the capabilities of present-day repetitive switching devices. Mass drivers and advanced laser inertial confinement fusion (ICF) systems now require energies of up to 100 MJ, whereas the requirements of some systems, such as the proposed Microfusion Test Facility laser, will approach 1 GJ. To keep the size of such systems within reasonable bounds, small capacitor packages of high energy density (50 kJ, with prototypes exhibiting densities to 250 kJ) and peak current capabilities of 750 kA per capacitor have been developed. However, development of switches capable of transferring high currents and large amounts of charge has not progressed with equal vigor, and the switching technology of the early 1960s cannot adequately meet the needs of present-day systems. Without doubt, the size and cost of switches impose increasingly severe limitations on new systems. A switch capable of simultaneously conducting 1 MA and 1000 C per shot will satisfy these goals.

The proposed Microfusion Test Facility laser, which this work supports, points out the limitations that present-day technology places on system size. LLNL studies show that cost effectiveness and reliability depend heavily on the availability of suitable switches. The switching system in the Nova laser facility uses 125 sets of paired ignitrons connected in series (250 total). Each set nominally switches a peak current of 100 kA and 50 C per shot. The Microfusion Test Facility would require 1250 sets of ignitrons (2500) at the present level of technology. A ten-fold increase in auxiliary systems (for example, monitors, triggering, and heat management) would also be required. However, a system built with 125 switches capable of handling 1 MA each could use the present Nova support systems with no increase in size.

A collaborative research effort is under way between LLNL, the manufacturers [Richardson Electronics and English Electric Valve (EEV)], and Texas Tech. University to design and manufacture ignitrons with the required capabilities.

Progress

Overview

Increased capacitive energy storage densities have steadily reduced the size of pulsed-power systems. Capacitors were once the largest system components, but in many present and projected applications, the volume of capacitors is equal to or smaller than the volume of switching. **Figure 1** plots the capacitance-switching volume ratio against energy density. At densities of 50 kJ and greater, further reductions in system size will depend increasingly upon a parallel reduction in switch volume, since switching now occupies more than 50% of the total volume.

The peak current capabilities of capacitors have long exceeded the capabilities of a single switch. In **Fig. 2,** we plot the number of ignitrons required to switch a capacitor at maximum current for a series of capacitors of increasing energy density. At the 50-kJ level, the cost of switching greatly exceeds the cost of energy storage.

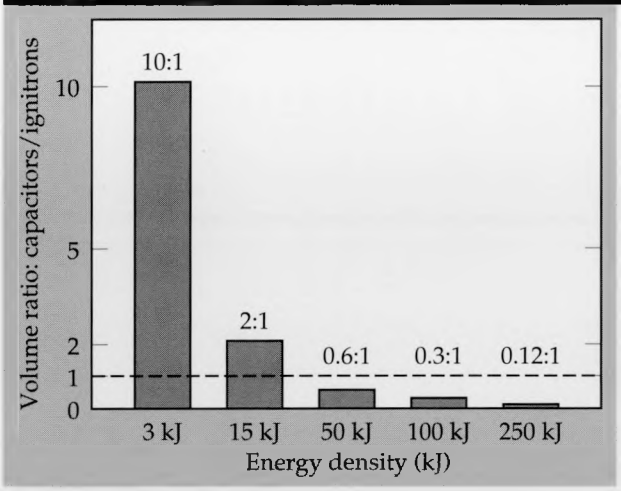


Figure 1. Volume ratio of density to switching (ignitrons for a 1-MJ, 1-MA capacitor bank, normalized to SCYLLAC capacitor volume and an installed Nova switch assembly. The 100- and 250-kJ packages exist as prototypes.

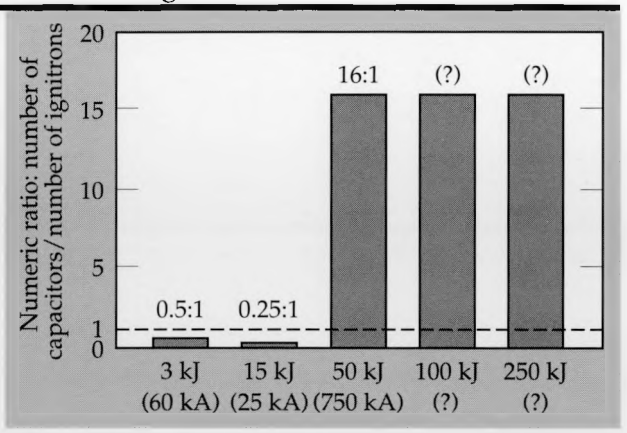


Figure 2. Number of series-parallel ignitrons required to switch a single capacitor at the maximum current capability of these capacitors. Question marks denote proprietory information.

Table 1. Required switch characteristics.

Repetitive operation: < 0.25 ppm
Minimum Lifetime: 500 shots
Simultaneous requirements:
Peak current: > 1 MA
Charge transfer: > 1,000 C
Pulse width: 1.0 ms
Voltage: 20 kV (@10 MJ)
Initial di/dt > 15 kA/μs

Our goal, then, is to develop a high-current, high-charge-transfer switch, which will be able to compete in volume and cost with modern capacitors. This project began in FY 86 with the development of a high-current test facility. In FY 86 and FY 87, we quantified performance levels and identified failure modes of commercially available size "D" and "E" ignitrons. Also, using data from previous tests and cooperating with the manufacturers, we produced an improved tube of conventional design that showed a 40% gain in performance. In late FY 87, we developed a new tube concept, the demountable hollow-anode (HA) ignitron, and in cooperation with industry in FY 88, we developed a commercial tube incorporating an HA structure.[1,2,3]

Development Goals

The performance gap between switches and state-of-the-art capacitors can be closed by improving the Nova switch by one order of magnitude in peak current capacity and by a factor of 20 in conducted charge. Thus, each switch must be able to conduct more than 2,000,000 C over its lifetime (>2000 shots). **Table 1** is a complete list of desired switch parameters. No switch that we know of can meet these requirements.

The switch we chose to develop, the ignitron (pronounced like the word "ignite"[4]), was the only device among those initially surveyed that could potentially meet these requirements. Commercially available ignitrons have been operated at one-third of the desired peak current and over 100% of the required charge transfer (although not simultaneously). The ignitron had the additional attractions of low cost, simplicity, and an existing manufacturing base.

A cross section of a standard size "E" ignitron is shown in **Fig. 3**. The size "E" tube package is typically 23 cm in diameter by 56 cm long and weighs 23 to 32 kg. The tube body is usually stainless steel and is equipped with an integral water jacket for cooling. Graphite anodes are normally found in the larger tubes ("D" size and larger). The tube body and the mercury pool form the cathode. In normal operation, conduction occurs preferentially between the anode and the mercury pool, which is the active part of the cathode. Almost no current should flow to the stainless-steel walls. The water jacket cools and recondenses mercury vaporized by the discharge and absorbs heat radiated by the anode.

Experimental Work

In FY 89, we concentrated on developing commercial versions of the HA ignitron. It was designed to solve several problems occurring in standard ignitrons, namely arc instability at high currents, wall attachment of the discharge, large arc drop voltages (which are undesirable in certain applications and which cause increased anode dissi-

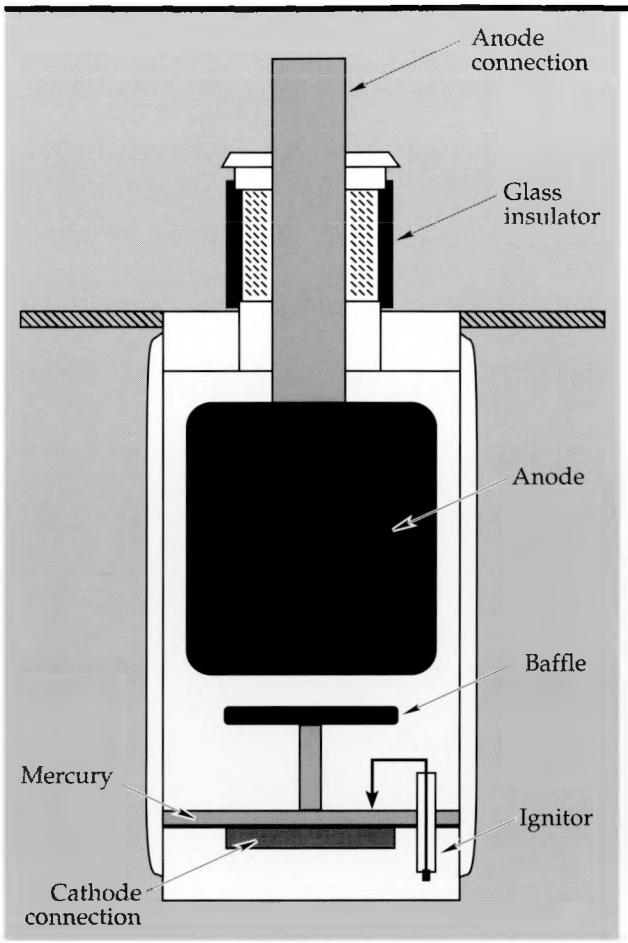


Figure 3. Cross section of a typical ignitron

pation and wear), and ignitor failures caused either by direct exposure to the discharge or by vapor-plating of metal evaporated off the cathode wall. Our laboratory prototype, the demountable HA tube, achieved an operating level of 950 kA and 280 C per shot without encountering these problems. The operational improvement can be seen in **Fig. 4**, in which the characteristic curves of a standard size "E" tube and the prototype HA tube are compared.

Using data from the demountable HA tube and a conceptual design for a sealed tube based on the HA geometry, Richardson Electronics produced a commercially packaged, graphite-anode HA tube, designated the NL-9000 (see **Fig. 5**). EEV is expected to follow suit with an HA tube of its own design.

Three NL-9000s were tested. The first NL-9000 was tested to the maximum level attainable in the bank, 770 kA and 270 C per shot. The tube drop curve, **Fig. 6**, exhibits a resistive slope of 220 μΩ, and a drop of 200 V at 770 kA. Some ignitor deterioration was noticeable at the higher current levels, but resistances were still over 250 Ω at 770 kA. The tube failed when a prefire occurred after eight 770-kA shots, destroying the ignitors. The second tube was tested up to 500 kA, when a broken weld caused it to fail. The third tube was also tested up to 500 kA and survived. We then began life-testing. This tube survived 728 shots at 500 kA and 200 C per shot, in continuous runs of up to 50 shots at a rate of one pulse per minute, before ignitor breakage caused it to fail.

We tested the latter two tubes with a "synthetic" test* circuit set to 10 kV and a recharge time of 35 seconds. The tube recovered completely within this time period, and no prefires were observed except in the third tube near the end of its life. The synthetic test circuit had the unexpected effect of improving the operation of the the NL-9000. This can be seen in **Fig. 7**, in which the tube drop for NL-9000, ser. #3, with and without a synthetic test circuit is plotted. When the synthetic test circuit is on, the characteristic curve is displaced downward by 20–30 V, and the tube drop and di/dt traces are much smoother. As a result, anode dissipation reduces by 20% at the 500-kA level. This amount is significant, but is expected to have only a modest effect on the life-test data. We believe the displacement of the curve is caused by the accelerated development of the discharge, as

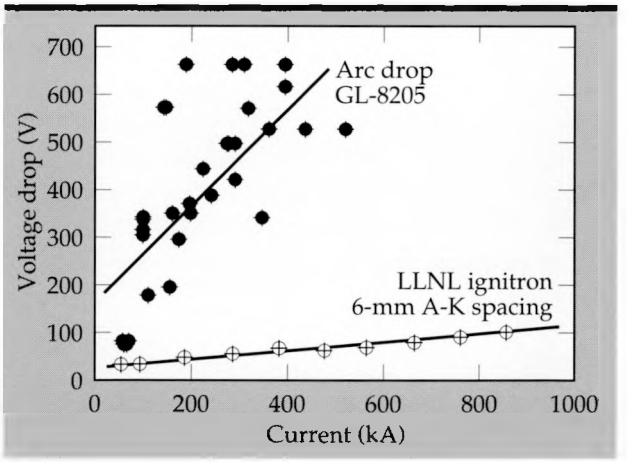


Figure 4. Tube drop curves of a standard size "E" tube, GL-8205, and LLNL ignitron at 6-mm A-K spacing compared. The LLNL ignitron has only a small fraction of the tube drop of the GL-8205 and a much stabler discharge (datum points for the LLNL tube are co-incident, 10 shots per point).

*Synthetic testing utilizes two banks, one to supply high voltage, and the other high current; the two banks are discharged simultaneously. The high-voltage bank, which is connected directly to the tube, tests recovery and hold-off. The high-current bank is isolated through a series switch. When the tube is triggered, both banks are discharged through the tube, simulating the effect of a single, large, high-voltage high-current bank.

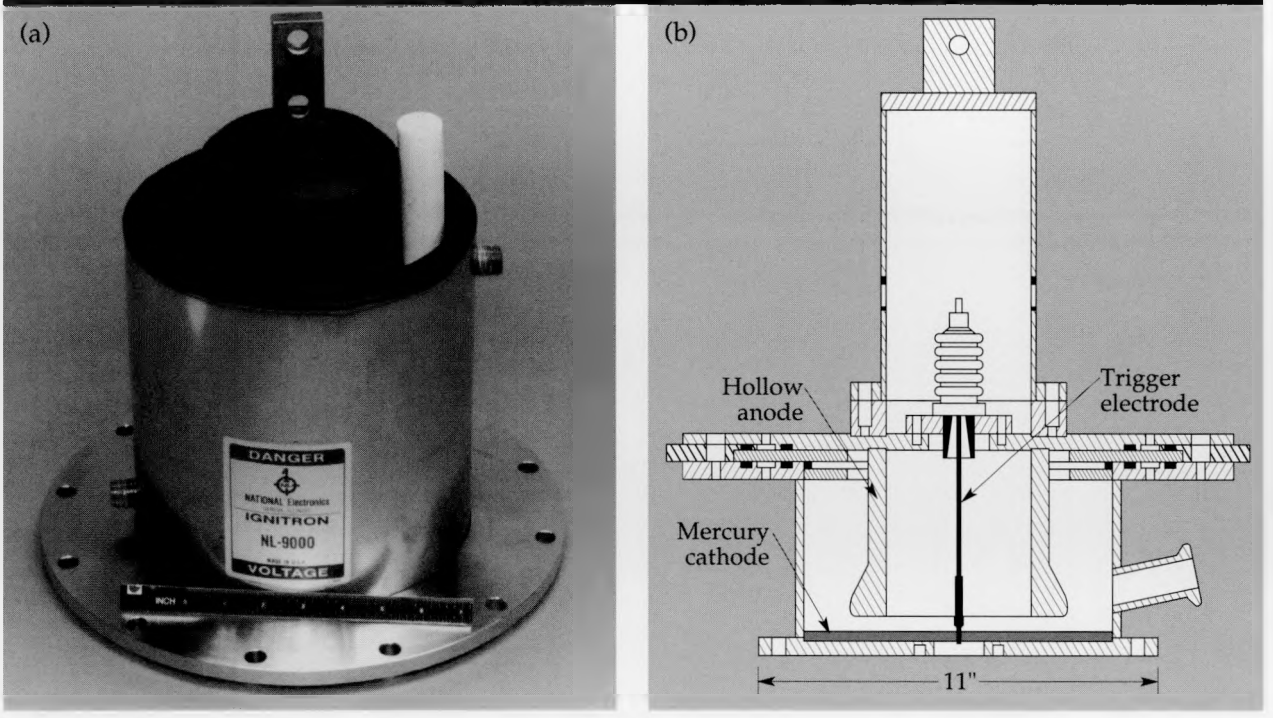


Figure 5. (a) Photograph of the NL-9000 ignitron. (b) Cross section of the Richardson Electronics NL-9000 hollow-anode ignitron.

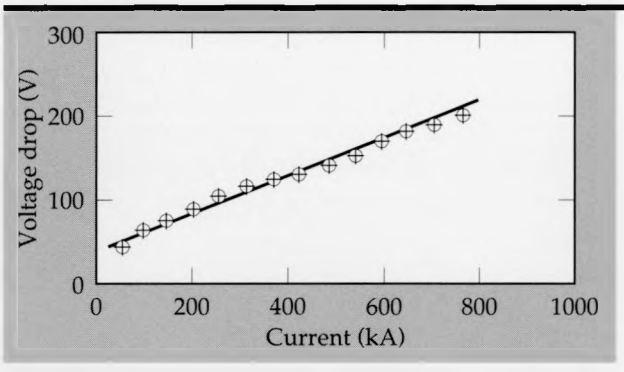


Figure 6. Tube drop plotted as a function of current for the NL-9000 ser. #1 ignitron, without synthetic test.

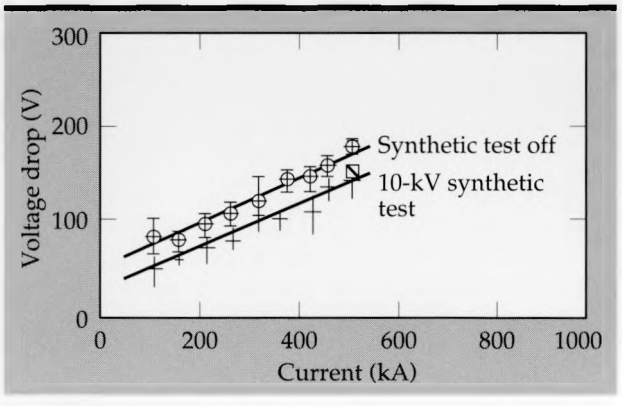


Figure 7. Tube drop plotted as a function of current for the NL-9000 ser. #3 ignitron, with and without synthetic test. The shaded bar superimposed on the synthetic test curve represents the 2-σ deviation of data over the life test of the tube (728 shots).

previously described by Cummings.[5]

Postmortem results on the NL-9000 indicate that 500 kA is the maximum current level that this tube can support and have a reasonably long life. At this level, the power dissipation at the surface of the anode is sufficient to cause graphite sublimation off of the anode. The tube is usable at higher current levels, but its operating life is greatly reduced. At levels approaching 770 kA, the anode begins to burn off of its connection, and its expected lifetime is in the tens of shots. To lower the dissipation on the face and connection of the anode at current levels higher than 500 kA will require larger anode structures.

Future Work

At 500 kA and 200 C per shot, the NL-9000 has reached the minimum level considered usable for the Microfusion Test Facility laser. However, more work needs to be done on the ignitor—still the weakest part of the tube—to reach the desired 2,000+ shot lifetime. Development of high-current and high-charge-transfer tubes will proceed in parallel with work on raising the NL-9000's reliability.

Our strategy for FY 90 will concentrate on developing a database for MTBF prediction in large systems by life-testing a statistically significant number of NL-9000 tubes. We will also continue to improve the NL-9000 series in parallel with life-

testing. We may construct a semi-demountable tube to aid in developing stronger and more reliable ignitor structures.

This year, we will also test a variant of the NL-9000, designated the NL-9000a, which has a hollow stainless-steel anode. Other tubes projected within this series may have molybdenum anodes and spark-gap ignitors.

Design work on larger [size "F"*] tubes, possibly designated as NL-9001s, will begin this year.

We believe that a 1,000-kA, 500-C per-shot capability will be achieved in a commercial tube within two years.

1. R. Kihara, *Evaluation of Commercially Available Ignitrons as High-Current, High-Coulomb Transfer Switches*, UCRL-96203, Lawrence Livermore National Laboratory, June 19, 1987; IEEE 6th Pulsed-Power Conference, Arlington, VA, June 29, 1987.
2. D.B. Cummings, R. Kihara, and K.S. Leighton, *High Current Ignitron Development*, IEEE Conference Record of the 1988 *Eighteenth Power Modulator Symposium* (IEEE Cat. No. 88 ch 2662–5), Hilton Head, SC, June 20–22, 1988 (IEEE, New York, NY, 1988), 122–127.
3. R. Kihara, D.B. Cummings, K.S. Leighton, A.P. Schulski, *Commercial High-Current Ignitron Development*, UCRL-100358 Rev. 1, Lawrence Livermore National Laboratory, (1989).
4. D.D. Knowles, *The Ignitron-A New Controlled Rectifier*, Electronics **(6)** 164–166 (1933).
5. D.B. Cummings, *Ignitron Discharge Growth During High Current Discharges*, IEEE Transactions on Communications and Electronics, **(82)** 512–523, (1963).

*Sizes are standardized by the manufacturer; the largest size currently available is size "E". We assume that the next larger size will be designated "F", but this is not an official size.

Magnetic Switch Modeling

Jan-Mark Zentler
Nuclear Energy Systems Division
Electronics Engineering

Our project objectives are twofold: to consolidate a knowledge base on magnetic switching and to develop modeling tools that accurately simulate the behavior of magnetic switches and associated circuits.

To help identify the features that a complete switch model must have, we surveyed how magnetic switches are used at LLNL and elsewhere. Next, we analyzed existing switch models and found that none would satisfy our intended applications, although some included features that could be incorporated into a suitable model. We then identified existing empirical and physics-based magnetic-materials models that could be incorporated into our switch model. We are working on a simplified form of the Chua-Stromsmoe model to simulate the behavior of typical switches. The model is being integrated into the Laboratory's SCEPTRE and NET2 circuit analysis codes. However, the existing circuit codes experience stability problems due to the highly nonlinear nature of magnetic models. We are collaborating with A. N. Payne of Fusion Energy Systems Division to develop a system-level modeling tool designed specifically for simulating circuits with many nonlinear elements.

As a supplementary effort, we are testing 2605CO Metglas ribbon (a common switch core material) to improve our materials-performance data base.

Introduction

A magnetic switch is essentially a saturating inductor (reactor) consisting of a ferro- or ferrimagnetic core around which current-carrying conductors are wrapped [see **Figure1(a)**.] Magnetic switches are especially well suited for applications that simulaneously require very high average power and high repetition pulse rates; this combination of features is not equally available through any other device. Large switches can deliver extremely high power. For example, magnetic switches in the master trigger system for the LLNL flash x-ray (FXR) accelerator can produce 2 GW, and experimental switches have been built elsewhere that can produce output power as high as 1 TW.

The switching action derives from the reactance. When a voltage (say, from a capacitor bank) is suddenly applied to the inductor, the reactance of the coil delays the flow of current for a brief moment, or "hold-off time." Then, upon core saturation, the reactor allows current to flow because the inductance drops to a very low value. A unique property of the magnetic switch is that it is not a "command switch," that is, it is never used as a voltage or current trigger, or in dc applications.

Using their high-power, high-repetition capabilities, two common and concurrent applications of magnetic switches are pulse delay and pulse compression. For pulse compression, multiple magnetic switches are connected in series to progressively narrow the width of a high energy pulse and thus to increase its power to very high values.

A more specific description of the operation of a magnetic switch can be given with the aid of **Fig. 1(b)**, a plot of magnetic flux density, B, within the core material as a function of magnetic flux intensity, H. (Such a plot is often called a "*B-H* characteristic curve.") The solid line represents the magnetization state of the inductor core during switching. The slope of the line is the inductance at any point in time.

An optimal switch would have low loss (narrow *B-H* loop), high "off-state" inductance, and low "on-state" inductance (square *B-H* loop). Additionally, a narrow *B-H* loop minimizes the "off-state" current leaking through the switch.

The volt-second product of the core determines the time to saturation, and hence the amount of time for which the reactor can hold off voltage. The precise hold-off time is a function of core material, core geometry, and number of turns, among other factors. After the current pulse is delivered, the core remains fully magnetized (*positive* remanence state), and therefore, a reset pulse of opposite polarity must be applied to the reactor to reverse the magnetization polarity (to *negative* remanence state).

The most widely used magnetic switching network is the series Melville line **Fig. 2(a)**, first proposed by W. S. Melville in 1951.[1,2] The Melville line is a series of resonant circuits, each of which consists of equal-value capacitors bridged by a saturable inductor. Each inductor is designed to saturate (thus passing current to the next stage) at the instant the capacitor preceding it reaches maximum charge. Because the pulse is compressed in each stage [see **Fig. 2(b)**] the Melville line is useful in such applications as free-electron lasers, which require high-power pulses at high repetition rates. The Melville line was little used for many years because of its inherent bulk and because competing spark-gap-switch technologies advanced rapidly. However, interest in the Melville line has been renewed because Strategic Defense Initiative applications demand switches that can operate reliably at high repetition rates. Magnetic switches have advantages in high reliability and repetition-rate capability not found in spark-gap switches.

Electrical pulse compression networks using magnetic switches are continually being developed in the LLNL Beam Research and Laser Isotope Separation programs. In the Beam Research Program, magnetic switches are used extensively in linear accelerators as spark-gap triggers and injector-grid drivers, as well as in power conditioning systems for linear-accelerator modules. Researchers in the Laser Isotope Separation Program use magnetic switches in flashlamp modulators for copper vapor lasers.

Magnetic switch development at LLNL has been almost wholly experimental. Workers optimize switch designs by refining prototypes—an expensive process. Consequently, switch construction is an art understood by only a few professionals. Calculational analyses have usually been limited to studies of simple inductive/capacitive networks.

Our project objectives are twofold: First, to consolidate the knowledge of individuals working on magnetic switching projects, both at LLNL and elsewhere. Second, using this knowledge, to develop modeling tools to accurately simulate the behavior of magnetic switches and associated circuits. With these tools we can expedite the design process, reduce development cost, and provide more refined switching networks. By helping us educate novice users on switch behavior, these tools also broaden the expertise on magnetic switching technology available at LLNL.

We have completed our first objective. Our current work centers on developing an empirical switch model based on the work of Leon Chua and Keith Stromsmoe at Purdue University. We are preparing to test switch core materials to acquire

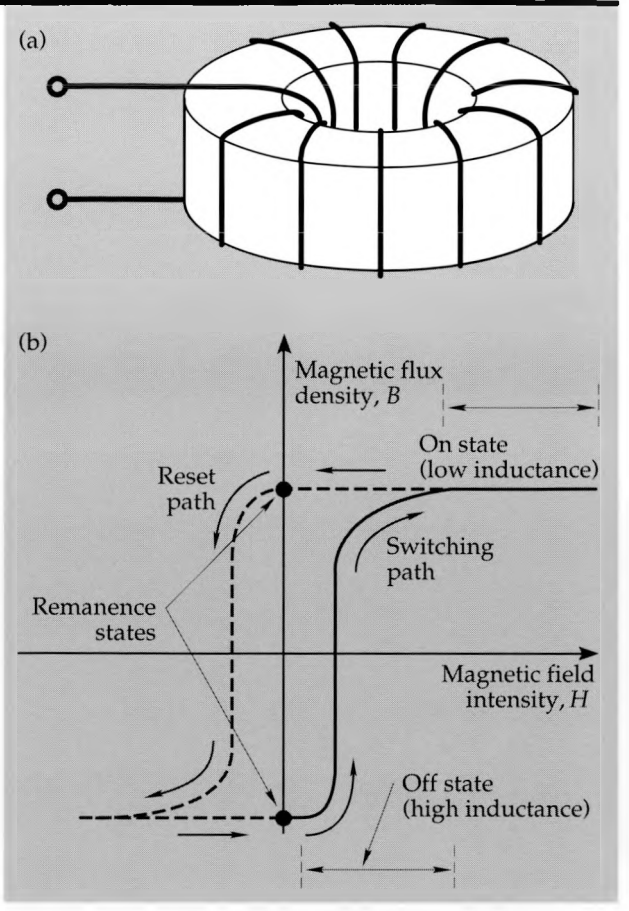


Figure 1. Theory of magnetic switch operation. (a) Simple switches can be constructed by wrapping a wire around a suitable core. (b) Characteristic *B-H* curve for a magnetic switch core. The switch is in the off state when the inductance (represented by the slope of the solid line), hence voltage hold-off capability, is very high. (The hold-off time duration is determined by the volt-second product of the magnetic core.) After each pulse, the core must be reset to negative remanence by applying a reset pulse (dashed line) with a polarity opposite that of the driving pulse.

data for calibrating the model. To facilitate its general use, the model is being integrated into the Laboratory's SCEPTRE and NET2 circuit-analysis codes. Also, we have been collaborating with A. N. Payne of Fusion Energy Systems Division to design a system-level modeling tool for simulating circuits that have many nonlinear elements, like, for example, magnetic switching networks

Progress

To obtain some idea of how switches are used and constructed, we surveyed researchers at agencies such as Sandia National Laboratories, Albuquerque, and Physics International, as well as researchers at LLNL. The results of this survey are

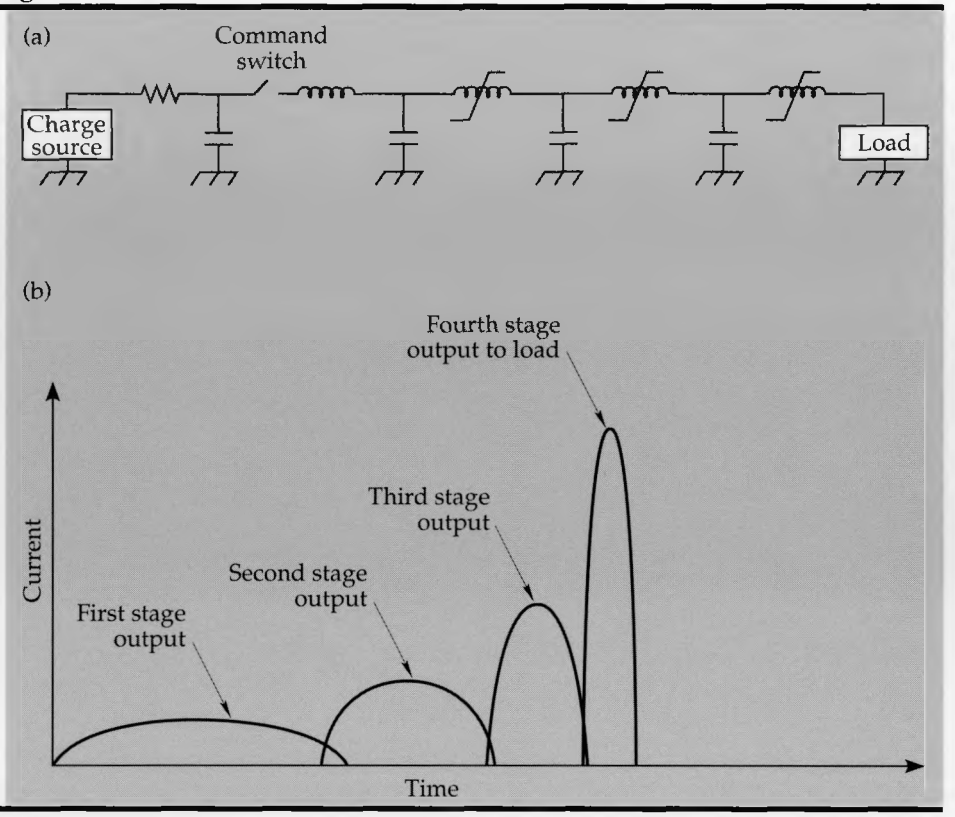


Figure 2. Series pulse compression network, or Melville line. (a) Each stage is a resonant circuit in which a magnetic switch bridges two capacitors (except for the last stage). Each switch is designed to saturate at the instant the capacitor preceding it reaches maximum charge. (b) The electrical pulse is compressed at each stage. Unless transformers are used, voltage remains nearly constant from stage to stage.

discussed in the "Introduction" section, above.

We next examined how existing models represent the behavior of magnetic switches and of the materials used to make them. With this information, we could then determine our approach for improving on existing magnetic switch models. These topics are discussed in the following three sections.

Critique of Existing Magnetic Switch Models

Various investigators have constructed detailed models of magnetic switches. Yet, none of these markedly improve upon the existing (but crude) two-piece inductance models.

Furthermore, the physics assumptions that form the basis of the few computer-based circuit-analysis codes written specifically for magnetic switching applications are insufficient for modeling the types of switch materials and operating conditions that we need to employ. For example, many of the models are based solely on the "sandwich domain" (magnetization) wall model, and investigators have shown that this particular wall-domain effect is dominant only near the end of the magnetization process, or at very high switching speeds. (The theory of *domain wall models* is discussed under a following subsection of the same name.")

We tried to identify commercially available circuit-analysis codes with saturable reactor models applicable to magnetic switching. None was found to include the dynamic effects (exemplified by *B-H* loop widening with frequency) that dominate switch behavior at high switching speeds.

Modeling Magnetic Material Behavior

Since we were unable to identify an existing magnetic switch model that is suitable for our applications, we decided our next step was to fully develop our understanding of magnetic material behavior. Thus, we initiated a study aimed at identifying relevant models of magnetic material behavior. We concentrated on descriptions of the magnetization physics of tape-wound cores, although ferrite models were also included. Sources investigated include F. J. Friedlaender's early research on domain nucleation and growth in nickel-iron alloys, the extensive work done by J. E. L. Bishop on all aspects of domain wall motion, D. Nitzan's ferrite modeling, and work performed by LLNL researchers R. W. Kuenning and S. D. Winter on sandwich-domain motion.

We found two classes of magnetic-material models suitable for use in a switch model: *empirical* and *physics based*. The physics-based models of interest are all of the same subclass known as *"domain wall models."*

Empirical models. Empirical models provide a

mathematical fit to experimentally obtained core performance data. The more useful models employ mathematical functions exhibiting the general form of a generic *B-H* loop. Magnetization physics can be used to provide the appropriate functions. Empirical models often contain some physics-based coding, in which case they are called "semi-empirical models." Fixed *B-H* loop models, a subset of the class of empirical models, are suitable only for modeling cores operating in quasistatic conditions or for crudely approximating single-frequency operation. However, fixed *B-H* models are often used in preliminary switch designs because of their simplicity. When developing an empirical model, one must minimize the number of model parameters that must be fit to data.

Domain wall models. In the magnetic switch core materials of interest (notably, thin amorphous metal tape and ferrite block), the material spontaneously divides itself into many domains, or regions of opposite magnetic alignment. The demarcation lines between adjacent domains are designated as "domain walls." (specifically, "180-deg Bloch walls")[3]. (Magnetic domain structure should not be confused with the grain structure existing in polycrystalline materials.)

Prior to the application of a magnetic field, the switch core material is fully and homogeneously magnetized in one direction (the *negative* remanence state). Upon application of a magnetic field, certain regions of the core material become progressively remagnetized into opposite polarity by the advancement (or "motion") of the domain walls.

Domain wall motion models have been postulated for both tape and ferrite cores. However, in tapes, the domain configurations are of three subclasses, or types, (1) surface domain, (2) bar domain, and (3) sandwich domain; and an examination of domain proliferation and growth over time shows that a comprehensive switch model must include wall-modeling elements for all three types.[4]

Figure 3(a) represents an amorphous metal tape subjected to dB/dt excitation. At the outset, the tape is fully magnetized in a single direction normal to the page. The interior lines represent time-lapse "snapshots" of a single domain wall as it progresses outward away from its source, the surface nucleation. These domains are semicylindrical in shape and grow in size with time. Bar domains are created either when two surface domains from opposite sides of the tape merge, or when a single domain grows to span the tape thickness. As the bar domain walls travel along the width of the tape, they bow in the center due to the influence of eddy currents induced in the material by the moving walls themselves. Magnetic field intensity decreases as wall bowing occurs. Eventually, the bowing becomes so severe that the walls are nearly parallel to the tape surface and become sandwich domains. Core saturation occurs as the process is completed. Under conditions of very high dB/dt excitation, the nucleated surface domains along each tape surface merge to form sandwich domains directly, thus bypassing the bar domain stage.

Figure 3(b) shows the magnetization time history of the tape in the form of a dynamic *B-H* characteristic curve. When the step excitation is ap-

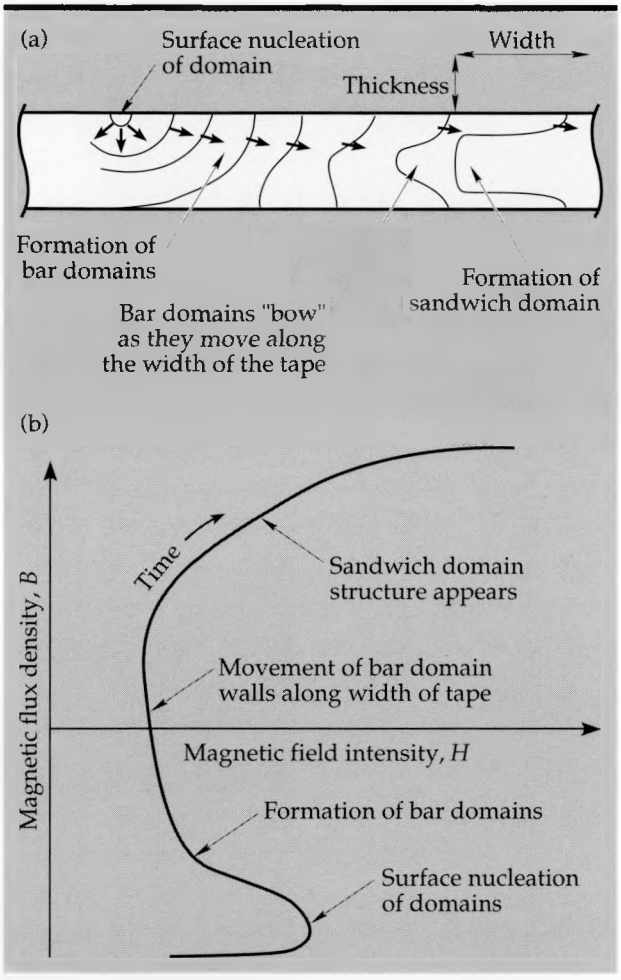


Figure 3. Illustration of domain growth over time. (a) An amorphous metal tape is subjected to step dB/dt excitation. (The view is an end-on cross section with tape width running left and right; thickness/width ratio is greatly exaggerated.) Initially, the tape is fully magnetized in a direction normal to the page, thus parallel to length of tape. (b) Time-line curve showing typical growth in magnetization. All three domain wall types—surface, bar, and sandwich—are present, indicating that a comprehensive switch model must include representations of all three domains.

plied, field intensity, H, rises rapidly, and then decreases as bar domains are created. Then, as the bar domains move along walls of tape, field intensity increases and continues to increase as sandwich domains appear. (The measurement equipment initially causes a finite rise in H. In theory, this rise would be instantaneous.)

Modeling Approach

Our research has indicated that, for all but the crudest simulations, the switch core must be radially zoned to represent the nonuniform magnetization that occurs during the switching process. Within each zone, the modeling of the magnetic behavior of the core material can be based on the average motion of the domain walls within the material. This average motion can be represented by either an empirical or a physics-based model. Ultimately, as empirical models become more sophisticated, they begin to rely on material physics to provide functional forms used in data fitting. Conversely, because of the complexity of the magnetization process, any physics-based model will require that some parameters be fit to data. We feel that the best model will be a combination of empirical and physics-based.

As a starting point in model development, we decided to implement a simplified version of the Chua-Stromsmoe empirical hysteresis model as a basis for a magnetic switch model.[5] The Chua-Stromsmoe model (referred to as the "C-S model") has the advantage of being simple in form, yet general enough to represent the very square *B-H* loop shapes characteristic of magnetic materials used in switching applications. For our purposes, we write the C-S model as:

$$H = f(B) + g\left(\frac{dB}{dt}\right) . \qquad (1)$$

The function f represents what is often called the anhysteretic curve and sets the fundamental shape of the *B-H* loop. We have developed a novel analytic form for the f function that is particularly well suited to modeling the square-loop materials:

$$f(B) = \frac{B}{\mu_s} + \frac{1}{\gamma}\left(\frac{1}{\mu_u} - \frac{1}{\mu_s}\right) \ln\left(\frac{Ke^{\gamma B} + 1}{e^{\gamma B} + K}\right) . \qquad (2)$$

μ_u, μ_s, γ, and K are constants which determine the shape of the anhysteretic curve.

The γ function provides *B-H* loop widening with increasing magnetization rate. We have been using a simple power series relation for the form of γ, but a more sophisticated form based on domain wall physics is being developed.

We integrated our version of the C-S model into the SCEPTRE and NET2 circuit analysis codes used at LLNL. In circuit form, the function f is represented by a nonlinear inductor and γ by a nonlinear resistor, as shown in **Fig. 4(a)**. Each inductor/resistor pair forms the material model for each radial zone of the switch core. The entire core is modeled as a series-connected network of these inductor/resistor pairs.

At present, we have limited our analysis to single-zone models, and have found that the parallel combination of two nonlinear circuit elements can cause numerical instabilities in both SCEPTRE and NET2. However, we were able to successfully simulate some simple circuits; for example, **Fig. 4(b)** shows the core *B-H* characteristic from SCEPTRE simulation of a capacitor being discharged across a magnetic switch. But we are concerned that SCEPTRE and NET2 will not be able to handle the more complex nonlinear circuits that we will need to simulate in the future. For that reason, we have been collaborating with A. N. Payne to develop a system-level modeling tool designed specifically for simulating circuits with many nonlinear elements. The first version of the code will be completed during the first quarter of FY 90.

Magnetic Materials Testing

We have endeavored to test magnetic materials used in magnetic switching cores to provide data for validation of our models and for use by designers of magnetic switches. Most of our testing has focused on Metglas because of its widespread use in the Beam Research Program at LLNL. We did an extensive search of the literature and found that most published data on the material was limited to measuring response to excitations at constant magnetization rate.

We have built a resonant capacitive discharge pulser to test material response to 1-coswt voltage profiles mimicking those experienced in magnetic switching networks. The pulser, shown in **Fig. 5**, is located in the LLNL pulsed power laboratory. It can subject a Metglas test core to average magnetization rates in the 0.1- to 10-T /msec range. A solenoid-driven tack switch is used to trigger the tester because of the wide range of charging voltages (0.1 to 10 kV) necessitated by the range of magnetization rates. Test cores are wound from 2-in.-wide ribbon to a radial thickness of 1/2 in. on a 6-in. mandrel. We are using a larger core size than used by most researchers, since this minimizes inaccuracies introduced by nonuniform magnetization across the core's cross section.

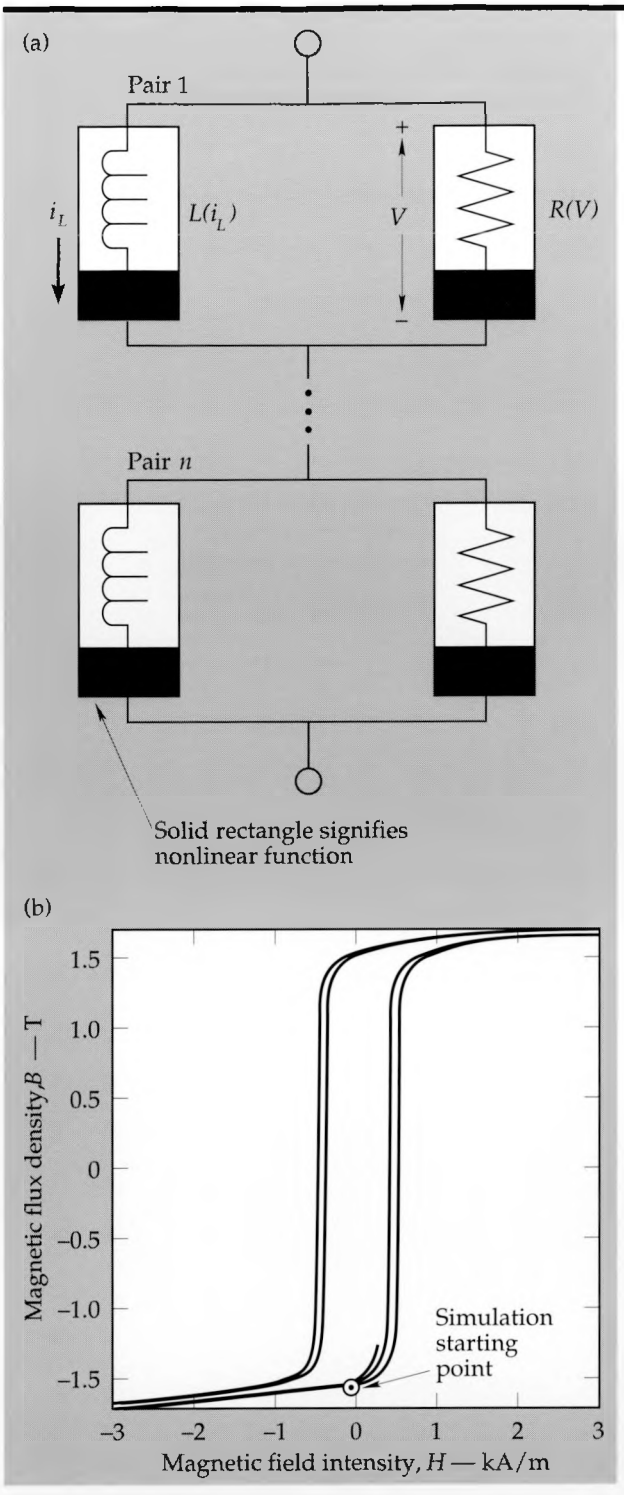


Figure 4. Circuit-block representation for switch modeling based on Chua-Stromsmoe empirical hysteresis model. (a) Each nonlinear inductor/resistor pair forms the material model for each radial zone of the switch core. The entire core is modeled as a series-connected network of these inductor/resistor pairs. (b) *B-H* core-magnetization characteristic from SCEPTRE simulation of a capacitor being discharged across a magnetic switch. As energy from the capacitor is dissipated, magnetization rate of the core material decreases, narrowing the *B-H* loop.

Future Work

We will continue development of the Chua-Stromsmoe-based magnetic switch model in FY 90 with the effort concentrated on producing a formulation more compatible with the SCEPTRE and NET2 circuit analysis codes.

We also plan to implement and evaluate additional switch models based on several other empirical *B-H* models. One of these is a model currently under development by M. L. Hodgdon at Los Alamos National Laboratory.[6] For another, we will construct a physics-based switch model using Bishop's formulas for magnetization from surface-, bar-, and sandwich-domain motion.[7] A simplified domain geometry based on the work of R. K. Avery of British Aerospace will be used to determine which type of domain motion is dominant.

Close collaboration will continue with A. N. Payne on design of a system-level modeling tool that is more adept at handling nonlinear switch models than SCEPTRE and NET2. We want to ensure that the tool includes the necessary features to handle the whole range of magnetic switch models, as well as other unique elements of magnetic switching networks and pulsed-power networks in general.

Tests of 2605CO Metglas cores under 1-cos*wt* voltage excitation will proceed in the LLNL pulsed power laboratory. We plan to modify the tester to measure material response to constant magnetization rates as well. This type of data can be found in the literature, so we will be able to compare our results with the work of others.

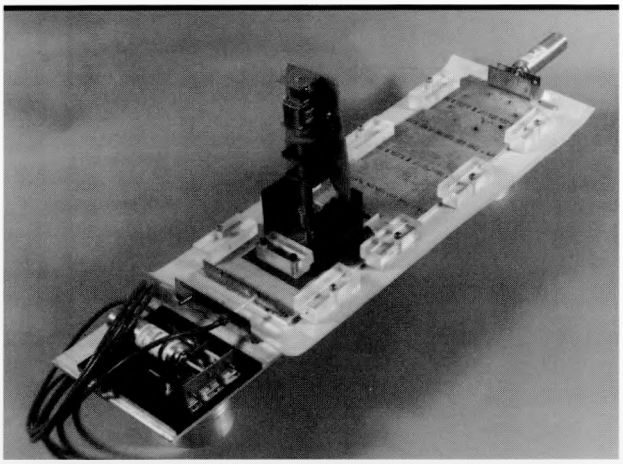


Figure 5. The resonant-capacitor discharge-type tester we employ to evaluate Metglas tape-wound cores at average magnetization rates from 0.1 to 10 *T*/msec. The solenoid-driven tack switch is seen at the center of the fixture. The core undergoing test is located underneath the parallel-plate transmission line.

1. E. Chu, "Design Considerations of Magnetic Switching Modulator," *Proc. 4th IEEE Pulsed Power Conf.*, p. 242 (1983).
2. W. C. Nunnally, *Magnetic Switches and Circuits*, Los Alamos National Laboratory, Los Alamos, New Mexico, LA-8862-MS (revised) (1982).
3. B. D. Cullity, *Introduction to Magnetic Materials* (Addison-Wesley Publishing Company, Reading, Mass., 1972).
4. R. M. Jones, "Step dB/dt Magnetization Losses in Toroidal Amorphous Ribbon and Polycrystalline Cores," *IEEE Trans. Magn.* **MAG-18**(6), 1559–1561 (1982).
5. L. O. Chua and K. A. Stromsmoe, "Lumped Circuit Models for Nonlinear Inductors Exhibiting Hysteresis Loops," *IEEE Trans. on Circuit Theory* **CT-17**(4), 564–574 (1970).
6. M. L. Hodgdon, "Applications of a Theory of Ferromagnetic Hysteresis," *IEEE Trans. Magn.* **MAG-24**(1), 218–221 (1988).
7. J. E. L. Bishop, "Modeling Domain Wall Motion in Soft Magnetic Alloys," *J. Magnetism and Magnetic Materials* **41**, 261 (1984).

High-Voltage, Subnanosecond Photoconductive Switching

Michael D. Pocha
Engineering Research Division
Electronics Engineering

Robert L. Druce and
W. Wayne Hofer
Nuclear Energy Systems Division
Electronics Engineering

Photoconductive switches offer many advantages in applications involving high power, high voltage, and high speed, such as direct generation of high power microwaves, Pockel's cell drivers, and calibration of high-voltage diagnostics. However, routine use of photoconductive switches in pulsed power and microwave generation will require improvements in open state voltage holdoff, device life, and efficiency of laser control. In an attempt to achieve these improvements, we are investigating photoconductive switching in gallium arsenide (GaAs) and indium phosphide (InP) for microwave generation and pulsed power applications. We have achieved holdoff fields in excess of 150 kV/cm in the open state. In the linear mode, we have successfully switched intrinsic GaAs with closing times of less than 300 psec at fields greater than 50 kV/cm and current densities greater than 2.5 kA/cm^2. We have achieved switch power gains (ratio of electrical power switched to required laser energy) of over nine in the linear mode. In the avalanche mode, we have switched fields of greater than 150 kV/cm (voltages to 35 kV) with rise times of 300 psec–2 nsec (depending on the device and charge field) and energy gains greater than 10^4. We have observed "lock-on" in intrinsic GaAs at fields consistent with those observed by other researchers. Theoretical modeling is an important part of our work. We have narrowed the physical mechanisms for avalanche and lock-on modes to three possibilities and plan to explore each in the next year.

Introduction

Photoconductivity, or modulating the resistance of a semiconductor material by light, is a very promising technology for high-power, high-speed switches. Solid state switches have many advantages over conventional gas and mechanical switches, including very fast switching speed, low trigger jitter, simple mechanical structure, optical trigger isolation, high thermal capability, and flexible geometry.

Three Modes

The most common mode of operation for photoconductive switches is the linear mode. Here, each photon creates one electron-hole pair, and both electrons and holes are free carriers of electricity. The resulting increase in conductivity is proportional to the number of photons absorbed. In the linear mode, closing time is limited only by the optical trigger speed, since carrier generation time is approximately 10^{-14} sec. With low fields, spontaneous recombination of carriers will open a photoconductive switch against voltage. Recombination times can be made shorter by, for example, irradiating GaAs with neutrons, giving opening times less than 500 psec.[1] One drawback to linear operation is the large laser energy required to generate high electrical powers in the linear mode. Power gains are theoretically limited to less than 100, and power gains of 10 have been difficult to achieve. As the field across the switch is increased above a threshold value, the switch goes into lock-on mode. In lock-on, the switch is closed in a manner identical to linear mode. The difference between the lock-on and linear modes is that the switch will not open spontaneously against high fields, making the device a closing switch. The disadvantage of lock-on is that the switch typically retains a residual field of 4–8 kV/cm, creating a high-dissipation state in the switch. At high fields, the switch will go into avalanche mode at low laser energies. In avalanche mode, the triggering laser must create only a few carriers, which are then multiplied by the field to induce high conductivity. Avalanche mode operation allows much higher power and energy gains than linear or lock-on modes.

Applications

Photoconductive switches have the potential to excel in applications requiring high speed and high voltage. Direct generation of high-power microwaves in the 500 MHz–5 GHz frequency range is an application for which photoconductive switches are particularly well suited. Photoconductive switches address the problem of efficiently generating an intense electromagnetic field and also the problem of launching the waves from an antenna structure. Photoconductive switches operating in the linear mode are an excellent way to generate a plane wave in a parallel plate transmission line. Modules of a parallel plate line can then be stacked to create a large-aperture radiating surface, increasing the radiating efficiency of the generator. As Pockel's cell drivers, photoconductive switches could provide fast rise times at desired voltage and current levels with minimal jitter and potentially high repetition rates.

Calibration of high voltage diagnostics is another very important potential application for photoconductive switches. It is very difficult to generate voltage pulses higher than 5 kV with subnanosecond rise times. Photoconductive switches have demonstrated pulse amplitudes in excess of 20 kV with rise times less than 300 psec. Pulsers using photoconductive switches would be a valuable tool in determining the response of high voltage diagnostics.

The fast recombination times and high thermal tolerance characteristic of GaAs and InP make them viable candidates for high repetition rate applications. Conventional pulsed power switches have difficulty operating at repetition rates greater than 5 kHz. Preliminary results indicate that GaAs photoconductive switches could be operated at repetition rates of 1 to 10 MHz if thermal management problems are overcome.

Progress

There are three important issues barring the routine use of photoconductive switches in pulsed power and microwave generation applications:

- Open state voltage holdoff,
- Efficiency of laser control,
- Device life.

We are addressing all three issues both experimentally and theoretically.

Experimental Work

We are performing experiments on dark voltage holdoff, switching speed, and switch gain (ratio of electrical output to optical trigger energy) for both linear and avalanche modes in the two experimental setups shown in **Figs. 1** and **2.** The characteristic impedance of the slab switch test setup (**Fig. 1**) is 100 Ω (parallel plate transmission line) while the impedance of the doughnut setup (**Fig 2**) is 50 Ω (microstrip). In both cases, the effective impedance seen by the switch is 50 Ω. The entire slab test setup can be inserted into a pressure vessel and pressurized to 50 psig. The switch area of the doughnut setup can be immersed in dielectric fluid. The electrical length of the slab test line is 3 nsec (1.5 nsec on either side of the switch). The electrical length of the doughnut test line can be varied from 500 psec to 40 nsec. The voltage is applied to the switch sample in a pulse of about 500 nsec duration in both the slab and doughnut switch configurations. The slab switch is controlled by an Nd:YAG laser with 100 psec pulse duration and 50 mJ maximum pulse energy capability. The doughnut switch is controlled with a 100 psec Nd:YAG laser or an Al:GaAs diode laser operating in the range of 800–900 nm with a pulse energy of approximately 20 nJ.

Dark Voltage Holdoff

Figure 3 shows the results to date of the voltage holdoff experiments. Several switch samples have been tested. This and other data show that a polyimide coating improves the voltage holdoff. Both coated and uncoated samples reached the goal of 100 kV/cm, but results are not consistent among different samples. Also, a breakdown always results in sample damage with subsequent performance degradation, making it difficult to judge the results of coating and etching operations. It is also clear that there is a residual voltage (corresponding to a field of 4–8 kV/cm, regardless of applied voltage) on the sample during breakdown reminiscent of voltages observed during lock-on, which could result in sample damage due to the resulting large temperature rise. Research is continuing in the area of surface flashover with polyimide coatings. In addition, we designed and constructed the new sample holder shown in **Fig. 4** to shape the electric fields away from the ends of the switch. The new holder was designed following the observation that there was consistent damage at the ends of the GaAs slab after dark breakdown. The effect of the new holder is still not fully established, since more

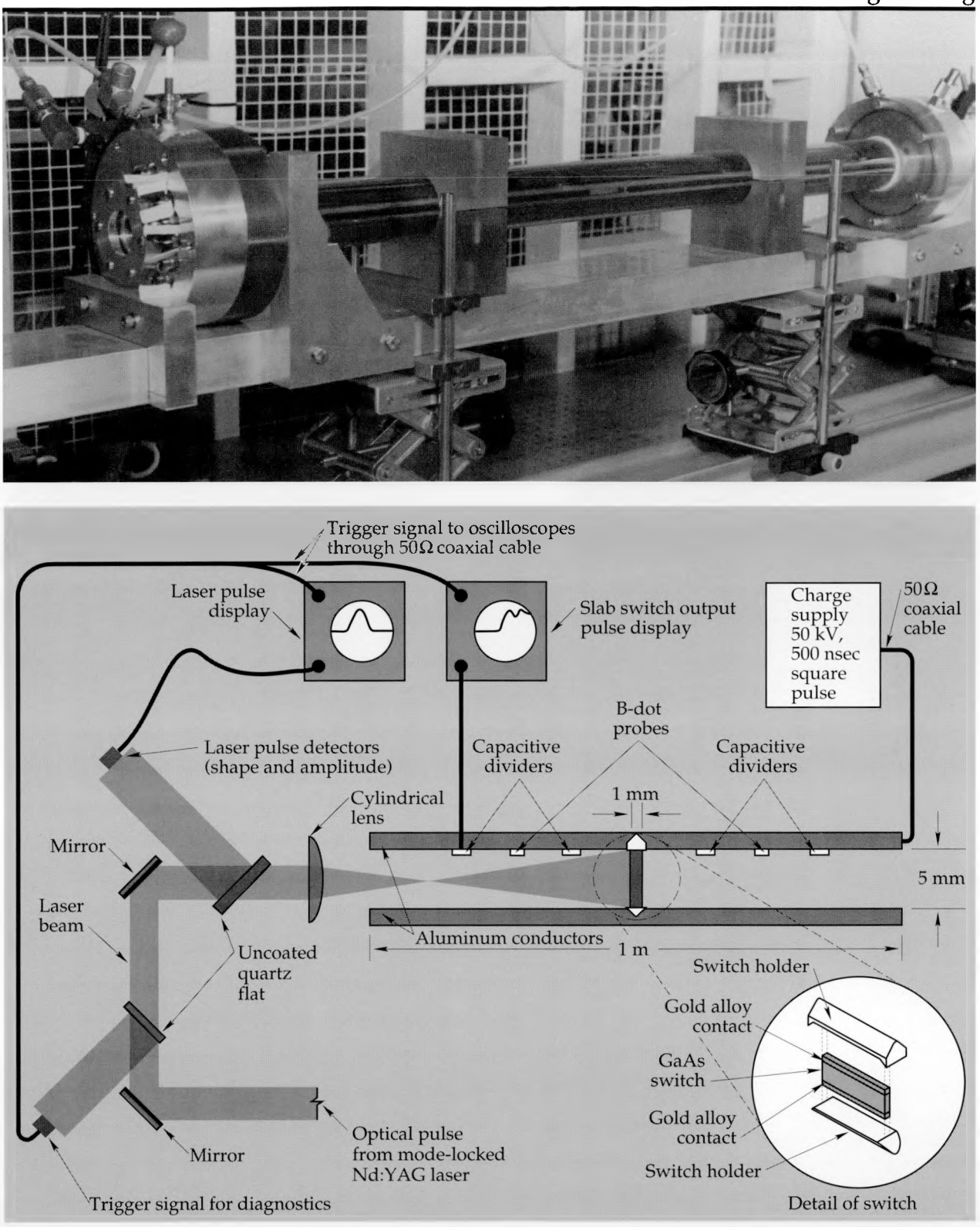


Figure 1. Experimental setup for the slab switch experiments. (a) Photograph of experiment showing pressure vessel, transmission line (note switch in the longitudinal center of the line), diagnostic penetrations of the pressure vessel, and laser beam shaping optics. (b) Drawing showing the complete system, including laser, charging supply, diagnostic chain to the recorders (including triggers), and the laser pulse monitoring system. The entire transmission line can be pressurized to 50 psig of SF_6 to prevent surface flashover of the switch. The charge voltage is variable from 10 kV to 50 kV. All high voltage and diagnostic components are shielded for electromagnetic interference protection.

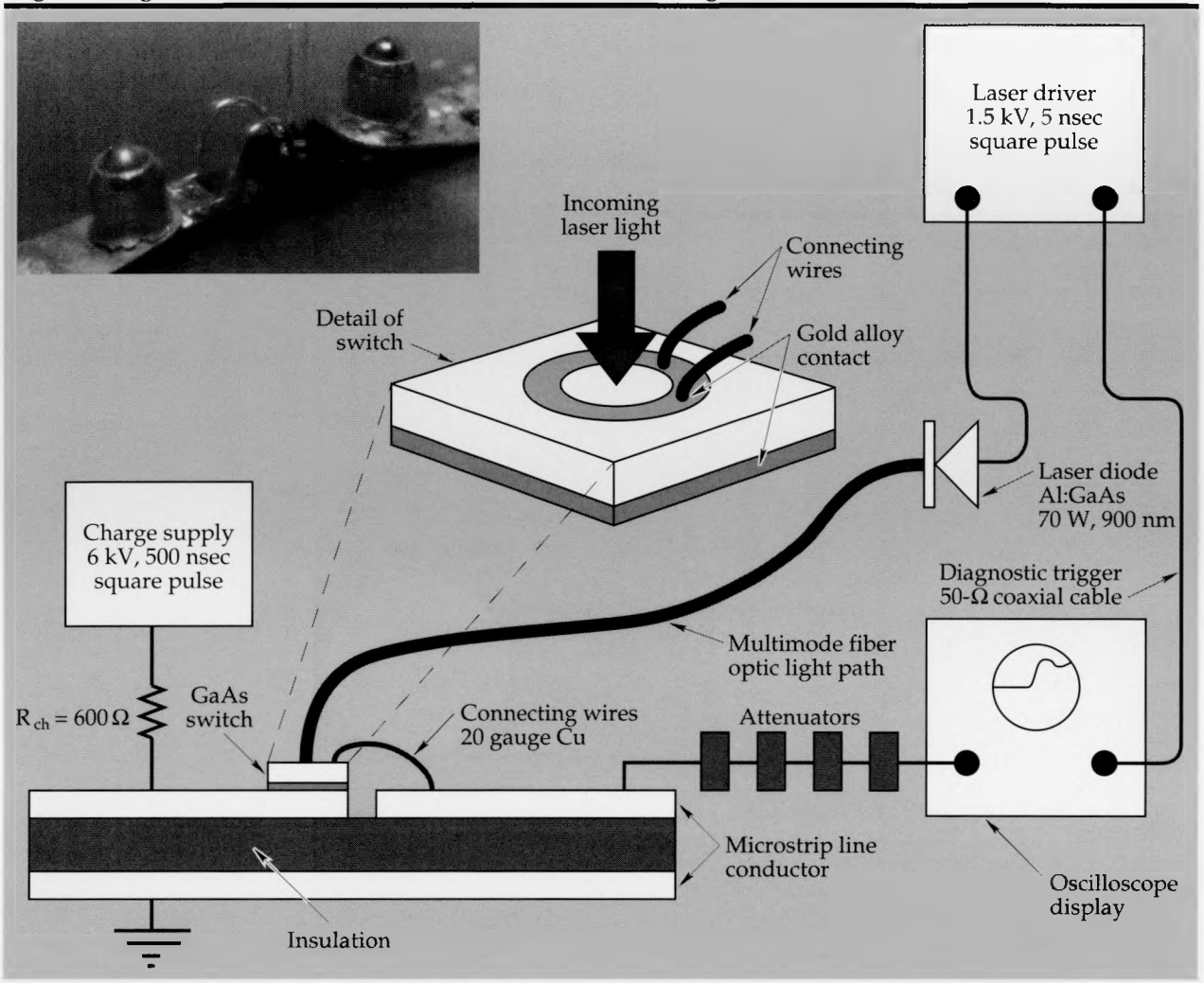


Figure 2. Experimental setup for the doughnut switch experiments. (a) Photograph of experiment showing the microstrip transmission line, the switch, and the optical fiber used to trigger the switch (the fiber is driven by an Al:GaAs laser diode). (b) Drawing showing the complete system, including the laser diode, charging supply, and diagnostic chain to the recorders.

data is needed to draw statistically significant conclusions.

Switching Characteristics

We have performed experiments in both slab and doughnut geometries to investigate switching speed, optical triggering efficiency, and device life. **Figure 5** shows a typical waveform obtained in the linear mode above lock-on, including the fast rise time associated with linear operation and the slow recovery associated with lock-on. The switch did eventually open (against voltage) during the shot shown in **Fig. 5**. At higher fields, the switch remains closed for the duration of the applied voltage. There is always a residual field of 4–8 kV/cm left across the device in the closed state of avalanche and lock-on modes which appears to be independent of applied voltage and current through the device. This residual field is reminiscent of the arc drop in a gas discharge, which is reasonably constant regardless of current and applied voltage.

The efficiency of laser control in linear and lock-on modes is still an issue. **Figure 6** is a comparison of power gain for a single switch and two switches stacked in the direction of laser propagation (1 mm each; effective thickness 2 mm for two switches). The absorption depth has been measured at approximately 4 mm so it is no surprise that increasing the thickness of the switch in the laser direction also increases the efficiency of laser control. Stacking three switches (an effective thickness of 3 mm) produced no further increase in efficiency (the gain decreased from that seen with two switches), which may be due to interference fringes at the

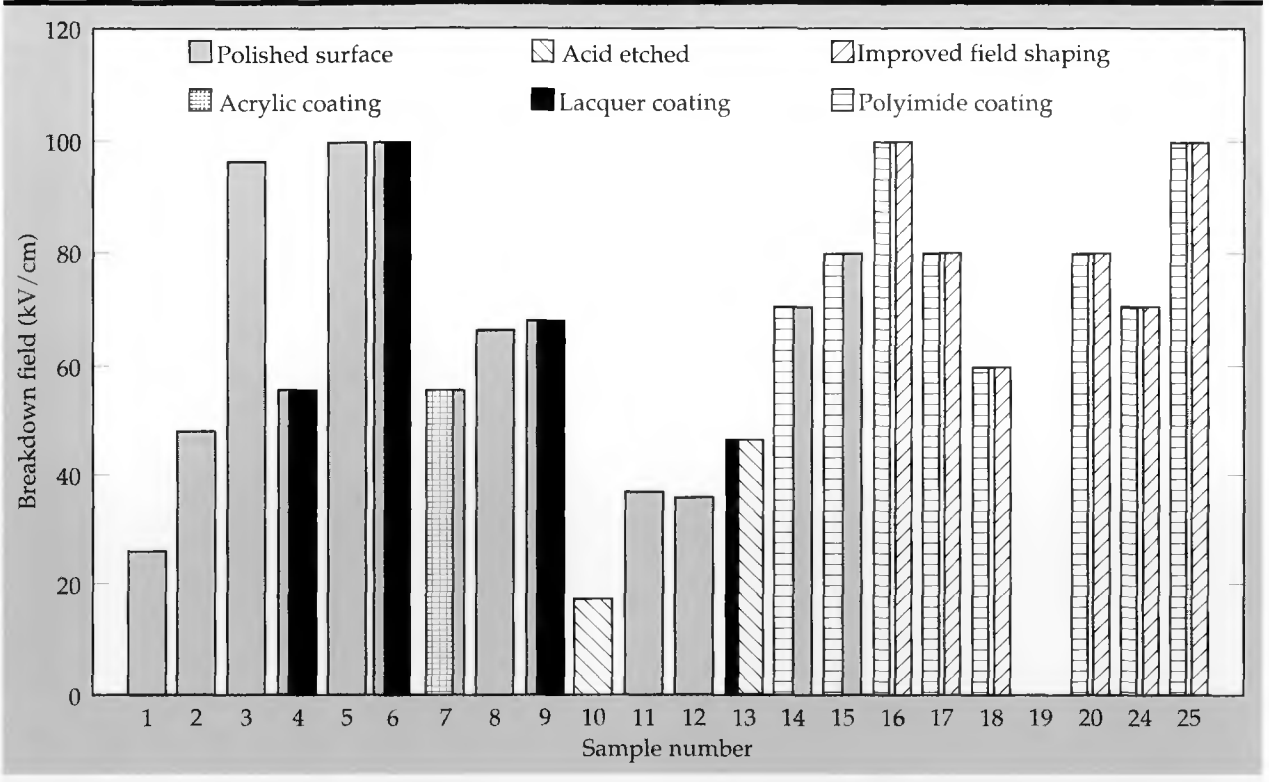


Figure 3. Dark breakdown voltage for tested samples, showing the random component of semiconductor physics responsible for dark breakdown. Note that several samples prepared in the same manner vary widely in breakdown voltage. All breakdown events are across the surface of the GaAs. To increase the breakdown voltage on the GaAs surface, we have tried several coatings with somewhat mixed results. The only coating that consistently increases breakdown voltage is polyimide. All samples are GaAs in an environment of 50 psig of SF_6.

three switch interfaces. The maximum gain attained to date has been 9 at 30 kV applied voltage.

Avalanche mode operation is very attractive for its high energy gain characteristics. **Figure 7** is a plot of typical waveforms obtained from slab switches operated in avalanche mode. The two waveforms shown in **Fig. 7** are representative of those found at low and high field and low and high laser energies. Note that there is an identifiable delay between the application of the laser and the avalanche pulse. This delay appears to be primarily a function of applied voltage and not a function of laser energy. We have successfully switched voltages to 35 kV (70 kV/cm field) with rise times of less than 2 nsec. We previously reported rise times as short as 500 psec in avalanche mode for the slab switches. A limitation in the voltage diagnostic has become apparent; the best 10–90% rise time for slab switches in avalanche mode is actually 1.25 nsec.

Experiments with shadowing the laser on the slab switch indicate that the mechanism responsible for closing the switch in avalanche mode is not simple avalanche. **Figure 8** shows a typical waveform for the doughnut switch. We have seen

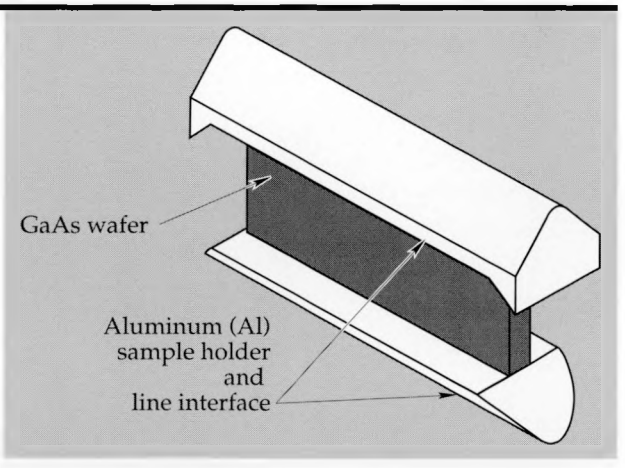


Figure 4. Drawing of improved sample holder showing "ears" used to shape the field away from the edges of the wafer. The SF_6 surrounding the switch has a very high field capability (>200 kV/cm). This sample holder was designed to take advantage of the very high field capability of SF_6 to "pull" the field away from the surface of the switch. Since all switch breakdowns are surface-related, reducing the field at the surface improves overall switch performance.

gains of over 10^4 for the doughnut switches with 20 nJ diode laser triggers. This device is normally operated in the avalanche mode at average fields of 150 kV/cm (6 kV applied voltage, 400 micrometer thick device). **Figure 9** is a comparison of the leading edge of a pulse triggered by the 100 psec Nd:YAG laser with a pulse triggered by the 5 nsec diode laser. **Figure 9** shows that the waveform is essentially independent of laser wavelength, additional evidence refuting the proposed mechanisms for avalanche mode operation.

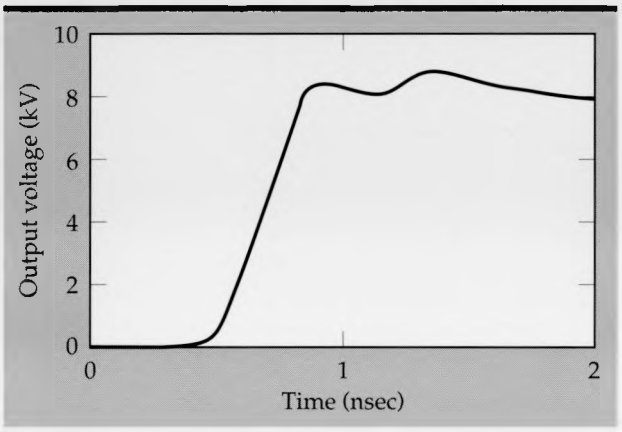


Figure 5. Typical lock-on waveform, showing closing switch characteristics with large laser energies. Note that the rise time follows the input laser pulse but that the fall time of the pulse is determined entirely by the circuit to which the switch is connected. This lock-on behavior is not fully understood but may be related to the filling of traps in the GaAs crystal lattice structure.

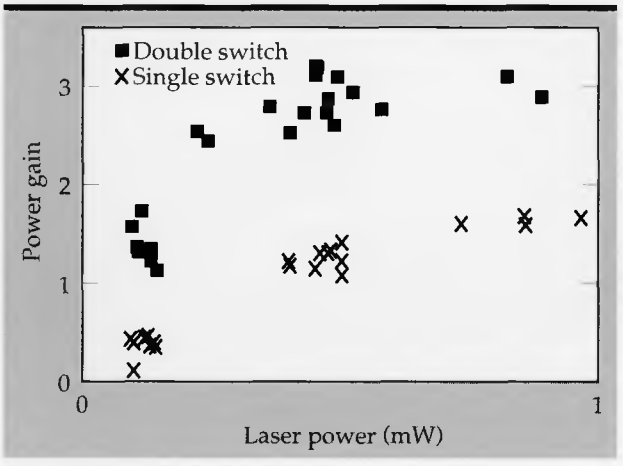


Figure 6. Plot of power gain vs laser power for 9 kV applied voltage on the slab switch. The double switch consists of two 1-mm switches mounted back-to-back in the direction of laser propagation, effectively doubling the laser path through the switch. Note that the double switch configuration is approximately twice as efficient (twice the switch gain) as the single switch. This increased efficiency indicates that the laser absorption depth is much greater than 1 mm, as we predicted from other research.

Device life is a problem in the linear, lock-on, and avalanche modes of operation. Device life now varies from a low of 50 shots to a high of over 10^4 shots. We have performed life tests on both slab and doughnut switches prepared with several processing technologies. Several doughnut switches were tested which had MBE-grown junctions added. Various layers (plastic and semiconductor) have been applied to tested switches on the theory that surface flashover is occurring after repeated stressing of the switch. To date, the sample-to-sample statistical variation has been a dominating factor in the life tests. Two devices prepared in exactly the same manner often exhibit drastically different lifetimes. It is clear from this result that more research is required in the area of GaAs switch processing technology.

Switching Mechanisms

A considerable portion of our effort this year has gone into modeling and theoretical analysis of the switching mechanisms in all three modes. The linear mode of operation is well understood. We can qualitatively, and in most cases quantitatively, explain the switching behavior observed. The lock-on and avalanche modes involve a variety of nonlinear effects which are still very difficult to explain. We have during the year narrowed down the possible mechanisms to three. These are field dependent trap filling, bulk avalanche by field enhancement in material containing a nonhomogeneous trap density, and current filamentation. We plan in the next year to examine these mechanisms both theoretically and experimentally to gain an understanding of the switching behavior.

Figure 10 illustrates the basic discrepancy between theoretical and experimental behavior. It is a typical plot of experimental and theoretical normalized output voltage vs the electric field across the switch before illumination, assuming uniform field distribution. It clearly shows a transition from linear mode to avalanche mode. The 20-60 kV/cm value is considerably lower than the theoretical transition assuming an avalanche breakdown field of greater than 200 kV/cm. Since avalanche multiplication is exponentially related to the electric field, it is difficult to explain the sudden change of mode at 20-60 kV/cm. Current filamentation has been suggested as the mechanism for this reduced transition electric field. However, current filamentation by itself does not create an enhanced electric field which would lead to avalanche multiplication, nor does it change the trap

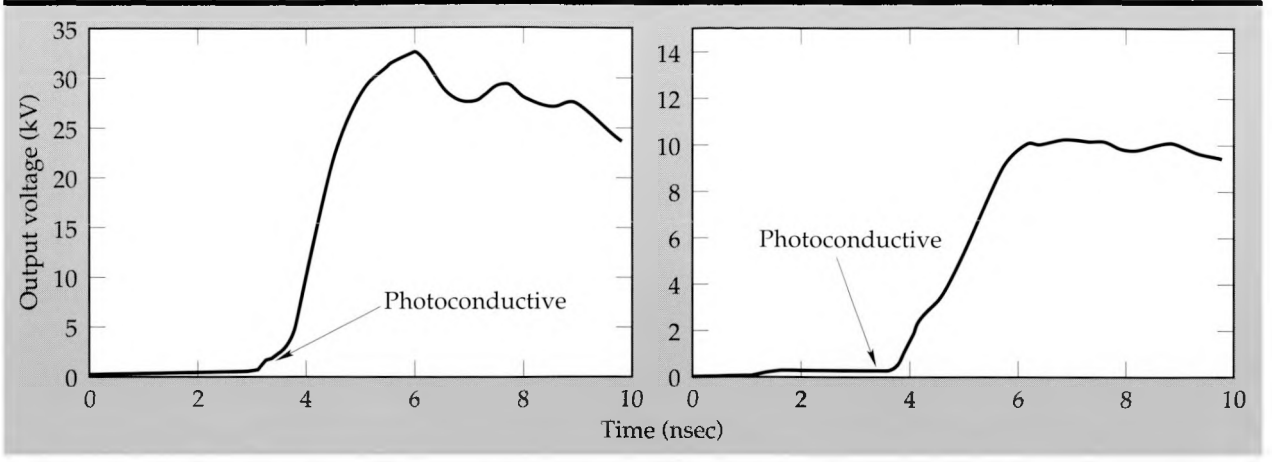


Figure 7. Plots showing typical avalanche mode waveforms at two voltages for the slab switch. Note the different vertical scales. The rise times are typically 1-2 ns in avalanche mode, depending on the charge field. We believe that the different shapes of the leading edges are due to the nonlinear carrier (electrons and holes) field-velocity curve in GaAs.

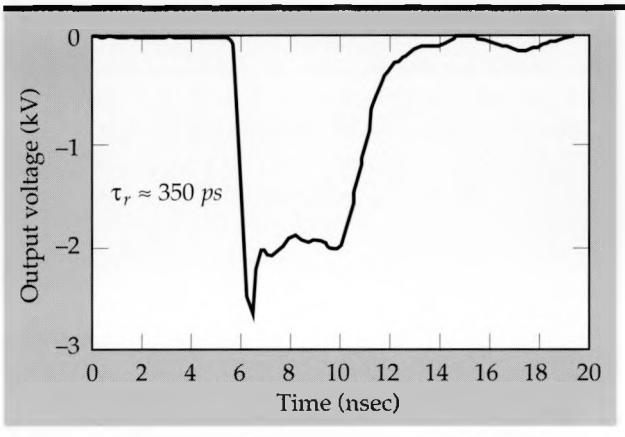


Figure 8. Typical doughnut switch waveform. The photoconductive pulse is very small on this plot. On an expanded scale, we can clearly identify the photoconductive pulse, which appears very similar to lock-on and which we believe is lock-on. The overshoot on the rise of the waveform is probably due to a small stray capacitance in the circuit used to drive the switch. We are now attempting to improve the driver dircuit.

density; it simply creates higher current density. Thus, if filamentation is a mechanism it must lead to either field enhancement or to trap filling, or perhaps thermal runaway, to cause the observed behavior. Since we do not see the avalanche or lock-on behavior in silicon, we conclude that thermal effects are not a likely explanation. The thermal capacity and thermal conductivity of silicon and gallium arsenide differ by less than a factor of three, and GaAs has a wider band gap, making thermal run away an unlikely mechanism. We can therefore concentrate our research on the two remaining mechanisms—field enhancement, and trap filling independently of the two-dimensional filamentation effect.

We have theoretically explored both of these mechanisms using both numerical and analytical calculations and have found that either mechanism is possible if the conditions are right. If trap filling is the mechanism it must be field dependent because of the transition field. A non-field-dependent trapping model does not result in a sudden transition from low output voltage to high output voltage, but rather a continuous change over the electric field range. If local field enhancement leading to avalanche multiplication is the mechanism, we find that electric fields cannot be high enough in a material that has a uniform density of traps between the electrodes of the switch. The electric field can build up to the critical levels for avalanche multiplication only if there is a nonuniform density of traps between electrodes.

Determining the dominant mechanism will require a coordinated experimental and theoretical approach to design experiments which can isolate the effects, perform the experiments, and compare the results to predicted behavior. In the next year we plan to use this experimental/theoretical approach in an attempt to determine the dominant mechanism. The possibility of filamentation will add another complication to be dealt with in designing and conducting the experiments.

Future Work

We are currently collaborating with Rockwell International on a project designed to generate microwaves with photoconductive switches in the 1–5 GHz region at very high powers. This project

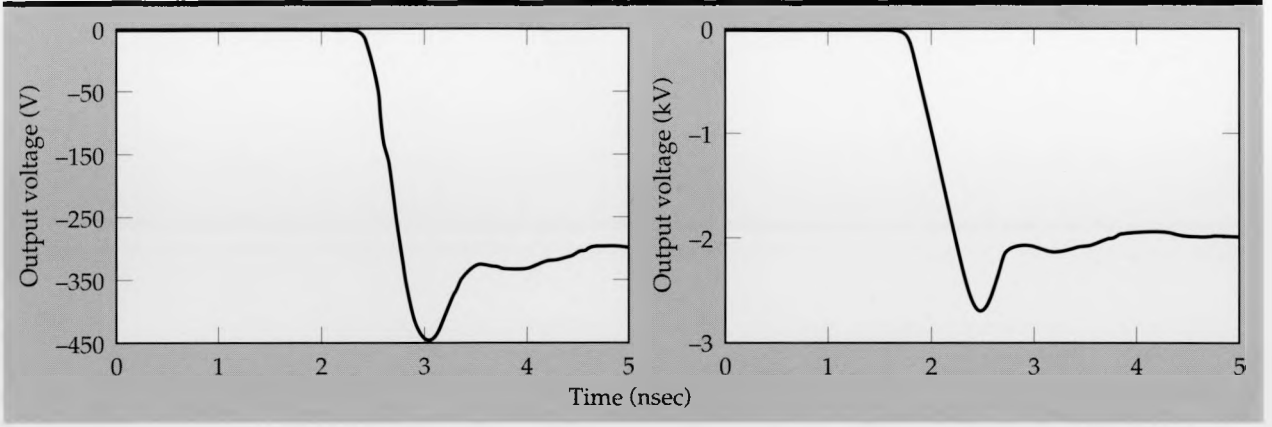


Figure 9. A comparison of the leading edge of a Nd:YAG triggered doughnut switch output pulse with a diode-laser-triggered, doughnut-switch output pulse. (a) Nd:YAG triggered. (b) Diode laser triggered. Note the similarity between the two pulse shapes even though the laser driving pulses are very dissimilar. The Nd:YAG laser delivers a 1.064 μm wavelength in a 100-ps-long pulse while the diode laser pulse is 900 nm, 5 ns. The absorption depth of the Nd:YAG laser is much greater than the diode laser (the Nd:YAG laser penetrates much deeper into the switch), refuting the theory that field compression is the dominant mechanism for avalanche switch operation.

is planned to span FY90 and beyond. In the area of switching research, we plan to continue efforts to improve laser control efficiency in linear and lock-on modes, improve rise time in avalanche mode, and increase the reliability and device life in all switching modes. Many aspects of this device engineering will rely on modeling to give us insight into device physics and possible improvement areas.

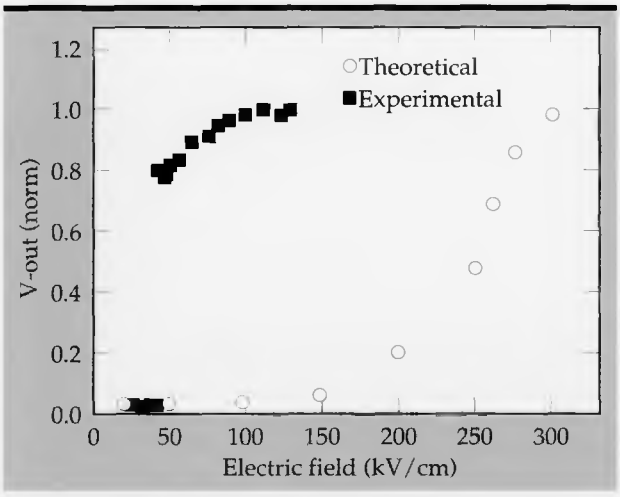


Figure 10. Typical plot of normalized output voltage vs open state electric field for a doughnut switch, using a uniform field approximation. This comparison of theory with experiment shows that additional physics must be incorporated into the present models to predict avalanche operation with complete accuracy.

Acknowledgments

The authors would like to thank a number of coworkers at LLNL who were part of our team this year and who were instrumental in obtaining the results reported here. Ken Griffin and Mike Wilson took responsibility for the slab switch testing and doughnut switch testing, respectively. Gerard Jacobson provided help with circuit design, packaging, and testing. Dan Okubo provided help with switch manufacture and mounting. Travis White and Clyde Dease contributed materially to the modeling effort. Ray Mariella grew MBE epitaxial layers.

1. C. L. Wang, M. D. Pocha, J. D. Morse, M. S. Singh, and B. A. Davis, "Ultrafast Photoconductor Radiation Detectors," Lawrence Livermore National Laboratory, Livermore, CA, UCRL-96004 (1987).

ENGINEERING RESEARCH AND DEVELOPMENT

Thrust Area Report FY 89

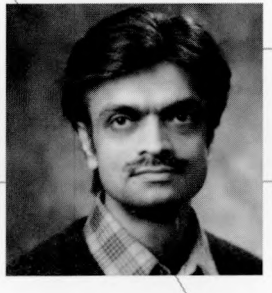
Nondestructive Evaluation

The goal of the NDE thrust area is to inspect parts and complex objects in order to determine their chemical and physical characteristics, flaws, and fabrication defects. In line with this goal, we support program initiatives (primarily those of R Program, Weapons, and the Laser Program), develop inspection tools and techniques, enhance the expertise of NDE personnel, and promote technology transfer.

In FY 89, we continued to develop the radiation and ultrasonic techniques used to collect quantitative inspection data and the signal and image processing tools needed to evaluate the data. Our primary contributions have been in the following areas:

- We are developing CT (computed tomography) scanners covering a broad range of energies and resolutions and have derived the associated reconstruction algorithms. The overall goal of this research is to improve the three performance parameters (spatial and contrast resolutions and system speed) that characterize CT imaging systems. In addition, we are addressing related topics such as elemental or effective-Z imaging, model-based imaging using *a priori* information, parallel processor architectures for image reconstruction, and scientific visualization of reconstructed data.
- We are developing a branch of x-ray optics called scanning x-ray microscopy in which interactions such as fluorescence, Compton scatter, and absorption are used to form an image. The range of our x-ray system extends to 20 keV. The overall object is to produce high-resolution images of the chemical composition or physical characteristics of materials in a short length of time (minutes to a few hours, rather than days). We also intend to apply the optics to related imaging techniques such as CT and to produce parallel beams for long-distance alignment.
- We are using ultrasonic and optical techniques to inspect fabrication-induced surface and subsurface damage in laser optic materials. Rayleigh-wave scattering is used to determine the extent of the damage. Frequencies are varied from 5 to 100 MHz. The results show that this method is a sensitive technique for locating the cracks and determining their depth. We are also using total internal reflection microscopy (TIRM) and photon backscatter (PBS) techniques as complementary approaches to characterize damage. We find that TIRM and PBS techniques can detect and size the flaws and determine their global distribution.

Satish V. Kulkarni
Thrust Area Leader



X-Ray Optics For Microimaging Scanning Microscopy and Other Applications

Richard W. Ryon
Engineering Sciences Division
Mechanical Engineering

We are developing x-ray optics primarily so that we can perform scanning x-ray microscopy using fluorescence and other interactions, such as Compton scatter and absorption. We also intend to apply the optics to related imaging methods such as tomography, radiography, and projection microscopy. The overall objective of the work is to produce high-resolution images of the chemical composition or physical characteristics of materials in a useful length of time; i.e., a few minutes to a few hours, rather than a few days to a few weeks. We also address a new need at the Laboratory to produce a parallel x-ray beam for long distance alignment applications.

The principal technique we are investigating is Kirkpatrick-Baez mirror optics. The mirrors are curved surfaces coated with alternating layers of a heavy and a light element (multilayers) that act as a diffracting medium. These mirrors effectively increase the efficiency of the system by gathering and focusing the x-ray photons produced by a conventional x-ray source. Most x-ray microscopy to date has been done with "soft" x-rays (<1 keV); however, higher energies are needed to fluoresce most chemical elements. Our first x-ray system focused energies of 8 keV. We are now extending our capability to greater than 20 keV.

The result of this R&D effort will be the ability to image the distribution of chemical elements on and beneath the surface of objects with resolutions heretofore unobtainable in the laboratory.

Introduction

Increasingly at LLNL, we see a need for imaging applications with resolutions of 1 to 10 μm using x rays with energies of approximately 20 keV (wavelengths of about 0.06 nm). By collimating beams from standard x-ray sources, we can measure areas of about 100 μm with sufficient accuracy.[1,2] To reach the resolution we seek, we must either use much more intense sources of x rays, such as synchrotrons, or improve efficiency by demagnification (gathering x rays over an extended area and focusing them to a very small spot). Recent advances in materials fabrication technology allow us to pursue demagnification using x-ray focusing optical elements.[3,4] By alternating layers of heavy and light elements such as tungsten and carbon, composite materials can be made with interlayer spacings of only a few atoms (that is, about 1.0 nm). The multilayer mirrors diffract at angles of incidence of 1 to 3 degrees, and thereby intercept a fairly high fraction of x rays from the source. (The solid angles are about 10^{-5} to 10^{-6} steradians when we demagnify the source by 10 times.) Two other LLNL researchers, Troy Barbee and Pete Biltoft, make our present mirrors on spherically curved surfaces because these surfaces are readily available. However, spherical surfaces have aberrations that can distort the focus of the beam. We plan to use elliptical surfaces to reduce aberrations and thereby improve the focus.

Progress

We have met our major milestones for FY 89. The Kirkpatrick-Baez system[5] was completed at Lawrence Berkeley Laboratory and installed in our lab. We designed and fabricated mirror sets for energies of 8, 17, and 22 keV. We wrote software for controlling the scanners and collecting data for step scanning. We are also modifying the software for fast slew scanning. Elliptical mirrors have been designed. The rotating target x-ray source and its safety system have been installed, and two alternative source installations are nearing completion. We successfully focused x-radiation to a point on an object and collected our first images of small-scale features (**Figs. 1 and 2**).

We have begun to apply the optical system to LLNL programmatic needs. The first application is to measure the rubidium distribution about the equator of a microsphere only 200 μm in diameter.

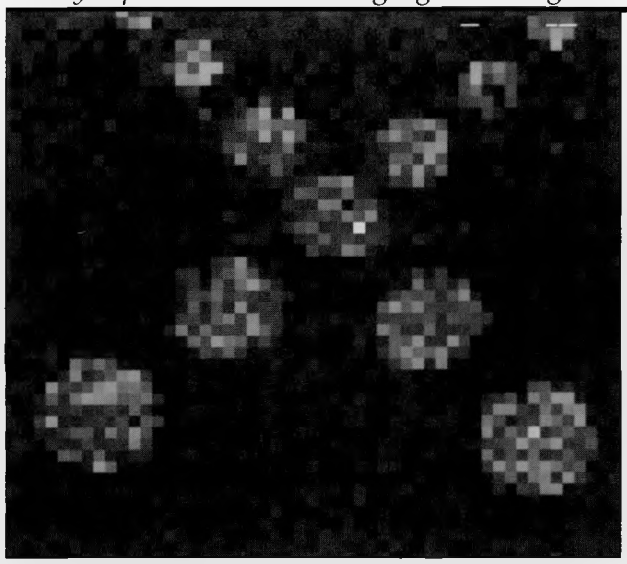


Figure 1. Laser-produced holes in erbia coating on steel used as an ultrasonic sensitivity test specimen. Hole diameters are 150, 200, 250, 300, 350, and 400 μm. Course scan, 2.2 × 2.0 mm, with 40-μm step.

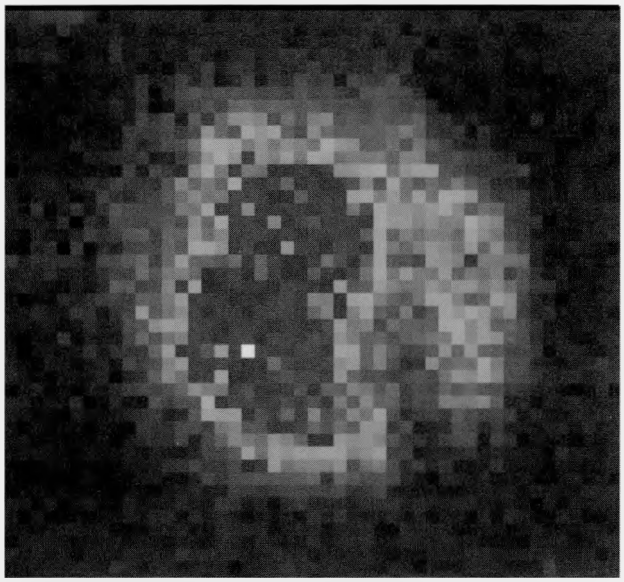


Figure 2. Close view of the smallest hole in Fig. 1, 0.26 × 0.23-mm scan with 5-μm steps.

We are still testing the power of the K-B microscope to excite rubidium fluorescence, which requires the use of 17-keV x rays. The second application is to produce a parallel x-ray beam in order to align an apparatus in a beam line at a distance of 670 m.

Future Work

In addition to applying the current system to programmatic needs, we will be improving it. For example, we plan to fabricate the first K-B elliptical mirrors. When we eliminate spherical aberrations in computer simulation, the ray tracing computer design work indicates that the focal spot size should reduce from approximately 25 × 30 μm to 9 × 19 μm when using a 100 × 100-μm x-ray source. In experiments, we will measure the focused beam's photon gathering efficiency, point spread function, x-y demagnification, and energy resolution. By comparing actual performance to the code design predictions, we will determine whether design or fabrication restricts resolution.

We also plan to investigate other configurations—clusters of hollow leaded-glass capillaries and shapes other than spherical. We now have clusters that consist of approximately 10^5 20-μm-diam hollow leaded-glass capillaries, which act as reflective collimators. The clusters are tapered so that each collimator points to the same spot. These clusters may prove to be a very efficient means of focusing x rays, although at some sacrifice in spot size. We also hope to investigate more efficient shapes, such as toroids, perhaps with ellipsoidal surfaces and with longitudinally varying spacing, in order to minimize aberrations and maximize the intercepted and focused radiation.

We plan to integrate our x-ray optics into other upgrade activities. An immediate application is tomography. By using a single mirror, we can produce a very fine "fan" of x rays and then use a linear detector array to rapidly acquire data. This experiment will be the first demonstration of tomographic imaging using x-ray optics.

Acknowledgments

Contributors include T.W. Barbee (C&MS), P.J. Biltoft (MFD), R.A. Day (ESD), P.C. Gabriele (ESD), R.L. Hurd (NESD), and K.G. Tirsell (L Div.).

1. J.H. Underwood, T.W. Barbee Jr., and C. Frieber, "X-Ray Microscope with Multilayer Mirrors," *Applied Optics*, **25**(11), 1730–1732 (1986).
2. R.W. Ryon, "X-Ray Optics for Microimaging Applications," *Engineering Research and Development Thrust Area Report FY 88*, **UCRL 53868-88**, 7-7 to 7-9.
3. M.C. Nichols and R.W. Ryon, "An X-Ray Microfluorescence Analysis System with Diffraction Capabilities," *Advances in X-Ray Analysis*, **29**, 423–434 (1986).
4. R.W. Ryon, H.E. Martz, J.M. Hernandez, R.A. Day, J.M. Brase, B. Cross, and D. Wherry, "X-Ray Imaging: Status and Trends," *Advances in X-Ray Analysis*, **30**, 35–52 (1988).
5. J.H. Underwood, "High-Energy X-Ray Microscopy with Multilayer Reflectors," *Review of Scientific Instruments*, **57**(8), 2119–2123 (1986).

New Techniques for Ultrasonic Detection of Surface and Subsurface Damage in Laser Materials

D. Kent Lewis and
Chol K. Syn
Engineering Sciences Division
Mechanical Engineering

Ultrasonic-wave studies were undertaken to see if Rayleigh-wave scattering could be used as a means of determining the extent of surface and subsurface damage in laser optics materials. Three transducer setups were employed: (1) a single transducer with a wide-angle lens (used closer than focal distance) for simultaneously measuring back and forward scattering; (2) a single transducer with a narrow-angle lens for measuring back scattering; and (3) two narrow-angle transducers used together for measuring forward scattering. Frequencies were varied from 5 to over 100 MHz. The results show that Rayleigh-wave scattering is a sensitive technique for determining surface and subsurface damage.

Introduction

We have performed ultrasonic-wave studies to see if scattering from Rayleigh surface waves could be used as a means of determining the extent of surface and subsurface damage in glass and crystal optics used in lasers. Rayleigh waves have been used previously to characterize surface-breaking cracks and effects of surface-hardening processes. Since Rayleigh waves exist only on and near a surface, and since their penetration is proportional to wavelength, they seemed a reasonable approach for detecting surface and subsurface damage in laser materials.

Our study methods employed three different acoustic-transducer setups for the detection of surface-wave scattering from both the back and forward directions. We used a conventional transducer setup to compare specular-wave results with our results.

Before discussing our three test setups, we must make note of some basic features of acoustic wave action that are central to our study approach: A smooth optic surface will cause specular reflection of acoustic radiation at all angles of incidence. However, only at the Rayleigh angle (≅31.3°) will a surface wave be generated. Furthermore, the same conditions that make generation of Rayleigh waves possible only at a specific angle also assure that these waves, traveling along the surface, will always reradiate at exactly this same Rayleigh angle. A limitation of conventional acoustic microscopy is that it can detect surface features only: It will not detect closed cracks that are perpendicular to the surface.

The basic test device is a focused acoustic transducer consisting of a piezoelectric disk, a silica transmission rod, and a concave spherical surface milled into the end of the transmission rod to form the lens. (See **Fig. 1**, which is a schematic cross section of one of the transducer designs used in our studies. The setup shown in **Fig. 1** is typical of that used in conventional acoustic microscopy.) Tests are conducted with the transducer assembly and test piece (or "sample") immersed in water. The lens is designed to refract the radiation to a focal spot in water such that an equal energy comes from any direction within the focusing cone.

The transducer setups and acoustic processes involved in our methods for wave generation and detection are:

(1) A single transducer with a wide-angle acoustic lens, aligned normal to the test surface.and positioned closer to the test surface than the focal distance (i.e., it is used "defocused") for simultaneously measuring back and forward scattering. See **Fig. 2(a)**.

(2) A single transducer with a narrow-angle acoustic lens, canted 31.3° off the test-surface vertical and positioned with focal point at test surface for measuring back scattering. See **Fig. 3(a)** "Transmitter/receiver".

(3) Two identical narrow-angle transducers as in case (2) above but set for measuring forward scattering. In **Fig. 3**, see the combination of **(a)**

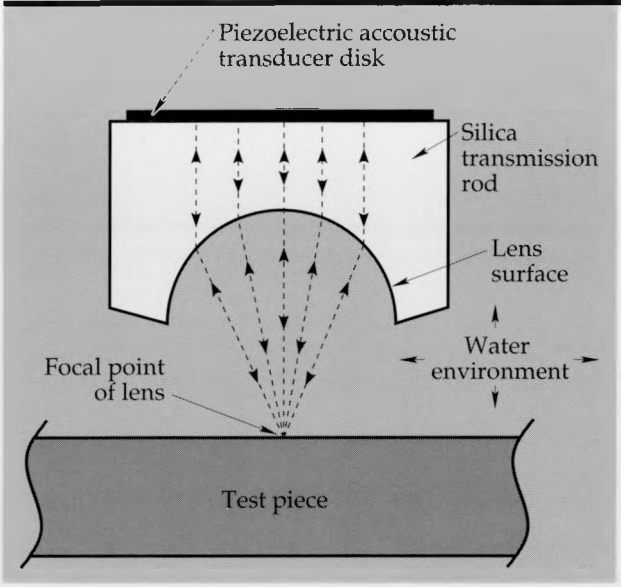


Figure 1. Schematic cross section of focused, acoustic transducer. Device consists of a piezoelectric transducer disk, a silica transmission rod, and a concave spherical surface milled into the end of the transmission rod to form the lens. The transducer is employed alternatively as a transmitter, a receiver, and as a transmitter/receiver. The lens is designed to refract acoustic waves to a focal spot in a water-immersion environment such that equal energy comes from any direction within the focusing cone. (This particular lens design is the wide-angle spherical-surface lens, the same as shown in Fig. 2.) The setup illustrated here is typical of that used in conventional acoustic microscopy.

"Transmitter" (only), and **(b)** "Receiver."

Setup (1) has the advantage of simplicity as well as the ability to detect flaws lying at any orientation. Setup (2) has ease of operation and higher signal level, although it is not sensitive to cracks lying perpendicular to the beam. Setup (3) detects both back and forward scattering, although alignment is more difficult.

For a single transducer oriented vertically to the test surface [setup (1) above] to be effective, it is necessary that the focused-beam cone angle be at least wide enough to contain waves at the Rayleigh angle. But basic spherical-lens theory tells us that the arc half-angle of the lens will be significantly greater than the angle of incidence. The optics design problem, then, was to come up with a lens with a sufficient curvature (a sufficiently wide arc angle) to refract waves at the Rayleigh angle. **Figure 2(a)** illustrates the proper angle that is required to achieve the incident Rayleigh angle of 31.3°.

An obvious, yet noteworthy, feature of this wide-angle setup is that those transmitted waves incident at the Rayleigh angle are circularly symmetic about the lens axis. The process is referred to as "large-aperture spherical-lens scattering inside the focal cone."

With the vertical orientation, the scattered waves that "return" to the lens along the pathways of the circularly symmetric Rayleigh-incidence

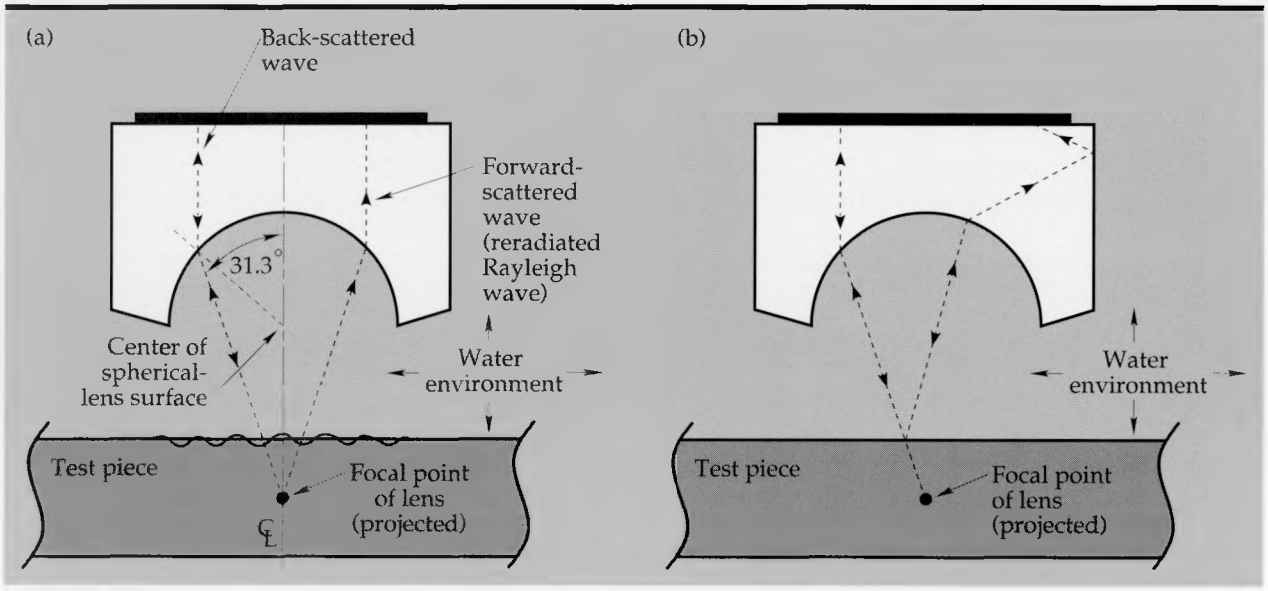


Figure 2. Acoustic transducer as illustrated in Fig. 1 but with test surface shown closer than focal distance (i.e., "defocused"). During tests, this particular design is aligned with its axis normal to the test surface, and is positioned closer to the test surface than the focal distance. This setup simultaneously measures back and forward scattering: (a) Pathways for back- and forward-scattered Rayleigh waves that are aligned with focal point. These waves are returned to transducer disk essentially unattenuated. (b) Typical oblique/zigzag pathway for specular reflection and those scattered waves not aligned with focal point. These waves are attenuated before reaching transducer disk.

waves [i.e., those that "appear" to originate at the focal point; see **Fig. 2(a)**] will be transmitted to the transducer/receiver essentially unattenuated. However, the specular reflections, as well as all other Rayleigh scattering, will be attenuated [see **Fig. 2(b)**]. The process leading to these results are discussed in the following two paragraphs:

When the focal point of a spherical lens is positioned on the test surface with the lens axis normal to the surface (**Fig. 1**), all reflected rays are returned to the lens such that they refract back into the rod parallel to its axis and are transmitted directly to the transducer/receiver (i.e., their incident and return pathways are symmetrically identical).

However, when the reflecting surface is closer to the lens than the focal point, the reflected rays do not return along the symmetically identical pathways, and therefore, no longer refract properly (for returning parallel to rod axis), but follow an oblique (and even zig-zag) course back to the transducer/receiver [**Fig. 2(b)**]. Thus the reflected field is phase shifted and greatly reduced in magnitude. For a semi-infinite solid, the only possibility in this case for proper refraction, such that incoming rays run parallel to the lens axis, is via the scattering from Rayleigh surface waves. Those scattered Rayleigh waves (both back and forward) approaching the lens on a pathway coincident with the circularly symmetric Rayleigh-incidence pathway will refract and transmit directly to the transducer [**Fig. 2(a)**]. Detection of the "symmetric" scattered waves are thus favored over the specular reflections and the "nonsymmetric" scattered waves.

The generation and detection of Rayleigh waves can also be done using either of the two setups shown in **Fig. 3**. In this case, the transducer employs a narrow-angle focusing lens and is used with its axis canted away from vertical by the Rayleigh angle and with its focal point positioned on the test surface. One setup uses just one transducer to detect back scattering [see **Fig. 3(a)**, "Transmitter/receiver"]. The other uses two transducers to detect forward scattering [see the combination of **Figs. 3(a)** and **3(b)**].

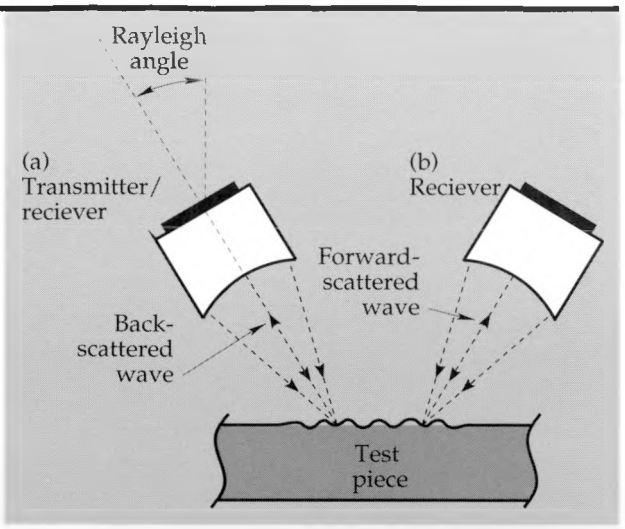


Figure 3. Setups for generation and detection of Rayleigh waves either using a single focused transducer [see (a) "Transmitter/receiver"] for measuring back scattering, or using two focused transducers [see (a) "Transmitter" (only) and (b) "Receiver"] for measuring forward scattering. Both tranducers have narrow-angle acoustic lenses, and are canted 31.3° (the Rayleigh angle) off test-surface vertical.

Progress

In our studies, three surface-wave acoustic processses (described in "Introduction") were investigated for detecting surface and subsurface damage: simultaneous back and forward scattering, back scattering, and forward scattering. Frequencies were varied from 5 to over 100 MHz.

Two sets of fused-silica disk samples were prepared as test pieces. For set 1, disks with unpolished surfaces were scratched using an indentation tool with different loads at different positions.

For set 2, disks with polished surfaces and existing subsurface cracks from manufacturing were etched with acid in an annular region around each of the disk's center. Etching was done to reveal surface cracks. From visual detection of the surface cracks in the etched region, we made an estimate of the subsurface cracking in the center region. We also obtained an estimate of the subsurface cracking from the results of concurrent surface-wave examination of both the etched surfaces and the scratched surfaces.

Calibration measurements were made using a single focused transducer to verify the predicted Rayleigh angle of 31°. The measured value (31.3°) was found to be in good agreement with the predicted value. Using this same setup, scans of the surface of two scratched samples (set 1) were done, and the resulting scattering maps showed that the largest three of five scratch marks were clearly visible. The smaller two did not show up.

We believe the smaller two scratch marks did not show up on the scans because of the high noise levels of the electronic equipment when operated at the sensitivity level necessary for Rayleigh-wave scans. A possible solution is to use our newly developed three-dimensional scan technique, in which a surface is scanned point by point using time averaging of the scattering-time signal. These

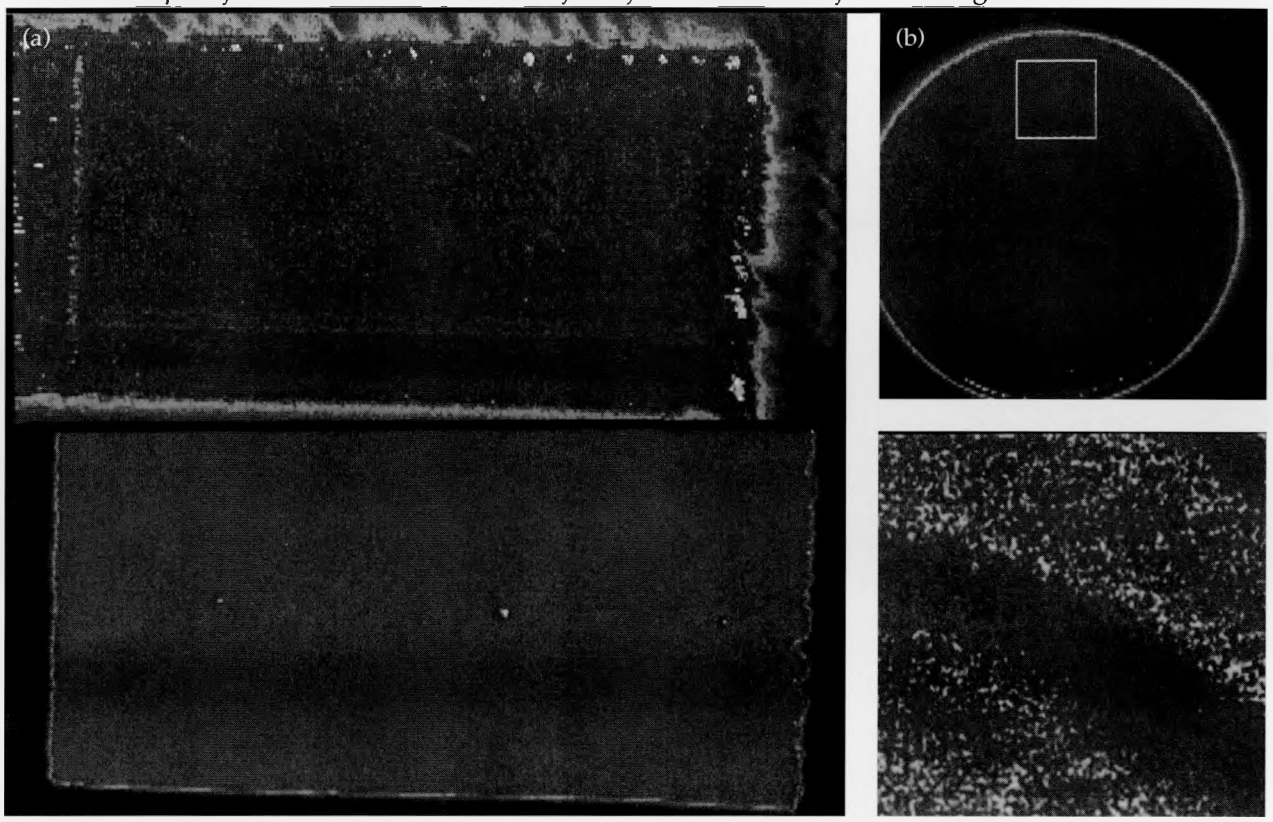


Figure 4. (a) Comparison scans of conventional acoustic microscopy (upper) to surface-wave microscopy (lower), using the same narrow-angle acoustic lens and the same etched surface. In both cases, the focal point was positioned at the surface. The lower scan used the lens at the Rayleigh angle for generating surface waves and detecting back scattering. (b) Scan of an etched sample made with a wide-angle acoustic lens normal to the surface and at a distance closer to the sample than the focal distance. Upper drawing is an overall view of etched region, and the lower is a detail of the same region.

much more sensitive records would then be time gated in the normal way and the magnitudes displayed in a conventional amplitude-scan manner.

In set 2, scans of the samples clearly showed the etch lines, but none of the predicted subsurface cracks was found. A subsequent scan of these same samples on the 100-MHz microscope at Sandia Laboratories, Livermore also failed to detect any of the cracks.

Figure 4(a) is a comparison of damage detection for an etched surface using conventional acoustic microscopy and surface-wave microscopy. Both scans were made using the same transducer, with a narrow-angle acoustic lens [see **Fig. 3(a)**]. The conventional scan (upper) used the lens normal to the surface and with the focal point at the surface, for detecting specular waves. The surface-wave scan (lower) used the lens canted at the Rayleigh angle, and also with the focal point at the surface, for generating surface waves and detecting back scattering. (Forward scattering results are not given here, since they merely confirm the back scattering results.)

Figure 4(b) is two views of the same scan of an etched sample made using a wide-angle acoustic lens normal to the surface and at a distance closer than the focal distance [see **Fig. 2(a)**]. The etched region is shown in an overall view (upper), and in a close-up view (lower).

The results show that detection of Rayleigh surface-wave scattering is a sensitive method of determining surface and subsurface conditions.

Non-Destructive Optical Characterization of Subsurface Damages in Laser Materials

Chol K. Syn
Materials Fabrication Division
Mechanical Engineering

We have used total internal reflection microscopy (TIRM) and photon backscatter (PBS) techniques to characterize fabrication-induced subsurface damage in transparent optics for use in high-power laser systems. We have shown that TIRM can detect and size subsurface cracks as small as a few micrometers in glass materials prepared by grinding and polishing. We have seen that the PBS technique can detect such surface and subsurface damages and can determine their global distribution.

Introduction

Grinding, polishing, and associated measuring and handling operations in the fabrication of optics, if improperly controlled, may induce surface and subsurface damages. While the surface damages are directly visible and easily removed, the subsurface damages are difficult to detect and can be easily overlooked. Subsurface microcracks larger than a certain critical size, if left undetected and unremoved in high-power laser optics, for example, may reduce the mechanical strength[1] and laser damage threshold[2] of the optics and jeopardize the performance of laser systems.

The current techniques for detection of subsurface damages—such as etching and taper polishing—are time consuming and destructive. Effective non-destructive detection techniques are needed in order to reduce the time and cost of optics fabrication without compromising the performance parameters. Thus, Ultrasonic/Optical Characterization of Subsurface Defects in Laser Materials was established as an Engineering Thrust Area Research project for FY 89. This report summarizes our work in the optical part of the project.

We have investigated two optical techniques for optics damage detection: total internal reflection microscopy (TIRM) and photon backscattering (PBS). The TIRM technique, devised by Temple,[3] provides highly magnified images of surface and subsurface defects and damages by capturing the light scattered from the damages when a laser beam is introduced under the surface at an angle greater than the critical angle and is totally reflected (see **Fig. 1**). The PBS technique produces a global map of the light scattered at a fixed angle when a laser beam is shone on the surface of an optic, and shows the distribution of surface and subsurface damages.

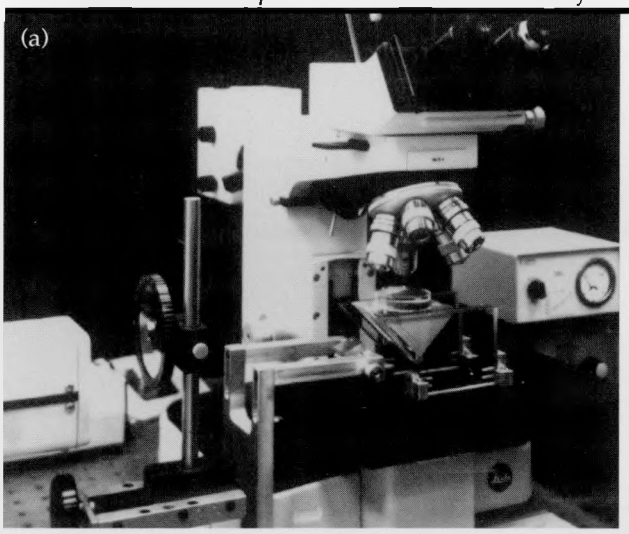
For our TIRM investigations, we have assembled a small TIRM attachment, installed it on an optical microscope, and examined various glass optics with known fabrication histories. For our other investigations, we have contracted with XMR, Inc., of Santa Clara, California, to utilize their expertise in the non-destructive PBS and destructive etching and decoration techniques they have developed.

XMR's etching technique, while similar to the conventional HF-HCl acid etching, adds proprietary ingredients to the etching solution to reveal some fine features not visible by conventional etching.[4] Their decoration technique utilizes the fact that a subsurface damage may cause a change in the electric charge distribution in the area around the damage. If such damage is close to the surface, the effect of changed charge distribution will be felt at the surface and may attract or repel charge-sensitive dielectric powder particles, such as carbon, much as iron particles line up along the field line of a magnet. Thus, application of a suitable dielectric powder may create a surface pattern revealing the subsurface damage.[4] The powder, however, may not be easily removed and may require a flash grinding or polishing run.

Progress

TIRM Investigations

The TIRM attachment (**Fig.1**) consists of a 2-in.-square prism mounted on the specimen stage of a Leitz optical microscope, a 15-mW He-Ne laser, and a 1-in.-square, laser beam-steering mirror. This arrangement allows us to switch instantly back and forth between the normal microscopy



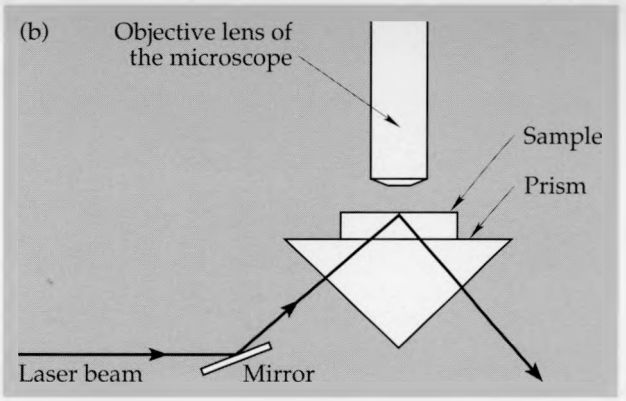
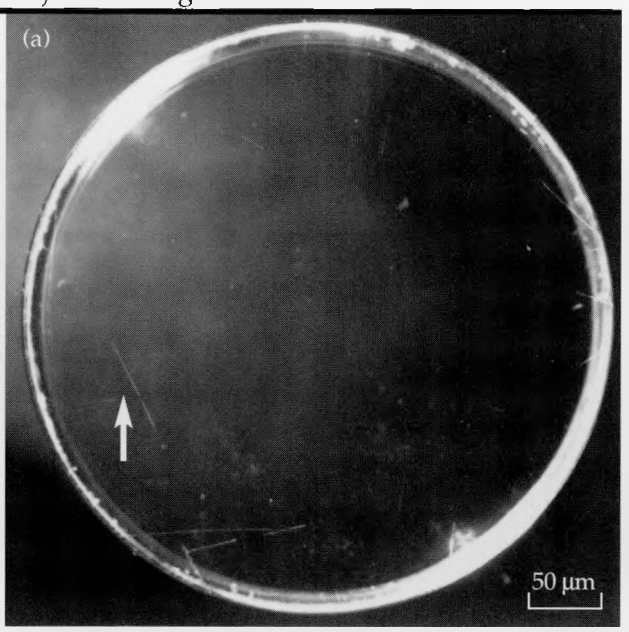


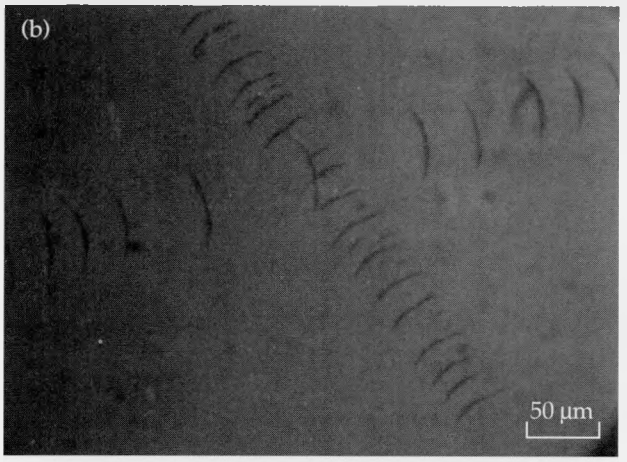
Figure 1. TIRM device and operating schematic. (a) Photograph shows TIRM attachment on a Leitz microscope. A sample sits on a prism mounted on the specimen stage of the microscope. (b) A laser beam is introduced through the prism and the sample to the underside of the top surface of the sample at an angle greater than the critical angle to create a total internal reflection condition. A few drops of refractive index matching liquid are applied for a good optical contact between the sample and prism. The light scattered by the subsurface damages is detected by the microscope objective lens and forms an image of the damages.

and TIRM or to use the two modes simultaneously.

The vertical clearance between the objective lens and the prism limits the maximum size of the optics sample to about 2 in. in diameter and 3/4 in. in thickness. Using this tool, we have examined fused silica, BK-7 glass, and other coated and uncoated samples.

Figures 2 and **3** illustrate subsurface cracks detected by the TIRM technique in a ground and polished fused silica sample. An annular etching 1/2 in. in width around the edge of the sample revealed several isolated and intersecting long scratches, as shown in **Fig. 2(a)**. **Figure 2(b)**—an expanded view of the intersecting scratches marked by the arrow in **Fig. 2(a)**—shows that these





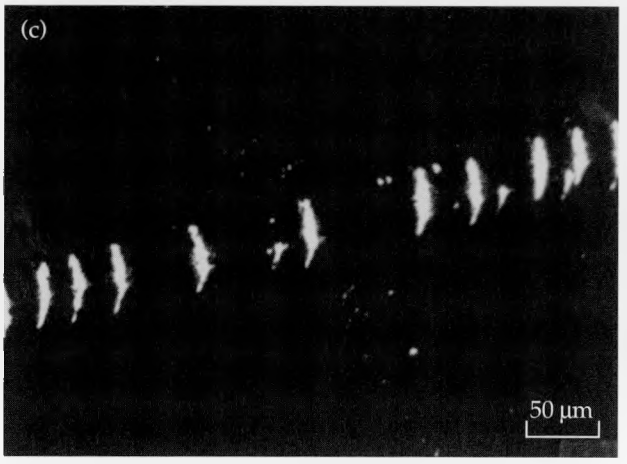


Figure 2. Examination of fused silica sample. (a) Photograph of the sample at low magnification shows scratches in the annularly etched area around the edge. (b) An expanded view in the Nomarski mode shows the intersecting scratches marked with an arrow in (a) and trails of transverse cracks; the same view is shown in (c) in TIRM mode.

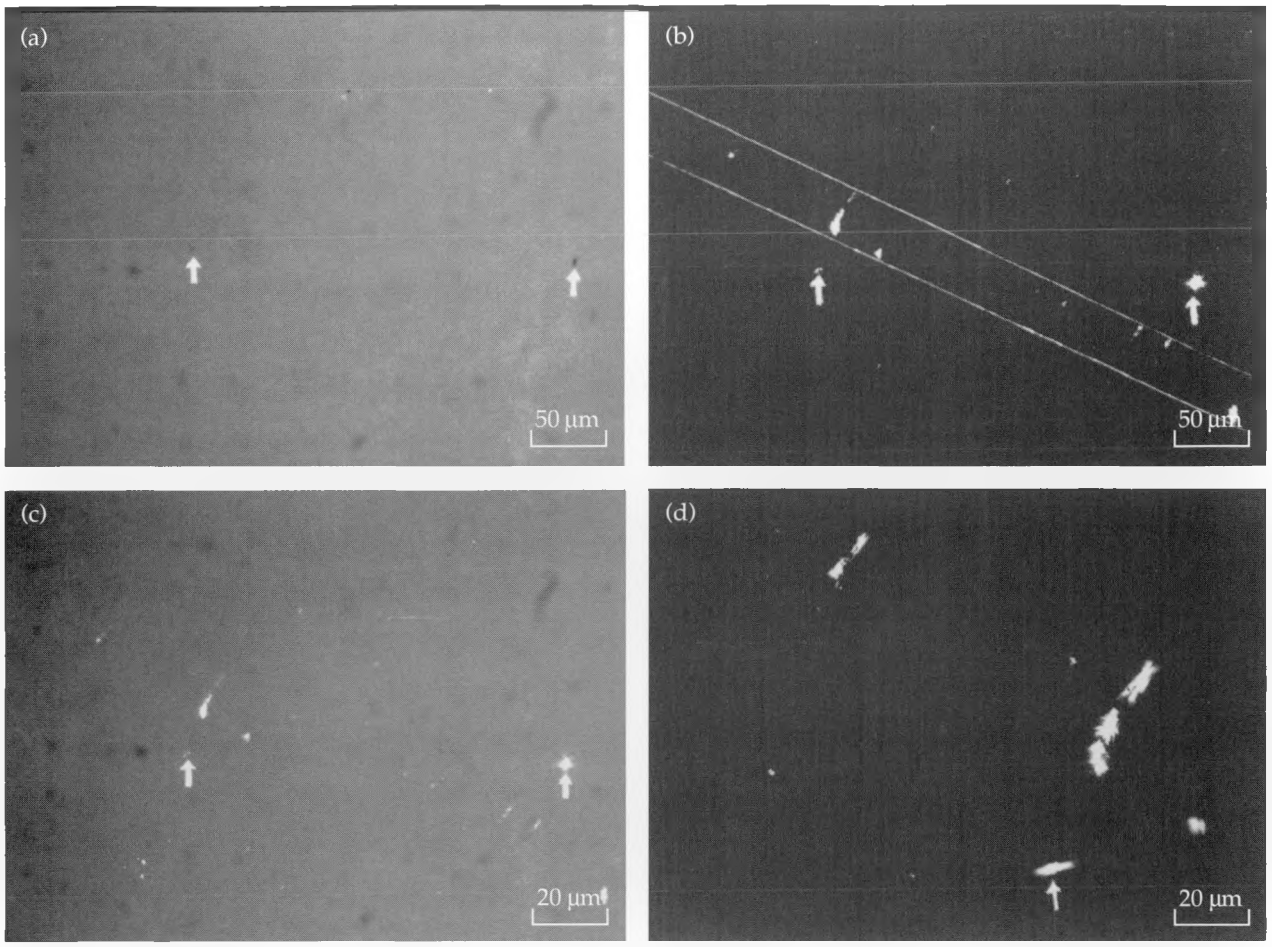


Figure 3. Examination of unetched area of same sample shown in Fig. 2. A trail of cracks is invisible in (a) Nomarski mode, and visible in both (b) TIRM and (c) Nomarski-TIRM imaging modes. The two arrows in each view represent the two small isolated artifacts or small pits serving as convenient reference points in matching views (a), (b), and (c). The trail of cracks is outlined in (b) by two parallel lines to show its width and direction. View (d) shows an expanded view of the area around the left-hand arrow in (b) at a magnification 2.5 times that of (b).

scratches are actually trails of transverse cracks that may have been formed by some oversized abrasive grits scratching and cracking the surface during the grinding and polishing process. A TIRM examination of such a trail of cracks opened to the surface by etching allowed us to establish the TIRM image characteristics of the cracks, as shown in **Fig. 2(c)**. **Figure 2(c)** also shows that the intensity of the TIRM images is rather strongly dependent on their direction relative to the laser beam path, which is from left to right in the photograph. Notice that the intensity of the crack trail running from the lower left to the upper right edge of the photograph is much stronger than is that of the trail running from the top left to the bottom right edge.

A search for subsurface cracks in the unetched central area of the same sample revealed several trails of such cracks, an example of which is shown in **Fig. 3**. **Figures 3(a)** and **3(b)** show the corresponding Nomarski and TIRM images of an identical area; notice that a trail of subsurface cracks which are visible in **Fig. 3(b)** are not visible in **Fig. 3(a)**. **Figure 3(c)** shows the same area in the simultaneous Nomarski-TIRM mode, and **Fig. 3(d)** shows the central-left area of **Fig. 3(b)** at a higher magnification. The arrows in **Fig. 3** indicate surface artifacts used as reference points in matching these micrographs, and the parallel lines in **Fig. 3(b)** outline the direction and the width of the trail of cracks between the lines.

The results shown in **Fig. 3** clearly indicate that the TIRM technique can detect and size subsurface cracks—as small as a few microns—which are not visible in the ordinary microscopy. However, the depth of such a crack, or more precisely the depth from the surface to the top of the crack and the depth from the surface to the bottom of the crack, cannot be measured from the TIRM imaging. Still, if the mechanism of the crack generation is known,

such a depth can be estimated from the surface-exposed length of the crack. If we can assume the Hertzian-type crack formation, we can say that the depth may be of the same magnitude as the length.[5]

Figure 4 shows the results of a TIRM examination of a vendor-supplied fused silica sample. While the quality of the surface of this sample seems much better than that of the sample described above, surface scratches associated with what seems to be a subsurface crack are observable. A faint scratch revealed in the Nomarski mode in **Fig. 4(a)** (not visible in this figure), is brightly revealed by TIRM mode examination, as shown in **Fig. 4(b)**. The variation of intensity and width along the scratch indicates that the image in **Fig. 4(b)** may represent a partially healed or closed crack. Two surface artifacts visible in the lower left side of **Fig. 4(a)** are clearly visible as bright spots in **Fig. 4(b)**, but other defects showing as bright spots in **Fig. 4(b)** are either barely visible or not visible at all in **Fig. 4(a)**, indicating that the TIRM technique may be an efficient means of inspection for surface artifacts and cleanliness. The dark-field technique available in some optical microscopes can enhance the visibility of surface scratches, but the TIRM technique seems far superior in terms of image quality.[6]

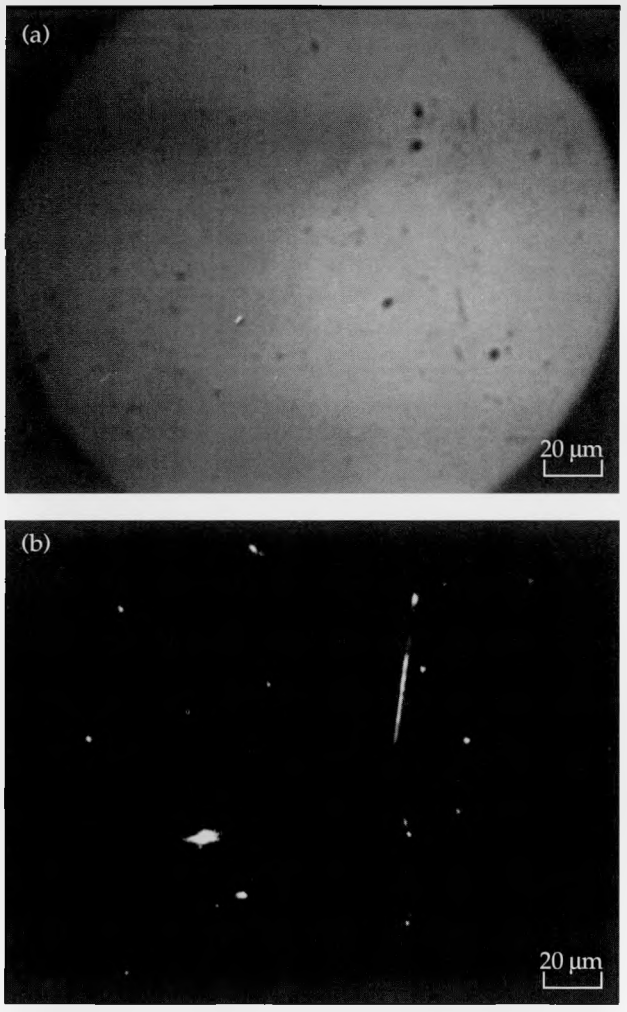


Figure 4. Examination of a vender-supplied, unetched fused silica sample. Views (a) and (b) respectively show a Nomarski micrograph and a TIRM micrograph of the sample. A faint scratch revealed in Nomarski mode (not visible) in (a) shows as a bright image in TIRM mode in (b). The intensity and width variation of the scratch in (b) indicate the possibility of a subsurface crack.

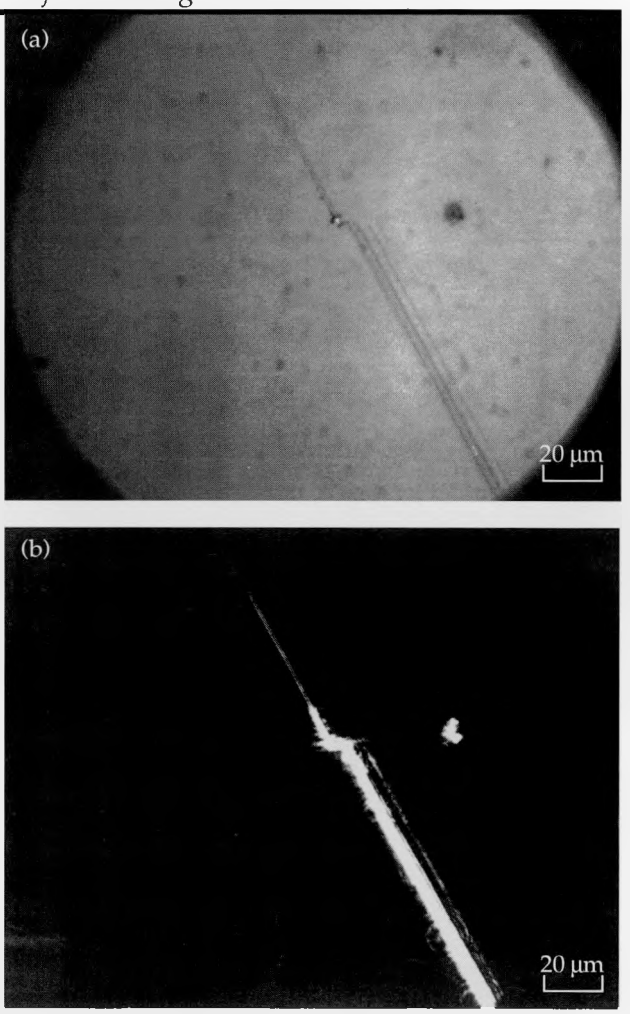


Figure 5. Examination of a vendor-supplied, multilayer, coated fused silica sample. Views (a) and (b) respectively show a Nomarski micrograph and a TIRM micrograph of the sample. Parallel scratches in the lower half of each view indicate a subsurface crack or delamination at the substrate-coating or interlayer interface.

Figure 5 records the examination of a vendor-supplied fused silica sample with a multilayer coating. **Figure 5(a)** shows parallel scratches hinged on a surface defect viewed in Nomarski mode; the corresponding TIRM image in **Fig. 5(b)**

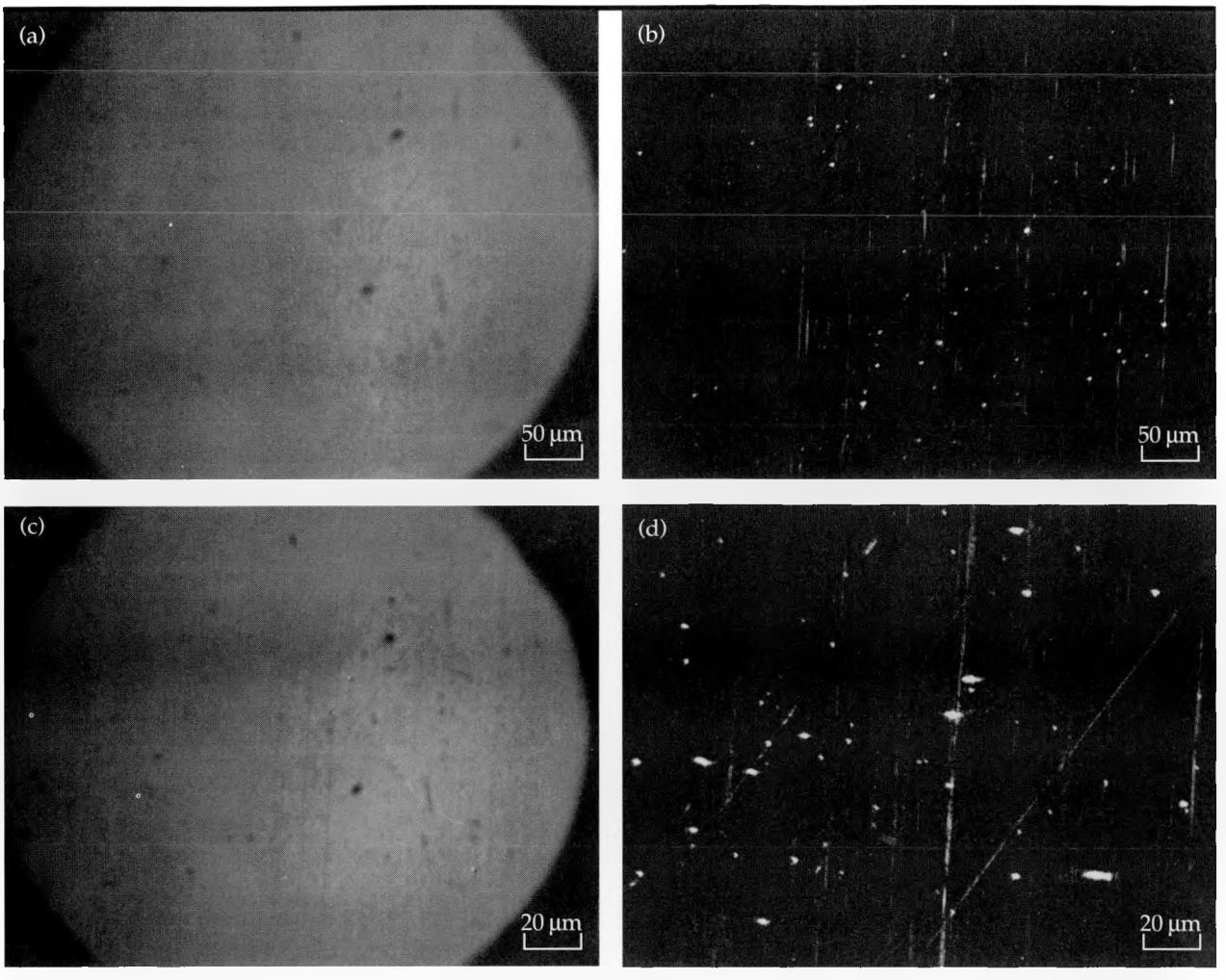


Figure 6. Examination of a ductile-ground BK-7 glass sample. Views (a) and (c) show Nomarski micrographs and (b) and (d) show TIRM micrographs of the sample. Views (c) and (d) show the center-right area of views (a) and (b) at 2.5 times magnification. Notice that in (d) a second grinding direction appears, indicating that a higher magnification reveals more details.

indicates that the coating within the parallel scratches in the lower side may be either connected to the subsurface crack or delaminated at the substrate-coating or interlayer interface.

Figure 6 records the examination of a ductile-ground BK-7 sample. **Figure 6(a)**—a Nomarski image of a surface area—reveals a few surface artifacts and barely visible grinding marks, while the TIRM image in **Fig. 6(b)** clearly shows the grinding marks. **Figure 6(c)** represents the center-right area of **Fig. 6(a)** at a higher magnification, and **Fig. 6(d)** is the corresponding TIRM image. In **Fig. 6(d)**, a second variant of grinding marks is visible. This result demonstrates the effectiveness of using the highest magnification possible in TIRM mode, as the objective lens is placed closer to the surface of the sample and thus captures more scattered light. (Alternatively, an objective lens with a shorter working distance may have the same light-capturing effect.)

The bright spots both on or off the grinding marks in **Figs. 6(b)** and **6(d)** could represent surface pits, embedded abrasive grits, or surface contaminants. Repeated cleaning eliminated some of the spots but also added new ones, indicating that some of them are really dust particles or cleaning artifacts. More sophisticated surface analysis, such as with a high-resolution/high-magnification scanning electron microscope and/or scanning Auger electron spectrometric microchemical analysis, may be able to identify these particles.

PBS Investigations

We supplied several optics samples to XMR, Inc., for subsurface damage characterization by nondestructive PBS and destructive etching and decoration techniques. The samples included a ductile-ground BK-7 glass, three vendor-supplied samples (one fused silica, one fused silica with

coating, and one zerodur), and three fused silica samples with different levels of known grinding damage, all having a surface finish of less than a nanometer. The BK-7 sample was the same one previously examined by TIRM (**Fig. 6**). None of the other samples had been so examined, though other specimens from the vendor-supplied batch had been, as shown in **Figs. 4** and **5**.

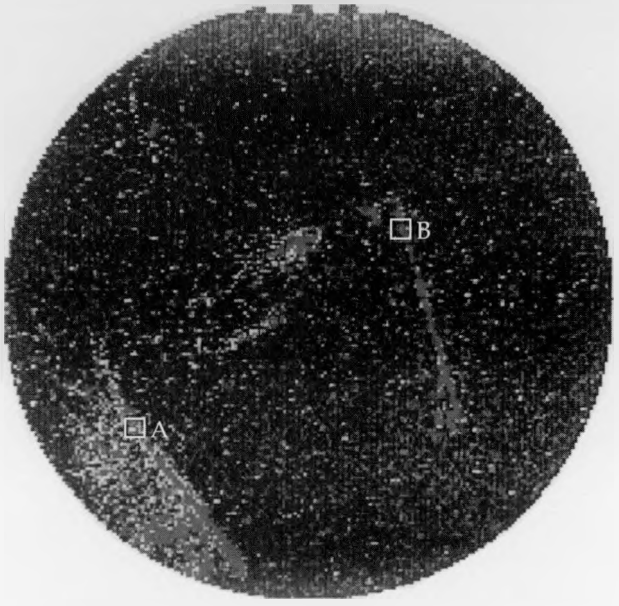


Figure 7. PBS map of a vendor-supplied coated fused silica sample. The map represents the 1-in.-diameter central area of the sample. The patterns of light contrast imply surface or subsurface damage or defects. (Areas marked "A" and "B" correspond to Fig. 10.)

The PBS investigation of the BK-7 sample showed an order-of-magnitude higher average light scattering than did other samples while exhibiting the directionality of the grinding marks, confirming the Nomarski and TIRM results. The vendor-supplied uncoated fused silica sample showed a very low average scattering, but its grinding direction was clearly discernible, indicating that subsurface damage, perhaps in the form of point defects or residual stress, exists.

Examination of the vendor-supplied coated fused silica sample, shown in **Fig. 7**, revealed a few clear macroscopic patterns in its PBS map which may represent some damage in the coating. The zerodur sample, when PBS scanned, could not be mapped, as the scattering was too strong. When TIRM examination was subsequently attempted on another zerodur sample, the light scattering was so strong that imaging was almost impossible, confirming the PBS result.

With PBS, the direction of the grinding in each of the three fused silica samples with the three different levels of grinding damage showed very clearly. The PBS map of the sample with the medium level of grinding damage is shown in **Fig. 8** for comparison with the map of the sample with the least level of damage, which is shown in **Fig. 9**. This latter sample had been polished longer than

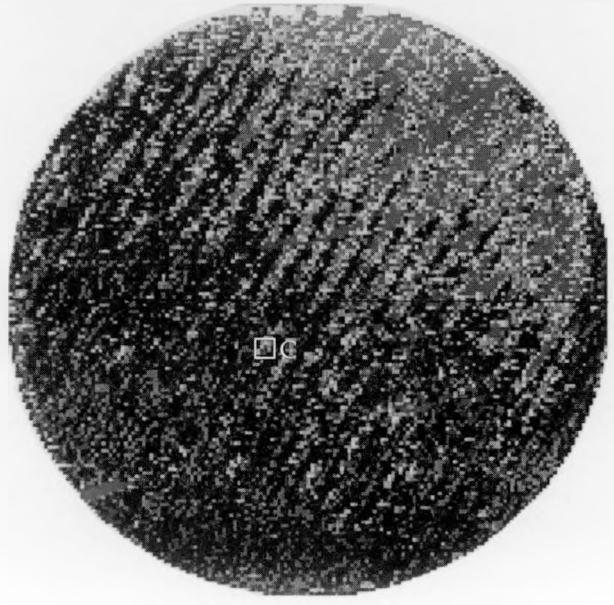


Figure 8. PBS map of a fused silica sample with medium level of grinding damage. Notice the grinding directionality. (Area marked "C" corresponds to Fig. 11.)

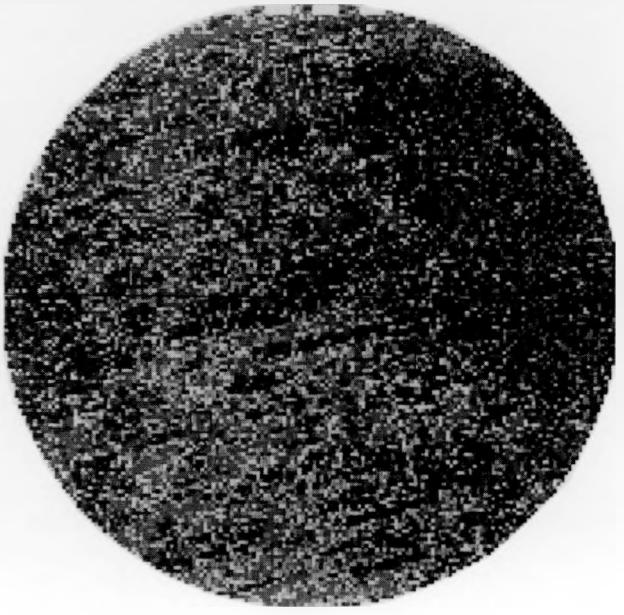


Figure 9. PBS map of a fused silica sample with the lowest level of grinding damage. The grinding directionality is still noticeable.

the sample shown in **Fig. 8** in an attempt to remove the grinding damage, yet its map (**Fig. 9**) clearly exhibits the direction of grinding, and the average level of the light scattering is higher than that of the medium-level sample (**Fig. 8**). These results indicate that the attempt to remove the grinding damage might actually increase the surface and subsurface damage.

It is not clear whether the grinding directionality patterns revealed in the PBS maps represent the long-range surface undulation in surface roughness or the subsurface damage structure, as it is difficult to separate the two effects. It may be necessary to measure the continuous surface profile, possibly over a 1-in.-square area, to identify the existence of such a long-range undulation. The etching and decoration techniques may then be able more clearly to reveal the nature of the patterns in the PBS maps.

Etching and Decoration

Following the PBS investigations, the BK-7 glass sample and the three vendor-supplied samples (fused silica, fused silica with coating, and zerodur) were etched on one half of the top surface and carbon decorated on the other half. The three fused silica samples with three different levels of grinding damage were not etched but were carbon decorated on the top surfaces. All of the samples were then examined by optical microscopy. The etching revealed subsurface damage in the form of clearly visible scratches and small pits in the areas of the samples corresponding to the high-intensity light-scattering areas in the PBS maps. The carbon decoration, with those samples whose PBS maps showed strong directionality of grinding, yielded a rather well aligned pattern of carbon particles indicating those grinding directions. In those samples (or sampling areas) whose PBS maps showed low-intensity light scattering, the carbon decoration generated rather fuzzy and subtle patterns which were difficult to recognize and interpret.

Figure 10 shows the etching on the vendor-supplied coated fused silica sample whose PBS map is shown in **Fig. 7**. The view shown in **Fig. 10(a)** corresponds to the area of high-intensity PBS marked "A" in **Fig. 7**. The etching reduced the thickness of the coating layer by about 10 μm, but obviously did not completely remove the coating, as the coating microstructure is still visible. Long scratches and small pits visible in **Fig. 10(a)** may be the remainder of a deep scratch or crack that may have been in the coating or in the substrate. **Figure 10(b)** shows the result of the decoration technique on the area of high-intensity PBS marked "B" in **Fig. 7**.

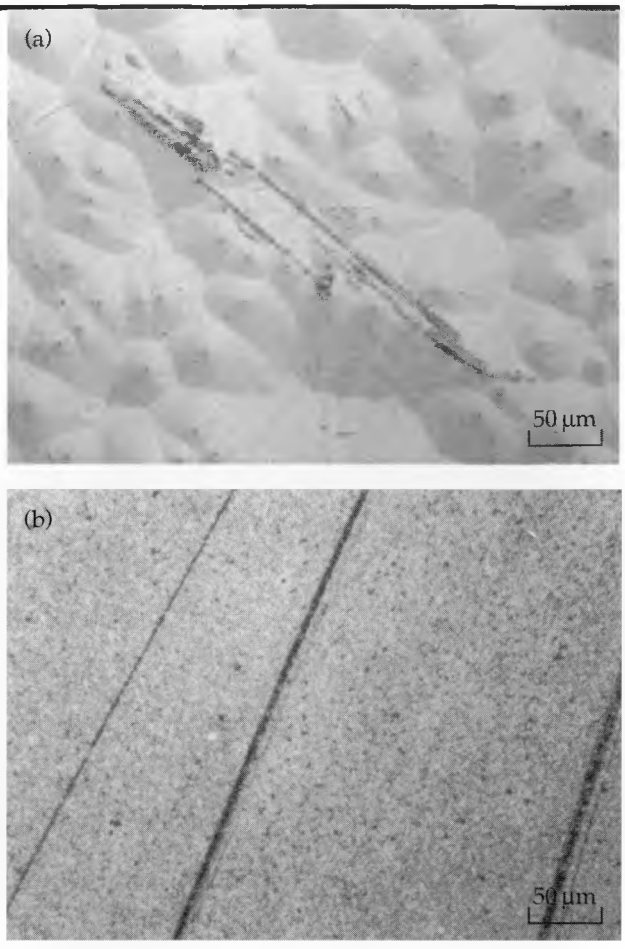


Figure 10. Nomarski micrographs of a coated fused silica sample. View (a) shows the sample after etching and corresponds to the area marked "A" in Fig. 7. View (b) shows the sample after carbon decoration and corresponds to the area marked "B" in Fig. 7.

The long and sharp line seems to indicate a long surface scratch or a long crack open to the surface rather than subsurface damage. **Figure 11** represents the carbon decoration performed on the fused silica sample whose PBS map is shown in **Fig. 8**. The area shown in **Fig. 11** corresponds to the area of high-intensity PBS marked "C" in **Fig. 8**. A rather faint decoration pattern of parallel curved bands visible in **Fig. 11** seems to indicate the grinding direction.

Conclusions

Our TIRM and PBS investigations yield the following conclusions:

- TIRM technique can detect subsurface cracks and measure their surface-projected sizes, but it cannot measure their depths from the surface and their submergent depths.
- TIRM enhances the visibility of surface damages

and artifacts and can be used as an effective technique to examine the surface for cleanliness.
- PBS can detect the global distributions of fabrication-induced surface and subsurface damages.
- Etching and decoration techniques can confirm the nature of the subsurface damages detected by TIRM and PBS techniques.

Future Work

More samples, including crystalline materials with differing fabrication histories, will be examined by TIRM and PBS to build up the database of subsurface damage characteristics and to further refine the TIRM technique. TIRM and PBS monitoring of the evolution of subsurface damage during the fabrication process of ground and polished fused silica or BK-7, or diamond-turned KDP (potassium dihydrogen phosphate) and other crystalline materials, will be performed to find a way to minimize or eliminate the subsurface damages.

A laser with a higher power output than that used in the present study is required for detection of fine defects and damages. If such a laser is obtained, the effects of the polarization of the incident laser beam described by Temple[3] and Jabr[6] will be investigated further to aid in characterizing submicron-sized defects and damages. Development of an inexpensive TIRM-type instrument for inspection and screening of optics as large as a few feet in diameter also is foreseeable. Such an instrument must be capable of fast scanning of the surface of the optics, and its development may be advanced by the experience gained in the laser scanning of platinum inclusions in the laser glass.

A natural extension of the TIRM technique may be the scanning optical tunneling microscope (SOTM), the basic concept of which has been around for a long time. The possibility of such an instrument has been greatly strengthened by the recent development of a scanning tunneling microscope. An SOTM will use the same total internal reflection mode of a laser beam incident from the underside of the surface to be scanned. The s-polarization of the incident laser beam will generate the surface evanescent wave, which will be tunneled by an optical stylus scanning the surface, while the scattering loss by the surface and subsurface defects will reduce the intensity of the tunneled light. Thus, a kind of bright field image of the surface and subsurface defects will be formed. On the other hand, the p-polarization component of the incident beam will not generate the evanescent wave; thus, the optical stylus will pick up only the scattered light from the same defects, forming a dark field image of them. It is difficult to predict the quality of the dark field images of an SOTM in comparison with the TIRM images, but the bright field images, with their potentially high resolution, make the SOTM an extremely attractive research subject as an alternative technique to TIRM.

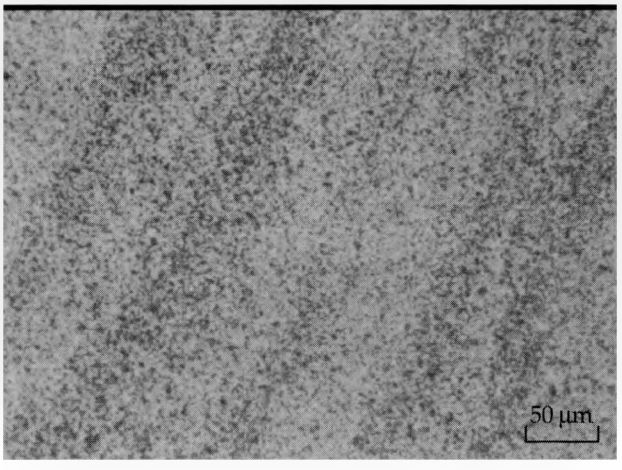


Figure 11. Micrograph, after carbon decoration, of a fused silica sample with medium level of grinding damage. This view corresponds to the area marked "C" in Fig. 8.

Acknowledgments

The author thanks Tony Demeris, who set up the TIRM attachment, and Ken Blaedel, John S. Taylor, and John R. Taylor, who shared the cost of the XMR, Inc., contract. Encouragement from Ted Saito and Satish Kulkarni is appreciated.

1. J. E. Marion, "Fracture of Solid State Laser Slabs," *J. Appl. Phys.* **60**, 69 (1986).
2. N. Bloembergen, "Role of Cracks, Pores, and Absorbing Inclusions on Laser Induced Damage Threshold at Surfaces of Transparent Dielectrics," *Appl. Opt.* **12**, 661 (1973).
3. P. Temple, "Total Internal Reflection Microscopy: A Surface Inspection Technique," *Appl. Opt.* **20**, 2656 (1979).
4. T. J. Magee and C. S. Leung, "Overview of Subsurface Damage in Brittle Materials—Evolution of Concepts and Current Status," in *Proceedings of Subsurface Damage in Glass* (American Society for Precision Engineering, 1989 Spring Topical Meeting, April 25–27, 1989, Tucson, Ariz.), p.15.
5. N. J. Brown, private communication based on B. Lawn and R. Wilshaw, "Indentation Fracture: Principles and Applications," *J. Mater. Sci.* **10**, 1049 (1975).
6. S. N. Jabr, "Total Internal Reflection Microscopy: Inspection of Surfaces of High Bulk Scatter Materials," *Appl. Opt.* **24**, 1689 (1985).

Computerized Tomography

Harry E. Martz
Engineering Sciences Division
Mechanical Engineering

Daniel J. Schneberk
Computations Division
Applications Systems Division

Stephen G. Azevedo
Engineering Research Division
Electronics Engineering

Michael F. Skeate and George P. Roberson
Nuclear Energy Systems Division
Electronics Engineering

We are developing several tomographic data-acquisition scanners along with associated computational techniques for image reconstruction, analysis, and display. Our goal is to nondestructively evaluate the internal structure of industrial objects associated with a broad spectrum of Laboratory-wide programs.

Introduction

Computerized tomography (CT) provides nondestructive two- and three-dimensional internal views of materials, components, and assemblies. Our primary concern is to build scanners to acquire CT data, and from these data to reconstruct an accurate representation of an object's internal structure. The required spatial and contrast resolutions are set by the specific application. Other relevant issues are minimizing the data acquisition time and organizing the computational job of reconstructing, displaying, and analyzing the CT image. The overall goal of our research is to improve three generalized performance parameters that characterize tomographic imaging systems:

- Spatial resolution—the sizes of various defects (flaws, cracks, voids, inclusions, etc.) detectable in the object being inspected. We would like to detect extremely small defects.
- Contrast resolution—the discernible density/elemental composition differences of various subregions in an object under inspection. We would like to display very slight differences in density and/or elemental composition.
- System speed—the time associated with data acquisition, processing and analysis, and display. We would like to gather, process, and display data very rapidly.

These three performance parameters can be visualized at the corners of a triangle with a central pivot (**Fig. 1**). This shows that in any given computerized tomography system, an improvement in any two parameters results in a worsening of the third.

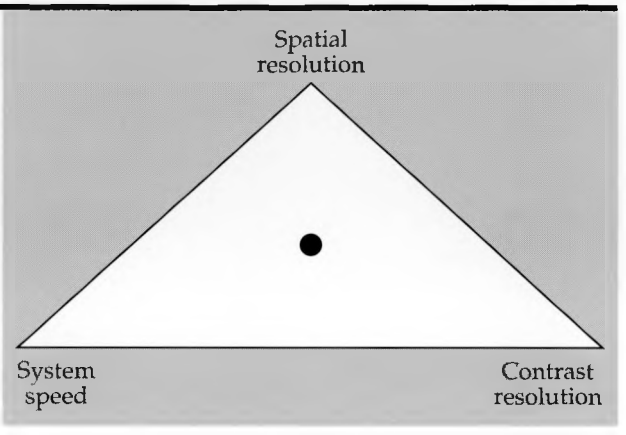


Figure 1. The three most important performance criteria, visualized at the corners of a triangle with a central pivot. This graphically reveals the fact that an increase in any two performance factors results in the decrease of the third.

Physical factors that limit our ability to improve these parameters include the randomness inherent in photon-counting, finite source and detector dimensions used in CT scanners, measurement noise, beam-hardening caused by energy-dependent absorption of polychromatic x-ray sources, scattering of the beam, limitations of the reconstruction algorithms, and inexact arithmetic used in computer programs.[1] Object-dependent factors that also affect our ability to improve these parameters include the object's physical dimensions, large differences in opacity within a single sample, and differences in elemental composition.

In addition, we are addressing several other research issues:

- Elemental or effective-Z imaging, based on physical parameters that are obtained by com-

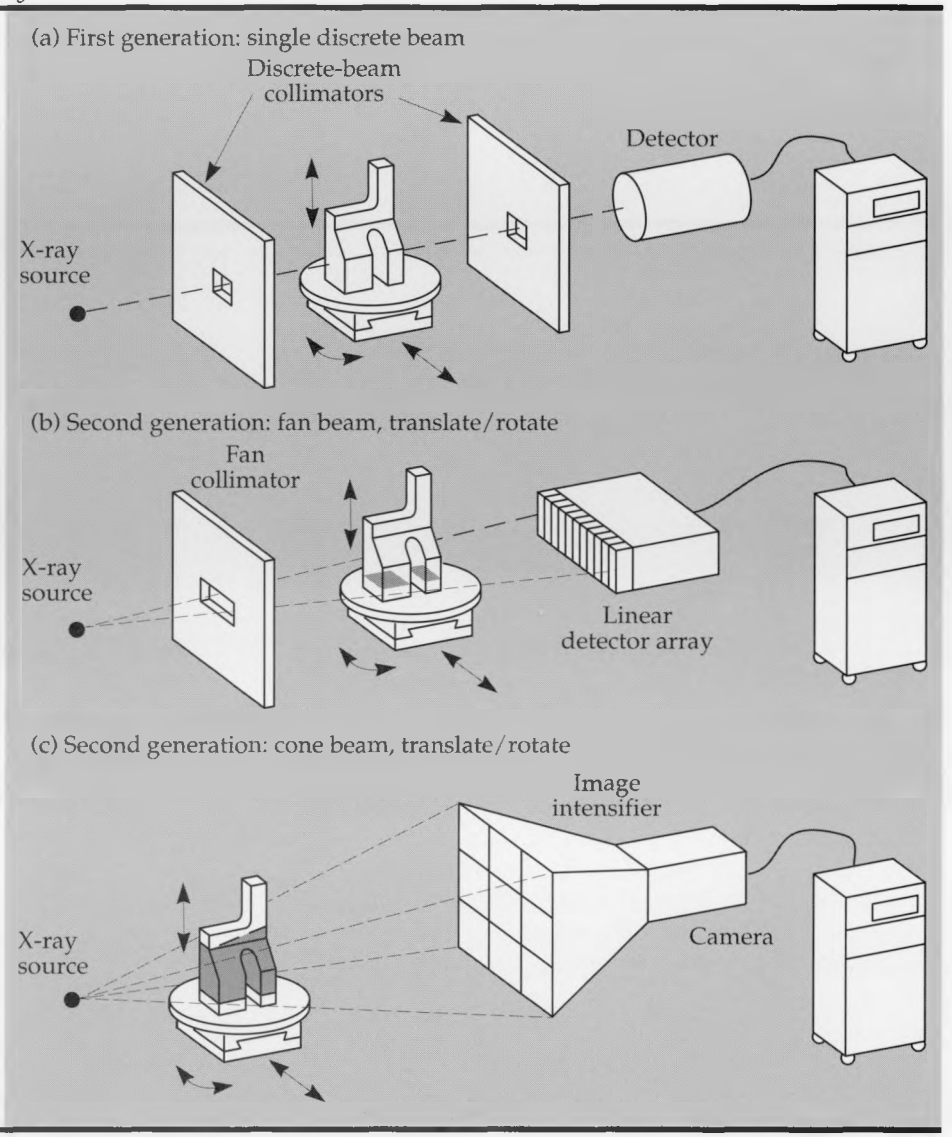


Figure 2. CT data acquisition geometries used in NDE: (a) Discrete-beam first generation translate-rotate design, (b) fan-beam design using a linear detector array, and (c) cone-beam design using a two-dimensional detector array. Both the second and third generation scanners can acquire data using either fan or cone beams. The primary difference between second and third generation scanners is in the latter there is no transverse or x-translational steps; the object or source/detector is only rotated.

paring two images recorded at two different energies.

- Model-based imaging using a priori information to improve image reconstruction from both complete and incomplete CT data sets.
- Parallel processor architectures for image reconstruction, through a joint research project with UC Davis.
- Scientific visualization of reconstructed data, exploring novel methods of rendering, displaying, and analyzing data.

In this report, we will describe the major progress achieved to date in our Laboratory-built CT scanners and some of the research issues involved in their continued development. We will also present the reconstruction algorithms and the advanced analysis and display facilities that we use to interpret and present our results. Last, we will indicate our future plans in the field of computerized tomography.

Progress

In this section, we describe our current CT scanners and some of the research issues involved in their continued development, as well as advances we have achieved in image reconstruction, display, and analysis. We also examine an application of one of our scanners (PBCAT) involving a novel technique for spatial discrimination of atomic number in a material.

CT Scanners

Commercial CT scanners built specifically for industrial NDE use have been available only since 1984.[2] These systems are expensive and are usually built to inspect specific objects. However, NDE problems at LLNL span a wide spectrum of object size, shape, and resolution requirements. Thus we

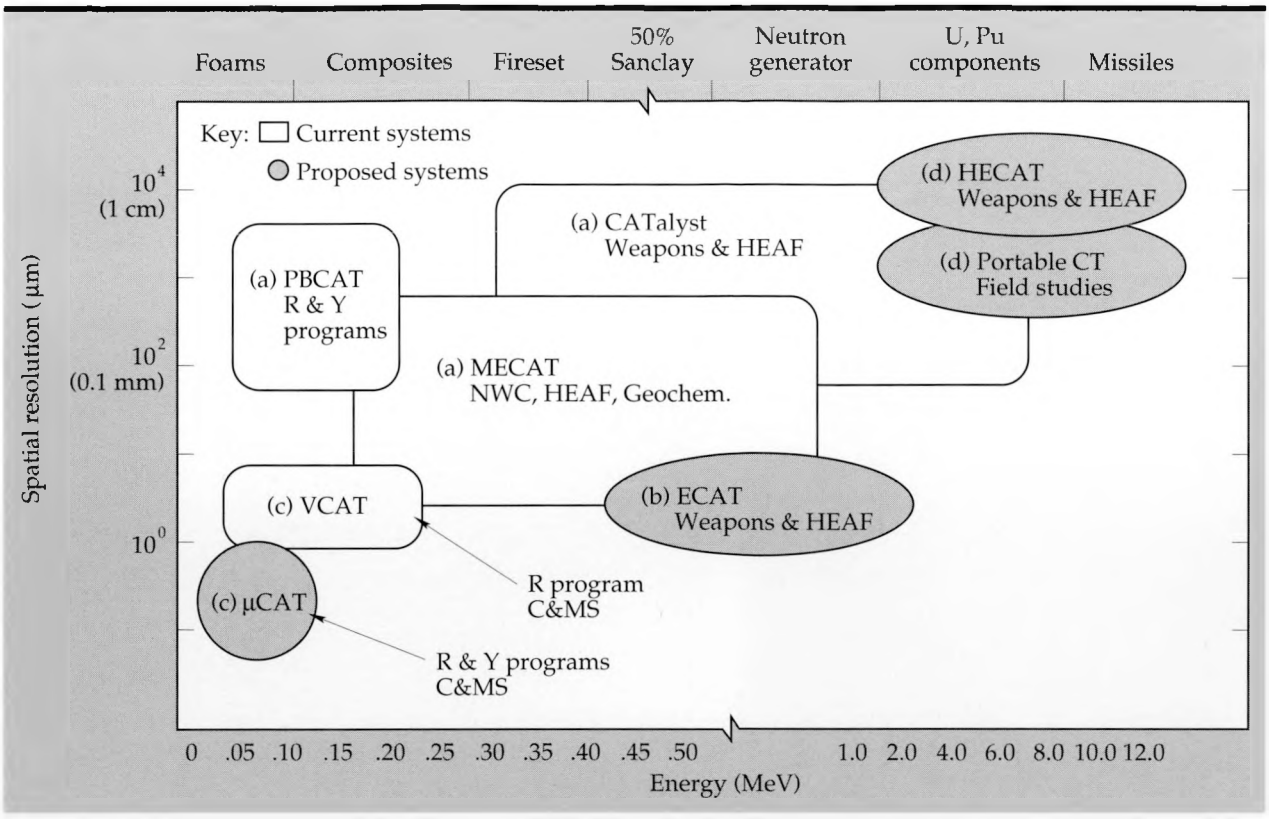


Figure 3. A schematic plot showing the current and future spatial resolution performance vs x-ray energy for our x-ray CT scanners: (a) three discrete-beam systems: Pencil-Beam CAT (PBCAT), a Medium-Energy CAT (MECAT)and a Transportable CAT scanner (CATalyst); (b) a fan beam system with a linear detector array: Explosive CAT (ECAT); (c) two CCD-based fan/cone-beam systems: Video CAT (VCAT) and MicroCAT (μCAT); and (d) a video-based fan/cone-beam system: High-Energy CAT (HECAT). The programs most likely to use these scanners are listed. In addition, possible applications are given along the top of the figure.

have sought to gain an expertise in CT while building a number of scanners internally to satisfy most LLNL CT imaging requirements.

The research and development effort in the NDE Section at LLNL over the last two years has succeeded in improving our CT imaging capabilities. Our CT project work has centered on three basic CT approaches, namely the first generation (pencil-beam), the second generation (fan-beam), and the third generation (fan- and cone-beam) scanners shown in **Fig. 2**. A more detailed description of these types of scanners was given in the 1987 Thrust Area Report and elsewhere.[3] We are now building seven x-ray CT scanning systems in our laboratories. These include three discrete-beam systems: Pencil-Beam CAT (PBCAT), a Medium-Energy CAT (MECAT); and a Transportable CAT scanner (CATalyst). We are also constructing a fan-beam system with a linear detector array: Explosives CAT (ECAT);* and two CCD-based fan/cone-beam systems: Video CAT (VCAT) and MicroCAT (μCAT); as well as a video-based fan/cone-beam system: High-Energy CAT (HECAT). These x-ray scanners are in various stages of design and assembly. Each system is constructed in a modular fashion so that improvements in any one component (e.g., better detector technology) can be easily retrofitted. The current resolution and energy capabilities of both our current and planned systems are shown in **Fig. 3**. In this report, we will discuss the three scanners that are closest to completion (PBCAT, VCAT, and ZCAT†) and provide data from each.

Pencil-Beam CAT (PBCAT)

The PBCAT system uses the first-generation translate/rotate CT design in which the object is

* The ECAT scanner is being built with funds from the Pantex/LLNL joint Project. This Project is using our CT scanners to study high explosives.

† The ZCAT scanner is not mentioned above since it was a temporary CT scanner built with borrowed equipment from MECAT and Real-Time Radiography to perform programmatic CT experiments.

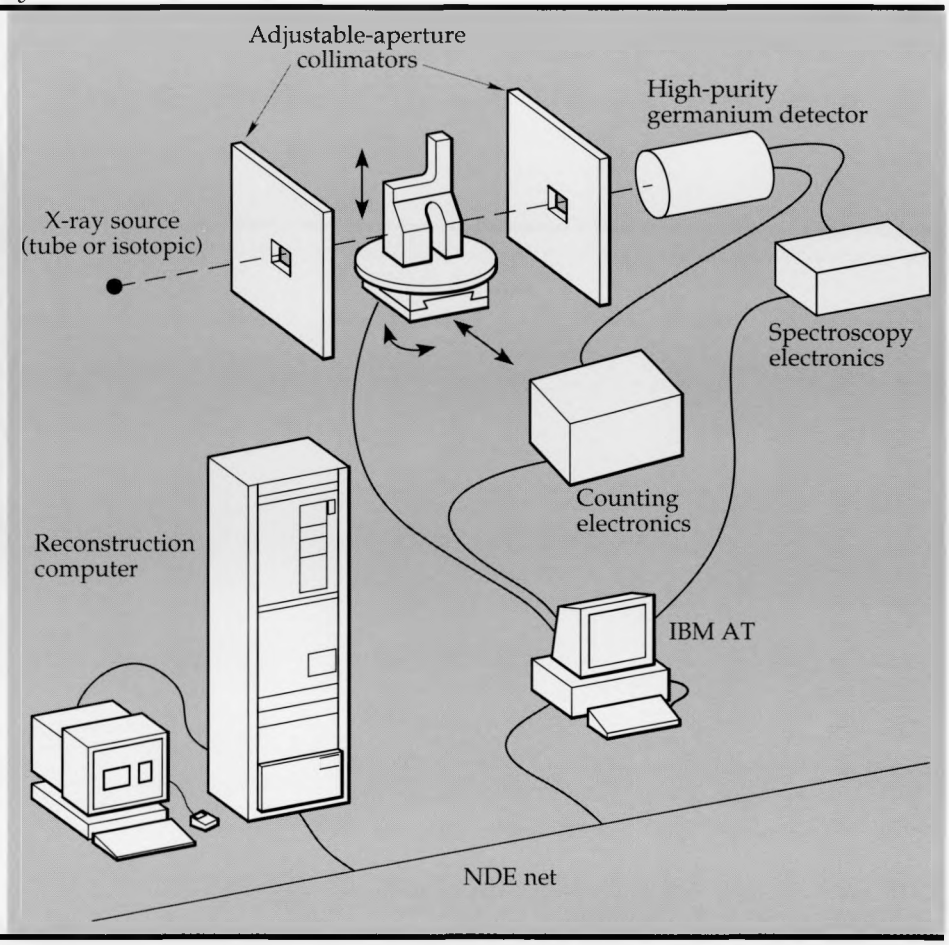


Figure 4. Schematic diagram of the PBCAT scanner showing the x-ray source, staging, detector system, and data acquisition and reconstruction computers. PBCAT uses the discrete-beam first generation translate/rotate design. The translational steps and rotational steps are obtained by movement of the object.

moved instead of the source-detector pair (see **Figs. 4 and 5**). Currently we are using radioisotopic sources such as ^{109}Cd (22, 25, and 88 keV) and ^{241}Am (60 keV). We can also use a rhodium anode microfocus (20- and 40-μm spot sizes) x-ray tube source for higher intensities at energies from 20–125 kVp. The detector is high-purity intrinsic germanium with a transistor-reset preamplifier. The data is acquired using standard x-ray spectroscopy hardware interfaced to an IBM PC. Further details on the PBCAT scanner hardware and software are given elsewhere.[4,5]

The main advantages of PBCAT are the lack of beam-hardening (for a monochromatic source) and the scatter rejection due to the well-collimated source and detector. The main problem with the system, particularly with monochromatic isotopic sources, is the time required for data acquisition. The smallest useful source and detector aperture size using either the ^{109}Cd or the ^{241}Am source is 0.25 mm × 0.25 mm, which yields count rates of roughly ten counts per second. At this rate an image with approximately 250 μm spatial resolution can take several days to acquire. The use of the microfocus x-ray tube solves this problem, giving us high count rates with even smaller collimators. However, the price is the introduction of beam-hardening errors due to the polychromatic x-ray beam and energy resolution.

The system can be used with x rays between 6 and 125 keV. It is mounted in a cabinet that limits the object size to about 0.1 m^3 (see **Fig. 5**). Quantitative CT images revealing the interior details of several objects of programmatic interest with spatial resolutions in the range of 250 μm have been achieved.[4–6] These objects include but are not limited to R-Program materials, weapon parts for LLNL and Sandia, and high explosives for Pantex. A representative CT image is shown in **Fig. 6**.[5]

Video CAT (VCAT)

VCAT is a camera-based CT scanner designed to examine low-density materials at high spatial resolutions. A block diagram of the VCAT system is shown in **Fig. 7** and a photograph is shown in **Fig. 8**. The x rays are provided by a microfocus x-ray tube (25–160 kVp, 100 W) which has a tungsten anode and a spot that can be focused down to 10 μm. A collimator mounted to the source tubehead can be adjusted to illuminate the test object with either a fan beam or cone beam of x rays. The test

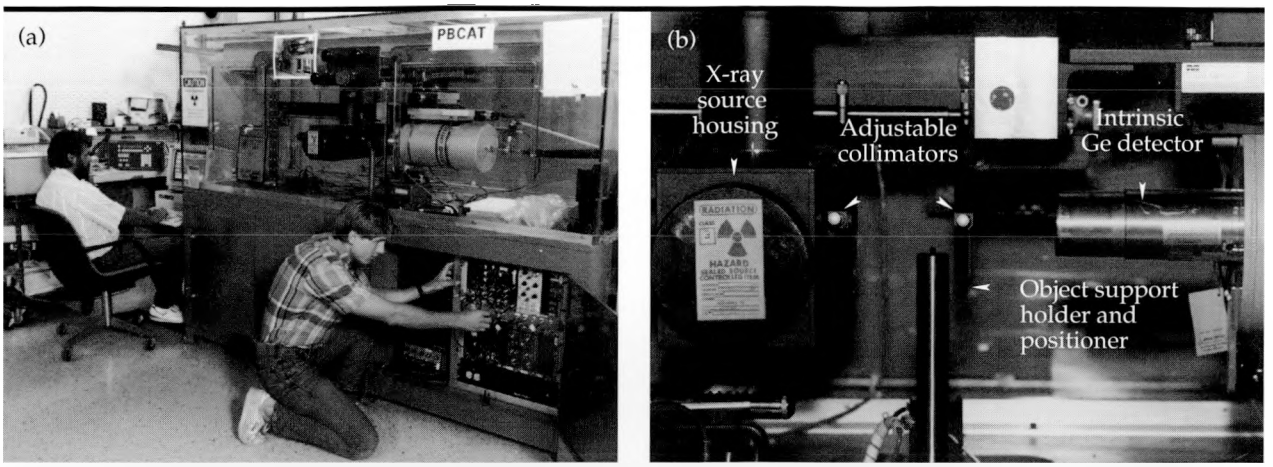


Figure 5. A photograph of the PBCAT scanner. (a) The entire scanner apparatus. (b) A closeup of the x-ray source housing (left), object holder (center rod), and detector (right). Double slit collimators are used on both the source and detector, and are adjustable from 0 to 3 mm² using micrometer calipers.

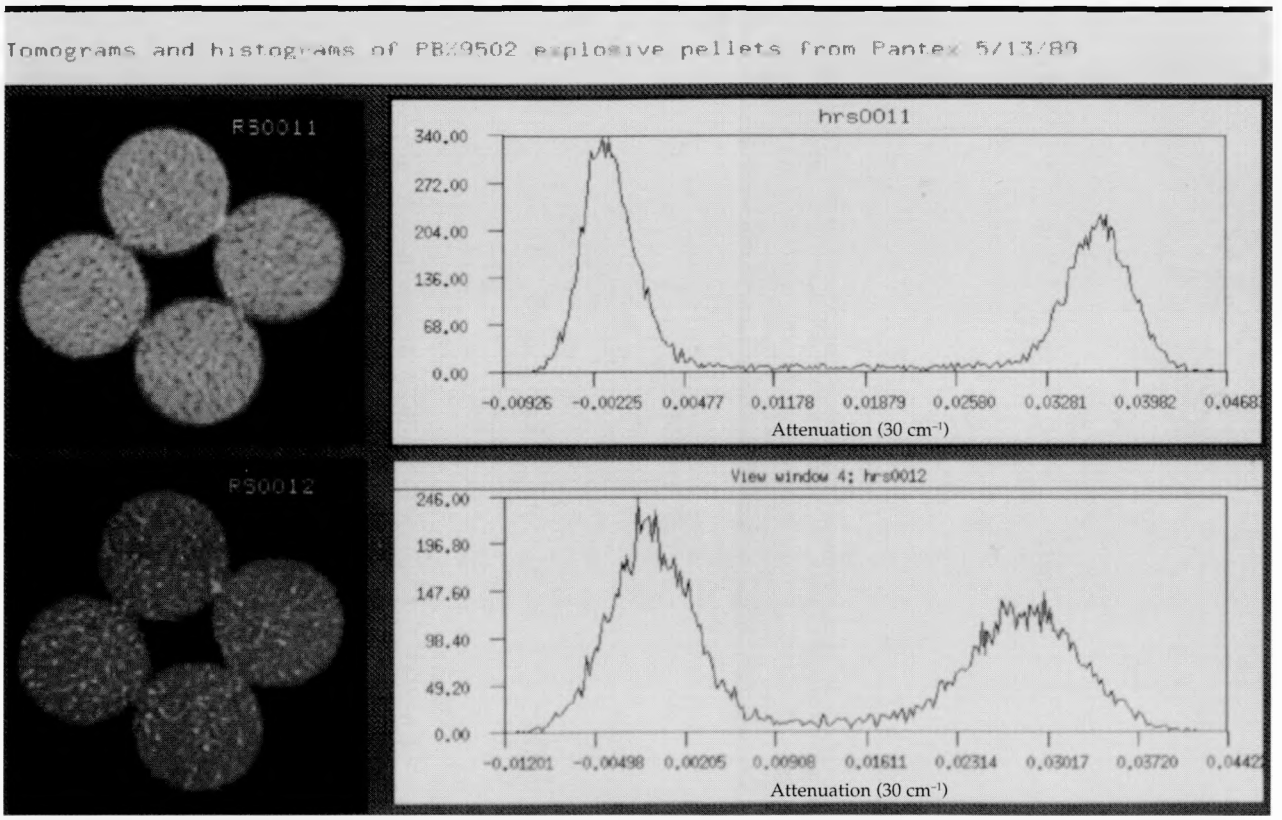


Figure 6. Reconstructed images of four small cylindrical PBX9502 high explosive pellets, each 0.5 in. diameter and 0.5 in. long. In the graphs, the vertical scale gives the number of values recorded at a given attenuation, while the horizontal scale indicates attenuation coefficient (30 cm⁻¹). The data was acquired at the center slice plane using a ^{109}Cd radioisotopic source. The 22-keV data is shown on the top left and the 25-keV data on the bottom left. The distribution of shading in the image represents attenuation variations in the pellet, which are related to changes in density and atomic number. The histogram of each image is shown to the pellet's immediate right. The histograms are a quantitative measure of the contrast resolution of our CT scanner.

object is supported by movable staging which allows five degrees of freedom: rotation (0.2 arcminute resolution); x, y, and z translation (2-μm resolution); and tilt. These degrees of freedom allow CT scans to be performed using either a second generation ("translate/rotate") or third generation ("rotate only") scan geometry. The entire system is controlled by a Sun computer workstation which

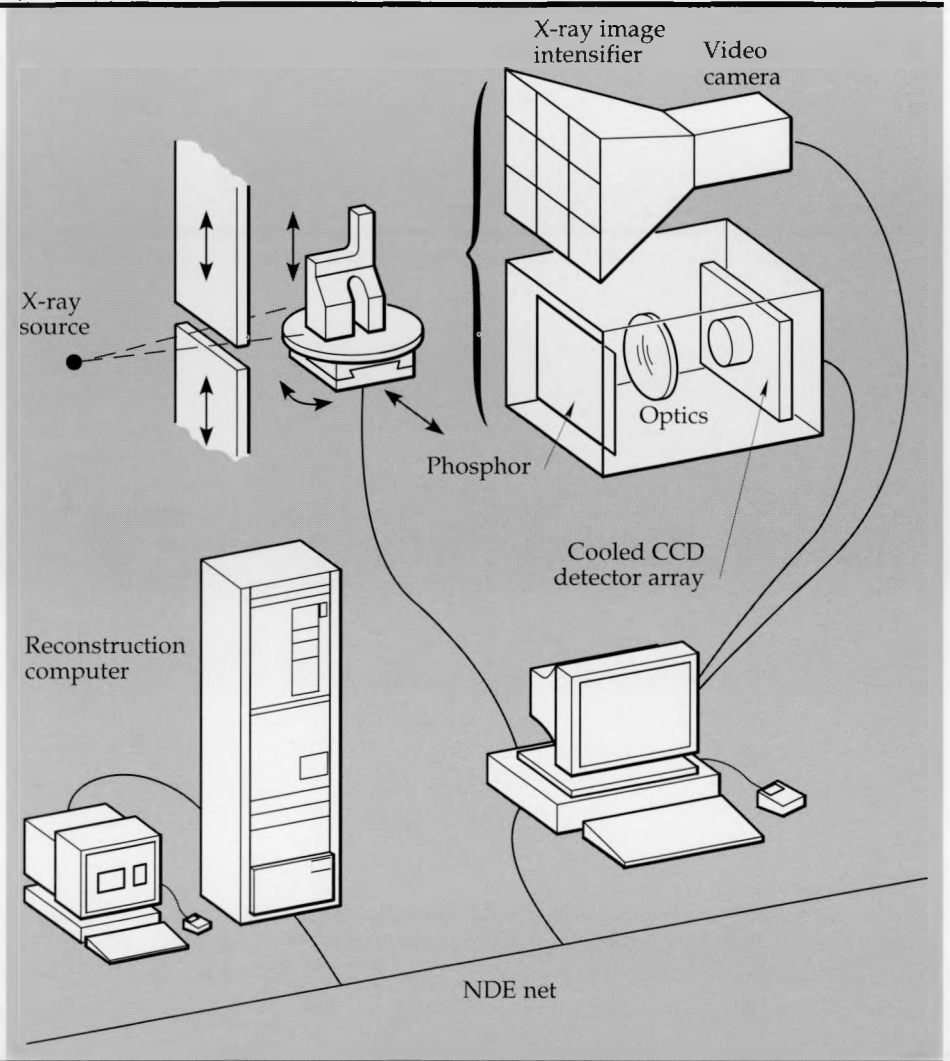


Figure 7. Schematic diagram of the VCAT scanner. VCAT uses two different detector systems to provide a flexible scanning capability. To quickly (15–20 min) acquire images with medium-spatial resolution (250 μm), a video-rate CCD camera/image intensifier detector system is used (top detector). When high-spatial resolution (100 μm) is required, a second detector system must be used. The high-resolution detector consists of a thin-rectangular scintillator and a light-integrating CCD camera (bottom detector). Two types of scintillating materials have been used: an optically polished cadmium tungstate crystal, and a 15-μm diameter glass fiber optic strand doped with terbium oxide and bundled into a plate.

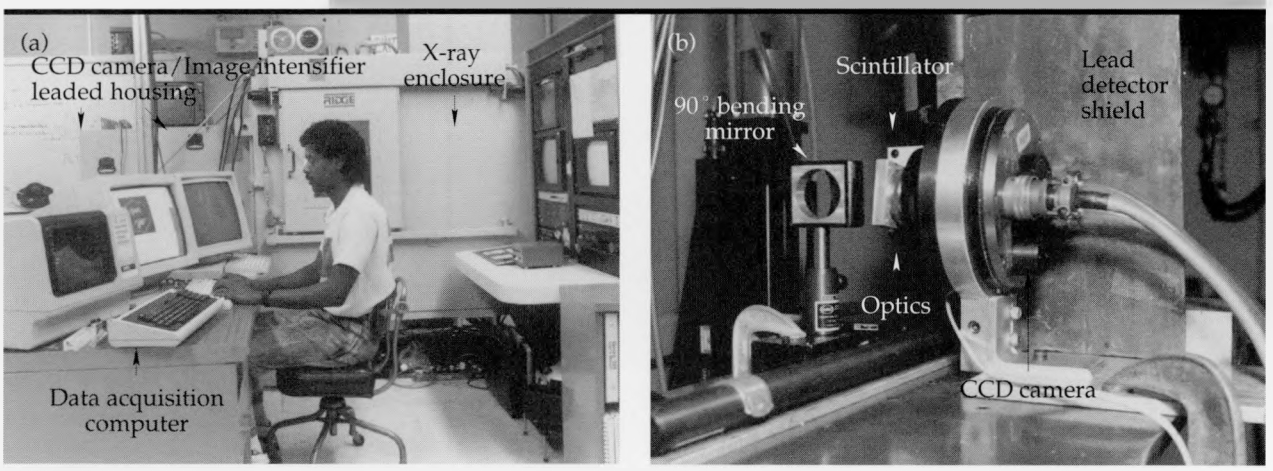


Figure 8. Photographs of the VCAT scanner. (a) The entire scanner apparatus. (b) A closeup of the CCD-camera/scintillator detector system (a preliminary breadboard design).

stores, processes, and displays the CT data.

VCAT uses two different detector systems to provide a flexible scanning capability. A video-rate CCD camera/image intensifier detector system is used to quickly (15–20 minutes) acquire images with medium-spatial resolutions (250 μm). When high-spatial resolutions (<100 μm) are required, a second detector system must be used. The high-resolution detector system consists of a thin-rectangular scintillator and a light-integrating CCD

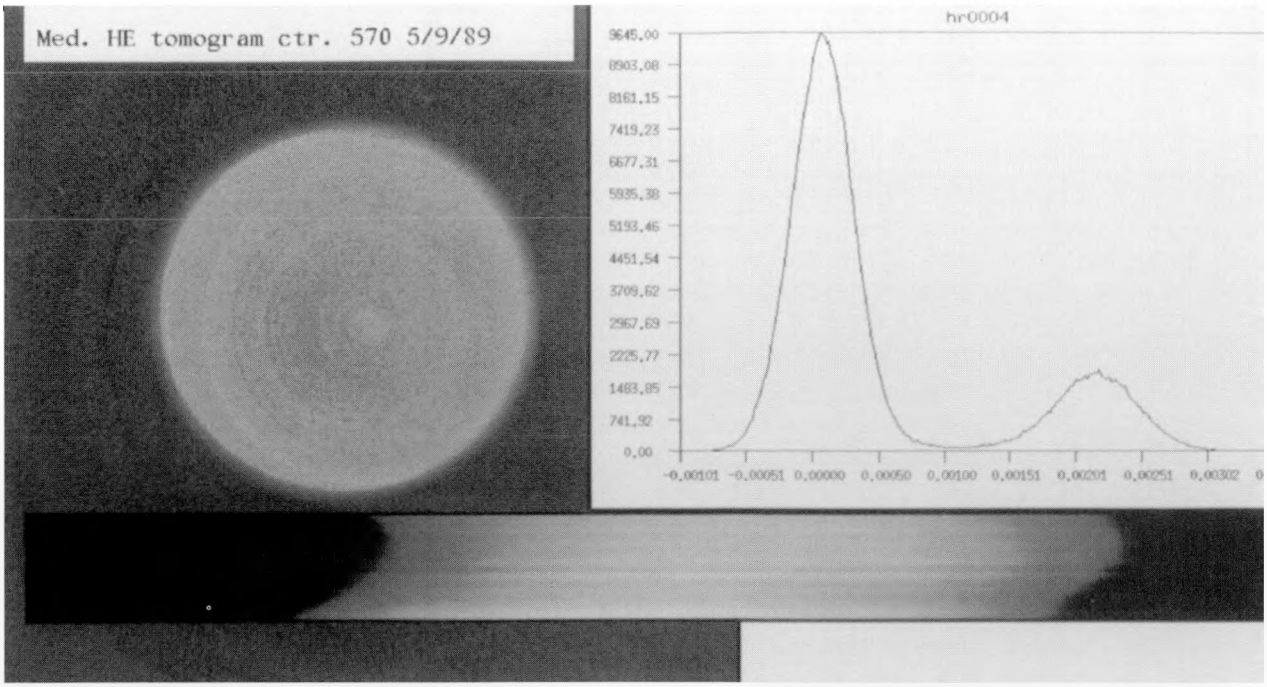


Figure 9. Reconstructed image of a medium PBX9502 high-explosive pellet (a cylinder 0.5 in. diameter and 1.5 in. long) at the center slice plane, using the VCAT scanner and the CCD-camera/scintillator detector system (left). The distribution of shading in the image represents relative attenuation coefficient variations in the pellet, which are related to changes in density and atomic number. In the plot, the vertical scale gives the number of values recorded at a given attenuation, while the horizontal scale indicates relative attenuation coefficient (pixels). The sinogram of this data is given in the image shown at the bottom.

camera(CCD camera/scintillator). Further details on the VCAT scanner hardware and software are given elsewhere.[4,7]

For both detectors, the microfocus x-ray tube allows us to use geometric magnification to increase the spatial resolution beyond that of the detector system (image intensifier/CCD, two line pairs/mm, and CCD-camera/scintillator, 17 line pairs/mm). The resolution ultimately depends on the object size, the focal-spot size of the tube source, and the resolution of the detector. The optimum magnification which provides the best achievable resolution in the object for a given focal-spot size will be determined for each detector.

The VCAT system provides x-ray images in the energy range from 25 to 160 kVp, the energy range of the microfocus source. It is mounted in a cabinet which allows objects of approximately 0.1 m^3 to be imaged. Resolutions in the range of 125 μm have been achieved. Because we can use either fan- or cone-beam scanner geometries as appropriate, data acquisition times are on the order of an hour for complete three-dimensional volumes. Qualitative CT images of the interior details of several objects of programmatic interest have been achieved.[4–6] These objects include but are not limited to R-Program materials, weapon parts for LLNL, and high explosives for Pantex. Representative CT images for the camera/image intensifier and CCD-camera/scintillator detector systems are shown in **Figs. 9** and **10** respectively.[5]

ZCAT

The ZCAT scanner was a temporary CT scanner, built with borrowed equipment to perform a CT study on concrete.[8] The ZCAT scanner was similar to the PBCAT scanner, since it was a first-generation translate/rotate CT system. However, the ZCAT scanner was designed to acquire data by translation of source/detector simultaneously (unlike the PBCAT scanner) as well as rotation of the object. A schematic electronic diagram and photograph of the ZCAT scanner are given in **Figs. 11** and **12**.

The borrowed equipment included the source, detector, and spectroscopy electronics from the MECAT scanner and staging from the Real-Time Radiography Upgrade Project. The x rays were produced using an ^{192}Ir (296, 308, 317, 468, and 604 keV) radioisotopic source and detected by a NaI(Tl) detector with gamma-ray spectroscopy instrumentation (**Fig. 11**).[8] The source and detector were collimated using fixed apertures of 2 mm on each.

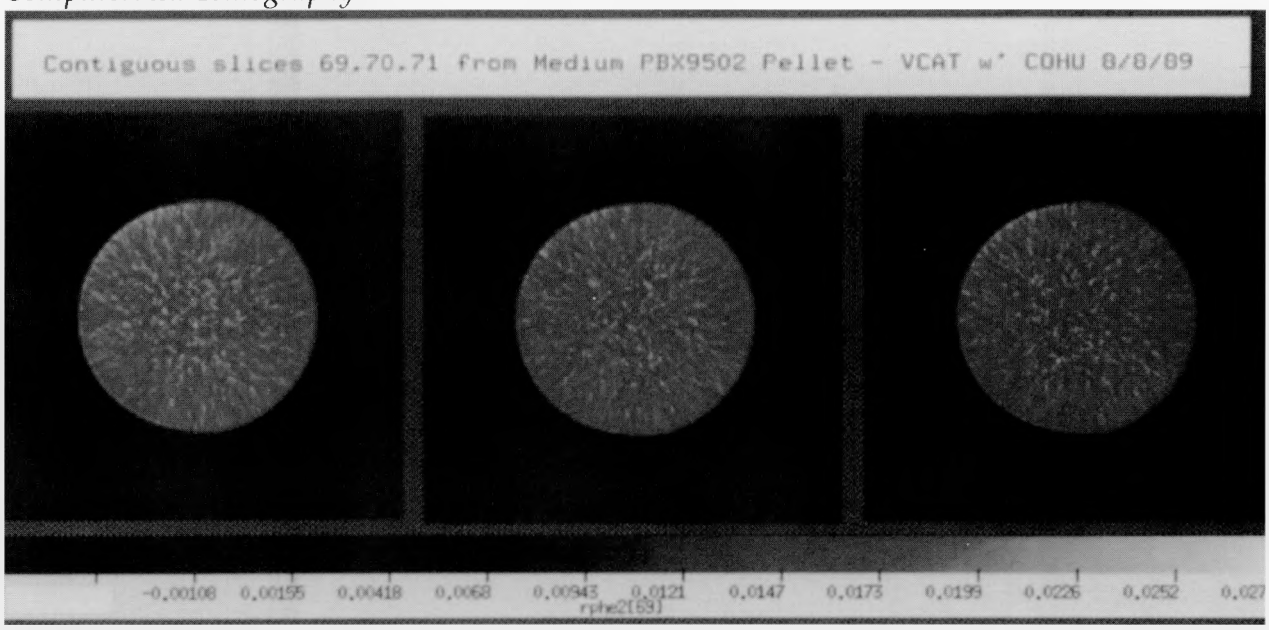


Figure 10. Representative reconstructed images of the medium PBX9502 explosive pellet, using the VCAT scanner and the CCD-camera/image intensifier detector system. All three slice planes are contiguous (69, 70, & 71 from the left to the right) and are located near the top of the pellet. The distribution of shading in the image represents relative attenuation coefficient variations in the pellet, which are related to changes in density and atomic number. The colorscale, shown here in black and white, is a relative measure of the attenuation coefficient in pixel number.

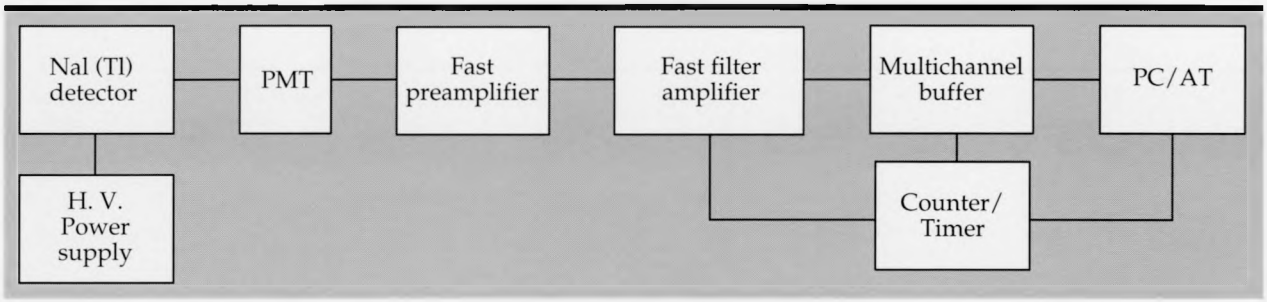


Figure 11. The temporary CAT scanner, ZCAT, x-ray electronics block diagram. The electronics were designed to acquire data by two different modes of operation: 1) pseudo-current integration mode with no energy discrimination when using the counter/timer, and 2) x-ray spectroscopy mode for medium-energy resolution at reasonably high count rates when the multichannel buffer is used.

The test object was supported by staging capable of handling 2,000 lb, which allowed three degrees of freedom: rotation, and x and z translation. The entire system was controlled by an IBM PC/AT computer. This computer was also used for data acquisition, processing and storage. Further processing, image reconstruction, display, and analysis were done on Sun workstations.

The ZCAT scanner could function using x rays over a broad range of energy, from several keV to ~1 MeV. It was housed in a large lead shielded room as shown in **Fig. 12(a)**. The mechanical fixturing used to support the source and detector limited the object size to about 0.2 m^3. Several CT images of various concrete objects were acquired. In one instance, an eight-inch concrete cube with embedded rebar was scanned. A representative image is shown in **Fig. 13**.[8]

CT Image Reconstruction, Analysis, and Display

Once the data from the CAT scanners has been collected, there are a number of steps that must occur before the investigator can receive useful information about the scanned objects. These steps include data preprocessing, slice reconstruction, image processing, and data display. These steps, and our progress in implementing them, are described here along with the research issues now under study.

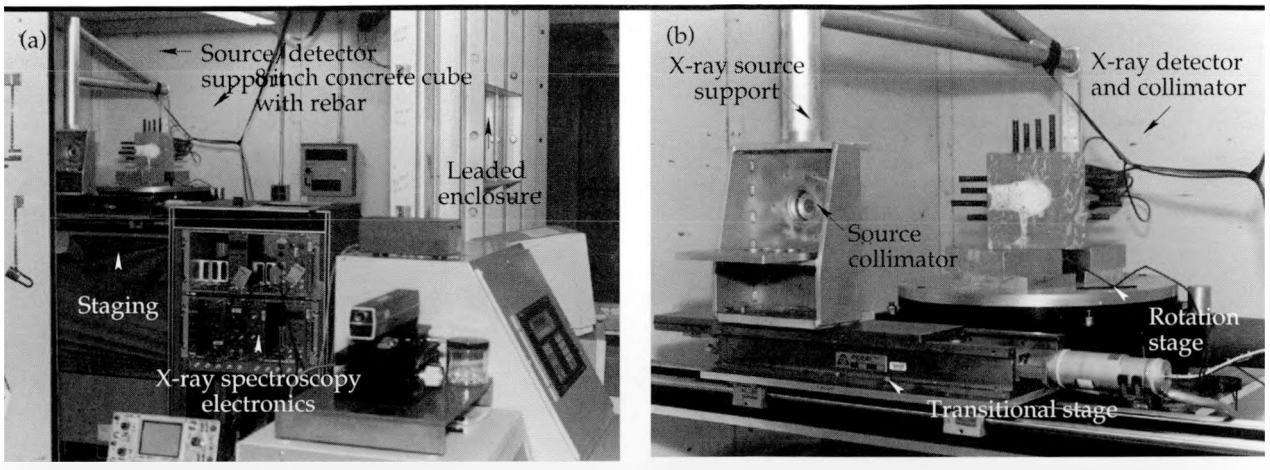


Figure 12. Photograph of the ZCAT scanner. (a) The leaded room used to house the scanner and associated electronics. (b) A closeup of the x-ray source (^{192}Ir) support, staging, and detector system. ZCAT is a discrete-beam first generation translate/rotate scanner. The translational step was obtained by simultaneous movement of the source/detector. The rotational step was obtained by moving the object.

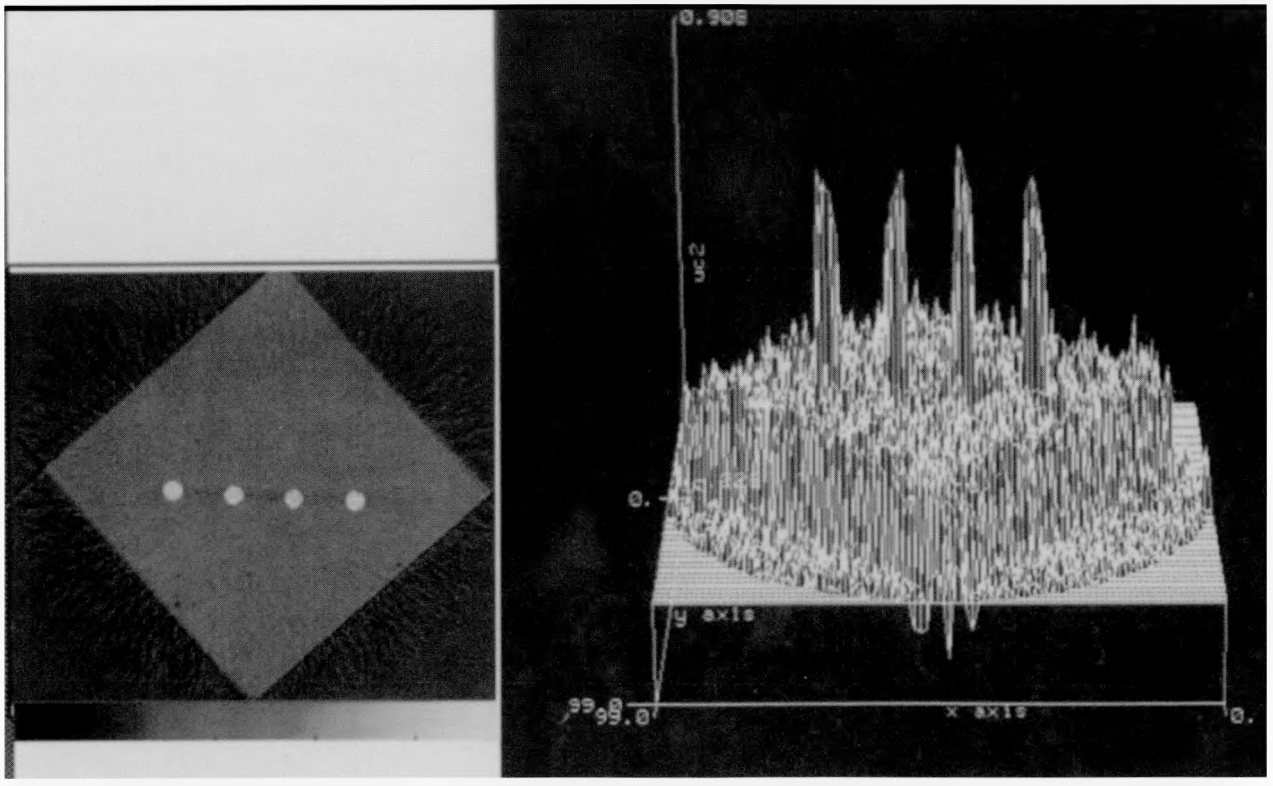


Figure 13. Reconstructed image of an eight-inch concrete cube with rebar (left). The distribution of shading in the image represents attenuation coefficient variations in the pellet, which are related to changes in density and atomic number. The attenuation coefficient scale is given in units of cm^{-1}. A perspective plot of this image is shown on the right.

Automated Reconstruction Processing

Before describing the research work done in CT imaging, we will briefly introduce each step. We also wish to point out that all steps are done modularly so that the entire process is automated, from acquisition to image hardcopy and report generation, to allow the maximum throughput of our scanners. At the same time, we have maintained flexibility to further process the data and explore new techniques. This required a set of standards and practices that are now in place,[9] but which will evolve further with each new scanner, computer, or display technique.

- Preprocessing—All scanners collect intensity profiles (or energy loss profiles in the case of

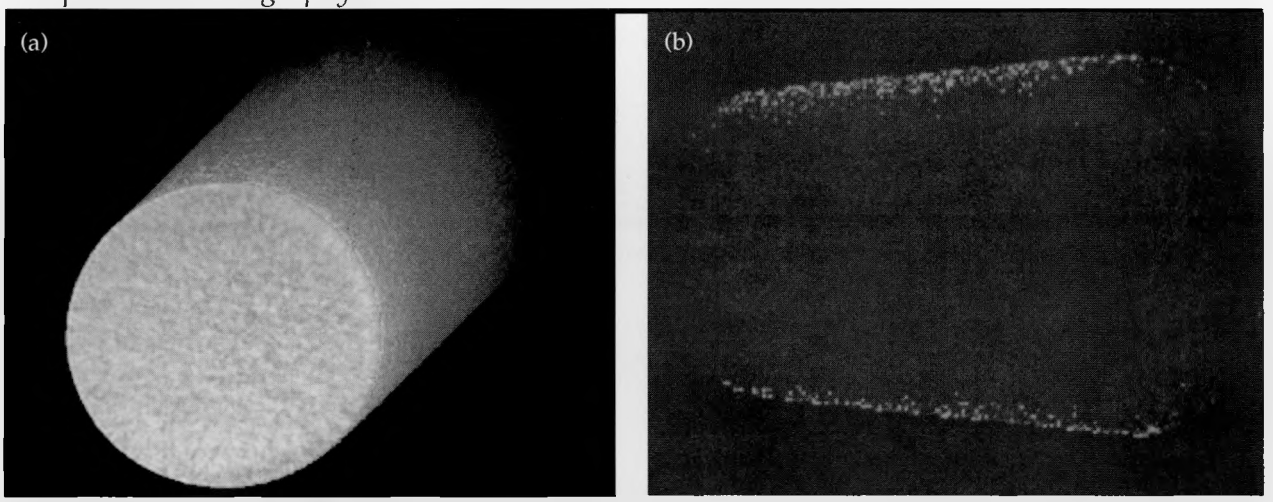


Figure 14. (a) Three-dimensional surfaced-rendered CT image of the medium high explosive pellet using all 102 slices acquired. Three representative slices are shown in Fig. 10. The distribution of shading in the image does not represent attenuation (or density/atomic number) variations in the pellet. (b) Three-dimensional volume-rendered CT image of the medium pellet. The distribution of shading in the image represents attenuation (density/atomic number) variations.

proton tomography) which must be scaled, corrected for imbalances, linearized, re-binned to account for scan geometry effects, and converted to line-integral (or ray-sum) measurements before entering the reconstruction codes. This preprocessing can be a complex matter for multi-element detectors such as the CCD camera used in VCAT.[7] Some preprocessing problems that we have addressed are centering errors, ring artifacts in third generation scanners, filtering issues, and background intensity corrections.[10,11] Line-integral data, scanner information, and all preprocessing parameters are stored in a data base for later retrieval.

- Reconstruction—After conversion to line integrals, the image must be computationally reconstructed to obtain an estimate of the linear x-ray attenuation coefficients. The methods for reconstruction from projections fall into two classes: continuous transform based techniques (e.g., filtered backprojection [FBP] or direct Fourier inversion) and discrete series expansion techniques (e.g., algebraic reconstruction).[1] We have implemented a number of these methods in our codes, so the user has easy access to them all through a menu. The user may also choose the computer that performs the reconstruction. The code automatically copies the data (if necessary) to/from that machine and carries out the operation. Machines that may be accessed include the local machine, a Sun with floating-point hardware, a Sun-4, a MIPS, an array processor, SPRINT,[12] and soon the OCF Cray and others (see the parallel architecture section p. 7-26).
- Image Processing and Display—For routine processing, analysis, display, and hardcopy of the reconstructed CT images, we use the VIEW signal and image processing system,[13] which provides general-purpose filtering, analysis, and display operations for multi-dimensional signals. The user interaction for multidimensional CT analysis has been designed around VIEW and is automated to allow rapid turnaround. More advanced volume display techniques are required when we have full 3-D data. We can interactively slice through planes of the data from any angle, or display the object as a reconstructed, shaded solid. Both of these techniques rely on the Pixar Image Computer for high-speed 3D data rendering. An example of a shaded volumetric surface rendered image of multislice CT data is shown in **Fig. 14**. We have the capability to generate videotapes for recording interactions with the volumetric data. New techniques that allow rendering on local workstations, at slower speeds, are also being used (see article in this Thrust Area report on scientific visualization). This work is done in conjunction with the Advanced Visualization Research Project (AVRP).

Model-Based Imaging

Industrial CT (ICT) applications present special problems not generally encountered in medical imaging. Most of these problems can be formulated as problems of under-sampling or limited data. We have developed and applied model-based techniques to different problems encountered in ICT scanners. Our effort divides into two

Table 1. Types of limited data in computed tomography.

Limited data	Description	Artifacts
Few projections		
Limited-angle	Good angular sampling over only a small range of angles (e.g., long objects)	Anisotropic distortion
Limited-view	Poor angular sampling over the entire object	Blurring, streaks
Few rays		
Spatial aliasing	Poor radial sampling	Blurring, streaks
Truncated projections	Outer rays missing (e.g., large objects)	Streaks, shading
Hollow projects	Inner rays missing (e.g., opacities)	Streaks
Motion	Movement in the object (e.g., flowing water)	Streaks
Inexact arithmetic	Discretization, finite precision, etc.	High-freq. noise
Measurement noise	Counting statistics, electronic noise, etc.	All of the above
Measurement error	Centering error, miscalibration, model mismatch, etc.	All of the above

areas of research: (1) the use of models for correcting reconstruction artifacts, and (2) the generation and use of synthetic projections for system modeling, algorithm testing, ring removal, and beam hardening corrections.

CT reconstruction algorithms produce poor images from sinograms where some ray sums or regions of ray sums are deficient or nonexistent. Unfortunately, this type of data is all too common in ICT. Such problems are known as incomplete-data or limited-data problems. Examples of these cases, listed in **Table 1** with their effects, include inability to complete access (360°) about the object (limited-angle), acquisition of only a few well-sampled projections (limited-view), insufficient sampling (in rays or angles), data dropouts, limited dynamic range of the detectors (resulting in hollow projections), large objects (resulting in truncated projections), and a variety of other noise sources. For production systems, these problems may be associated with an interest in reducing data acquisition time—e.g., region-of-interest (ROI) scanning of local areas in large objects, or coarse scanning for rough, fast inspection. The model-based techniques we discuss can be used on most of the above problems.

For the hollow projections case, no photons are detected over some subset of each projection due to a high-density (radio-opaque) portion of the object. We are not interested in imaging the opaque object, which we know exists, but are more interested in the less-dense regions surrounding it. The standard FBP algorithm produces artifacts in the image (listed in **Table 1**) due to hollow projections. An example of these artifacts for the hollow projections problem can be seen in **Fig. 15**. A lexan phantom with various notches and holes was scanned using VCAT in a third-generation scan geometry and reconstructed into the image shown in **Fig. 15(a)**. High-density wires were inserted into two holes of the same phantom and scanned at the same source energy to produce the image in **Fig. 15(b)**. The source energy was such that the wires were not penetrated. Notice the severe streaking in the reconstruction due to lack of information about the slice, particularly between the two wires where the two sets of three small holes are not discernable.

Fortunately, we can exploit the known *a priori* information into an iterative model-based algorithm to address this problem.[10] This method is based on a domain-iterative extrapolation scheme, in which we iteratively transform from the projection to the spatial (image) domain and back, applying known constraints in each domain. The loop continues until convergence is achieved to some criterion. For the lexan phantom of **Fig. 15**, we used the known outer boundary of the object, the expected attenuation range of lexan, and the location of wires. The reconstruction results after eight iterations using the extrapolation algorithm is shown in **Fig. 15(c)**. Although streaks still exist, the small sets of holes are clearly visible.

In order to make model-based imaging possible, we have developed an extensive CT simulation capability. This software uses models of objects to

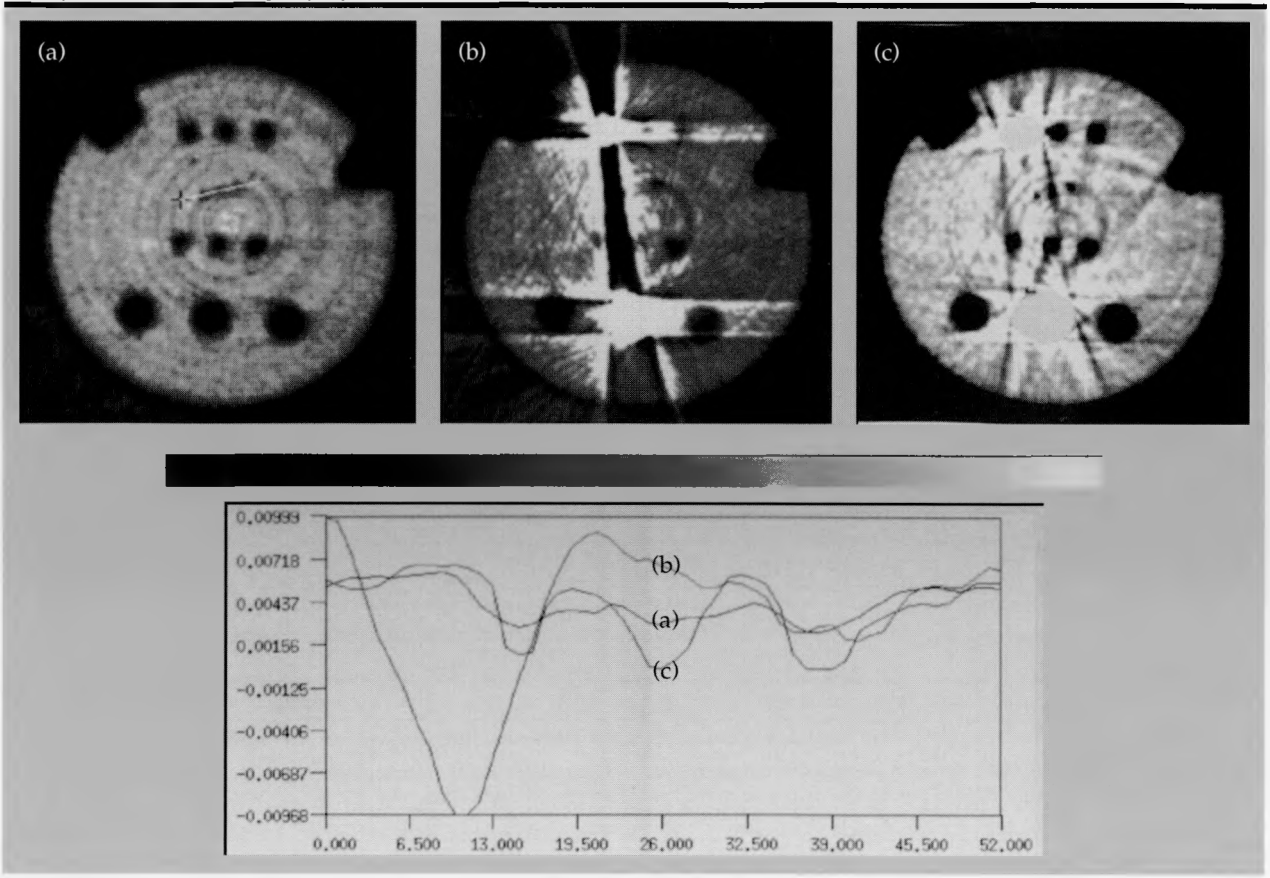


Figure 15. An example of artifacts related to the hollow projections problem. (a) A lexan phantom with various notches and holes was scanned using our third-generation VCAT scanner and reconstructed using the FBP algorithm into the image shown. (b) The same phantom with high-density wires inserted into two holes was scanned at the same source energy. The source energy was such that the wires were not penetrated. Notice the severe streaking in the FBP reconstructed image due to lack of information about the slice, particularly between the two wires where the two sets of three small holes are not discernable. Given the outer boundary of the phantom, the expected attenuation range of lexan, and the location of the wires, the extrapolation FBP algorithm can be used to replace the missing ray sums to improve the image reconstruction. (c) The resultant reconstructed image after eight iterations. Although streaks still exist, the two sets of three small holes are clearly visible. (d) ID profiles of the smallest set of three holes. Clearly the ID profile for the image (c) results in the most accurate resolution.

compute object images and projections through them. This capability has been used not only in the iterative reconstruction techniques just described, but also for model-based ring removal, optimum energy calculations, beam hardening corrections, and photon scatter analysis. An example of the ring removal technique is shown in **Fig. 16**. We are also exploring the possibility of including random photon modeling using Monte Carlo codes such as MCNP, TART, or COG for simulating the CT experiments. This may in the future lead to computer-aided nondestructive testing that is incorporated into the original CT design process.

Parallel Reconstruction Architectures

One of the greatest challenges of image reconstruction is the extreme computational load involved in reconstructing multiple slices from CT scanners such as VCAT. Typical reconstruction times on standard serial computer systems range from a few minutes to more than an hour per slice; when we have hundreds of slices to reconstruct, these times are prohibitive. In addition, the model-based algorithm described above requires many successive FBP and forward transform calculations, which can make this procedure very time-consuming.

Fortunately, the form of the forward transform is similar to that of backprojection, and it was shown to be readily implemented in parallel.[14] We have demonstrated image reconstruction on LLNL's SPRINT parallel processor, and near-linear speedup has been shown; even better performance will be required for three-dimensional or model-based CT.

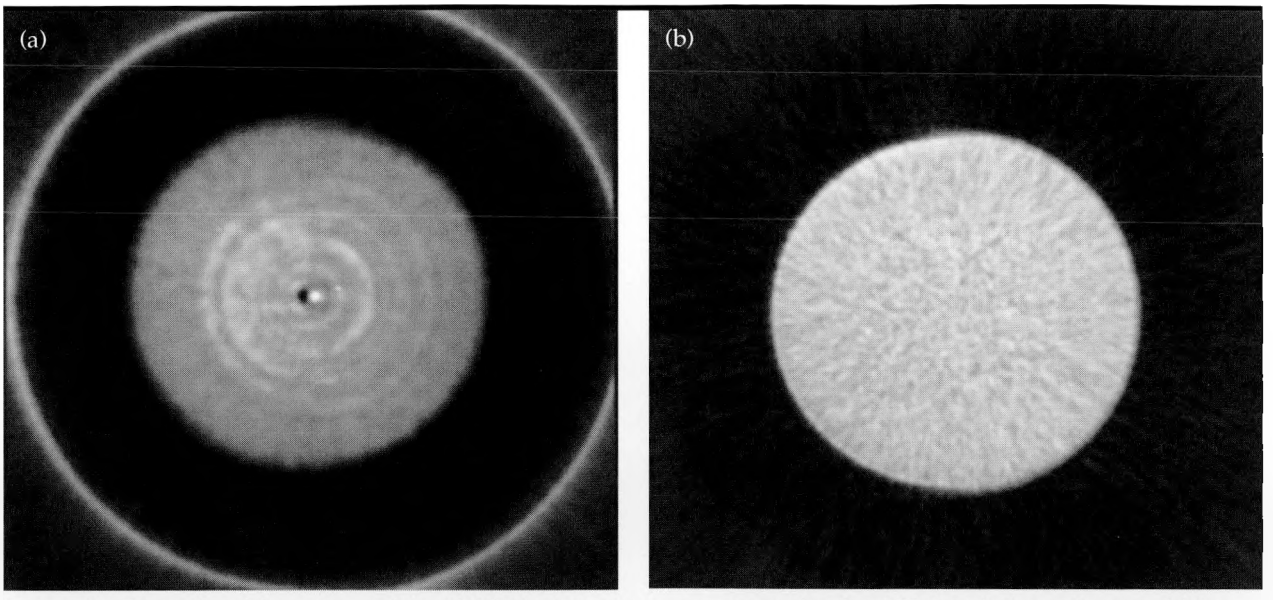


Figure 16. Example of ring-removal, in which the images were computed using CT simulation software. (a) Image with spurious ring artifact. (b) Image with false ring removed.

The Radon Transform Computer project, a joint project between UC Davis and LLNL, is lead by the CT group and co-funded by the NDE thrust area through the CT project, R-program, and Engineering Research (RISE thrust area). In this project, the ideas gained from the SPRINT research are being applied to build a prototype computer for extremely fast (100 times faster than SPRINT) reconstruction times.[15] Significant progress has been made, and the first prototype will be completed in FY 90.

Our research and the joint research with UCD led us to another joint research effort between us and ARACOR. This joint effort is supported by a California Competitive Technology Program (CCTP) grant for technology development and transfer of cone-beam (Konoscope) reconstructor software and hardware.

Elemental Analysis and Effective Atomic Number (Z) Imaging

It is often useful to map out the elemental or effective atomic number (effective-Z) distribution within materials, components and assemblies, instead of plotting results in terms of density or linear attenuation coefficient. We have implemented current medical techniques, and developed our own approaches to different types of elemental analysis. Our work may be divided into two areas: (1) employing and extending the techniques for calculating the saturations (weight fraction) of different elements in a chemical matrix, and (2) developing and implementing techniques for calculating effective-Z images, as applied to CT tomograms.

In collaboration with UC Davis Land, Air, and Water Resources (LAWR), we performed empirical studies that examined the distribution of different elements (air, water, and oil) introduced into a matrix of glass beads. This was considered a relevant sample object for their soil diffusion modeling efforts in the San Joaquin Valley. In collaboration with Z-division we have developed new "table-based" techniques for calculating effective-Z images for arbitrary compounds. Our work has involved the following tasks: 1) research and development of the technique, 2) calculation of the error variance for this approach as a function of the accuracies of the ray-sums, 3) translation of the optimal energy results for single-energy scans into the multiple energy setting, and 4) testing of the technique with simple gauging experiments and CT phantoms. Once validated, these techniques were applied to different types of programmatic inspection.[16] For example, we have calculated effective-Z images to study PBX9502 high explosive materials, part of our collaborative work with PANTEX.[5] In **Fig. 17**, we present a Z-image of four small PBX9502 pellets, accompanied by a histogram of the image. A detailed analysis of this work is presented elsewhere.[5,16]

Future Work

In FY 90, our primary goal is the completion of several stand-alone CT scanners. Each scanner will be equipped with our full set of automated reconstruction and analysis tools. We are currently developing correction methods to reduce the effects

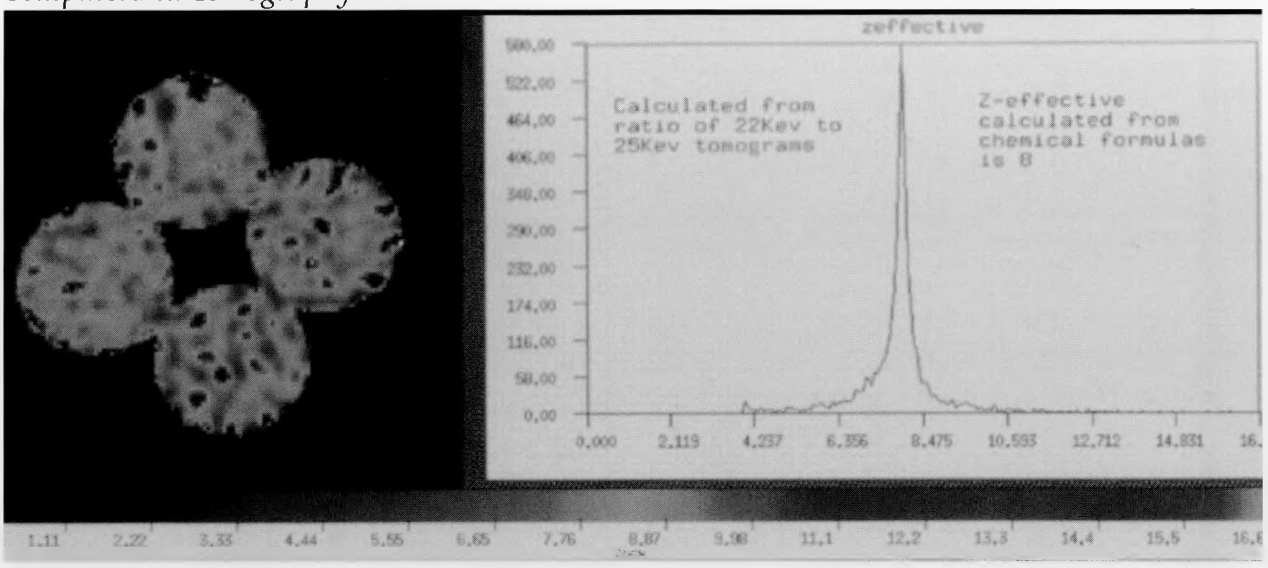


Figure 17. A representative Z-image of four small PBX9502 pellets. An effective-Z histogram of this image is shown on the right. In the histogram, the vertical scale indicates the total number of attenuation values recorded at a particular z (effective atomic number), while the horizontal scale indicates values of z.

of photon scattering, source/detector blurring, and beam hardening. We will extend our capabilities in three-dimensional imaging of programmatic materials, parts, and assemblies; and we are pushing for higher resolutions, higher speeds, and more accurate quantitative results. The CCTP project with ARACOR will be instrumental in extending these capabilities. We will continue to investigate reconstruction techniques for incomplete data, and will begin to apply these techniques to solve programmatic problems. An important new area is the investigation of alternative CT imaging modalities such as neutrons and heavy particles, which could furnish an entirely new way of looking at materials.

Our programmatic work will continue this year in support of R-Program, Z-Division, PANTEX, Weapons (LLNL; Sandia, Livermore; and Atomic Weapons Establishment, England), Treaty Verification, and Geochemistry. In the coming fiscal year, we will support new efforts in the area of materials and waste management, High Explosive Applications Facility (HEAF), Nuclear Chemistry, and in transferring CT technology to the Nuclear Weapons Complex (NWC).

Our interaction with other research projects at LLNL will continue this year. The Radon Transform Computer project will reach a critical stage when a working prototype of a two-dimensional reconstructor board and the first VLSI reconstructor element are completed. Our interactions with the Advanced Visualization Research Project will also continue to provide improved three-dimensional visualization of various Laboratory objects. We will be working closely with the Image Recovery and VIEW efforts in Engineering, as well as the other CT efforts in Chemistry, R-Program, Sandia, and Lasers.

Acknowledgments

The authors would like to thank Derrill Rikard, Linwood Hester, and Walter Kendricks for their contributions to this project.

1. G. T. Herman, *Image Reconstruction from Projections: The Fundamentals of Computerized Tomography* (Academic Press, New York, 1980).
2. J. Stanley, ARACOR, Sunnyvale, CA, private communication (1988).
3. S. G. Azevedo, H. E. Martz, and J. M. Brase, "Computed Tomography for Nondestructive Evaluation," *Energy and Technology Review*, Lawrence Livermore National Laboratory, Livermore, CA (October, 1988).
4. H. E. Martz, S. G. Azevedo, J. M. Brase, K. E. Waltjen and D. J. Schneberk, "Computed Tomography Systems and Their Industrial Applications," Lawrence Livermore National Laboratory, Livermore, CA, UCRL-98492 (1988), and accepted for publication by *Appl. Radiat. Isot.* (in press).
5. H. E. Martz, G. P. Roberson, M. F. Skeate, D. J. Schneberk, S. G. Azevedo, and S. K. Lynch, "High Explosives (PBX9502) Characterization Using Computerized Tomography," Lawrence Livermore National Laboratory, Livermore, CA, UCRL-ID-103318 (1990).
6. H. E. Martz, G. P. Roberson, M. F. Skeate, and S. G. Azevedo, "Computed Tomography Upgrade: ESD

Final Upgrade Report for FY 89," Lawrence Livermore, National Laboratory, Livermore, CA, to be published.

7. H. E. Martz, M. F. Skeate, D. J. Schneberk, and S. G. Azevedo, "Design, Performance, and Application of a CCD Camera-Based CT System," Lawrence Livermore National Laboratory, Livermore, CA, UCRL-101263 (1989), and *The American Society for Nondestructive Testing Topical Conference on Industrial Computerized Tomography Topical Proceedings*, (July 25–27, Seattle, WA, 1989), pp. 34–38.
8. H. E. Martz, G. P. Roberson, M. F. Skeate, D. J. Schneberk, and S. G. Azevedo, "Computed Tomography Studies on Concrete Samples," Lawrence Livermore National Laboratory, Livermore, CA, to be published (1990).
9. S. G. Azevedo, H. E. Martz, M. F. Skeate, D. J. Schneberk, and G. P. Roberson, *"Computed Tomography Software and Standards"*,Lawrence Livermore National Laboratory, Livermore, CA, to be published (1990).
10. D. J. Schneberk, S. G. Azevedo, H. E. Martz, and M. F. Skeate, "Sources of Error in Industrial Tomographic Reconstructions," accepted for publication in *Materials Evaluation ASNT J. Am. Soc. for Nondestructive Testing*, Columbus, Ohio (1990).
11. F. L. Barnes, S. G. Azevedo, H. E. Martz, G. P. Roberson, D.J. Schneberk, and M.F. Skeate, *"Geometric Effects in Tomographic Reconstruction"*, Lawrence Livermore National Laboratory, Livermore, CA, to be published (1990).
12. A. J. De Groot, S. G. Azevedo, D. J. Schneberk, E. M. Johansson, and S.R. Parker, "A Systolic Array for Efficient Execution of the Radon and Inverse Radon Transforms", Lawrence Livermore National Laboratory, Livermore, CA, UCRL-98393; and Proc. SPIE, Advanced Algrithms and Architectures for Signal Processing III, Vol. 975 (1988).
13. J. M. Brase, V. J. Miller, and M. G. Wieting, *The VIEW Signal and Image Processing System*, Lawrence Livermore National Laboratory, Livermore, CA, UCID-21368 (1988).
14. S. G. Azevedo, A. J. DeGroot, J.M. Brase, H. E. Martz, A. K. Jain, K. W. Current, and P. J. Hurst, "Tomographic Image Reconstruction using Systolic Array Algorithms," in *Review of Progress in Quantitative Nondestructive Evaluation (NDE) 1988*, D. O. Thompson and D. E. Chimenti, Eds. (Plenum Press, New York, NY, 1989) pp. 415–422.
15. S. G. Azevedo, J. M. Brase, H. E. Martz, A. K. Jain, K. W. Current, and P. J. Hurst, "A Radon Transform Computer for Multi-Dimensional Signal Processing," *Proc. ICASSP* (May 1989), pp. 1457–1459.
16. D. J. Schneberk, H. E. Martz, and S. G. Azevedo, "Calculation of Effective Atomic Number from Multiple Energy Tomographic Data," Lawrence Livermore National Laboratory, Livermore, CA, to be published.

ENGINEERING RESEARCH AND DEVELOPMENT

Thrust Area Report FY 89

Remote Sensing, Imaging, and Signal Engineering

Signal and image processing have always been important for supporting existing LLNL programs but now these technologies are becoming central to the formation of new programs. Exciting new applications such as radar remote sensing of ocean surface conditions and x-ray laser imaging are allowing us to participate in the development of new Laboratory programs. We have been very active in working with the Laboratory IR&D program as well as existing programs to define new directions.

We must also maintain and continue to build our technical base in signal and image processing in support of existing programs through such applications as diagnostic image processing and seismic signal processing. Outstanding examples are new deconvolution and image reconstruction techniques (see articles by D. Goodman and J. Berryman) which are being applied to problems in Nuclear Test and Non-Destructive Evaluation. In this fast-moving research area, we must continually update the tools which are used by a variety of programs at LLNL to evaluate experimental data.

Recent research work in RISE has concentrated on the development of new imaging technologies to solve problems in national defense and answer questions at the frontiers of science. Our work in Radar Ocean Imaging (see article by Fitch and Brase) applies new techniques in statistical signal processing to fully utilize information from advanced radar systems. These methods will allow ocean surface wave spectra to be observed at higher resolution than ever before.

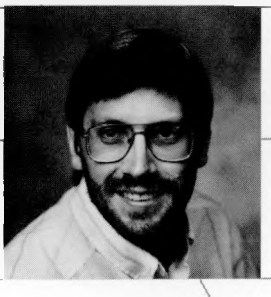
The performance of the current generation of optical telescopes is limited by atmospheric tubulence which degrades resolution in both defense and astronomical applications. Through our work in Atmospheric Compensation Algorithms (see article by Lawrence and Goodman) promising results have been demonstrated using higher-order spectral techniques. Interest in this work is high both within and outside the Laboratory.

LLNL is a world leader in developing lasers in the x-ray region of the electromagnetic spectrum. RISE investigators have been working with LLNL physicists to develop imaging applications for these devices. We are currently investigating the possibility of imaging DNA structure at resolutions unprecedented in living cells.

Over the past several years, RISE has developed a series of computer software systems for signal and image processing. These systems provide portability among the many computer systems used at LLNL and give us a platform for transferring the results of specific research and development projects to application areas. Our major signal and image processing system, View, is used by several major Laboratory programs and has been distributed to over 350 university, industry, and governmental sites.

Work in RISE involves a diverse set of sciences and technologies ranging from optical physics to microbiology to advanced computer architectures. Collaboration with other Engineering Thrust Areas, such as NDE and Computational E&M, and with other Laboratory departments, such as Physics and Biomed, is central to our continuing work in innovative imaging and signal processing applications.

James M. Brase
Thrust Area Leader



Robust Reconstruction Methods for Acoustic and Seismic Tomography

James G. Berryman
Engineering Research Division
Electronics Engineering

We have developed a new theory of traveltime tomography, showing that it is possible in principle to solve this very difficult problem by reformulating it as a problem in nonlinear constrained optimization. We have also developed a sophisticated new set of computer codes to apply these ideas to sound wave propagation through the earth, or through machine parts that need to be tested nondestructively. The main new idea is based on the following simple observation: Fermat's principle of least traveltime states that first arrivals follow ray paths with the smallest overall traveltime from the point of transmission to the point of reception. This principle determines a definite convex set of feasible slowness models—depending only on the traveltime data—for the fully nonlinear traveltime inversion problem. The existence of such a convex set allows us to transform the inversion problem into a nonlinear constrained optimization problem. In a new iterative reconstruction algorithm, the minimum number of nonfeasible ray paths is used as a figure of merit to determine the optimum size of the model correction at each step. The numerical results show that the new algorithm is robust, stable, and produces very good reconstructions even for high contrast materials where standard methods tend to diverge.

Introduction

The classical methods of tomography provide a means for reconstructing a two-dimensional function from a set of line integrals. For medical x-ray tomography, such line integrals are provided by measurements of the amplitude attenuation for straight rays passing through the body. When backprojected along the known ray paths, the attenuation data provide a picture of an inhomogeneous density distribution, which can then be displayed and interpreted for purposes of diagnosis.

Tomography has many uses outside the field of medicine, including applications in electron microscopy, acoustical and optical tomography, and radio astronomy. In geophysical applications involving the whole earth, or in local reconstruction problems such as borehole-to-borehole scanning with electromagnetic or seismic probes, the assumption of straight ray paths is often a poor approximation. Nevertheless, sophisticated data processing schemes such as geophysical diffraction tomography, which uses both wave amplitude and phase in the reconstruction process, are known to work very well when a starting model of sufficient accuracy is available. Thus, there is reason to believe that the reconstruction problem with ray bending could in principle be solved if some sensible procedure for finding such a starting model (other than trial and error) could be found. Recent progress towards finding an iterative algorithm for computing just such a model from traveltime data is described in this report.

Progress

Traveltime tomography reconstructs a slowness (reciprocal wave speed) model from measured traveltimes for first arrivals. The locations of sources and receivers are assumed to be known, but the actual ray paths are not known and must be determined along with the model slowness. Fermat's principle—that the path taken is the one of least traveltime—has been used extensively in forward modeling, *i.e.*, given the slowness model Fermat's principle determines the ray paths. However, Fermat's principle may also be applied in an entirely different way during the reconstruction of the slowness model using traveltime data, as we have shown.

To set notation, let t be the measured traveltime vector such that $t^T = (t_1, \ldots, t_m)$, where t_i is the traveltime along the i-th ray path (a superscript T implies the transpose). We form our model in two

dimensions by dividing the rectangular region enclosed by our sources and receivers into rectangular cells of constant slowness. Then, s is the model slowness vector $s^T = (s_1, \ldots, s_n)$, with s_j being the slowness of the j-th cell. For forward modeling, s and t are related by the equation

$$M s = t \ , \qquad (1)$$

where M is an $m \times n$ matrix whose matrix elements $l_{i,j}$ are determined by the length of the i-th ray path as it passes through the j-th cell. **Equation (1)** simply states that the total traveltime along a ray path is the sum of the traveltimes through each of the cells traversed by the ray.

One paper presenting a detailed analysis of the standard approaches to solving the inverse problem associated with **Eq. (1)** using weighted least-squares methods has been published this year.[1] New insight into the fully nonlinear traveltime tomography problem is contained in another paper.[2] In essence, we show that the forward problem represented by **Eq. (1)** may be replaced by the m constraints

$$(M s)_i \geq t_i \ . \qquad (2)$$

This substitution follows from Fermat's principle: the first arrival necessarily follows the path of minimum traveltime for the model s. Thus, **Eq. (2)** must be satisfied by any ray-path matrix M if s is the true model and therefore any model that violates **Eq. (2)** along any of the ray paths is not a feasible model. An exact solution to the inversion problem is found if and only if all of the inequalities in **Eq. (2)** become identities for some choice of model slowness vector s.

Another important concept is that of "feasibility violation number" $N_M(s)$. For any combination of ray-path matrix M, slowness vector s, and measured traveltimes t, it is straightforward to determine the number of rays that violate the constraints of **Eq. (2)**. This number is clearly equal to zero in the feasible region. Furthermore, it is a monotonically increasing function of distance from the feasibility boundary—once one of the hyperplanes of **Eq. (2)** is crossed, we never cross it again if we keep moving in the same direction. Thus, this number is cheap to compute and gives us a rough idea of how close we are to the feasibility boundary.

The key ideas behind our new algorithm may now be summarized as follows: Given a set of transmitter-receiver pairs and any given model slowness s, Fermat's principle may be used to find the ray-path matrix M associated both with s and with any slowness γs (where $\gamma > 0$) in the same direction as s. An optimum scale factor γ may be found by doing a weighted least-squares fit to the traveltime data. The best weights to use are described in detail in two papers.[1,3]

Having found the optimum slowness $s_b = \gamma s$ in the given direction, we next attempt to improve the model by finding another direction in the slowness vector space that gives a still better fit to the traveltime data. As many others have done, we first compute a damped least-squares solution s_μ. Next we note that both of the points found so far are guaranteed to lie in the nonfeasible part of the vector space—at least one and generally about half of the ray paths for both of these models will have traveltimes shorter than those of the measured data. Furthermore, although the point s_μ gives a better fit to the traveltime data, this fit is certainly spurious to some extent because it is based on the wrong ray path matrix; the ray path matrix used in the computation of s_μ from s_b is the one that was correct for slownesses along the direction s_b. Thus, both of the points we have found so far are off the feasibility boundary and the second point s_μ is of questionable worth also because its value was obtained in an inconsistent manner.

Recall also that the solution of **Eq. (1)**, if one exists, must lie on the feasibility boundary. Thus we would like to use s_μ and s_b to help us find a point on this boundary that is optimum in the sense that it is as consistent as possible with the ray path matrix M, with the measured traveltimes t, and with the feasibility constraints. The fact that traveltime error may be reduced by moving in the direction of s_μ may still give us an important clue as to the best direction to move in the vector space, *i.e.*, we may want to move in the direction $s_\mu - s_b$ but perhaps we should stop before we arrive at s_μ. How far then should we move in this direction?

If we consider **Fig. 1**, we are reminded that the feasible region is convex. Therefore, there may exist a point between the points s_μ and s_b that is closer to the feasible region than either of the two end points. If we could find this point s_l and then scale up to the point in the same direction lying on the feasibility boundary, then we would find s_f in the figure. In principle, it is possible to find the point on this line closest to the feasibility boundary. But it is much easier to compute the feasibility violation number $N_M(s)$. As we move along the direction $s_\mu - s_b$ from s_b, we generally find that this number achieves a minimum value at some intermediate point. This point of minimum $N_M(s)$ is the point s_l in the figure.

It is possible to prove that all three of these points are distinct unless they represent an exact solution to the inversion problem. So unless we have already solved the problem, these three points form a triangle and the size of the triangle gives us an estimate of how far we are from a solution.

These ideas have all been repeatedly confirmed in a large number of reconstructions on synthetic examples.[2,3] **Figure 2** shows three examples of typical results (middle column) obtained using our new reconstruction algorithm on a structure with a low speed anomaly (20%, 50%, or 100% lower than m. These results are compared to the best results of a standard damped least-squares algorithm (left column) and to the target model (right column). The reconstructions shown in **Fig. 2** were performed on a model structure with 8×16 cells using 320 rays, including 256 rays from left to right and 64 rays from top to bottom. The most startling example (not shown here) is one in which the wave speed contrasts are 100%. In this case, the standard damped least-squares algorithm becomes so unstable after about twenty iterations that the results become singular; by contrast, our new algorithm is completely stable and produces the result shown in **Fig. 2** after about ten iterations. The solution remains stable after hundreds of iterations if we force the algorithm to continue iterating. The new algorithm has been successfully tested on many other synthetic examples and also on real seismic and electromagnetic data.

The visual comparison in **Fig. 2** is not really "fair" to our method, in the sense that we have chosen *typical* results from the convergence set of this algorithm and have compared them to the *best* results for the standard algorithm. Such a comparison can be made for the synthetic examples because we know the correct answer and therefore can tell when the algorithm starts to diverge. For real problems we never have this luxury. With high contrasts and noisy data, the damped least-squares method does not converge and, lacking *a priori* knowledge, we never know when to terminate the iteration sequence. By contrast, the new algorithm converges quickly to a solution in its convergence set.

We have shown that Fermat's principle plays an essential role in wave speed reconstruction via traveltime tomography. Not only does this principle determine the ray paths once a slowness vector is given, but it also determines which slowness vectors are feasible and nonfeasible. The resulting problem in nonlinear constrained optimization[4,5] is very hard to solve directly, since a lengthy computation of the convex bounding surface using Fermat's principle and accurate ray tracing techniques is required. However, robust methods of reconstruction[3–8] have been found that make use of these ideas without mapping the location of the bounding surface explicitly.

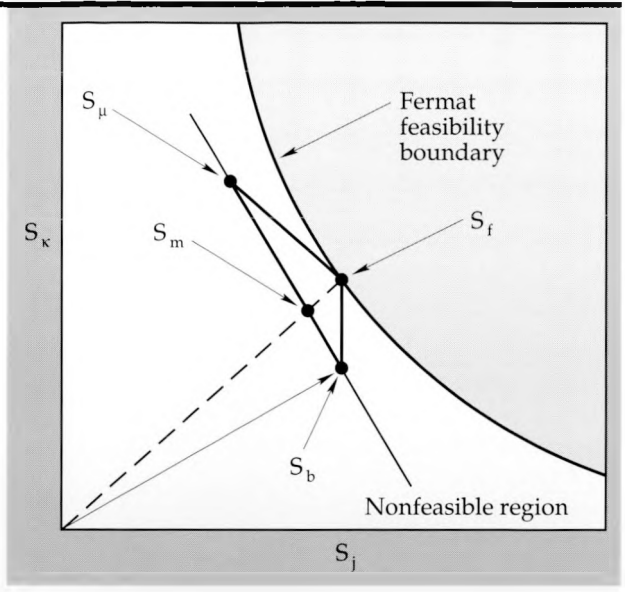


Figure 1. Schematic representation of the transformed slowness vector space of *s*, including the special points s_b and s_f, and the hyperplane in which they lie. Feasible vectors have associated traveltimes at least as great as the measured traveltimes. A sensible slowness model for the current choice of ray paths is the unique feasible point closest to the hyperplane containing s_b and s_f. We may think of this figure as a snapshot of the steps involved in one iteration of the reconstruction algorithm.

Codes are now operational for two-dimensional reconstructions of sound wave speed between two boreholes, and also for all possible view angles, thus making the work relevant to Laboratory acoustic tomography applications. We have also applied these methods to real field data and have started extending the codes to three-dimensional reconstructions.[7]

Future Work

Programmatic applications of these ideas in the next year will focus on three areas: (1) acoustic tomography for non-destructive evaluation (NDE) applications, (2) seismic tomography for the Earth Sciences Department and waste cleanup activities, and (3) ocean acoustic tomography for applications to acoustic anti-submarine warfare (ASW) activities. Generalization of the feasibility constraint

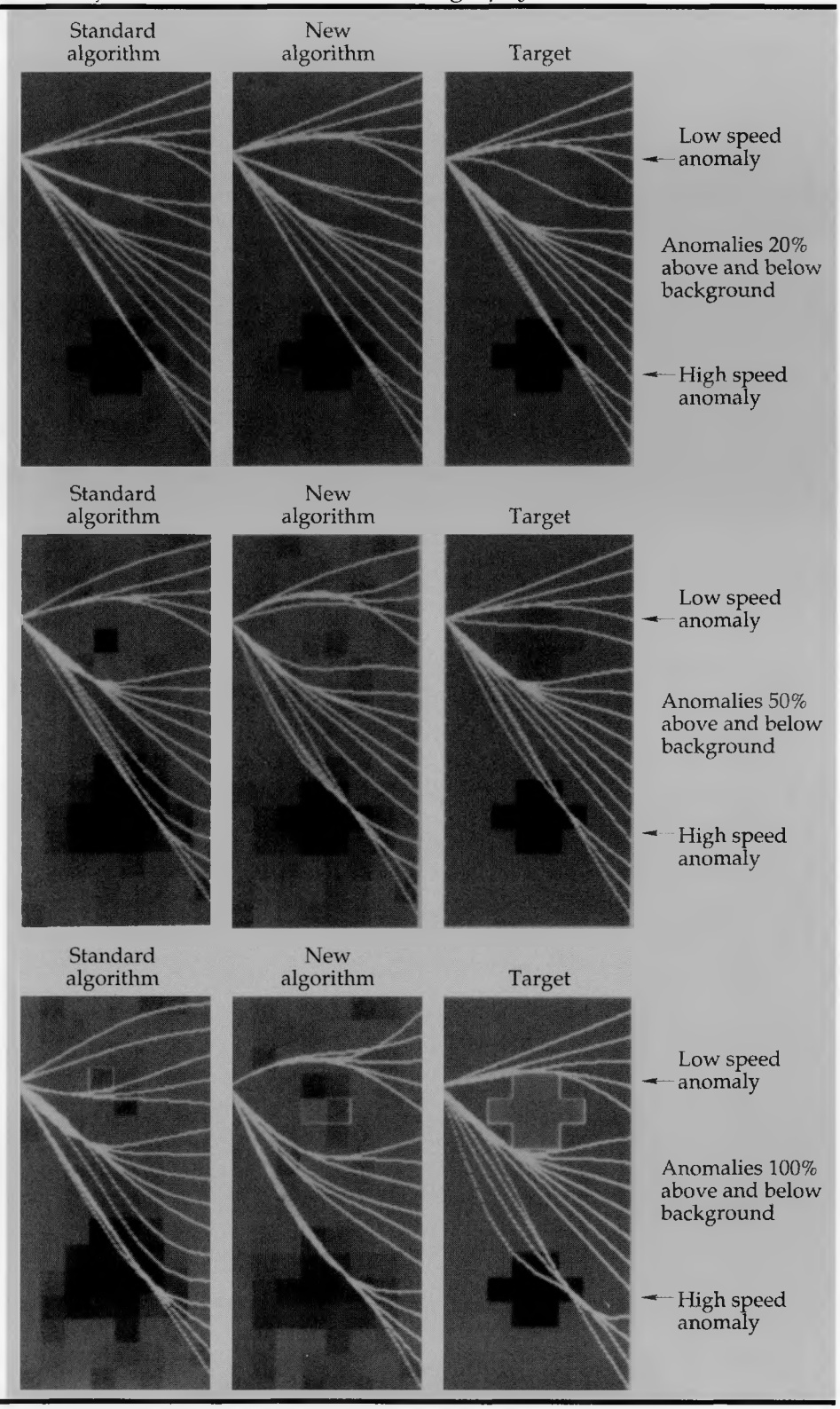


Figure 2. Comparison of new reconstruction algorithm with standard least-square algorithm for several different targets. The robust new algorithm converges in a well-behaved fashion without any need for *a priori* knowledge of target characteristics. Superimposed on the images are samples of some of the bent ray paths either used in the reconstructions or obtained in the forward calculations for the traveltime data.

ideas to other types of inversion problems such as electrical impedance tomography[6] is also well underway, and is expected to start making contributions to inversions of field data during the coming year. Finally, we can further enhance the reconstruction methods for acoustic and seismic applications by introducing geophysical diffraction tomography for distorted (non-plane) waves. This approach requires more data of higher quality both in amplitude and phase, but the resulting reconstructions will be of correspondingly higher quality as well.

1. J. G. Berryman, "Weighted least-squares criteria for seismic traveltime tomography," *IEEE Trans. Geosci. Remote Sensing* **27**, 302–309 (1989).
2. J. G. Berryman, "Fermat's principle and nonlinear traveltime tomography," *Phys. Rev. Lett.* **62**, 2953–2956 (1989).
3. J. G. Berryman, "Stable iterative reconstruction algorithm for nonlinear traveltime tomography," *Inverse Problems*, to appear February, 1990. Also, Lawrence Livermore National Laboratory, Livermore, CA, UCRL-100787, (March 1989).
4. J. G. Berryman, "Traveltime tomography and nonlinear constrained optimization," *Proceedings of the Twenty-second Annual Asilomar Conference on Signals, Systems, and Computers*, Oct. 31–Nov. 2, 1988, Pacific Grove, CA (Maple Press, San Jose, CA, 1989), Vol. 1, pp. 54–57; also Lawrence Livermore National Laboratory, Livermore, CA, UCRL-99831, (October 1988).
5. J. G. Berryman, "Seismic traveltime tomography and nonlinear constrained optimization," MIT Earth Resources Laboratory, Third Annual Founding Members Workshop on Seismic Wave Propagation and Inversion in Heterogeneous Media, August 1, 1989.
6. J. G. Berryman, "Imaging fluid flow in rocks using electrical impedance tomography," invited talk at the Minisymposium on Theoretical and Practical Aspects of Electrical Impedance Tomography, SIAM Conference on Mathematical and Computational Issues in Geophysical Fluid and Solid Mechanics, Houston, Texas, September 25–28, 1989.
7. J. G. Berryman "Seismic crosshole tomography and nonlinear constrained optimization," talk presented at SIAM Workshop on Geophysical Inversion, Houston, Texas, September 27–29, 1989.
8. T. J. Yorkey and J. G. Berryman, "An iterative algorithm for traveltime tomography between boreholes," Lawrence Livermore National Laboratory, Livermore, CA, UCRL-100770, (March 1989).

Radar Ocean Imaging

J. Patrick Fitch and James M. Brase
Laser Engineering Division
Electronics Engineering

We have been investigating the use of air- and space-borne radar systems to image ocean-surface features resulting from ship-generated internal waves.

Introduction

Ship-generated wakes were seen with imaging radars as early as 1978, using the space-borne radar system Seasat. Some of the wakes seen with this L-band radar were surface wakes resulting from the central turbulent wake or Kelvin wake, and others were thought to be surface effects due to internal waves propagating on a submerged interface of two different water densities (a pycnocline). However, there was no verifying or "sea truth" data for the images from Seasat. Various simplifying assumptions regarding the imaging process, specifically the use of Bragg scattering models, provided an apparently consistent model for predicting the signal-to-noise ratios of ocean images for Seasat.

In 1984, the Synthetic Aperture Radar Signature Experiment (SARSEX)[1] incorporated ships and moorings for collecting sea truth about naturally occuring internal waves off the east coast of the United States. This experiment used an aircraft with synthetic aperture radar (SAR) imagery, obtained simultaneously at two different wavelengths (L and X band). The NASA space shuttle imaging radar SIR-B was also used to image the area, but this effort was accompanied by less extensive sea-truth data than was available for the aircraft flights. When the techniques used to analyze Seasat data were applied to predicting the outcome of SARSEX, it was believed that the X-band image would have a significantly lower signal-to-noise ratio than the L-band image. However, the actual outcome of the experiment showed nearly identical signal-to-noise ratios for the two bands.[2] There clearly were more questions to answer.

The internal waves used in the SARSEX experiment were generated by tidal flows over bottom topography. The resulting internal waves, and consequently the surface effects, are much larger in amplitude than the expected internal wave from a ship. In order to investigate the physics behind the ship-generated internal wave surface effects, a controlled experiment with minimal internal wave background was designed to be undertaken at two sheltered bodies of water in Scotland, Upper and Lower Loch Linnhe. Three experiments of varying complexity have been held at the Loch Linnhe sites. The most recent and most extensive experiment was held in August, 1989, and consisted of two trial periods: one week in the upper loch, and a second week in the lower loch, with an interim week for moving and recalibrating equipment. A host of sensors was deployed in the lochs to measure water parameters, including current, temperature, density, and surface slope statistics, as well as meteorological conditions such as wind velocity, temperature, and humidity. The surface effects due to the ship-generated internal waves were visible in SAR imagery, collected by four separate aircraft at six microwave bands with multiple polarizations.

In order to characterize the underlying physics of the imaging process, it is necessary to understand the nature of the internal wave generated by a surface ship, the propagation characteristics of the internal wave on the pycnocline, the effect of the internal wave near the surface, and the interaction of the surface currents with the internal wave and the wind. All of these factors interact to generate the surface conditions which scatter the radar field. The work is multidisciplinary, incorporating theory and methodology from hydrodynamics, scattering models, and multidimensional signal processing. In our first year of funding we have emphasized the development of analytic and numerical models for describing the scattering phenomenology, together with multidimensional signal processing and analysis of airborne SAR data.

Progress

The project has developed a numerical simulation capability for modeling microwave scattering from ocean surfaces. We can calculate scattering amplitude and phase from a first-order expansion of the magnetic field integral equation, an analytic

description of the power spectrum of the ambient ocean, a vector for the wind speed, and a surface-current model for the internal wave. Bragg and Kirchoff approximations can be modeled by this simulator. By examining the temporal and spatial correlations of the scattering, we have been able to predict performance for various detection methods. Conventional electromagnetic codes would be difficult to use in this application because the scale length of the physical scatterer (the ocean surface) has a representation requiring many grid points in finite difference codes. The large number of grid points (wavelengths per resolution element) makes the required computation overwhelming.

The simulation now models real-aperture radar systems, which means that the radar footprint is taken as the size of a resolution element. We plan to extend the simulator to model synthetic-aperture radar (SAR) systems, keeping in mind that the 3-dB footprint in a SAR system is much larger than the final resolution obtained by post-processing the collected data. This simulator would then be interfaced with our hydrodynamic simulator to yield a prediction of system performance. Obviously this type of simulation, once validated by experimental data, would be very useful for surveying the large parameter space associated with radar ocean imaging systems. The resulting recommendation for an optimal radar system design would specify wavelength, incidence angle, altitude, polarization, and velocity, and would represent a substantial savings in experimental time and cost. The final simulation would also provide a means of evaluating proposed systems in terms of the operating conditions under which the system can identify signals of interest. There is a need for continual communication among experimenters, theorists, and system designers. Our goal is to use numerical simulation

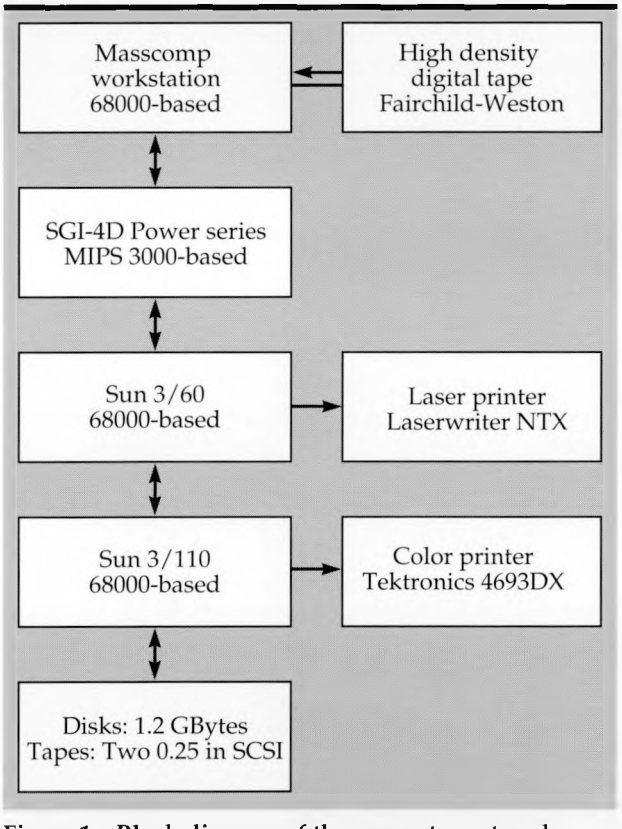


Figure 1. Block diagram of the computer network system which was fielded at Prestwick Airport for processing synthetic aperture radar (SAR) data from the NASA-JPL DC-8, including a data transcription system for high-density digital tapes, a multiprocessor compute server for reconstructing images from the raw radar data, and two workstations for interactive data analysis, display, and generation of hardcopy.

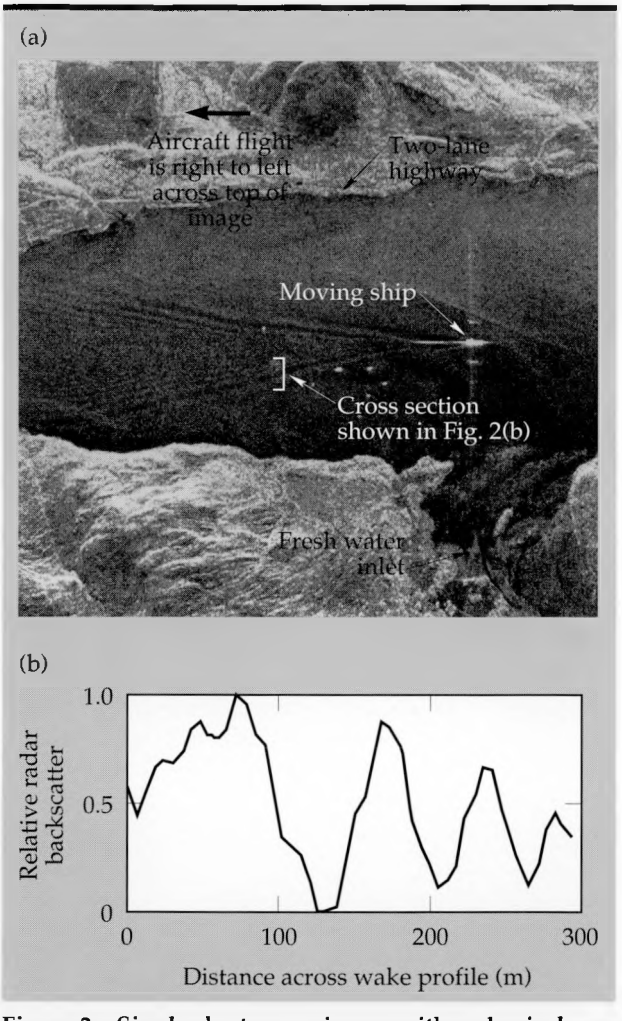


Figure 2. Single slant-range image with a physical scale of approximately 5 km azimuth by 2.6 km range, as reconstructed from raw data in about 10 minutes on the multiprocessing system: a) L-band image with vertical transmit and vertical receive polarization collected by the NASA-JPL DC-8 radar at a 33 degree incidence angle; b) average of 50 slices through one arm of the wake to show the surface profile.

of radar scattering to provide a timely and cost-effective means of searching the parameter space.

An initial investigation into the required hydrodynamics produced a relatively simple analytic model based on potential flow theory. This model has been used to predict thermocline displacement due to a ship's hull as well as near-surface water currents for input to the radar scattering simulations. Due to the expected inaccuracies of this model, we have begun development of more sophisticated numerical models of the internal wave generation and surface modulation processes.

In FY 89 the project has also developed a SAR system for processing data from the NASA-JPL DC-8 Airborne SAR. This ground-based system is used for post-processing the collected data, and includes integration of the high-density digital tape unit, data reformatting software (developed in cooperation with JPL), and several image reconstruction algorithms. A block diagram of the computer system which was fielded at the Loch Linnhe 89 experiment is given in **Fig. 1**. This system provides full-resolution images for all polarizations and was designed as an adjunct to the JPL onboard, near-real-time, low-resolution system, which produces a single SAR image for each pass over the experiment site, although data for up to 12 images is collected. An image reconstructed in the field by the LLNL system is shown in **Fig. 2** with the surface effects due to the internal wave clearly visible. Analysis of the twenty highest-priority aircraft passes is underway. Note that this analysis represents 240 SAR images, each requiring substantial computational time.

Future Work

We plan to extend the FY 89 results to include higher-order effects in the scattering models and wave propagation constraints, as well as aircraft motion compensation in the SAR processing. In addition, a more advanced hydrodynamic simulator is under development, and better post-processing of the SAR imagery is being developed to use pattern recognition techniques for incorporating spatial information about the wakes. Also, small-scale laboratory experiments are being planned for validating various components of our current models. We also intend, if possible, to participate in the joint US-UK Loch Linnhe 90 experiment. In general, the project plan is to calibrate our models with data collected from field experiments.

Acknowledgments

The radar ocean imaging work is supported by LLNL internal Director's Initiative and Engineering Research funding and is part of the Imaging and Detection Program under R. Twogood. Funding for the high-density digital tape was provided by DARPA. The success of the project could not have been achieved without the outstanding technical contributions of T. Yorkey, M. Wieting, M. Newman, B. Johnston, and D. Chambers.

1. SARSEX Interim Report, JHU/APL STD-R-1200 (May 1985).
2. D. Holliday, G. St-Cyr, and N. Woods, "Comparison of a new radar ocean imaging model with SARSEX internal wave data," *Int. J. Remote Sensing* **8**(9), 1423–1430 (1987).

Atmospheric Compensated Imaging

Taylor W. Lawrence, Dennis M. Goodman, and J. Patrick Fitch
Laser Engineering Division
Electronics Engineering

We have developed and implemented algorithms that can remove the blurring effects of atmospheric turbulence from telescopic images. These algorithms will permit astronomical objects and satellites to be imaged at the diffraction limit of any given telescope. We have successfully applied these algorithms to both simulated data and data from a terrestrial imaging experiment. These image-sharpening algorithms should have a wide application in both astronomical and defense imaging problems.

Introduction

A telescope of aperture D, imaging at a center wavelength λ, has a theoretical angular resolution of λ/D, the limit imposed by diffraction due to the finite-size aperture. However, ground-based telescopes must view objects through the atmosphere, which is a turbulent, randomly inhomogeneous medium. Local temperature gradients in the atmosphere create random phase fluctuations in the propagation path that can be characterized by a correlation scale length r_0. A normal, long-exposure image contains many realizations of this random fluctuation, with the net effect being an overall blurring of the image and a resulting angular resolution of λ/r_0. Thus, the resolution of the telescope is degraded by a factor of D/r_0, and is limited by the correlation scale of the atmosphere, not by the size of the telescope. For example, the world's largest telescopes, such as the 5-m Hale telescope on Palomar Mountain, have a theoretical diffraction-limited resolution of about 0.1 microradians for imaging at visible wavelengths. However, the correlation length of the atmosphere for visible imaging is typically 10 cm; consequently the resolution of these giant telescopes is degraded by a factor of 50 (5 m/10 cm). The resulting resolution is 5 microradians, no better than most backyard astronomers' telescopes!

This limitation, which applies to all ground-based telescopes, severely hampers our ability to image distant objects. However, recent advances in short-exposure-imaging technology, in conjunction with new computer post-processing algorithms, could provide a method for imaging astronomical objects and artificial satellites to the diffraction limit imposed by the telescope aperture, instead of the limit imposed by the local atmospheric correlation length. In addition to the immediate impact on existing large telescopes, sparse optical arrays hundreds of meters in diameter could be constructed to image higher-altitude satellites and could revolutionize ground-based astronomy. We have taken a significant step in proving that this method is feasible.

Our technical approach is based on nonstandard imaging techniques which make use of the statistical information in short-exposure images. Specifically, the turbulence-induced effects of the atmosphere on the object image can be "frozen" if the image exposure time is reduced below the correlation time of refractive index fluctuations (typically 1 to 10 msec, depending on wind velocity). Short-exposure images captured in this way are a random superposition of diffraction-limited images of the object. Each short-exposure interference pattern is called a speckle image, and is analogous to speckle observed with coherent laser illumination. Procedures for extracting object information from a sequence of speckle images are collectively referred to as "speckle interferometry."[1] These techniques have been successfully used by the astronomical community to image star clusters. However, such algorithms have not previously been used for imaging extended objects through the atmosphere.

Progress

In FY 88 we began work on a comprehensive numerical simulation of the speckle imaging process, and put together the hardware necessary to carry out a horizontal-path imaging experiment using a 0.5-m telescope and a CCD-based speckle camera. In addition, we implemented state-of-the-art image reconstruction algorithms and applied them to both the simulated and experimental speckle images. With this setup, we obtained high-resolution images of a satellite transparency illumi-

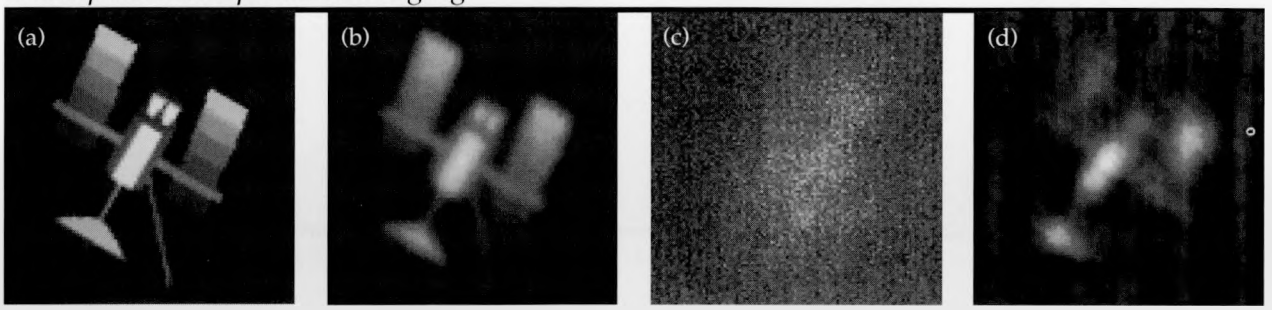


Figure 1. Application of our algorithm to data from the horizontal path imaging experiment, scaled to represent the imaging of an actual satellite at an altitude of 3000 km. (a) Computer-generated satellite transparency, actual size 3 cm. (b) Calculated, diffraction-limited image obtainable with optically perfect 0.5-m telescope and no atmospheric degradation of the image. (c) Actual speckle image of the target transparency, obtained at a range of 1.2 km through turbulent sea-level atmosphere, with a 0.5-m telescope, and exposure time of 5 msec. (d) Result of using our algorithm to process 1000 speckle images, demonstrating feasibility of speckle imaging algorithm for use in satellite surveillance.

nated with light levels representative of low-earth-orbit (LEO) objects (those having an orbital range of a few hundred kilometers). In FY 89, we turned our attention to fainter objects, namely those satellites with a range of a few thousand kilometers in high-earth-orbit (HEO), and found it necessary to use a hybrid algorithm developed here at LLNL in order to obtain a satisfactory reconstruction. The main obstacle that had to be overcome was the additive detector noise characteristic of CCD electronic amplifiers. Our algorithm successfully dealt with this problem, and we ultimately demonstrated high-resolution imaging of HEO satellites. Details of the simulator, experiment, and algorithm development are given below.

The computer-based atmospheric simulator allows input objects to be propagated through an isoplanatic (i.e., shift-invariant or linear) atmosphere with arbitrary r_0. The distorted wavefront is then received by a telescope with a pupil function defined as the superposition of a number of different-sized circular apertures with arbitrary obscurations. In addition, the simulator includes noise sources characteristic of Poisson photon statistics and non-ideal photon detection (i.e., noisy CCD electronics). Speckle images of a point reference and objects previously used in the horizontal path experiment were generated, and compared quite well with both theory and the actual images obtained in the experiment. A thorough validation of the simulator was completed by comparing reconstructed images from the simulator with those from the experiment over four orders of magnitude in light level.

The horizontal path imaging experiment utilized a 50-cm, f/8 telescope with a target board placed at a range of 1.2 km. A speckle imaging system was assembled on the back of the telescope, consisting basically of a low-noise (7 e^-/pixel) 512 × 512 CCD array controlled by a micro-Vax-based data acquisition system. The foothills of Mt. Diablo, east of San Francisco Bay, were chosen as the site for the experiment, since this area afforded the easiest access to a range of unobstructed atmospheric path lengths from 1 to 3 km. A path length of 1.2 km was ultimately chosen to give the appropriate level of atmospheric turbulence. Data from an unresolved point reference on the target board indicated turbulence levels in the range $D/r_0 = 15$ to 25. In addition, the shape of the average power spectrum of the point reference was compared with theory to confirm that the turbulence had a statistical character sufficiently similar to that seen in vertical imaging (i.e., it was Kolmogorov). Reconstructed images from data collected during the experiment demonstrated the feasibility of imaging satellites at HEO, and also confirmed the validity of the atmospheric physics models used in the simulator.

During the first phase of our algorithm development here at LLNL, we sought to develop certain state-of-the-art algorithms that have attracted interest in the astronomical community.[1,2] These algorithms are: (1) the Labeyrie technique for Fourier magnitude estimation, (2) the bispectral (or triple correlation) technique for Fourier phase estimation, and (3) the Gerchberg-Saxton-Fienup Fourier iteration technique for the imposition of object support and positivity constraints. We applied a combination of these techniques and found that the resulting algorithm was successful at reconstructing objects representative of LEO satellites.[3] However, when we turned our attention to fainter objects representing HEO satellites, we realized the critical importance of bias effects due to both photon noise and additive noise in the CCD. These bias effects cannot be removed merely by processing a larger number of speckle frames. Instead, the bias terms

must be estimated separately and subtracted. Although methods had been developed by the astronomical community to deal with photon bias, there is no published methodology on ways to estimate the additive bias. We therefore carefully analyzed the effects of additive noise, and derived an analytical expression for its contribution to both the power spectrum and the bispectrum. We modified our algorithms to compensate for these bias effects, and ultimately demonstrated high-quality reconstructions of dim targets representative of HEO satellites. **Figure 1** shows an example of an image reconstructed from 1000 speckle frames collected with our CCD camera. To appreciate the power of the algorithms we have developed, one need only compare the relatively featureless, noisy speckle frame (c) with the outcome of the reconstruction process (d). On comparison of the final product (d) with (b), we see that we have indeed achieved diffraction-limited detail.

Future Work

Our main goal for FY 90 is to apply our algorithms to speckle images of actual satellites obtained with our CCD speckle camera mounted on the 1.6-m telescope at the Air Force Maui Optical Tracking Station (AMOS). We hope to obtain reconstructed images of LEO satellites out to ranges of 500 km, and to compare our results with the performance of the adaptive-optics-based Compensated Imaging System on the same beamline. In addition, we will continue to explore new ways to improve our reconstruction algorithms.[4] Current efforts include using least-squares methods for phase recovery, and the incorporation of "a priori" information such as positivity and boundedness of the object. Our ultimate goal is to demonstrate diffraction-limited imaging of objects out to geostationary orbit (38,000 km).

1. F. Roddier, "Interferometric Imaging in Astronomy," *Physics Reports* **170**, 97–166 (1988).
2. G. R. Ayers, M. J. Northcott, and J. C. Dainty, "Knox-Thompson and Triple-Correlation Imaging Through Atmospheric Turbulence," *JOSA A* **5**, 963 (July 1988).
3. T. W. Lawrence, J. P. Fitch, and D. M. Goodman, "Image Reconstruction Using the Bispectrum," *Proceedings of the Asilomar Annual Conference on Signals, Systems, & Computers* (Oct. 1988).
4. D. M. Goodman, "On the Derivatives of the Homomorphic Transform Operator and their Application to Some Practical Signal Processing Problems," *Proc. IEEE*, to be published April 1990.

New Algorithms for Deconvolution and Identification

Dennis M. Goodman
Laser Engineering Division
Electronics Engineering

Extraneous linear dynamics introduced by electromagnetic, acoustical, and optical sensors can seriously distort experimental data. The accuracy of such data can often be greatly improved by using computer processing to remove these distortions. For this purpose, we have written signal processing computer codes that use both parametric and nonparametric methods for processing many types of signals in the time domain. Time-domain methods for identification and deconvolution are more flexible than their frequency-domain counterparts. With the advent of faster computers that permit efficient implementation, time-domain approaches have become the methods of choice for solving these problems.

Introduction

Two signal processing problems occur frequently in a wide variety of scientific experiments. In the first problem, one is given noisy measurements of the response of a device to a known input signal. One must then develop a general description of the device's input-output behavior so its response to an arbitrary input signal can be predicted. In the second problem, one is given a description of a device's input-output behavior and noisy measurements of its response, and one must then determine the nature of the input signal. Often it is reasonable to assume that the device is linear, so that its input-output behavior is completely determined by its impulse response. In this case, the first problem is called impulse response identification, and the second is called deconvolution.

Both problems come into play when attempting to remove the unwanted effects of a measurement system from experimental data. First, the system is characterized by applying a known input, measuring the output, and estimating the system's impulse response. Second, the identified impulse response is deconvolved from the experimental data. Typical measurement devices that introduce spurious linear dynamics which must be removed before the results of an experiment are analyzed include optical systems, acoustic transducers, and electromagnetic probes. The model for both deconvolution and identification is shown in **Fig. 1;** the equation relating measured output to input is

$$\{y(n)\} = \{h(n)\} * \{x(n)\} + \{e(n)\},$$

where $*$ denotes convolution. Because convolution is commutative, methods for determining $\{h(n)\}$ from $\{y(n)\}$ and $\{x(n)\}$ also apply to determining $\{x(n)\}$ from $\{y(n)\}$ and $\{h(n)\}$. Here we will discuss impulse response identification, but our comments also apply to deconvolution.

At first glance, solving this problem appears easy: merely divide the Fourier transform of the output by the Fourier transform of the input, and then inverse transform to get an estimate of the impulse response. Unfortunately, the impulse response estimation problem is usually ill-posed because the input has almost no energy at some frequencies, and dividing transforms produces a wildly varying, high-magnitude estimate. Another problem with frequency domain approaches is that they assume the data records to be zero at both ends; if this is not the case, the estimate may be highly biased.

The solution to the ill-posed difficulty is to use *a priori* information to constrain the estimate. Many of these constraints can be imposed with either

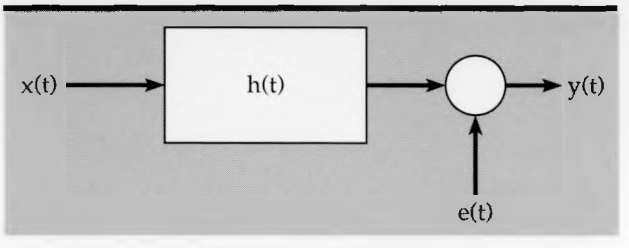


Figure 1. Model for the deconvolution and identification problems. y(t) is the observed output of the device, corrupted by measurement noise e(t). For the deconvolution problem, the impulse response h(t) is known, and the input x(t), must be estimated. For the identification problem, x(t) is known and h(t) must be estimated.

time- or Fourier-domain methods, but time-domain methods are more flexible and have the critical advantage that they do not force a zero signal assumption at both ends of the data. Their principal drawback was the additional computer time required, but faster computers now make time-domain methods attractive. The goal of this development project was to implement parametric and nonparametric time-domain methods. Parametric time-domain methods constrain the estimate by assuming that the impulse response derives from a model described by a relatively small number of parameters. The estimate of the impulse response is then obtained by adjusting the parameters to fit the data. Conversely, nonparametric methods do not constrain the impulse response to a particular model, but require it to satisfy more general constraints on its energy or frequency content. Parametric approaches are called for if a model of the impulse response is available, but may be useful even if one is not. Conversely, nonparametric approaches are more flexible and can be applied to a wider class of problems. A more detailed introduction to this subject has recently been published.[1,2]

Progress

Parametric Methods

Because difference equation models are the sampled-data analog to differential equations, they are the most useful parametric models. Many real-world devices can be adequately described by a low-order, constant-coefficient differential equation. By fitting a difference-equation model to the input-output data records of such a device, one can obtain not only a good estimate of its impulse response but also estimates of the poles and zeros of its corresponding transfer function. Indeed, these transfer function parameters may be the goal of the experiment. Furthermore, if the assumed model structure is correct, parametric modeling can extrapolate transfer function behavior to parts of the frequency spectrum where the input signal has no energy. During the last six years, the NLS output-error system-identification program[3] has been applied successfully to transient data collected from a wide variety of experiments. Recent applications have included electromagnetic,[4] acoustic,[5] and seismic[6] problems both within and outside LLNL.

This year we added subroutines to calculate Cramér-Rao lower bounds for the accuracy of transfer function parameters estimated by NLS. A frequent and surprising outcome of difference equation modeling is a situation in which the parameter variances are much higher than the variance of the estimated impulse response. This seemingly paradoxical behavior is explained by the fact that the impulse response of a system can be relatively insensitive to the parameters of its corresponding transfer function. On the other hand, there are also cases where the impulse response is very sensitive to the transfer function parameters; for these cases the parameter variances will be relatively low. This unexpected behavior often leads to misinterpretation and controversy when an experimenter attempts to attach physical significance to the poles and zeros of an estimated transfer function. The Cramér-Rao bound is an excellent tool for helping experimenters understand how much confidence they should have in their data. This year we have used the bounds calculated by our subroutines to interpret the results of applying difference equation modeling to a classical electromagnetic estimation problem.[7]

Nonparametric Methods

A difference-equation model is not appropriate for many deconvolution and identification problems. In particular, difference equations often do a poor job of modeling systems with both dynamics and delays; a transmission line is a good example of such a system. An obvious nonparametric impulse-response estimate is the sequence $\{\hat{h}(n)\}$ whose convolution with the system input is a least-squares (minimum energy) approximation to the output, *i.e.*, $\{\hat{h}(n)\}$ minimizes $\|\{y(n)\} - \{\hat{h}(n)\} * \{x(n)\}\|$. This sequence is the minimum-variance unbiased estimate. However, if the problem is ill-posed, the minimum variance will be very large, and the estimate will oscillate wildly, behaving exactly like the estimate obtained by dividing transforms.

Because the unconstrained least-squares estimate is too large and oscillates wildly, a common approach is to apply *a priori* information by minimizing the fit to the observed output, subject to a constraint on the estimate. Such an estimate is the $\{\hat{h}(n)\}$ that solves

$$\begin{array}{ll} \text{minimize} & \|\{y(n)\} - \{h(n)\} * \{x(n)\}\| \\ \text{subject to} & L[\{\hat{h}(n)\}] \leq K, \end{array}$$

where $L[\{\hat{h}(n)\}]$ is a weighted measure of the energy in $\{h(n)\}$ or its derivatives. There remains the problem of selecting K, which determines the amount of constraint. In FY 87 we implemented an algorithm that minimized the Mallows C_L statistic to determine the value of K, and we observed that

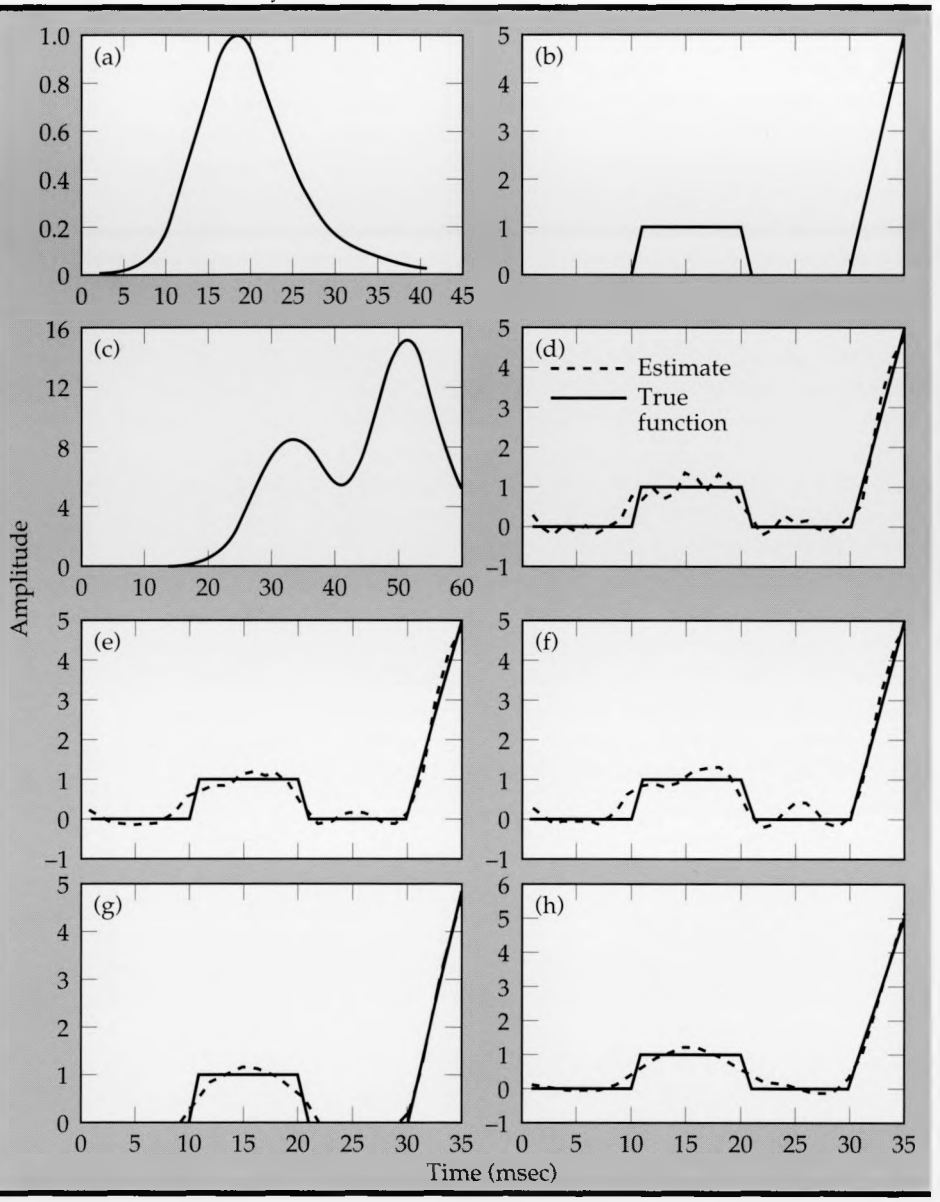


Figure 2. Example of the behavior of nonparametric algorithms for an identification problem. (a) Known input. (b) Unknown impulse response. (c) Resulting noise-free output. (d), (e), and (f) Nonparametric estimates of the impulse response, using constraints on (respectively) impulse response energy, its first derivative, and the energy in its second derivative. (g) Estimate obtained with a positivity constraint. (h) Estimate obtained with a conjugate gradient algorithm.

this approach gave excellent estimates of the impulse response.[8] This year we improved the numerical properties of that algorithm, and because the C_L statistic requires an estimate of noise energy, we extended the algorithm to use generalized cross-validation when such an estimate was not available.

For many practical problems the unknown signal represents intensity, energy, or magnitude and so must be nonnegative. Imposing a nonnegativity constraint on the estimate can greatly lower its variance. Solving the minimization problem of the previous paragraph subject to this additional constraint is a quadratic programming problem; we have developed software to solve this problem and obtain nonnegative estimates. Unfortunately, however, the statistical approaches we have used to determine K either have not been derived or are difficult to implement for nonnegative estimates. For this case we used the discrepancy principle to determine K; this is an intuitive method requiring the data matching error to be the same as the expected noise error.

Although the nonparametric techniques we have discussed yield high quality estimates, they are computationally intensive and may not be appropriate for interactive situations. In order to obtain faster solutions while retaining the advantages of time-domain methods, we have developed an algorithm based on conjugate gradients. This algorithm is an iterative approach that constrains the estimate by stopping after only a few iterations. We use the discrepancy principle to determine when to terminate the iteration. Because the computational cost increases as N^2, where N is the data length, rather than as N^3 as is the case for most time do-

main methods, the conjugate gradient algorithm is much faster.

We have also investigated the behavior of certain approaches to the blind deconvolution problem in which an input signal must be deconvolved from a system response that is either unknown or random. We discuss our conclusions in an upcoming publication.[9]

We have applied these nonparametric algorithms to several important problems at LLNL. These include interpretation of nuclear test data, pulse shaping for the NOVA laser, and characterizing the response of probes used in electromagnetic transient problems. An example showing the behavior of the three nonparametric algorithms we have developed is shown in **Fig. 2.** The input is shown in **Fig. 2a,** the unknown impulse response in **Fig. 2b,** and the resulting (noise-free) output in **Fig. 2c.** Note that we have truncated the output at the 60th sample; if a frequency domain identification method were to be applied to this data set, this truncation would produce severe ringing in the estimate of the impulse response. Note also that this is a difficult estimation problem because convolving the smooth input with the impulse response smears out the the sharp transitions in the impulse response and makes them difficult to recover from the resulting smooth output. **Figures 2d, 2e, and 2f** show the result of using generalized cross validation and constraints on the energy in (respectively) the impulse response, its first derivative, and its second derivative. **Figure 2g** shows the estimate obtained with a positivity constraint, and **Fig. 2h** shows the estimate obtained using the conjugate gradient algorithm.

Future Work

In FY 90 we will develop better statistical methods for determining the amount of constraint for the positivity and conjugate gradient methods, and we will continue to improve the numerical behavior of all our algorithms. We will develop performance measures for all our methods; this will include extending Cramér-Rao bounds for NLS to time and frequency domain measures of the impulse response variance, and developing variance and sensitivity measures for the nonparametric methods. Finally, we will continue to seek practical applications for these techniques within LLNL programs.

1. J. V. Candy, G. A. Clark, and D. M. Goodman, "Transient electromagnetic signal processing," *Time Domain Measurements in Electromagnetics*, E. K. Miller, Ed. (Van Nostrand Reinhold, New York, N.Y., 1986) Ch. 12.
2. D. G. Dudley and D. M. Goodman, "Transient identification and object classification," *Time Domain Measurements in Electromagnetics*, E. K. Miller, Ed. (Van Nostrand Reinhold, New York, N.Y., 1986) Ch. 13.
3. D. M. Goodman, NLS, "A system identification package for transient signals," Lawrence Livermore National Laboratory, Livermore, CA, UCID-19767, March 1983.
4. D. G. Dudley and D. M. Goodman, "System identification in electromagnetic scattering," Tutorial Paper, accepted for publication in *IEEE Trans. Antennas and Propagation*, to appear in 1990.
5. J. D. George and D. M. Goodman, "Application of the output error system identification method to the calibration of underwater acoustic transducers," *Proc. NATO Advanced Study Institute on Underwater Acoustic Data Processing*, Kingston, Ontario, July 1988.
6. M. D. Denny and D. M. Goodman, "A case study of the seismic source function: Salmon and Sterling reevaluated," submitted to *J. Geophysical Research*, 1989.
7. D. M. Goodman, H. Tran, D. G. Dudley, and R. R. Weyker, "The Cramer-Rao bound in electromagnetic inverse problems," *Progress in Electromagnetics Research Symp.*, Boston, July 1989.
8. U. K. Bhargava, R. L. Kashyap and D. M. Goodman, "Impulse response identification with deterministic inputs using nonparametric models," *IEEE Trans. Acoustics, Speech, and Signal Processing*, July 1987.
9. D. M. Goodman, "On the derivatives of the homomorphic transform operator and their application to some practical signal processing problems," to appear in *Proceedings of the IEEE*, April 1990.

Computational Recovery of Holographic Images

Thomas J. Yorkey, James M. Brase, J. Patrick Fitch, and Sean K. Lehman
Laser Engineering Division
Electronics Engineering

Gregory A. Clark
Engineering Research Division
Electronics Engineering

We have investigated the computational recovery of images from holograms. A set of simulation and reconstruction techniques has been developed for x-ray laser holography, and experiments in visible-light holography have been performed that validate their results. We have begun work on techniques for enhancement of resolution based on maximum-entropy image-restoration techniques.

Introduction

The Laboratory has issued a *Directors Initiative* to explore the possibility of using the LLNL-developed x-ray laser to produce three-dimensional holographic images of the distribution of chromosomes in a living cell. The only previously available method for looking at living cells has employed optical wavelengths, which yield a much lower resolution than would be available with x-rays. Although electron microscopy has a resolution potential higher than methods employing x-rays, the test samples must be dehydrated and doped with heavy atoms prior to examination.

To carry out the *Director's Initiative* and to prepare ourselves for the holographic x-ray laser imaging experiment, we have undertaken a project with the following goals:

- Develop a modeling ability for the design of an x-ray-laser imaging experiment.
- Explore advanced image-restoration concepts and algorithms to improve holographic-image resolution.
- Apply x-ray-laser holography, at 30-nm resolution, to the three-dimensional imaging of live cells.

Because complex biological objects of interest exist in an equally complex, three-dimensional environment, it is difficult to estimate the accuracy (and therefore the usefulness) of x-ray holographic images of these objects using the usual measures of resolution, such as Rayleigh criteria. To overcome this difficulty, we are approaching the design of x-ray holography systems by developing computational techniques for both the simulation of the formation of an x-ray hologram, and the subsequent computational reconstruction of the holographic image in three dimensions. Simulation allows us not only to establish resolution limits for the imaging of complex objects, but also to vary the parameters in a controlled manner and thus to optimize imaging system designs.

In the project's first year, we have concentrated on the following two efforts: (1) hologram simulation and image reconstruction techniques using wave-propagation algorithms, and (2) maximum-entropy image restoration techniques. We have developed and validated techniques for simulation and computational reconstruction of both Gabor and Fourier-transform holograms.

Validation of both the simulation and the reconstruction techniques was accomplished by scaling the wavelength and then applying the techniques at visible-light frequencies.

As an initial test of our simulation and reconstruction techniques, we have performed contrast-enhancement computer-modeling experiments using a 300-Å protein fiber in water with gold tags. Since the scattering properties of protein are not much different than water, we attached gold tags at known locations to make the protein stand out.

The finite size limitations of a detector, plus other practical considerations, reduce the final resolution. Next year, we will bring together the results of our current efforts and concentrate on resolution enhancement of reconstructed three-dimensional images, specifically, complex biological objects. Using *a priori* information, about the object may allow us to regain some of the lost resolution.

Progress in X-ray Laser Holography

Hologram Simulation

Our diffraction-based simulation treats the scattering volume (the test object) as a sequence of layers in the propagation direction z (see **Fig. 1**). In each layer, the index of refraction, n_i, is specified as a function of transverse coordinates x and y. We assume that $n_i(x,y)$ is constant through the layer thickness Δz.

The test object is illuminated by a plane wave, and the resulting wave is propagated from layer to layer. The attenuation and phase shift of the propagating wave due to the material in a single layer is given by the transmission coefficient:

$$t(x,y) = \exp[-k\beta(x,y)\,\Delta z]\exp[jk\alpha(x,y)\,\Delta z] \quad , \qquad (1)$$

where

k is the wave number of the illumination,
j is $\sqrt{-1}$,

and where the index of refraction is given as:

$$n_i(x,y) = 1 + \alpha(x,y) + j\beta(x,y) \quad . \qquad (2)$$

Note in **Eq. (1)**: $\exp[-k\beta(x,y)\,\Delta z]$ is the attenuation and $\exp[jk\alpha(x,y)\,\Delta z]$ is the phase shift.

At the beginning of nth layer, the incident field is multiplied by $t_n(x,y)$ and then propagated through distance Δz by convolution:[1]

$$U_n{+}1(x,y) = U_n(x,y)t_n(x,y) \;**\; Q(x,y) \quad , \qquad (3)$$

where

U_n is the field at the start of the nth layer,
** indicates two-dimensional convolution,
and where the quadratic phase function is given as:

$$Q(x,y) = \frac{\exp(jk\,\Delta z)}{j\lambda\,\Delta z}\exp\left[j\frac{k}{2\Delta z}(x^2+y^2)\right] \quad , \qquad (4)$$

where λ is the wavelength.

This process is repeated in turn for each layer of

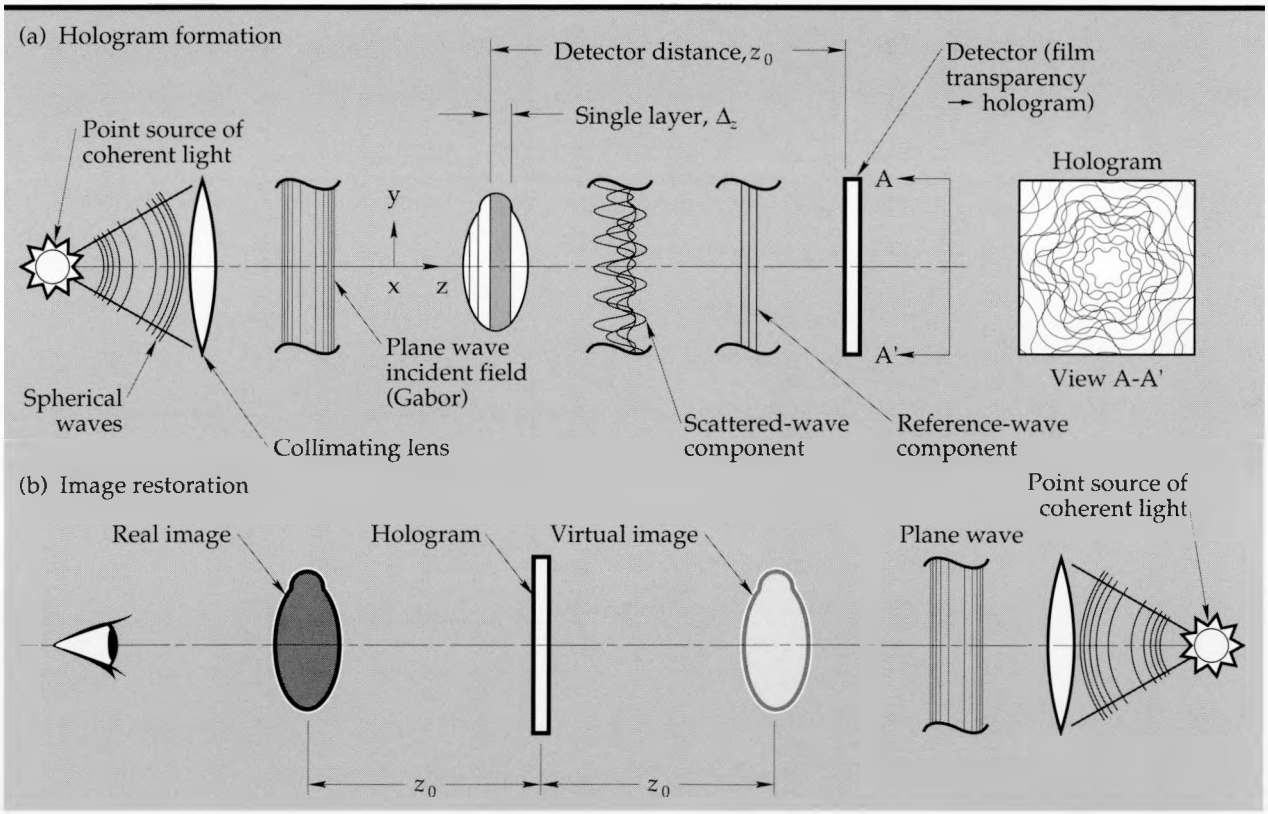


Figure 1. A cross-sectional view of the basic holographic setup and process using a plane-wave incident field for a single-view hologram: (a) To form a Gabor hologram of a thin test object, we illuminate it with a coherent light source and capture the diffraction pattern (the hologram) on a photograhic film transparency. (b) To reconstruct the object's image from the hologram, we illuminate the hologram from the opposite direction, using the same coherent source. In the image restoration process, the resulting field amplitude has both a virtual and a real image. The virtual (or twin) image causes distortion and limits the image resolution. This entire process can be computationally modeled and simulated.

the scattering volume. Then a final propagation step calculates the diffraction field at the detector. The detected hologram intensity is then calculated.

The effects from two sources of limited resolution must be included in the simulation: finitely limited detector resolution and finitely limited incident-beam illumination intensity. We already have some limited detector resolution arising from the discretization imposed by the convolution implementation, which uses discrete Fourier transforms. To study the effects of limited detector resolution, we initially calculate the hologram intensities at a higher resolution and then low-pass filter the resulting diffraction pattern. In this way, we can easily create, and thus evaluate, the limiting effects by varying the detector resolution with a single diffraction pattern.

Limited illumination intensity can reduce achievable resolution because of the resulting Poisson-distributed quantum noise. To generate a quantum-limited hologram, we first calculate the hologram intensities as described in the above paragraph. The relative intensities are then scaled to mean photon numbers using

$$m(x,y) = aN_i I(x,y) \qquad (5)$$

where

a is the detector element area,

N_i is the incident photon-flux density (in photons/cm^2), and

$I(x,y)$ is the simulated hologram intensity.

The final quantum-limited hologram is then formed by using $m(x,y)$ as the mean for realizing a Poisson-distributed random variable for each detector position.

Hologram Computational Reconstruction

Reconstruction from a single hologram is done using diffraction techniques similar to those used for simulation. The hologram is illuminated by a plane wave for the Gabor geometry. The resulting field is then propagated to a detector using **Eq. (3).** The detector distance can be incremented such that the entire scattering volume (the object of interest) is reconstructed.

When reconstructing a single hologram in three dimensions, the resolution in the longitudinal direction is typically much worse than in the transverse direction. The widths of the point-spread function, h, for the longitudinal and transverse directions are, respectively, $\sim\lambda/(NA)^2$ and $\sim\lambda/NA$, where NA is the numerical aperture of the imaging system. Since the NA is the ratio of the aperture radius to the propagation distance, this system has poor depth perception. (Typically NA is much less than 1.) A possible solution we are exploring to improve longitudinal resolution is to illuminate the object from many angles—that is, to employ x-ray holographic tomography.

For a weakly scattering object (that is, for an object that looks not much different than the background), a hologram contains information about the object corresponding to a spherical surface in the Fourier transform of the object (the Ewald sphere[2]). To reconstruct the object with good three-dimensional resolution, we must fill in the Fourier space out to the limiting spatial frequency defined by the illumination wavelength. The basic reconstruction procedure consists of propagating single holograms through the object volume and summing the resulting low resolution images.[3,4] For weakly scattering objects, fundamental questions remain on the number of angles required for any given resolution and the extent to which *a priori* knowledge can be used in the imaging process.

This procedure is shown graphically in **Fig. 2. Figure 2(a)** shows that many holograms are made of the same object at different angles. Mathematically, this corresponds to capturing information about the object along semicircles in the Fourier domain, as **Fig. 2(b)** shows. To reconstruct the object, each hologram is illuminated as shown in **Fig. 2(c).** Since Fourier space was not completely covered, the reconstruction will be imprecise.

Validation Using Visible-Light Experiments

Our development of computational tools for simulating and reconstructing x-ray holograms is closely linked to an experimental program in visible light holography. The experiments are scaled from x-ray wavelengths of 5 nm to visible light wavelengths of 500 nm. The resolution goal scales from 30 nm to 3 μm. We have obtained good agreement between experiment and simulation, and have demonstrated a capability to computationally reconstruct both Gabor and Fourier-transform holograms made with visible light.

Simulations Using Gold Tags for Contrast-Enhancement

Since, protein fibers do not scatter much light at x-ray wavelengths, we investigated putting gold tags on the fibers. As an initial test of our simulation and computational reconstruction techniques, we simulated holograms of a 300-Å-diameter protein fiber in water. **Figure 3(a)** shows the fiber labeled with 500-Å gold spheres to increase its x-

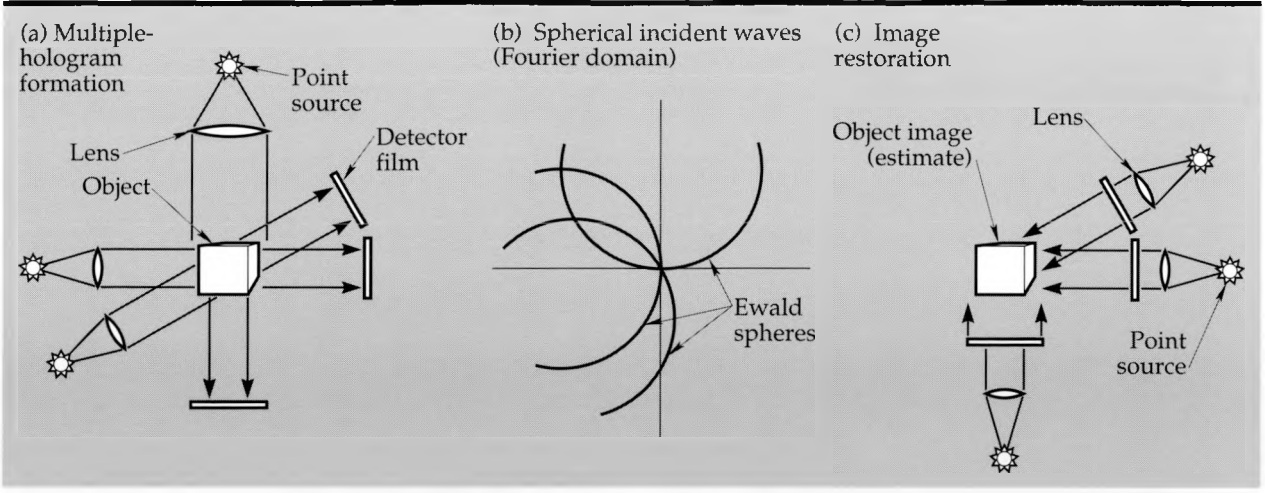


Figure 2. Plan view of the basic setup and three-step procedure for producing a full three-dimensional holographic image using spherical-wave (Fourier domain) incident fields and multiple views: (a) Many holograms are made of the same object at different angles. (b) Mathematically, this corresponds to capturing information about the object along semicircles in the Fourier domain. (c) To reconstruct the object, each detector is illuminated as shown.

ray scattering efficiency. **Figure 3(b)** is a simulated hologram of the unlabeled stand made with a source intensity of 4.4 mJ/cm^2 (10^{14} photons/cm^2). The hologram does not contain much information about the protein because the protein is such a weak scatterer. Therefore, this image would not be useful for structural studies. **Figure 3(c)** is a reconstructed image of the gold-labeled strand. The gold tags, which can be clearly located, provide useful structural information.

Progress in Maximum-Entropy Image-Restoration Techniques

Maximum-entropy deconvolution algorithms are designed to remove the effects of blur and noise from a signal or image. We have used maximum-entropy deconvolution algorithms to restore distorted signals and images. The concept of maximum entropy comes from information theory where entropy is a mathematical measure of information content. When entropy (information) is maximized, the resulting image has the most—and therefore, ideally, the correct—information about the signal.

Given a measured signal containing the effects of blur and independent noise, and given our knowledge of the point-spread function that produced the blur, the problem is to estimate the original, uncorrupted image signal.

When the entropy is maximized, the estimate of the original image is its best. As a constraint, the algorithms assume that the image is nonnegative, which aids in producing good restorations. Tests with simulated data show that the maximum-entropy algorithms generally produce better estimates of the uncorrupted image than any of the

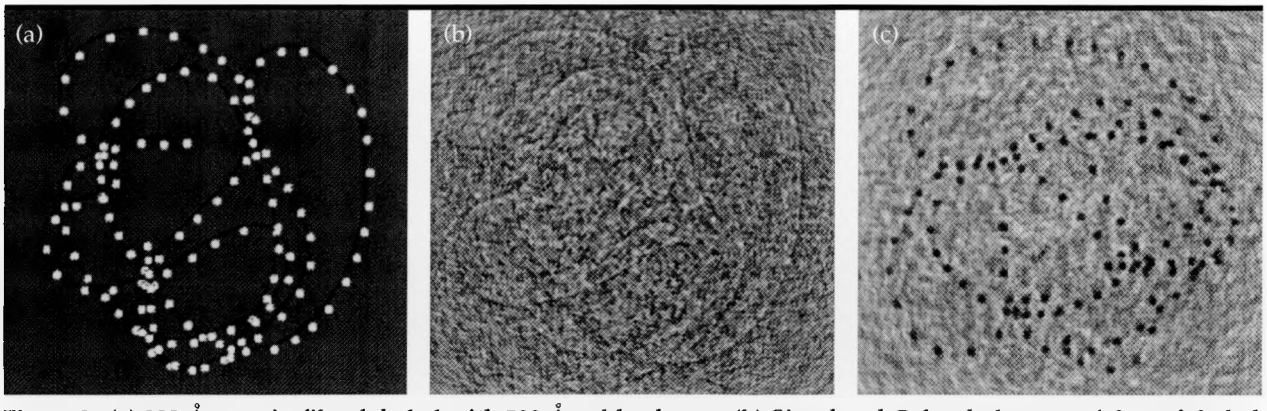


Figure 3. (a) 300-Å protein fiber labeled with 500-Å gold spheres. (b) Simulated Gabor hologram of the unlabeled strand made with a low source intensity. (c) Reconstructed image of gold-labeled strand, which shows that tagging can be used to increase visibility.

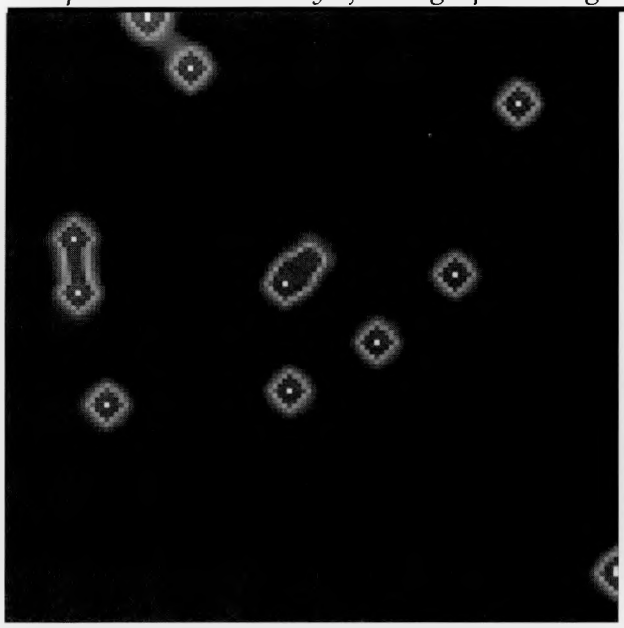


Figure 4. A simulated image of tagged DNA molecule blurred with a low-pass filter. The image of the molecule is distorted. When components of the molecule are located close to one another, they are not easily resolved.

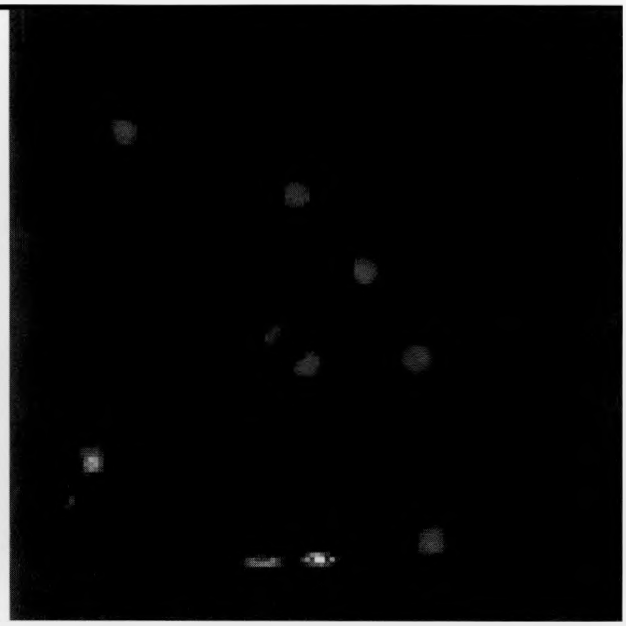


Figure 5. Restored image produced by processing the image in Fig. 4 with a two-dimensional, maximum-entropy algorithm of Frieden.[5] The blurring has been greatly reduced, and the iodine tags are now resolvable.

least-squares algorithms.

Figures 4 and **5** show the results of processing a simulated iodine-tagged-DNA image degraded by a low-pass filter. **Figure 4** is the blurred image, and **Fig. 5** is the restored image produced by processing the **Fig. 4** image with a two-dimensional, maximum-entropy algorithm of Frieden.[5] The restored image shows greatly improved resolution of the individual components.

Future Work

We will expand upon the application of these techniques to a systematic study of the resolution enhancement in x-ray imaging of biological objects.

1. J. W. Goodman, *Introduction to Fourier Optics* (McGraw-Hill, San Francisco, 1968).
2. E. Wolf, "Three-Dimensional Structure Determination of Semi-Transparent Objects from Holographic Data," *Opt. Comm.* **1**(1), 153–156 (1969).
3. A. J. Devaney, "A Filtered Backpropagation Algorithm for Diffraction Images," *Ultrasonic Imaging* **4**, 336–350 (1982).
4. A. J. Devaney, "Structure Determination from Intensity Measurements in Scattering Experiments," *Phys. Rev. Lett.* **62**(20), 2385–2388 (1989).
5. B. R. Frieden, "Restoring with Maximum Likelihood and Maximum Entropy," *J. Opt. Soc. Am.* **62**(4), 511–518 (1972).

ENGINEERING RESEARCH AND DEVELOPMENT

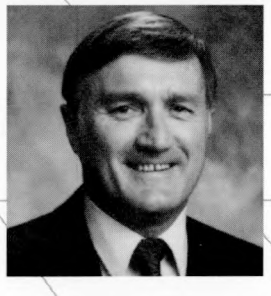
Thrust Area Report FY 89

New Initiatives

New Initiatives research and development encourages idea that are not covered by the thrust areas. Our task is to seek innovative and highly leveraged concepts that will have a large payoff. This past year we funded projects in four technologies:

- The x-ray imaging project developed a technique in which an opaque sphere was used to diffract x rays.
- The artificial intelligence work developed an expert system to assist in the control of a particle beam accelerator.
- We extended our previous work in parallel processing by expanding and upgrading the Sprint computer to enable us to run very large electromagnetic codes.
- The scientific visualization and data management projects wrote new software algorithms that enhanced our ability to "see" data produced from experiments and computer simulations.

Edward A. Lafranchi
Deputy Associate Director for Engineering



SPRINT 2 as a Computing Engine for Electromagnetic Simulations

W. Travis White III, Bryan S. Lawver, Anthony J. DeGroot, and Erik M. Johansson
Engineering Research Division
Electronics Engineering

This year we upgraded the computer hardware of our SPRINT parallel-processing system with faster processors, more memory per processor (1 Mbyte), and a faster and more versatile interprocessor connection board. We tested the improved system, which we call SPRINT 2, by using it to execute the TSAR three-dimensional, finite-difference, time-domain electromagnetics code. By adapting the architecture of SPRINT 2 to match the spatial grid of TSAR, we created a computing engine for solving electromagnetics problems. On one 4-million-word problem, SPRINT 2 executed TSAR about 40% as fast as a single Cray X/MP processor, yet our machine can be built for less than 0.5% of the cost of a Cray. To demonstrate the usefulness of SPRINT 2 on problems of practical interest to LLNL, we modeled the leakage of a microwave pulse into the autopilot/battery section of a conventional missile, the results of which are presented here.

Introduction

The Engineering Department at the Laboratory has long-standing requirements to solve large electromagnetics (EM) problems. These include analyzing electromagnetic pulse (EMP) effects and traveling waves, characterizing antennas, and estimating radar cross sections. Traditionally, LLNL has relied on supercomputers such as the Cray X/MP to solve such problems. Almost always, computer performance has limited the size of the problems that could be solved. If someone could make a less expensive computer (costing, say, $20,000 to $100,000) that could solve EM problems with tens of millions of unknowns as fast as a supercomputer (costing $10 million to $20 million), the Laboratory would benefit significantly.

One promising alternative to supercomputing is parallel processing (also called "concurrent processing"), a method of computing in which two or more tasks are done simultaneously rather than sequentially.[1] In 1985 we began designing and building an array of processors to serve as a test bed for new parallel processing schemes. The array, called SPRINT, consisted of a reconfigurable network of 64 transputers, each equipped with 128 kbytes of memory.[2] A specially designed and totally electronic interprocessor connection panel permitted almost arbitrary architecture (paths of communication) among individual processors in the network. An early and important result was that with a suitable pattern of interconnections, some algorithms such as simple matrix multiplication actually executed faster on SPRINT than on a Cray X/MP. Tailoring the architecture of a multiprocessor computer to match specific algorithms provided an extra degree of freedom in optimizing performance as compared with most single-processor computers. We experimented with a variety of algorithms and architectures and documented the results.[3-7]

One numerical method that seemed particularly well suited to SPRINT was finite-differencing. Finite-difference computer codes typically have simple algorithms and logically regular grids. In particular, efficient time-domain EM codes often use finite-difference methods. Given the general importance of EM computations to the Laboratory and the suitability of finite-difference methods to parallel processing, we began to consider making a SPRINT-based computing engine for solving EM problems.[8]

In 1988, we implemented on SPRINT a simplified version of the in-house finite-difference EM code TSAR. Due to the limited memory per processor, this simplified version of TSAR lacked important features necessary for useful simulations of EMP effects—features such as the use of radiation boundary conditions, the ability to use an EM wave originating outside the computational grid, and the ability to monitor arbitrary field compo-

nents at arbitrary times and locations throughout the grid. The only way to incorporate this missing software was to expand the hardware.

In this article, we discuss three points. First, we describe how we modified SPRINT by upgrading the hardware. Then we summarize our implementation of the TSAR code to make, in essence, a computing engine for solving EM problems. Finally, we present the results of a large computation—the coupling of an external microwave pulse into a missile hull—to show the usefulness of the SPRINT system in EM applications of interest to the Laboratory.

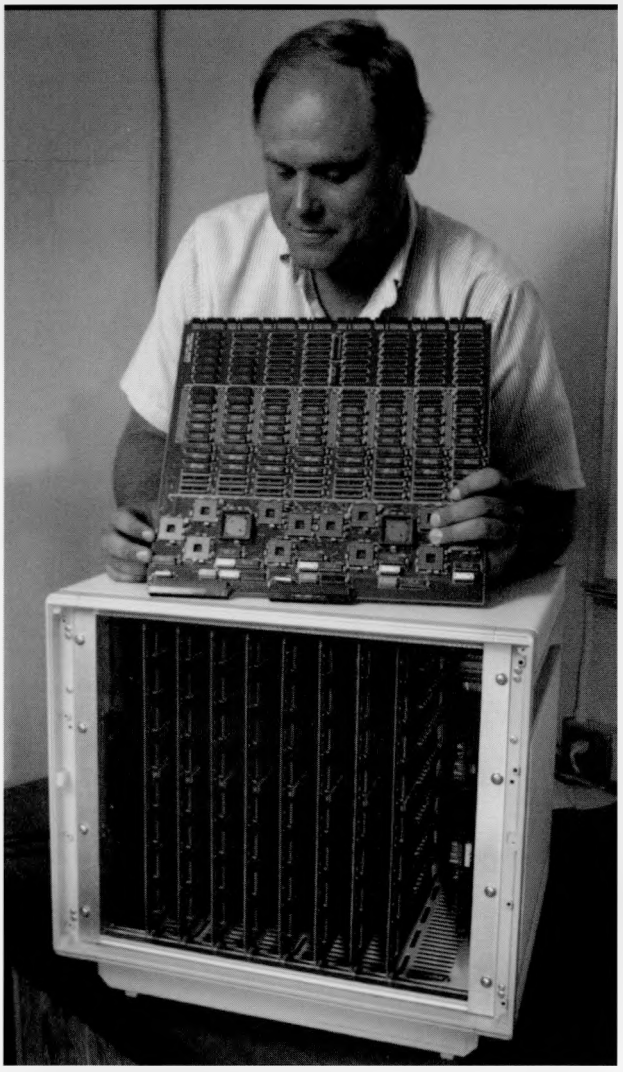


Figure 1. The SPRINT 2 chassis box contains a power supply and eight boards with eight processors per board. The engineer is holding a partially filled processor board, to show the layout of the components. This LLNL-designed system uses less than 200 W of power but runs about as fast as one processor of a Cray X/MP in many useful applications. Not shown are the host computer (a MicroVAX II) and the ancillary hardware used for storing and displaying data.

Progress

SPRINT 2

This year we completely redesigned and rebuilt the SPRINT into a system we call SPRINT 2. We replaced all of the transputer processors with faster, floating-point T800 transputers and increased the memory eightfold to 1 Mbyte per processor. Each T800 transputer is roughly equivalent to 1.5 VAX 11/785 computers in terms of computational speed and word size. Using a 32×32 switch, we redesigned and rebuilt the programmable interprocessor connection system to permit more flexible patterns of nodal connections and achieve faster communications among processors. With the new design, we can now effectively rewire the network in less than a microsecond so that the topology can be shifted quickly to any arbitrary network of 64 four-port computers. We also designed a plug-in board for future implementation of high-speed direct graphics (40 Mbytes/sec).

As in SPRINT, each processor in SPRINT 2 has its own memory module. There is no central shared memory. Each transputer in SPRINT 2 has four data links, a limitation imposed by the manufacturer. For any pair of processors to share data, they must communicate via the data links. The data links are asynchronous, full-duplex links capable of transferring up to 20 Mbits/sec. In the language of computer scientists, SPRINT 2 operates as either a single-instruction-multiple-data (SIMD) machine or as a multiple-instruction-multiple-data (MIMD) machine. Processors operating in the SIMD mode execute the same instruction simultaneously but on different data. Processors operating in the MIMD mode can execute independent instructions. MIMD systems afford better division of labor than SIMD systems for a fixed number of processors. Thus, on a per-processor basis, SPRINT 2 shows more performance than, say, a Connection Machine, which is a strictly SIMD device.

A MicroVAX II serves as a controller for the system. The MicroVAX partitions a given problem among the processors, assigns the network topology to the system, and performs limited post-processing of results from SPRINT 2 if desired. In addition, program development is performed on the MicroVAX.

Figure 1 shows a photograph of SPRINT 2. The unit, excluding the host computer, occupies about 3 ft^3 (0.1 m^3) and requires no more power to operate than a pair of 100-W light bulbs. The low power consumption results from the extensive use of CMOS chips. The improved design has attracted

interest outside LLNL, and the Laboratory is investigating the possibility of transferring this technology to private industry.

In cost and performance SPRINT 2 compares favorably with other parallel processors that are commercially available. Generally speaking, both performance and cost scale linearly with the number of processors in a system. Thus, SPRINT 2 is orders of magnitude faster and more powerful than the simple one- or four-processor boards that can be purchased commercially to boost performance of personal computers built by IBM or Apple. On the other hand, it is correspondingly less powerful than, say, Sandia Laboratory's 1024-node hypercube[9] or a 65,000-node Connection Machine. Advantages that SPRINT 2 has over the big machines are its low cost and, for us, its ease of use. Standard Fortran and C codes can be implemented easily on it. When configured in such a way as to optimize performance for a specific class of problem, SPRINT 2 becomes a computing engine for solving such problems.

Implementing the TSAR Electromagnetics Code

TSAR is a finite-difference code that solves Maxwell's equations of electrodynamics as a function of time in three spatial dimensions.[10] It uses a cubic lattice of cells and a popular scheme of spatiotemporal sampling in which, for each cell of the grid, one electric field component is sampled at the midpoint of each edge of the cell, and one magnetic field component is sampled at the midpoint of each face of the cell. The electric and magnetic fields are updated at alternate half-steps in time. An important feature of this arrangement is that the updating of a given field component requires information only from nearest-neighbor cells in the grid.

The nearest-neighbor communication and the nature of the offsets in sampling are the keys to implementing TSAR on SPRINT 2. The local communication ensures that all grid points assigned to an individual processor can be updated on a given time step, including those few points on the boundary of the processor's domain, which can be updated with data obtained from neighboring processors. If global communication were required, each processor would have to communicate with every other processor. However, the SPRINT architecture is such that only nearest-neighbor communication is permitted between processors. Global communication would require information to be relayed across the network rather than passed directly, which would drastically slow the performance of the system.

Adapting TSAR to run on SPRINT 2 required very little specialized programming, since most of the code was written in standard Fortran. Portions of the code that were not standard were revised. To partition a problem among the transputer arrays, we wrote some programs that first subdivided the grid into uniform pieces and then broadcast identical copies of the compiled version of TSAR to each transputer.[11] For SPRINT 2, with its greatly increased memory, we reinstated the features of TSAR that the limited memory of the earlier SPRINT had forced us to omit.

We tested the performance of our SPRINT 2 implementation of TSAR on several EM problems that we had previously executed on other computers. These tests demonstrated that the code worked correctly. On a 4-million-word problem, we found that the full code required approximately 0.8-0.9 msec per time step per field component per grid cell to execute (including input/output). This is roughly 2.5 to 3 times slower than running TSAR on a single processor of a Cray X/MP. For a given total problem size, the execution time diminished in proportion to the number of processors being used. Also, for a fixed problem size per processor, the total problem could be increased without requiring more real time to execute. Both of these statements reflect the so-called "linear speedup" that indicates near-ideal operation of a parallel processing system, and they show that the time spent on interprocessor communication was small compared to the time spent computing.

Figure 2 illustrates how a grid might be partitioned among several processors. As suggested by the figure, the electric fields at the left boundary of each processor are sent to the adjoining processor to update the neighbor's boundary values of magnetic fields at half-integer values of the time step. Conversely, the magnetic fields at the right of each processor are sent to the adjoining processor to update the neighbor's electric fields at integral values of the time step. The figure depicts a grid that has been cut like a loaf of bread into slabs. In practice, we find it more efficient to cut the grid like a potato into french fries. Therefore, we partitioned our problems into an 8×8 array of long, rectangular blocks. Furthermore, due to internal symmetries in the TSAR code, we always operated with an even number of points per processor.

Modeling Microwave Penetration Into a Missile

To demonstrate the usefulness of our computer system and software on a problem of practical interest, we modeled the penetration of a micro-

wave pulse into the autopilot/battery section of a retired conventional missile, a Standard ARM RGM-66D2X from General Dynamics.[12] Computed quantities included induced currents, power leakage, and instantaneous field contours. The modeled section contained a battery pack, autopilot electronics, gyroscopes, a wiring harness, and a few smaller components. To simulate the rest of the missile hull, we used two flat-bottomed, hollow, cylindrical end caps abutting the modeled section like open coffee cans with their bottoms against a pipe. A one-cell-wide circumferential gap separated each end cap from the autopilot/battery section. The overall size of the object, including the two end caps, was about 3 ft in length by 14 in. in diameter. Spatial resolution of the model was 0.3 in. on a simple cubic lattice of half a million grid points.

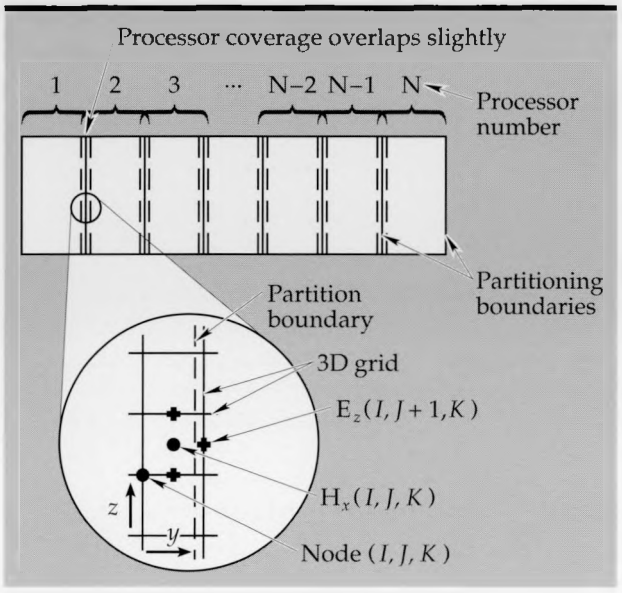


Figure 2. Example of partitioning a grid for the TSAR code. Cell (*I,J,K*) is an arbitrary cell at the grid boundary common to two processors. The tangential electric and magnetic fields (E and H, respectively) must be shared by both processors. The expanded view of the boundary shows that the electric and magnetic field components associated with cell (*I,J,K*) are offset from the corner of the cell, node (*I,J,K*), as discussed in the text.

Figure 3 shows the autopilot/battery section of the hull (without the end caps) and the wiring harness. From the illustration, one can appreciate the geometric complexity involved in modeling real hardware. With present technology, the process of digitizing a complex object often takes as much human labor as the process of running an EM code. The figure, which is computer-drawn from engineering blueprints, depicts the parts of the object in a fairly realistic way with smooth surfaces and nonuniform spatial sampling. The representation seen by our EM code has a more jagged, stair-stepped appearance than **Fig. 3** due to the cubic arrangement of sample points.

Our simulated microwave pulse shape was the time derivative of a simple Gaussian function, and its duration was about 0.4 nsec. The plane wave struck the missile section squarely from the starboard side, with the electric field polarized parallel to the axis of the hull. This provided maximum coupling to the interior of the missile. Except for the two circumferential slots between the end caps and the body, all apertures in the autopilot section were sealed shut, permitting us to investigate com-

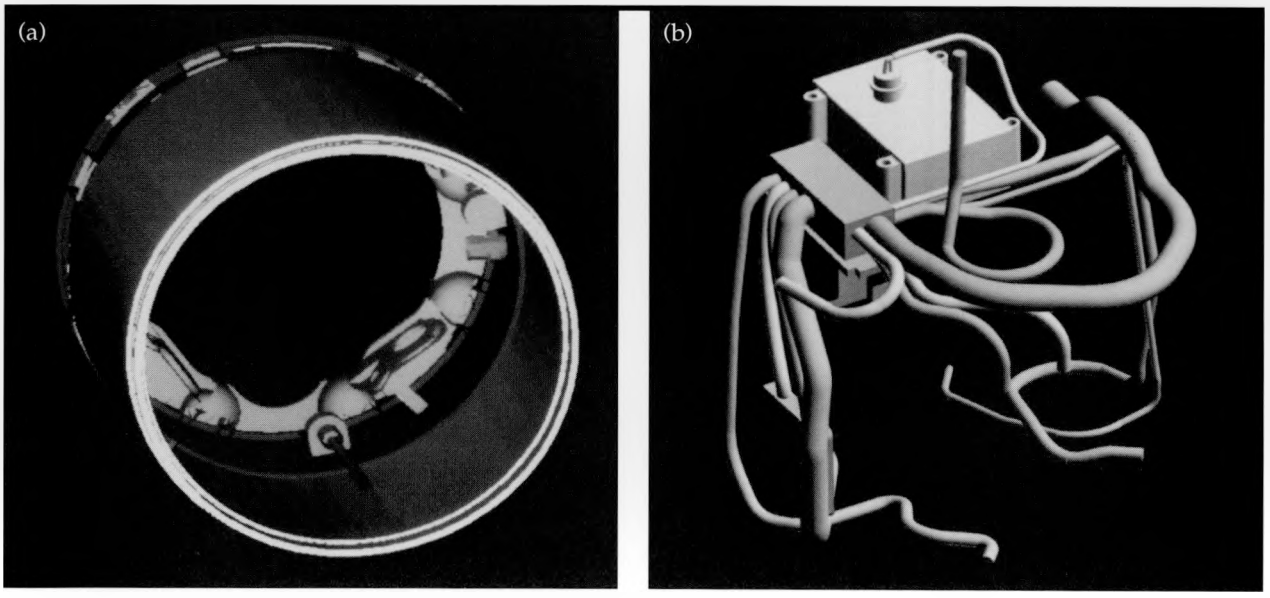


Figure 3. View of selected components of the autopilot/battery section of a standard ARM missile. (a) Hull. (b) Wiring harness.

putationally the general effects of microwave leakage at body joints. Since the main purpose of this work was to demonstrate the usefulness of a method of performing EM calculations rather than to perform a detailed simulation, we treated only linear coupling. We normalized all results to an incident field of unit peak electric intensity and disregarded effects such as air breakdown near sharp edges, which can be important at high field strengths.

Figures 4 and 5 show the response of a wire inside the cavity. In applications such as estimating vulnerability of systems to EMP effects, it is necessary to compute how much current an EM wave that leaks into an enclosure will induce on an internal wire. For this reason, it is important for EM codes to have at least a crude capability to model wires within bigger systems. The wire in this example protruded 0.3 in. into the cavity from the aft end cap and was L-shaped, with the longer portion (7.8 in.) directed horizontally toward the incoming beam. **Figure 4** shows the calculated instantaneous current induced at the base of the wire. The incident microwave pulse is shown in the inset. The wire tended to ring thousands of times longer than the duration of the incident pulse, due to a combination of undamped cavity and wire resonances. These resonances appear as narrow spikes in **Fig. 5,** which shows the spectral response function in amperes per unit electrical field amplitude (m/ohm) obtained by Fourier-transforming the input pulse and the instantaneous current shown in **Fig. 4.**

Figure 6 shows the net instantaneous power entering the cavity and the accumulated field energy stored in the cavity as functions of time. To determine the incident power, we integrated the incoming electromagnetic flux (Poynting's vector dotted into the inward normal) around the two circumferential slots. We sampled Poynting's vector, which is the vector cross-product of the electric and magnetic field vectors, at points coincident with magnetic-field sample points and used linear interpolation to compensate for the spatial and temporal offsets between the electric and magnetic field components. Specifically, we spatially averaged the electric field and temporally averaged the magnetic field. As the figure indicates, the electromagnetic energy in the cavity rose quickly, due to the fast input pulse, but then it dissipated slowly. This is because the model included no strong damping mechanisms such as ohmic losses that result from skin currents on the imperfectly conductive surface of the hull. The slow bleed-off of electromagnetic energy is consistent with the long ringing of the wire shown in **Fig. 4.** We noted that

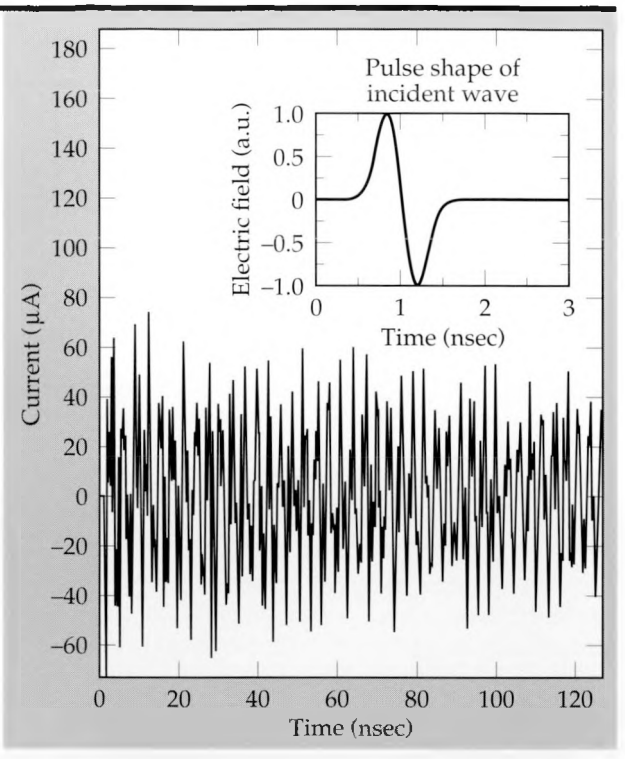


Figure 4. Calculated current induced in a wire inside the missile case by microwave leakage through the circumferential gap between each end plate and the hull of the autopilot section. The incident pulse assumed for the calculation (see inset) has a peak amplitude of 1 V/m. The calculated ringing induced by the subnanosecond pulse persists for microseconds due to the model's lack of explicit damping mechanisms such as skin loss. Current is measured at the base of the wire.

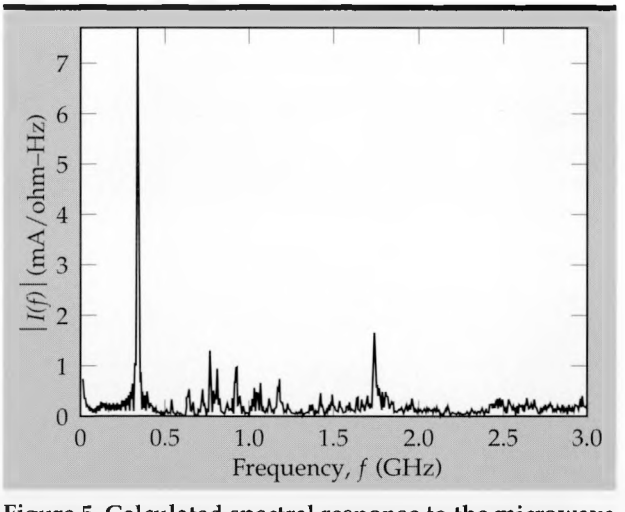


Figure 5. Calculated spectral response to the microwave pulse by a wire inside the standard ARM missile. The resonant spike at 0.34 GHz is slightly below 0.36 GHz, the fundamental resonance of a straight wire of identical length protruding from a perfect ground plane.

the calculated induced currents on the exterior of the missile tended to persist a long time, too. This persistence occurred despite the fact that the external skin currents can radiate their energy into free space and so would be expected to damp quickly. As a result of the persistent external skin currents, energy tended to leak into the cavity even after the microwave pulse had long vanished. This delayed microwave penetration appears as weak oscillations in **Fig. 6** in the plot of instantaneous power entering the cavity of the missile. We suspect that the persistent external currents of the calculation are a numerical artifact due to imperfect boundary conditions.

Figure 7 shows a pair of contour plots that represent lines of constant axial electric field as computed by our code. The contours are stepped logarithmically in bels (dB/10). **Figure 7a,** a top view, shows a plane parallel to the direction of propagation of the incident beam. **Figure 7b,** a side view, shows a plane perpendicular to the incident beam. The large blank areas between contours are the conductive portions of the object, which the field cannot penetrate: the wire harness, the battery, and the H-shaped hull formed by the shell of the autopilot section and its two end caps.

All of these TSAR computational results on the ARM missile required about 7 hr of processor time on SPRINT 2, which was equivalent to about 2.9 hr of processor time on a single processor of the Cray X/MP using the TSAR program. We did some of the postprocessing in parallel with the main computation by using the MicroVAX host computer of SPRINT 2. Because these postprocessing computations required less work than the main calculation, they did not slow the performance of the transputer array. The rest of the postprocessing, such as Fourier-transforming and plotting the data, we did later on a separate VAX 6330 computer.

The biggest bottleneck in using SPRINT 2 for large EM problems proved to be in entering the

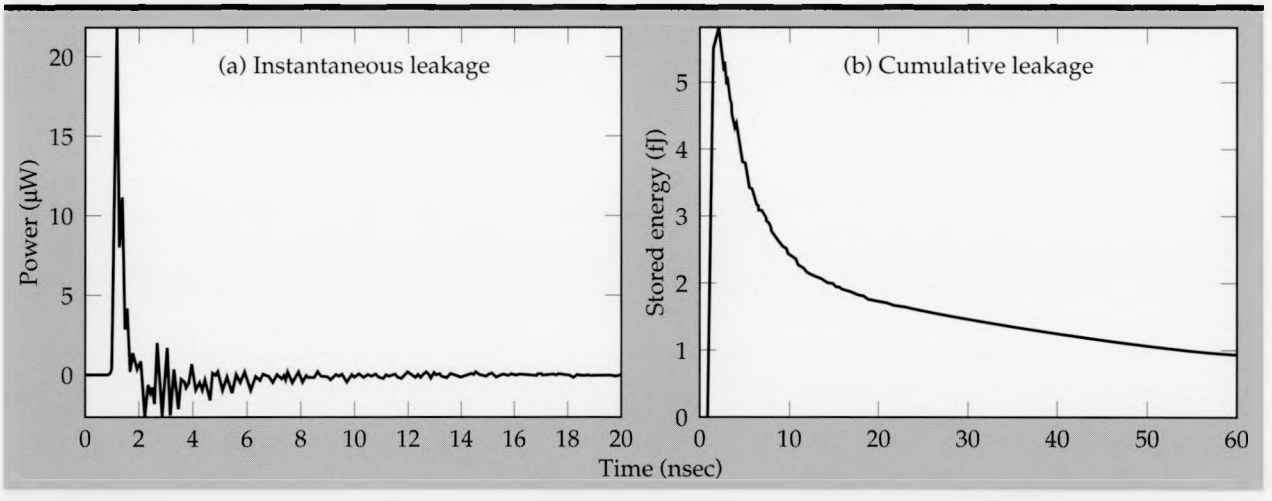


Figure 6. (a) Calculated instantaneous power leaking into the missile section, represented by the spatial integral of the electromagnetic energy flux (Poynting's vector) around the slots at the end caps of the missile section. (b) Calculated electromagnetic energy inside the missile section, represented by the time integral of (a).

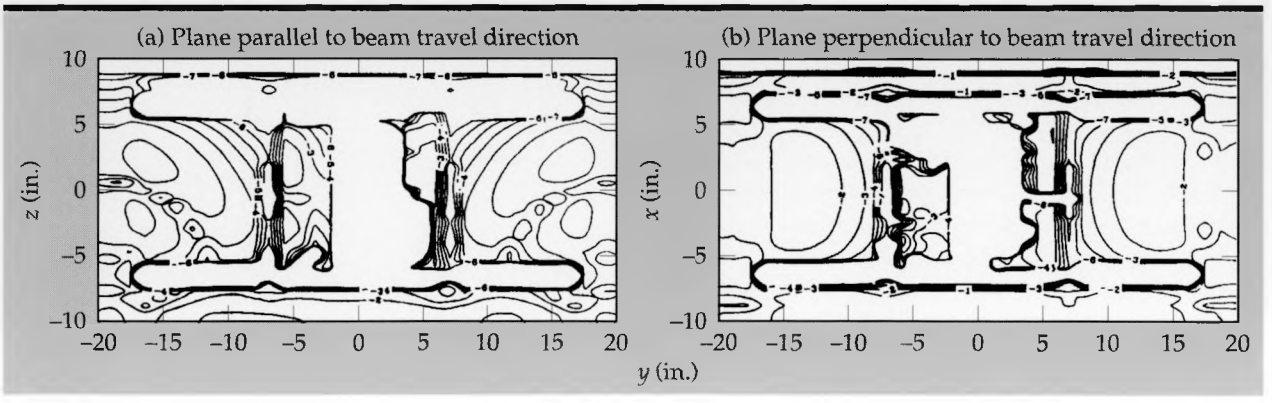


Figure 7. Contours of constant axial electric field: (a) In a plane coaxial with the missile and parallel to the direction of travel of the incident microwave beam. (b) In a plane coaxial with the missile and perpendicular to the direction of travel of the incident beam.

initial data. Our missile required 300,000 integers in order to define its structure. The MicroVAX host could read the input data and partition it among the 64 processors of SPRINT 2 at a rate of only about one point per 6 msec, or 30 min for the entire input array. We believe we can reduce the initialization time by a factor of three by revising our input algorithm.

The output flowed much faster than the input. This was partly because the amount of data coming out of the processors on any given time step was small compared to the amount of data required to initialize the processors. At most, we extracted 8000 out of a possible 4 million values per time step. The time required for output in this "worst case" was five to ten percent of the total job time.

Future Work

As the performance of microprocessors improves and the cost drops, the general trend toward using massively parallel processing is expected to extend to many other types of technical problems in addition to numerical simulations—for example, to database management, data analysis, and command and control. To explore the new processing techniques, there will be a continuing need for a test bed that will easily allow us to convert existing software to a multiprocessing environment. In addition, because the architecture of parallel processor arrays continues to evolve, fast and versatile ways to reconfigure the interconnections among processors will be necessary. SPRINT 2, with its 64 transputers, readily reconfigurable interconnection network, and Fortran and C compilers, is an ideal test bed already available to users at the Laboratory for testing future concepts in parallel processing.

In FY 90 we plan to continue using SPRINT 2 as a test bed, not only for EM problems but also for a variety of other parallel computations. Problems we plan to attack include advanced modeling of molecular dynamics for H-Division, cataloguing and matching of human gene patterns for Bio-Med Division, and possibly simulation of solid-state electro-optical devices for Engineering Research Division. Hopefully, this work will provide a stepping-stone toward "Cray on a wafer" technology of the future, with hundreds or thousands of processors sharing a 4-in.-diam silicon wafer, performing billions of operations a second at near room temperature.

We also plan to improve the SPRINT 2 hardware next year by adding high-speed (40-Mbyte/sec) video buffers that will allow us to pipeline output from the processors to a graphics device as fast as results are calculated. Finally, we will build two different SPRINT-like arrays of transputers—one to perform tomographic image analysis, and one to demonstrate the feasibility of test-ban treaty verification.

In summary, from 1985 through 1988 we designed and built SPRINT and showed that it could provide a cost-effective alternative to supercomputing in selected areas such as image analysis and computed tomography. In 1989 we upgraded SPRINT to SPRINT 2 and showed that it could compete favorably with supercomputers as a computing engine for EM problems. In the future, we plan to continue to explore new applications for inexpensive arrays of dedicated parallel processors.

1. S. R. Colley, "Parallel Solutions to Parallel Problems," *R&D*, 42–45 (November 21, 1989).
2. A. J. De Groot, E. M. Johansson, J. P. Fitch, C. W. Grant, and S. R. Parker, "SPRINT—The Systolic Processor With a Reconfigurable Interconnection Network of Transputers," *IEEE Trans. Nucl. Sci.* **NS-34**(4), 873–877 (1987).
3. A. J. De Groot, *Bit-Level Systolic Arrays*, Ph.D. Dissertation, University of California, Davis (June 1989).
4. A. J. De Groot and S. R. Parker, "Systolic Implementation of Neural Networks," *Proc. Soc. Photo-Optical Instrumentation Engineers, High Speed Computing II*, Vol. 1058 (January 1989).
5. A. J. De Groot, S. R. Parker, and E. M. Johansson, "Image Processing Using the SPRINT Multiprocessor System," to appear in Proceedings of the IEEE International Conference on Systems Engineering, August 1989.
6. A. J. De Groot and S. R. Parker, "Systolic Processing for Large-Scale, Real-Time Computations," May•June 1989 issue of *Energy and Technology Review*, Lawrence Livermore National Laboratory, UCRL-52000-89-5•6, pp. 21–32.
7. W. G. Hoover, C. G. Hoover, I. F. Stowers, A. J. De Groot, and B. Moran, "Simulation of Mechanical Deformation Via Nonequilibrium Molecular Dynamics," Proceedings of the CECAM Meetings on Microscopic Simulation of Macroscopic Flows, University of Brussels, Brussels, Belgium, August 21–25, 1989.
8. W. T. White III, A. Taflove, J. C. Stringer, and R. F. Kluge, "Design and Implementation of an Efficient Finite-Difference, Time-Domain Computer Code for Large Problems (U)," Proceedings of the Third National Conference on High Power Microwave Technology for Defense Applications (Secret), Kirtland Air Force Base, New Mexico, December 1–5, 1986.

9. J. L. Gustafson, G. R. Montry, and R. E. Benner, "Development of Parallel Methods for a 1024-Processor Hypercube," *SIAM J. Sci. Stat. Comp.* **9**(4), 1–32 (July 1988).
10. R. R. McLeod and W. T. White III, *Temporal Scattering and Response Software User's Manual Version 1.0*, Lawrence Livermore National Laboratory, M-211 (September 1987).
11. B. S. Lawver and A. J. De Groot, "Partitioning a Finite Difference Code for a Local Memory Multiprocessor," Fourth Conference on Hypercubes, Concurrent Computers, and Applications, March 1989.
12. R. J. Gardner, Jr., and R. B. McDonald, *Final Report for the Generic Missile Section Tests*, Vols. I and II, General Dynamics Pomona Division, TMS6-344-86-019 and -020 (August 11, 1986).

Scientific Database Project

Everett W. Wheelock
Nuclear Energy Systems Division
Electronics Engineering

The Scientific Database Project has developed a system architecture that uses a relational database to integrate the elements of data management: data acquisition, analysis, display, and archiving. The system architecture can be applied to almost all Laboratory experiments where large amounts of data are acquired and manipulated, yet may be easily customized for specific programmatic requirements. In FY 88 we developed a prototype system that could be modified for use on a variety of computer systems. In FY 89 we produced a template that is now being used in several Laboratory programs.

Introduction

The new generation of digital data acquisition systems being used at the Laboratory are increasingly powerful, often able to acquire more information in a single day than can be easily analyzed in a month. Tracking and manipulating such large amounts of data have become increasingly onerous, raising such problems as:

- Cryptic data file names
  - "UFD1S2.DAS"
- A lack of connections between related pieces of data
  - Raw data on a floppy disk
  - Equipment settings in a hand-written logbook
  - Analytical results stored on a different computer
- No easy way to sort and select data for analysis
- No data backup/archiving system.

The goal of the Scientific Database Project is to develop a system architecture that relieves this data management bottleneck by integrating data acquisition, analysis, display, and archiving within the framework of a relational database.[1] This system provides a flexible and powerful method of organizing and retrieving the data generated by various Laboratory experiments.

The application program that implements this general architecture is called "SciDB" and its principal elements are:

- A relational database called INGRES, which is used as a central data server
- Links to VIEW and SIG, signal and image processing codes
- Links to data acquisition systems
- A pulldown menu interface designed for ASCII terminals, which can be extended to the new generation of bit-mapped graphics displays.

SciDB System Organization

The SciDB prototype currently runs on a VAX organized as an independent set of modular subsystems. **Figure 1** shows the major subsystems and linkages. Access to the system is provided by an easy-to-use pull-down menu. The core database is implemented in INGRES, a commercial relational database management system. Information flow to and from the core database is handled by custom linkage routines. These links automatically convert data into a format that can be read by external data acquisition systems and analysis codes.

The majority of the code is written in C to help ensure cross-system portability. This and the modular design permit us to modify and upgrade the system easily as well as to customize it for specific applications. Care has been taken with the design of the menu to ensure that new and occasional users may perform standard operations without being overwhelmed by the system's complexity. The interface also gives experienced users access to the powerful features of the relational database for special operations. On line help is available at all times.

SciDB Data Model

Figure 2 shows a basic data model used to represent information within the database. In the SciDB system, an "experiment" is the collection of information pertaining to a single event. This generally consists of the equipment configuration, equipment settings, administrative information (experiment name, date, description, etc.), and the data collected from the event. The experiment configuration consists of a set of "datapaths." Datapaths are a block diagram description of the topology of a single source of information concerning an event; i.e., the route "travelled" by a single "chan-

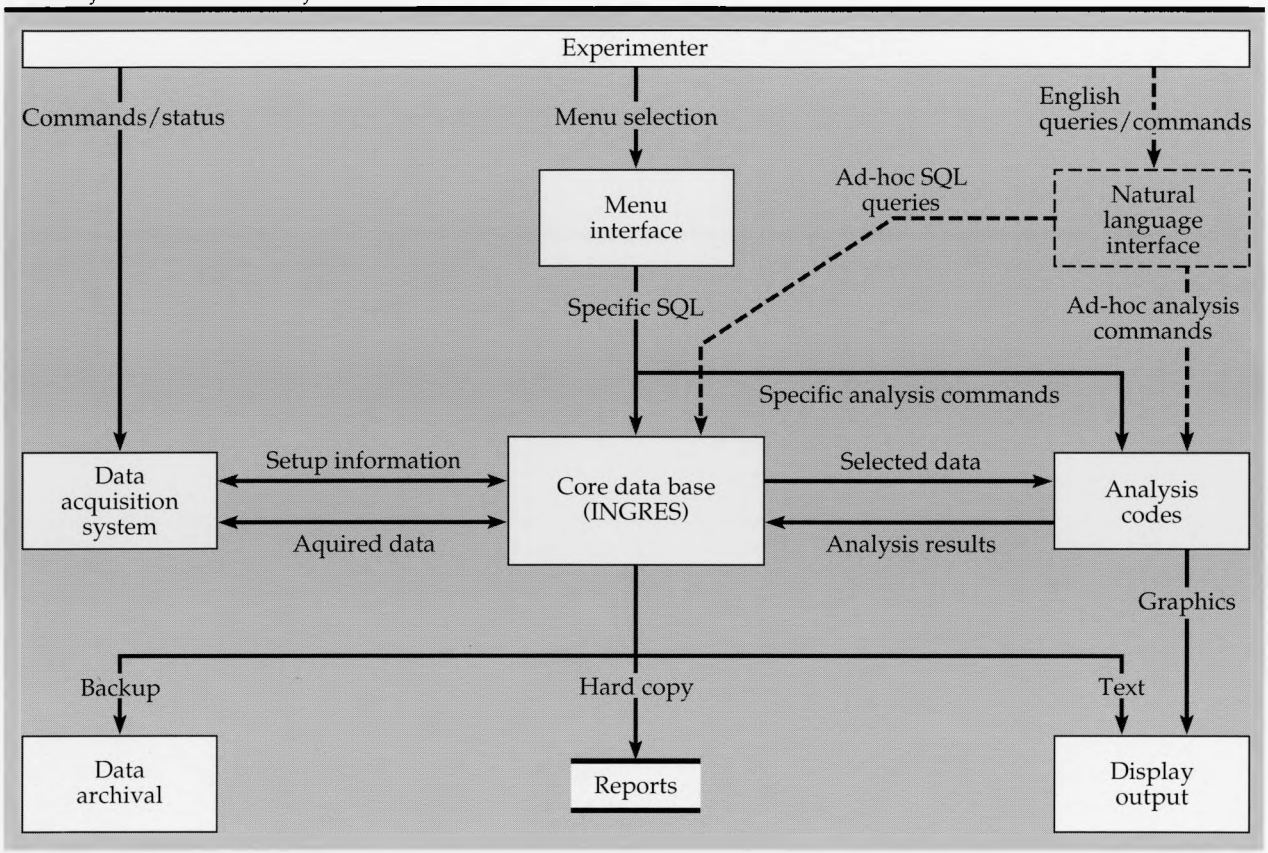


Figure 1. SciDB System Architecture. The menu shown at the top center of the figure accesses INGRES, a commercial database. The data that flows to and from INGRES is handled by custom linkage routines. These links convert the data into a format that can be read by data acquisition systems and analysis codes. The dotted arrows and circle on the right of the figure indicate a future natural language interface to the system.

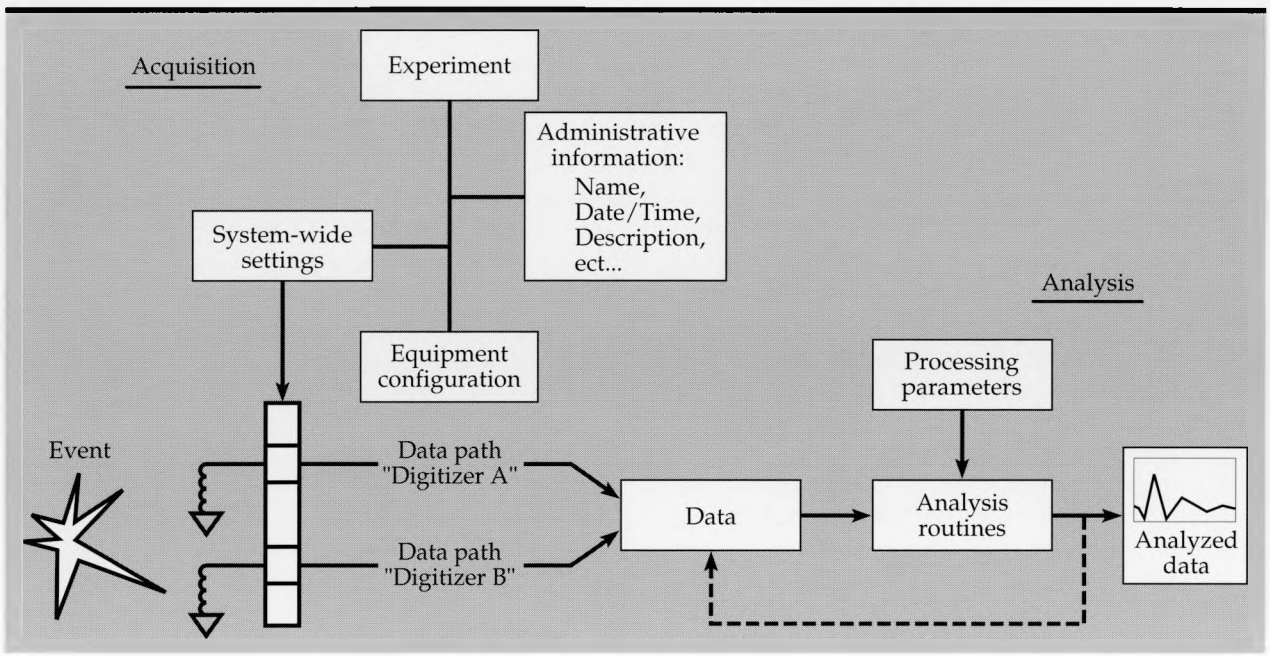


Figure 2. SciDB Data Model. The figure shows conceptual groupings of information that are represented as tables in the core database. Analyzed data remains linked to original raw data and experiment information as indicated by the dotted arrow.

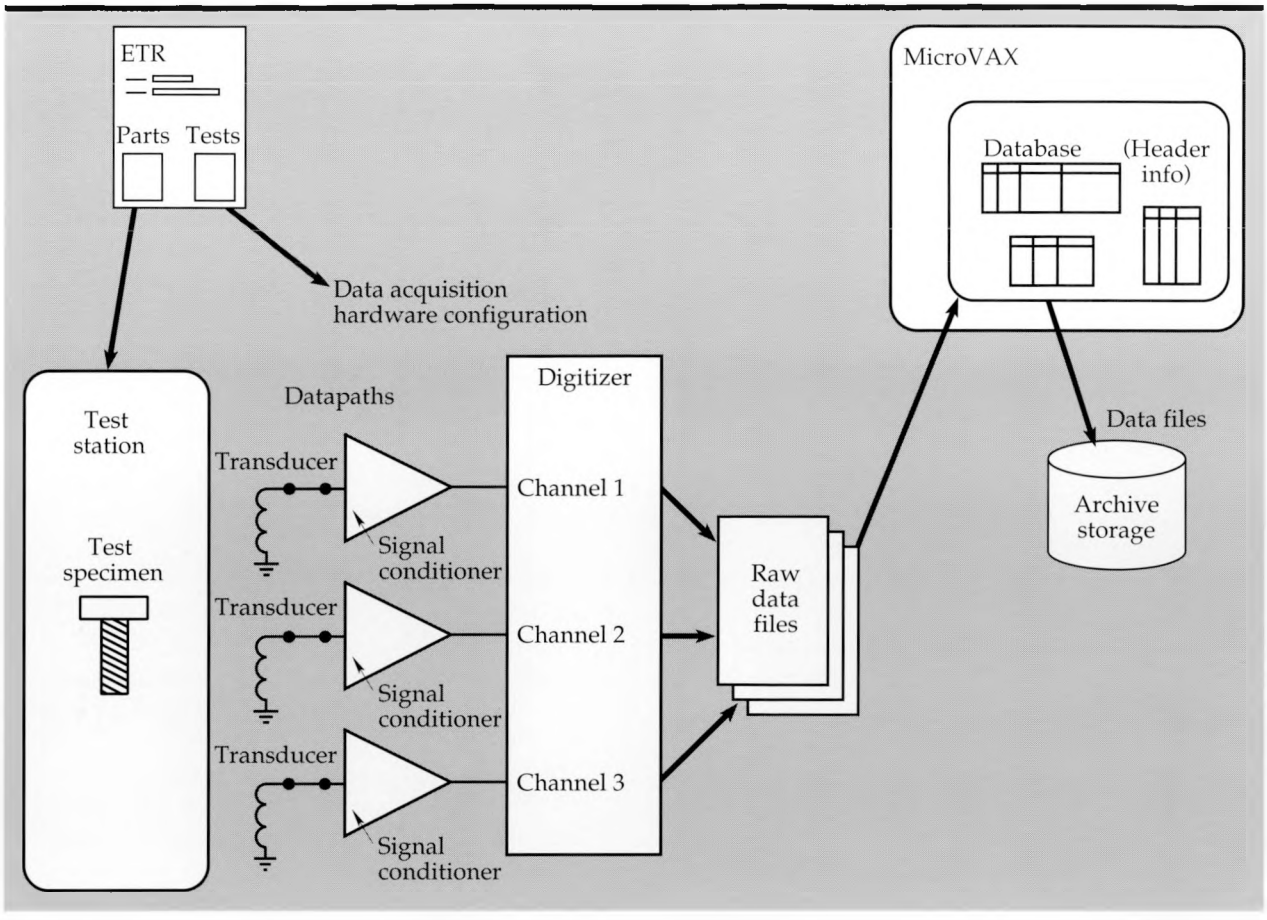


Figure 3. Conceptual model for the materials test and evaluation database (MTE/DB). Engineers begin by using the database to generate an engineering test request (ETR), which lists the specimens to be tested and the tests to be made. A technician then calls up the ETR from the database and configures the hardware that will record the raw data. Next the tests are made and the database is used to transfer the raw data from the test station to the MTE section's MicroVAX, where datafile header information is stored in database tables. The datafile itself is copied to archive storage. Throughout this process the MTE database tracks the specimens, tests, hardware configuration, and datafiles for later searching, sorting, and analysis. Subsequent analysis data can be stored in the database and linked back to the original ETR, test, or specimen.

nel" of raw data through transducers, cables, digitizers, etc. As the figure shows, some items in the datapaths (for example, digitizers) may require parameter settings. These settings are tracked within the database and are linked to all the raw data produced from a given experiment. Data selected from the database may be analyzed and the results returned to the database linked to the original raw data.

In relational database terms, the data model is decomposed into a set of normalized relations (tables) which may be operated upon by relational algebra implemented through the database query language. The two query languages offered with INGRES are QUEL and SQL. These database queries may be embedded in custom programs, on-screen forms, or may be entered directly in an ad-hoc fashion. The base relations are designed in such a way that the data model may be extended in a straightforward manner, without extensive database redesign. This ability allows us to easily customize the system for specific projects, while preserving the generic application.

Operation

As an example of actual use, an experimenter may call up information on a previous shot from the database and use that as a default for a new shot. After making any necessary parameter modifications to the default, the experimenter can automatically format and transfer the setup to the data acquisition system, where the new experiment is performed and raw data collected. This raw data is

then automatically imported into the database where it can be selected for analysis by a variety of analysis codes. All analysis results can be stored in the database, with "threads" linked to the original raw data and experiment information. At any time, the results of database operations may be output as hard copy, using a powerful report generator.

Progress

In FY 89 two Laboratory groups began to use our relational database. The first group uses our system to analyze data accumulated in the Emperor and Anechoic Chamber facilities. The second group tracks raw waveform data through various analysis steps mainly involving the signal processing program SIG.

Future Work

We are currently designing a database **(Fig. 3)** to support Materials Test and Evaluation, a group in Engineering that studies the mechanical properties of materials. This work marks a significant expansion of our database support; specifically, it will be the first time that we will have used ORACLE rather than INGRES as the core database. ORACLE is a structured query language (SQL)-compliant commercial relational database with features similar to INGRES; it is increasingly being used at LLNL and within the DOD community. The two are the main contenders in the commercial relational database market and we now have the ability to support both of them.

This database also has a significant advantage in that users who are not database specialists can use their Mac personal computers as front end devices to access user interface, graphics, and local database processing; meanwhile the VAX back end will continue to perform complex database queries and batch processing.

Summary

The SciDB system provides numerous advantages to experimental programs:

- Speeds up the process of acquiring and analyzing data through the use of defaults and prior setups.
- Makes the overall acquisition/analysis process easier by linking separate parts of the system into an integrated whole.
- Reduces the chance for human error by automating time-consuming manual data format translation.
- Maintains data integrity by tracking and archiving all the information related to an experiment.
- Perhaps most importantly, it provides a means of identifying data interrelationships, which might otherwise go unnoticed, by organizing information clearly and logically in a relational database.

The system is becoming a valuable component of the standard data acquisition/analysis cycle. It will support two of the main contenders in the commercial database market, INGRES and ORACLE, and is being extended to support personal computer front ends.

1. E. Wheelock, "Scientific Data Management," an article in the Thrust Area Report FY 88, Lawrence Livermore National Laboratory, Livermore Calif., UCRL-53868-88 (1988).

An Expert System for Tuning Particle Beam Accelerators

Darrel L. Lager, Hal R. Brand, and William J. Maurer
Engineering Research Division
Electronics Engineering

We have developed an expert system that acts as an intelligent assistant to operators tuning a particle beam accelerator. The system incorporates three approaches to tuning:

- Duplicating within a software program the reasoning and the procedures used by an operator to tune an accelerator. This approach has been used to steer particle beams through the transport section of LLNL's Advanced Test Accelerator and through the injector section of the Experimental Test Accelerator.
- Using a model to simulate the position of a beam in an accelerator. The simulation is based on data taken directly from the accelerator while it is running. This approach will ultimately be used by operators of the Experimental Test Accelerator to first compare actual and simulated beam performance in real time, next to determine which set of parameters is optimum in terms of centering the beam, and finally to feed those parameters to the accelerator. Operators can also use the model to determine if a component has failed.
- Using a mouse to manually select and control the magnets that steer the beam. Operators on the ETA can also use the mouse to call up windows that display the horizontal and vertical positions of the beam as well as its current.

Introduction

Particle beam accelerators are large, complex systems that must be operated by people rather than machines to be effectively controlled. When machines have been used to control such systems in the past, they have frequently failed, usually because the conventional approach of feedback control using a numerical model of the system has failed. This failure occurs because the system is strongly nonlinear, continually changes due to component failures, involves physical phenomena for which satisfactory models have not been derived, and may be so complex that it is simply too expensive to derive the model.

Despite these problems, individual operators are able to run such complex systems effectively. The operators appear to have a set of small, rule-of-thumb (heuristic) models for the various components of the system, and to know how to manipulate those components to achieve the desired control.

In an attempt to duplicate the expertise of operators of large accelerators, we have developed an expert system called MAESTRO (Model and Expert System Tuning Resource for Operators) that models the procedures involved with tuning such accelerators and with fixing them when components fail (a frequent occurrence). MAESTRO is a software program that blends physics models of the system and operator heuristics. We chose the MAESTRO acronym to emphasize the metaphor of a conductor unifying and coordinating the activities of control, diagnostics, physics models, and post-run analysis—all activities that are critical to the success of physics experiments. Traditionally those activities would have been separated, and different people within different groups would have been responsible for them. However, one of the advantages of MAESTRO is that is imposes unity and consistency on the system while it is also helping operators tune the system more efficiently. As an example of the latter, operators can make much better informed control decisions because they can observe the simulation while the machine is running. Also, physicists can quickly gain insight into a particular phenomenon and make appropriate changes to the model because discrepancies between the model and the machine are more readily apparent.

Particle Beam Accelerators

A particle beam accelerator is a device that accelerates electrically charged atomic or subatomic particles, such as electrons, protons, or ions to high en-

ergies. Laboratory researchers use accelerators to study high-energy physics problems that arise in the Weapons, Beam Research, and Magnetic Fusion Programs. We have focused our efforts on two accelerators, the Advanced Test Accelerator (ATA) located at LLNL's Site 300, and the Experimental Test Accelerator (ETA) located at the Laboratory itself.

The ATA is the more complex of the two and consists of an injector, an accelerator, a transport section, an emittance selector, a tuning dump, an achromatic jog (a-jog), and a wiggler. The injector produces a pulse of electrons and injects it into a beam pipe, an evacuated pipe a few inches in diameter running the length of the machine. (A beam pipe with its associated components is called a beam line.) The accelerator section increases the energy of the electrons in the beam pulse up to about 50 MeV. An alignment laser guides the beam through the accelerator section and delivers it to the transport section. The transport section steers and focuses the beam into the emittance selector which "strips off" the electrons with undesired energy. A bending magnet directs the beam either to the tuning dump which absorbs the energy in the beam while the operator is tuning, or to the a-jog which moves the beam onto the wiggler beam line. The wiggler couples energy from the electron beam into another laser beam, greatly increasing the energy in the laser beam. A typical experiment is to deliver an electron beam into the wiggler and investigate its characteristics for producing gain in the laser beam.

The ETA is somewhat simpler than the ATA and consists of an injector, an accelerator, a transport section and a wiggler. One of the differences between the two is that the ETA accelerator increases the energy of the electrons in the beam pulse to only ~ 6 MeV, without using laser guiding. Another difference is that the wiggler couples energy into a microwave beam, increasing the energy of the beam. In a typical experiment the ETA produces high-power microwaves that heat a magnetic fusion plasma. Both accelerators have a variety of diagnostic devices that determine the beam's position within the pipe and its energy distribution. Both systems also have ancillary components (vacuum pumps, safety systems, cooling systems, and pulsed power systems) in their beam lines, further increasing their complexity.

Tuning

The goal when tuning an accelerator is to produce a beam that has the desired temporal and spatial energy and charge distributions required for a given experiment. The optimum distribution or "beam profile" is depicted in **Fig.1**, where the beam pulse traveling down the beam pipe is on center within a few millimeters, all the electrons have uniform axial velocity and zero transverse velocity, and all are uniformly distributed within the pulse. In reality the beam has nonuniform velocities and is subject to a variety of instabilities, causing it to disperse at the nose and tail. (A common instability is a corkscrew, named for the spiral shape assumed by the beam.) Serious instabilities cause the beam to break up or strike the wall of the beam pipe.

Figure 2 shows the devices used for tuning the beam line. The beam bugs shown at the left of the figure are diagnostic devices for determining the beam position and current profile. A beam bug produces three oscilloscope traces as the beam pulse passes through it, one for the current, one for the x-position, and one for the y-position. The operator determines if the beam has the desired axial charge distribution and total charge by observing the current waveform. The position waveforms give the beam position along the horizontal (x) and vertical (y) transverse axes of the pipe. Video cameras record the light emitted when the beam strikes foils inserted into the beamline. The resulting images give information on the transverse distribution of the beam pulse and are used to determine the beam focus.

Steering magnets change the direction of propagation of the beam. It requires a pair of steering magnets a certain distance apart to displace the axis of propagation without changing the direction. The first changes the angle of propagation and the second compensates for the angle introduced by the first. Quadrupole magnets focus the beam and have the usually undesirable side effect of steering the beam when it enters off-center. A general rule of thumb is to keep the beam on-center through the quadrupoles to suppress the unwanted steering.

Complications

The tuning task is complicated by things the operator can readily adapt to, but are difficult for computers to deal with. Since the machine is made up of a large number of components, and since many of them are highly stressed state-of-the-art designs operating at high voltage and high current, there are random failures in the hardware every day. Common occurrences are power supplies failing and the alignment laser drifting out of position. Less often magnets will short- or open-circuit. The operator handles these by periodically observing the status of the machine. If they cannot cor-

or grows more complex, as the operators learn more of its idiosyncrasies, or as more is understood about its physics. In the first approach, called "cloning the operator," we encode as faithfully as possible the procedures and reasoning followed by the operator. This was the approach taken for tuning the ATA because it was not possible within the allotted time to develop an accurate model for the beamline. A disadvantage of using only this approach is that operators may not be able to develop successful procedures for the far more complex machines being designed, such as the superconducting supercollider.

A second approach, model-based tuning, simulates the beam propagating through the sections of the machine. The most notable aspect of this approach is that the model is hooked into the actual system and is running at the same time the machine is running, so the simulation is based on measurements that are being made by the machine at the moment the machine is making them. This approach allows the operator to compare the simulation with the actual beam position, to choose an optimum set of parameters for centering the beam, and to download those parameters onto the real machine at a substantial savings in time. The disadvantage is that it may not be economically feasible to develop a sufficiently accurate simulator, either because the machine may be too complex or because it changes too often due to component failure.

The third approach is to tune the system manually, but provide the operator with more powerful tools for tuning the machine. For example, this might take the form of displays derived from the raw data or different interfaces that make it easier to control the machine. The goal is to achieve a blend of these approaches that minimizes the time required to tune and maximizes the time available for performing physics experiments. Each of these approaches is discussed below.

Actually, no matter which approach we stress, we will always need the manual approach because operators will almost always have more knowledge of the machine than is economically feasible to encode into a computer program. Moreover, they can take into account information that may not be readily accessible to the computer such as the "sound" the machine makes when it is not running quite right. They may also be able to instantly diagnose a failure because they remember what happened when that same situation arose on a machine they were tuning 20 years ago.

Cloning the Operator

The operators of the Lab's accelerators go through a set of procedures when their machines are first powered up. The procedures reflect two different kinds of reasoning by the operator. In the first instance, the operator is concerned with the overall tuning of the machine ("global strategy"). Global strategy is made up of many lower-level "local strategies" that are each concerned with the tuning of a small section of the machine.

When operators are performing a local strategy, they reason about components upstream (opposite the direction of propagation) and downstream from a chosen diagnostics device. For example, as shown in **Fig. 3**, when steering through the transport section, the operator observes the trace data and determines there is a position error at a bug. He reasons about the devices downstream and decides whether it is desirable to correct the error at this bug or to ignore it and examine the next bug. He reasons about devices upstream to decide which can be used to correct the error. He usually chooses a steering magnet and changes its field strength until the position error is zero.

In pseudo-English form this strategy can be described as:

Walk the bugs of the transport section moving downstream from the one nearest the injector —

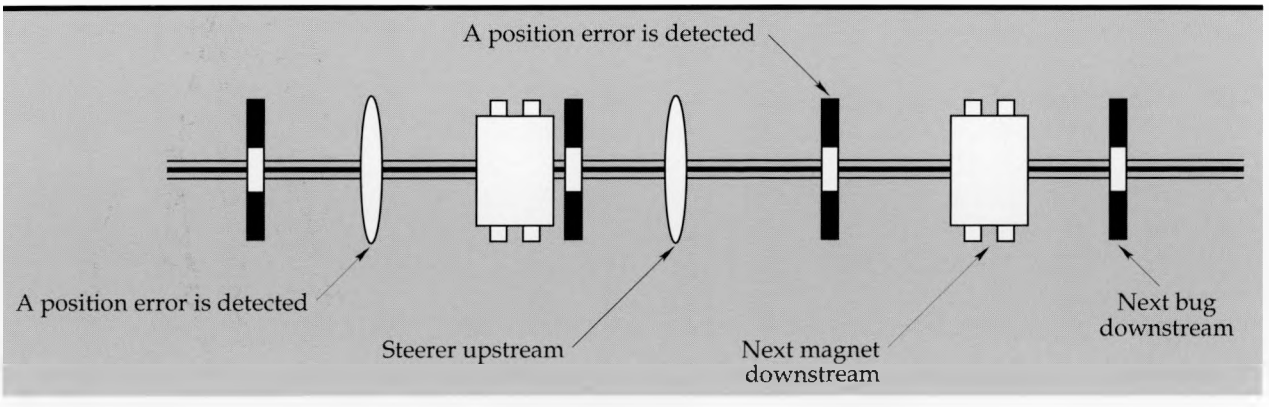


Figure 3. The operator reasoning about devices upstream and downstream from a bug.

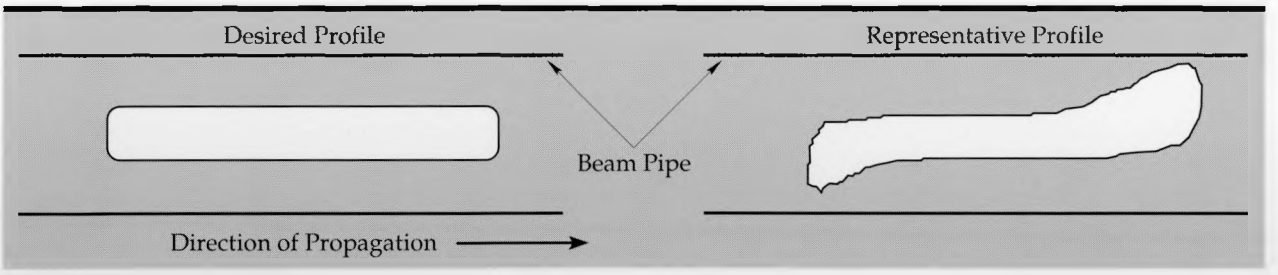


Figure 1. Beam profile while tuning.

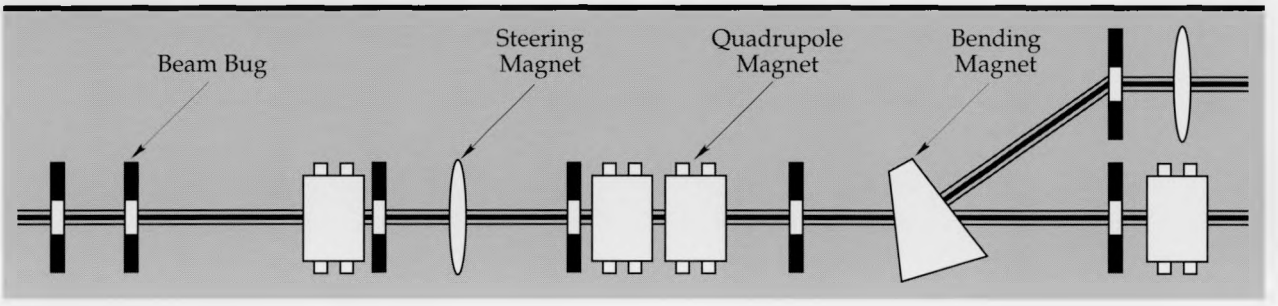


Figure 2. Components affecting tuning.

rect the problem, they modify their tuning strategies and attempts to tune anyway until the problem is fixed.

The operator must also deal with unreliable diagnostics data. There is a high shot-to-shot variation in the characteristics of the trace data. There are also bad shots when the machine misfires that must be ignored entirely. As a result, the operator observes several shots and determines an "average" of the beam characteristics. He will also devote one beam bug to watching the beginning of a section so he can determine if changes in the bug he is observing are caused by his tuning efforts or by changes in the beam characteristics at the input.

Since these accelerators are research tools the hardware is often reconfigured. Generally, the changes involve adding or deleting components, moving existing ones to different places, or changing the relationships among them (e.g., connecting magnets to different power supplies). Another way the configuration can change is that during a run the beam may be directed down different beam pipes to do a desired experiment. The operator is usually told of the changes at the beginning of a shift and readily adapts to the differences. However, the expert system had to be designed with an interface that could tell it about the changes and could verify that the MAESTRO representation reflected the true machine.

Another complication the operator easily adapts to is that the machine is not built exactly as designed. Beam bugs may be installed with their measurement axes rotated away from perfectly vertical or horizontal. The operator simply realizes this and instead of just using the x-axis knob to change beam position they will vary both x and y to do it. They will also move a knob and if the beam steers in the wrong direction, they will merely move the knob the other way. The components are also imperfect so they all need calibration curves. The operator easily deals with these problems because much of the tuning is aimed at achieving a position equal to zero (on center). However a model needs more precise knowledge than an operator of the calibrations, positions, and orientations of the devices if it is to compute a desired change in one step (an operator can determine a change experimentally).

Finally, the machine does not measure the same way from one day to the next, even when all the hardware is unchanged. Various critical aspects of the machine are impossible to measure with sufficient accuracy. For example, the ion gauges that measure the vacuum profile along the beam path are not sufficiently accurate. As a result the benzene profile is not the same as before and the propagation of the beam through the benzene is slightly different. The physics of transport of a beam through benzene are not well understood, so it is extremely difficult to compensate for the measurement system by modeling. The operator has evolved tuning strategies that are relatively insensitive to these phenomena.

Expert System Techniques

MAESTRO blends three distinct tuning approaches to achieve better performance than any single one alone. A trade-off among these approaches can be made as the machine is modified

has been more of an opportunity to get onto the system and model the various components. Consequently, this fiscal year we are much closer to a tuning model of an accelerator.

Making use of such a model involves two distinct phases. The first is the "commissioning" phase[3] during which the simulator is matched with the real accelerator by measuring what effect each component has on the beam. This phase forces proper bookkeeping because the effect of each component on the beam trajectory must be calculated accurately for the model to work. Any rotations, tilts, offsets, and miscalibrations must be eliminated or incorporated into the model during this phase so the simulator and real machine produce the same beam trajectory.

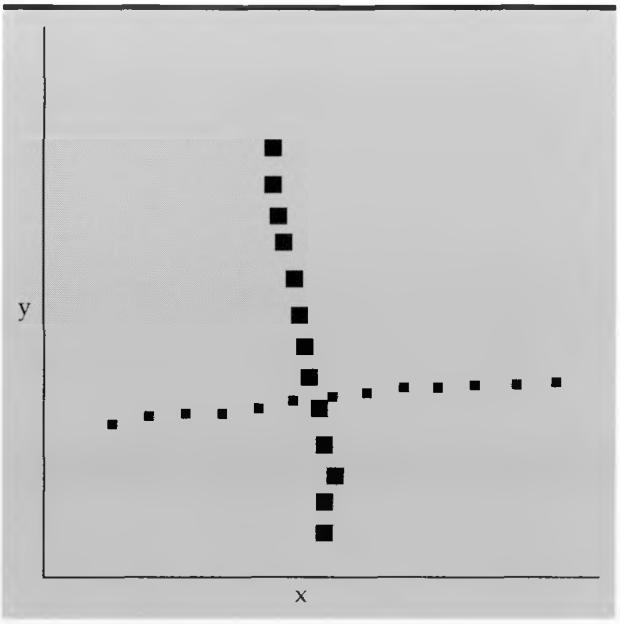


Figure 4. Variation of x and y beam positions as the vertical and horizontal steering magnets are swept through a range of values.

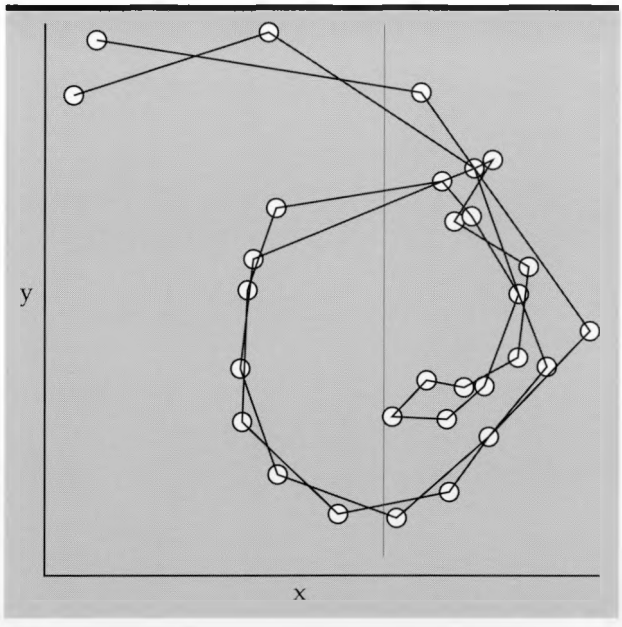


Figure 5. Variation of x and y beam positions as the solenoidal field is varied through a range of values. The solenoidal field causes the beam to spiral as it passes through the magnet. As a result, varying the field causes the locus of points representing the beam position to assume a spiral shape. The spiral becomes tighter as the field increases.

As part of the commissioning phase we have developed procedures to measure various beam parameters. For example, we measure the effects that steering magnets have on beam position by sweeping horizontal and vertical steering magnets through a range of values, producing the cross-shaped pattern in **Fig. 4**. Similarly, we measure the effects that solenoid magnets have on focus by sweeping through a range of settings, producing the spiral-shaped pattern shown in **Fig. 5**. We have started the commissioning phase on the injector section of the ETA accelerator and have chosen the smallest subset possible: one magnet followed immediately by a beam-bug. Once the effects of these components have been measured, we will commission the remaining injector magnets and then proceed down the accelerator beginning with the first ten-cell set. Once we are finished with the accelerator, we will tackle the remaining beam lines.

As discussed above, during the commissioning phase we bring the simulation model and the accelerator into relative agreement. In the second, operational phase we can actually begin to tune the accelerator and diagnose failure. To tune the accelerator, the first step is to turn the system on and measure actual beam position in the accelerator. Given these measurements, the initial beam energy, and the commissioned model we can estimate the launch conditions of the beam—its initial position and transverse velocities (denoted as x, x′, y, y′). These conditions are estimated by fitting the measured data to simulated data and then varying the launch conditions to get the best fit. Finally, we can use a nonlinear parameter estimator (part of the simulation model) to select tuning parameters that will center the beam. Obviously, a very large number of tuning parameters are available. Initially, an experienced operator will select a subset of these for tuning, and iterate the process with different subsets. Eventually, the operator's knowledge and experience will be encoded into an expert system that will select the subset of tuning parameters and will decide how much iteration is necessary.

When the simulation model and the accelerator

if the position error at a bug is too large and the error can't be ignored
then reduce the error to zero by tweaking the steering magnets upstream
if there are none, align the laser
if there are two with no quadrupole between them, use them as a pair
if the first upstream steerer is too close, use the second one upstream
otherwise use the first upstream steering magnet.

Note that this strategy is relatively unaffected by component failures, since a failed component is simply deleted from the beamline and is no longer considered as an upstream or downstream component.

The global strategy is concerned with properly applying local strategies, based on the present state of tune of the machine. The guiding philosophy is "focus and steer the beam but don't put the beam into the wall." Given that philosophy, the operators first check each major section of the accelerator to see if the beam has arrived from the previous section. If it has not, the operators go back to the previous section and use a coarse steering technique to get the beam through the section. Once the beam has made it all the way to the tuning dump (that portion at the end of the accelerator that can absorb the full energy of the beam without disrupting the system or damaging components), the operator goes back to the begining of the transport section and meticulously centers the beam while monitoring the beam current at the tuning dump.

We have encoded the local strategies followed by the operator in a representation we call a Monitored Decision Script (MDS). An MDS is an extension of the notion of scripts introduced by Schank.[1] Scripts are used to represent step-by-step procedural knowledge. The canonical example is the script for eating at a restaurant — first you are seated, then you order, then you eat, then you pay, and then you leave. These things have to be performed in order, but exactly what is done at each step is decided when you visit a particular restaurant. We use the Script part of an MDS to represent the procedures involved in performing the local strategies. The Decision part of the MDS name comes from the decisions that must be made when errors are detected (for example, when a magnet doesn't respond to a request to change its field because the supply connected to it has failed). The Monitor portion of the name comes from monitors that are periodically examined to check the health of the system. This emulates the behavior of the operator, when he periodically checks to see if, for example, the laser alignment has drifted off position.

To implement the global strategy, we have written a number of separate MDSs that are grouped according to what state the accelerator is in; that is, whether the accelerator is in the start-up, coarse, or fine-tuning states. One of these states is listed in a so-called pre-condition field of each MDS. Each MDS also contains a post-condition field, within which is listed the result of successfully executing the MDS. For example, if a particular MDS has a pre-condition field that lists initial start-up, then the post-condition field would state that the beam has reached the tuning dump. We have also developed an AI program called an inference engine to manipulate the MDSs and perform the global strategy. The engine first checks to see which state the accelerator is in; next, it matches that state with one of the MDSs that lists the state in its pre-condition field; next, it selects one of these MDSs and executes it. Executing the MDS puts the machine in a new state (not necessarily the one listed in the post-condition), then the cycle of match, select, and execute is repeated until the accelerator is tuned.

We applied the MDS and inference engine programs to the problem of centering the beam in the transport section of the ATA accelerator. Our initial approach was to decompose the steering problem into two simpler, decoupled problems; namely, centering the beam in x, then y. We thought this was possible because our system had "knowledge" of the beam bug rotation angles and other information that would enable decoupling the x and y steering. Unfortunately this approach was unsuccessful because there was coupling between the vertical and horizontal steerers that we were unable to represent given our time constraints. Instead we modified our approach to more faithfully incorporate the strategy used by the operator. This was performed in two phases. In the first phase we diagnosed a pair of steerers and determined which more strongly influenced the x position and which the y. In the second phase we centered the beam using the steerers by repeatedly halving the error in x, then y, using an optimization algorithm.[2] This approach successfully automated the centering of the beam.

Model-Based Tuning

For the ATA discussed above, we chose to clone the operator as the best way to automate the system, given the complexity of the accelerator and the limited amount of machine time available to us. For the ETA we chose to model the system because the ETA is less complex than the ATA and there

go out of relative agreement (due to component failure), the model can be used to find those regions within the accelerator where there is still agreement. For a single component failure, there will be two regions of agreement roughly surrounding the failed component. From this information, a list of suspect components can be compiled. With the aid of an expert operator (or eventually, an expert system) a number of possible failures can be proposed. The nonlinear parameter estimator is then used to estimate both the magnitude of the proposed error and the improvement in the model's predictive ability, given the proposed error and the data currently available from the accelerator sensors. With this information, many proposed errors can be rejected because either the magnitude of the error is beyond reasonable limits, or the improvement in the model by the addition of the error is too small.

In the ideal case, one is left with only one reasonable error. In the other cases, one is normally left with only a few possibilities. Beam redirection experiments can then be performed to further isolate the accelerator-model discrepancy. Selection of these experiments will initially be done by experienced operators, but will eventually be included within the expert system's capabilities. When the problem is finally found, then the accelerator can be fixed and/or the model can be updated with the (fit) error information, and tuning can proceed as before.

In conclusion, model-based tuning promises to put more science into the art of tuning and lead to a more rigorous understanding of the machines. The model enables operators to "see" the effects of their tuning in the regions between (not covered by) sensors. Given sufficiently accurate models, it should be possible to determine a set of parameters that will change the present state of the machine to a desired (tuned) state in a single step. Experiments can be performed off-line using the model to determine, for example, the effects of adding new components to the beamline at far less expense than would be the case if the real machine were used.

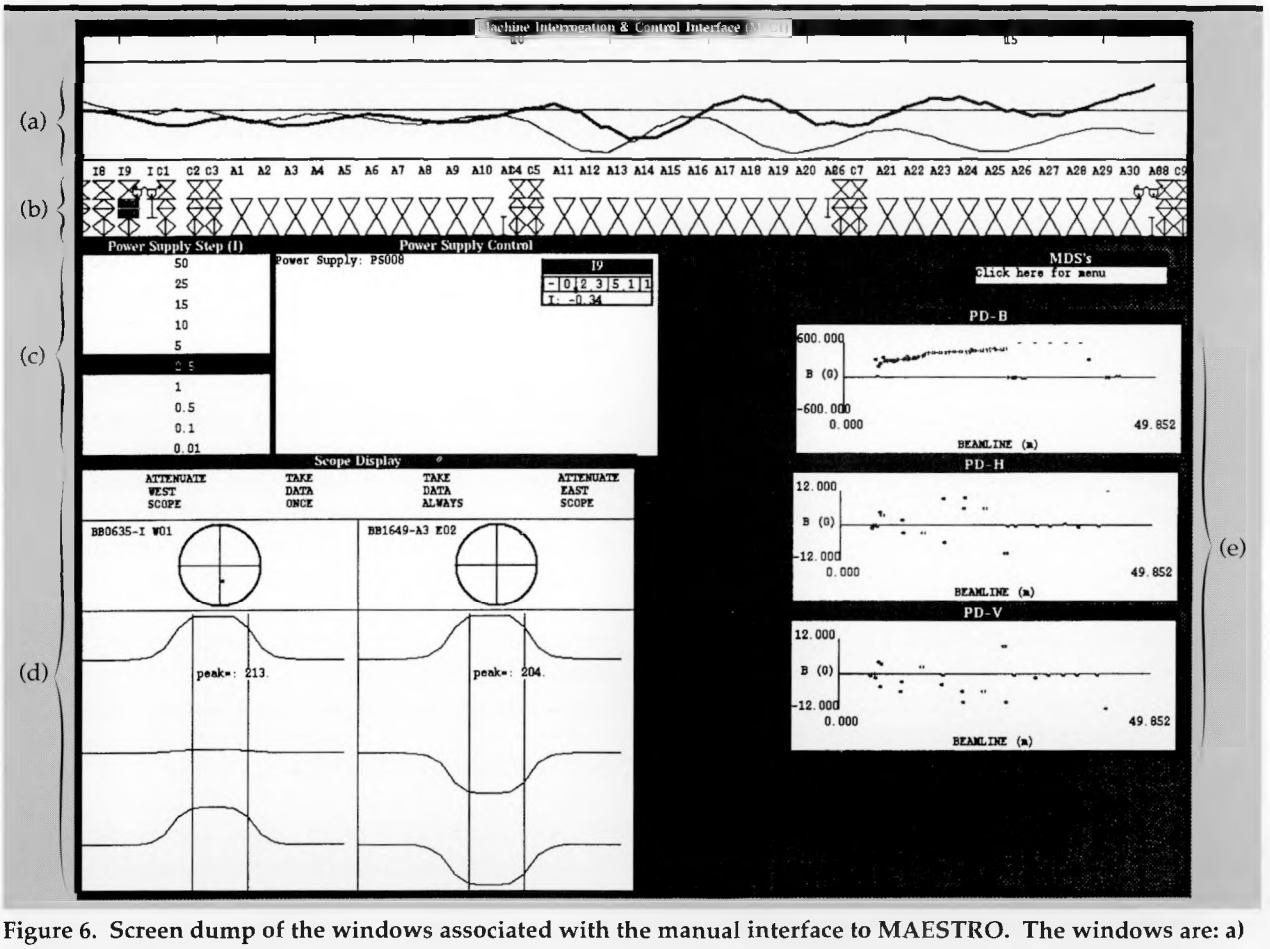


Figure 6. Screen dump of the windows associated with the manual interface to MAESTRO. The windows are: a) simulated beam position vs distance down beamline. The x position is the thin line and y is the thick line; b) icons depicting the components according to their positions in the beamline; c) power supply control window; d) scope display showing the I, x, and y waveforms from the selected beam bugs; e) field strength vs distance down the beam line for the solenoids, horizontal steerers, and vertical steerers.

Manual Interface

The manual interface for MAESTRO is shown in the screen dump in **Fig. 6**. The interface consists of several windows that become visible on the screen as necessary. The Machine Interrogation and Control (MICI) window is the main one for interacting with the system. It consists of two panes, one above the other. The top pane displays the output from the simulator, showing the horizontal beam position as a thick line and the vertical position as a thin line. The lower pane shows a set of icons depicting the components in the beamline. The locations of the icons and their shape are derived from the information describing the beamline in the MAESTRO knowledge base. Components added to the beamline are automatically included in the MICI display once the information has been added to the knowledge base.

The operator controls the magnetic fields by positioning a cursor over an icon with a mouse. Clicking over a vertical steering magnet icon, for example, causes power supply control windows to become visible below the MICI window (**Fig. 6**). By clicking the mouse as the cursor is positioned in the windows we can increase or decrease the current in the appropriate power supply by a given increment.

The Scope Display window is used to control and display data acquisition. Clicking the mouse over the appropriate label in the window causes data to be acquired and displayed as a set of three traces in the window: the time history of the beam current, x position, and y position. The actual location of the beam within the pipe is derived from the trace data after accounting for sensor misalignment and is displayed in the circular "bulls-eye" displays in the upper part of the window.

There are other windows that display additional derived data. A "position display" window shows a value vs position down the beam line. A "bug-walk," for example, makes position displays that show peak current, x position, and y position at the locations of the beam-bug position monitors along the beamline. Similarly a "magnet-walk" shows the fields at the centers of the magnets vs their location in the beamline. Three magnet walks appear in **Fig. 6** showing the solenoidal fields (labeled PD-B), the horizontal steering fields (labeled PD-H), and the vertical steering fields (labeled PD-V) vs distance down the beamline.

There are also windows for displaying historical data. Clicking the mouse over a beam bug icon causes a "shot history" window to appear. By clicking the mouse over buttons on the window the operator can view all the oscilloscope traces acquired by that bug during that day's run.

Future

We are planning to improve the shot history mechanism so that it can not only manipulate past shot data but also past machine configuration data. For example, we want to have the ability to re-do the signal processing with different control parameters. We also want to be able to ask such questions as, "During the last three months what was the highest current magnitude measured when the machine had the long collimator installed?" Our approach is to develop an unstructured database based on the artificial intelligence representation scheme known as a semantic network.

We are also developing the ability to acquire image data from cameras observing the beam striking foils inserted into the beamline. We will apply image understanding techniques (the ability of an AI program to interpret or judge a visual image) to determine the position and focus of the beam.

Summary

The MAESTRO software environment was developed to function as an intelligent assistant to an operator who must tune complex systems such as particle-beam accelerators. It incorporates three approaches to tuning. The "cloning the operator" approach uses an inference engine and the MDS representation to encode the strategies and reasoning followed by the operator. The model-based approach makes use of a beamline simulator and a nonlinear least squares parameter estimator to first "commission" the model and then determine optimum tuning parameters. The third approach lets the operator perform tuning manually and provides him with displays that easily let him determine the machine status. Finally, a history mechanism lets the operator view past data to compare the present tune with ones previously obtained.

Acknowledgments

The authors wish to gratefully acknowledge the contributions of Doyle Rogers and John Clark, the tuning experts. Their ability to make order out of chaos is truly amazing.

1. R. C. Schank and R. P. Abelson, *Scripts, Plans, Goals, and Understanding*, Lawrence Erlbaum, Hillsdale, NJ: (1977).

2. P. G. King, and S. N. Demming, "UNIPLEX: Single Factor Optimization of Response in the Presence of Error," *Analytical Chemistry*, Vol. 46, No. 11, pp 1476–1481 (September 1974).

3. M. Lee, and S. Clearwater, "GOLD: Integration of Model-Based Control Systems With Artificial Intelligence and Workstations," Workshop on Model-Based Accelerator Controls, Brookhaven National Laboratory, Upton, New York (August 1987).

Interactive Volumetric Visualization

Brian K. Cabral and Chuck W. Grant
Engineering Research Division
Electronics Engineering

During FY 89, we developed two volume-visualization computer programs, VRENDER and CELL-TRACER, for displaying three-dimensional (3-D) scalar field data. These tools enable researchers throughout the Laboratory to visualize both experimental and simulated 3-D data. While the two tools use completely different technologies, they both render transparent views of 3-D scalar fields. These transparent volumetric views are critical to understanding the internal structures of 3-D data sets.

Introduction

Scientists at LLNL are beginning to measure and model 3-D phenomena. Traditional 2-D surface plots or cutaway views of this type of data provide a very limited insight into the actual 3-D structure of the data. Using the traditional 2-D techniques, scientists are forced to build 3-D visualizations in their heads, a task which may be all but impossible due to the complexity or unfamiliarity of the data.

The Advanced Visualization Research Project (AVRP) has developed two true 3-D or volumetric visualization computer programs. These algorithms allow the user to see into 3-D volumes by controlling the graphical rendering processes, including both the orientation of the data to be displayed and the color and transparency of various parts of the scalar field. In this way, the user can peer into various parts of the data. Once the user finds a particular rendering that best visualizes the data, an animation can be readily computed and recorded.

VRENDER and CELL-TRACER are both written in the C programming language, and run on a large variety of machines. Both programs accept very general types of 3-D scalar field data. The machine-independence and generality of the programs make them readily applicable to a large number of scientific uses.

Progress

In the past, all of the volumetric rendering done at LLNL was performed using expensive, one-of-a-kind systems. This made it highly impractical for programmatic efforts to systematically exploit this technology. In FY 89, we have taken the first steps to alleviate this problem. VRENDER and CELL-TRACER, our two new machine-independent volume rendering tools, go beyond previous capabilities not only in terms of portability, but also in terms of performance and visualization power.

Why did we develop two different volume rendering programs? The two programs, while producing similar results, use completely different rendering algorithms, which embody different compromises between performance and generality. As we will see, VRENDER renders volumes much faster than CELL-TRACER and any other known proprietary volume renderer. CELL-TRACER, on the other hand, is more general and makes it easier to experiment with new rendering techniques and different data types. The two programs are complementary weapons in our arsenal of data visualization techniques.

VRENDER

VRENDER uses volumetric compositing techniques to render semitransparent images of 3-D data on orthogonal grids. The volumetric data is read in and traversed in a back-to-front order as determined by the volume's orientation with respect to the viewer. The orientation is user-specified, as are the color and opacity values to be used in the rendering process. As the data is traversed, it is colored and shaded; this process is controlled by the user-defined parameters. The resulting colored data values are then composited into the image, and algorithms then compute the next nearest voxels or data elements. This process is done entirely with integer arithmetic, resulting in a very fast algorithm. **Figure 1** shows an example of this process, a volumetric image of red blood cells.

VRENDER uses a number of powerful volume manipulation routines, which when combined as described above result in an extremely portable

and fast volume rendering tool. The same technique can be used as a preprocessing step to put shadows into the volume. The power of the techniques used in VRENDER open a vast array of possibilities that have yet to be explored for dealing with volume data.

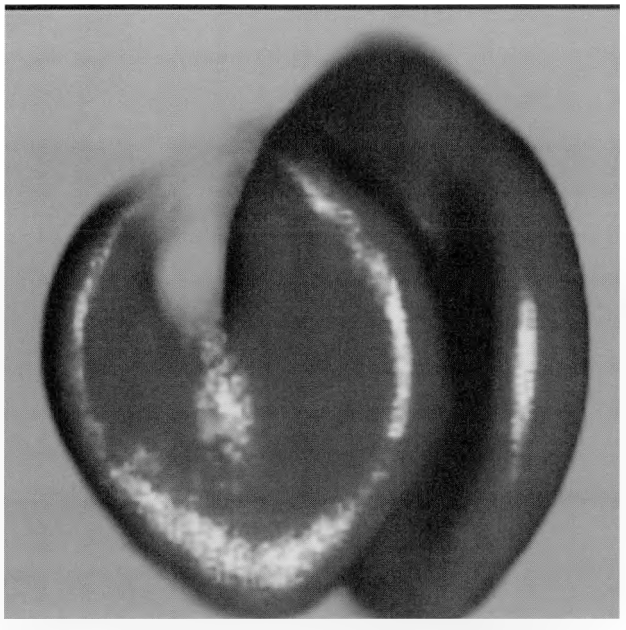


Figure 1. Volumetric image of red blood cells processed by VRENDER. This algorithm rapidly renders color and opacity values and even specular highlights. Data gathered by F. Waldman; rendering performed by C. Grant and M. Allison.

CELL-TRACER

CELL-TRACER uses ray tracing techniques to create semitransparent images of 3-D data on an orthogonal grid. The name is derived from the algorithmic technique used to render the data. CELL-TRACER shoots a ray into the bounding box of the volume of the data. It then traces from cell to cell (a cell is made up of eight vertices) along the path of the ray (see **Fig. 2**). As it traces the ray, it gathers color and transparency information as a function of the data at the various points along the ray. These color and transparency values are summed until the ray either escapes the volume or until the resulting transparency value is opaque. This integration process is very similar to x-ray radiography. The difference is that CELL-TRACER permits color and transparency to be altered as a function of the raw data. This allows the scientist to highlight areas of interest and then color them in a way that is particularly meaningful to the problem being studied. **Figure 3** shows a volume rendering of a computational fluid dynamics problem.

Since CELL-TRACER does all of its work in floating point arithmetic, it is slower than VRENDER. It does, however, provide extra flexibility. For instance, perspective comes "for free" by the very nature of the ray tracing process. Extra accuracy is achieved by using floating point arithmetic, and interpolation can be done on the fly if more visual resolution is needed. Lastly, the ray tracing

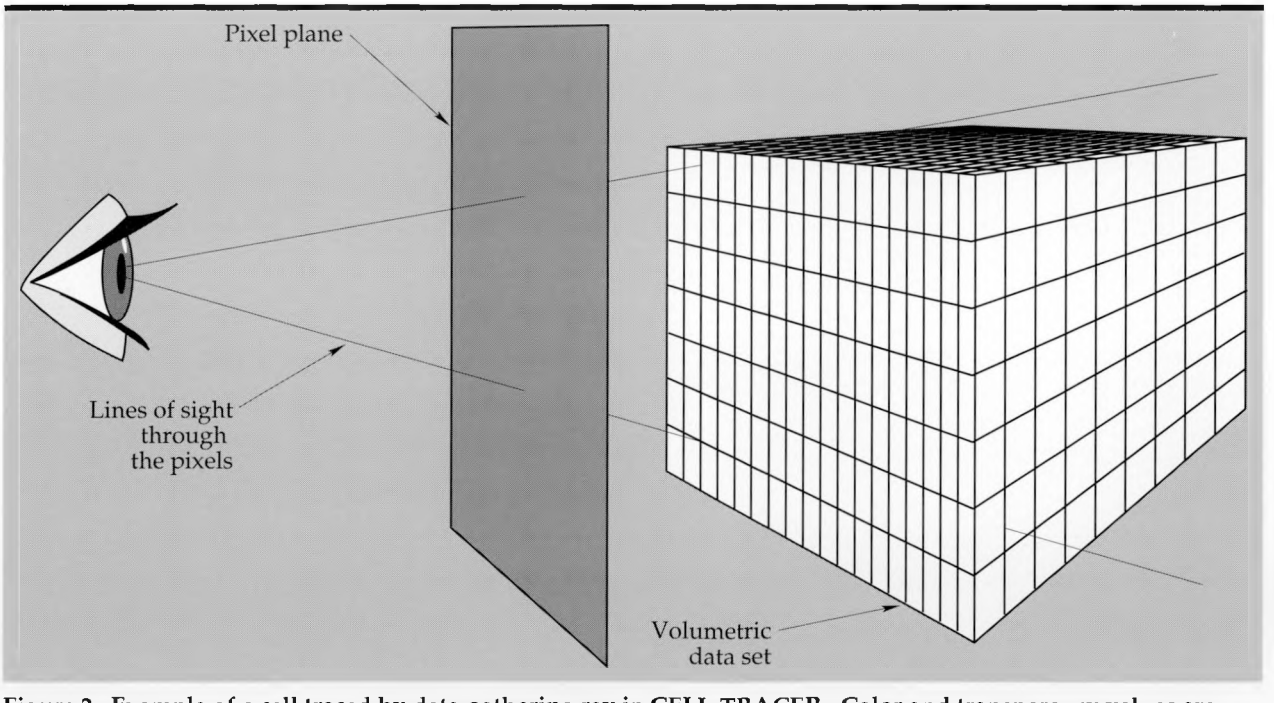


Figure 2. Example of a cell traced by data-gathering ray in CELL-TRACER. Color and transparency values are summed over all cells traversed by a ray to assemble information to be integrated into an image.

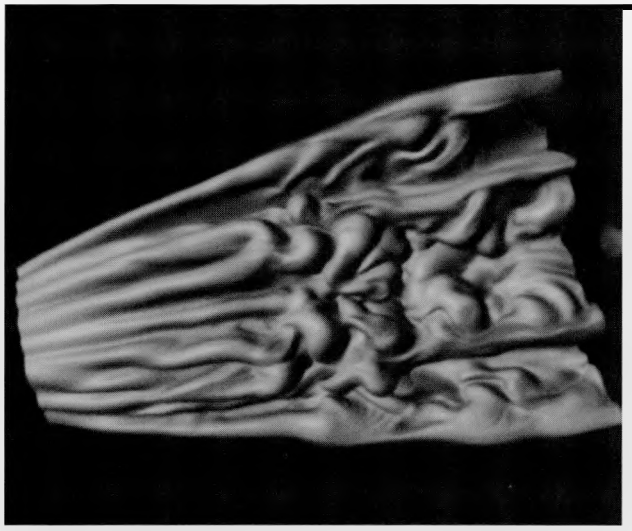


Figure 3. A CELL-TRACER volume rendering of a computational fluid dynamics problem. Information from a very large array of data points is assembled into an easily-understood image, describing the behavior of a complex physical problem. Data courtesy of John Bell.

technique lends itself more readily than VRENDER to algorithmic experimentation. For instance, it is easier to extend ray tracing to operate on other data types, such as solid geometry. Such extensions are more difficult with the VRENDER algorithm.

Future Work

There are two directions for our future work in the area of volume visualization. First, we plan to make the tools easier for investigators to use without expert intervention. Second, we intend to extend the techniques to broader classes of data and to make the tools run faster.

At this time, running VRENDER and CELL-TRACER requires a fair amount of expertise. To make these tools more readily accessible to the average user, we need to turn them into a finished product. During FY 90 we expect to begin the final development of VRENDER and CELL-TRACER. This will include integrating the two tools through use of a common user and data interface, and preparing a user's manual for both products. The net result will be a new and powerful set of visualization tools not now available in the public sector.

We expect during FY 90 to extend the volumetric rendering techniques to vector fields and potentially nonrectangular grids. We also expect to create both a fine- and coarse-grain parallelization of one or both of these codes to increase the throughput for lengthy, time-consuming animation sequences. The net effect of these efforts will be improved visualization capabilities for many of the LLNL scientists studying 3-D phenomena.

Scientific Visualization Development

Stephen G. Azevedo and Charles W. Grant
Engineering Research Division
Electronics Engineering

James M. Brase, Jerome E. Krammen, and Melvin G. Wieting
Laser Engineering Division
Electronics Engineering

We have developed interactive display algorithms for computer architectures that allow us to display multidimensional data more effectively. We have also made it easier for programmatic users to produce their own visualizations.

Introduction

The Scientific Visualization Project processes and interprets multidimensional data generated by an extremely broad range of phenomena – from the massive amounts of seismic data generated by earthquakes or weapons tests, to the electromagnetic waves generated in magnetic fusion experiments, to the signals generated by an ultrasonic pulse traveling inside a part. Because the data contained in such signals are not always easy to interpret, one of our goals is to improve our ability to extract information from the signals. To this end, we have developed algorithms for the display and visualization of experimental data. The information gained by interpreting the data has helped Laboratory researchers analyze, for example, the reliability and vulnerability of a part or system, analyze seismic data to separate nuclear weapons detonations from earthquakes, and align laser beams.

During the course of this project we have developed a number of techniques for manipulating multidimensional data:

- Converting formats
- Filtering
- Enhancing the contrast
- Reformatting (rotation, scaling, translation)
- Correcting distortion
- Producing final hardcopy (film, viewgraph, or videotape).

In FY 89 our goal was to make these techniques more accessible to programmatic users by having simplified versions of the algorithms ported to VAX, Macintosh, TAAC-1, and other common computers.

Progress

We are currently improving our methods of visualizing various types of volumetric data; i.e., data with values defined on a three-dimensional (or higher), regularly spaced, densely sampled grid. The data value at each grid point can be a scalar or vector quantity. To manipulate the data, we have used VIEW,[1] a signal processing code, in conjunction with the following systems:

- Pixar imaging computer (PIC)
- TAAC graphics accelerator on a Sun workstation
- UNIX/X-windows workstations (e.g., Sun, DEC, Stellar, Silicon Graphics).

One of the difficulties in visualizing volumetric data on a two-dimensional display is the reduction of dimensions. Some type of projection of the data must be used. Once projected, near data values may overlay and obscure farther data values. The data values of interest may be well within the interior of the volume. We have developed techniques to make important features of the data stand out without being obscured by the massive quantities of uninteresting data points. Another basic difficulty of volumetric data is that the quantity of data which defines the volume tends to be very large. Volumes with tens of millions of grid points are typical. This large amount of raw data makes interactive processing difficult.

The techniques for effectively visualizing volumetric data can be divided into two classes: display and rendering. "Rendering" is the process of taking the data and generating a picture. "Display" is the process of presenting the rendered pictures to the user (usually interactively and rapidly).

Display techniques include:

- Generating single images,
- Interactively "slicing" volumes at any oblique angle
- Producing precomputed "film loops" of rotating volumes.

We have rendered three-dimensional data sets in the following ways:

- Opaque volumes
- Transparent volumes

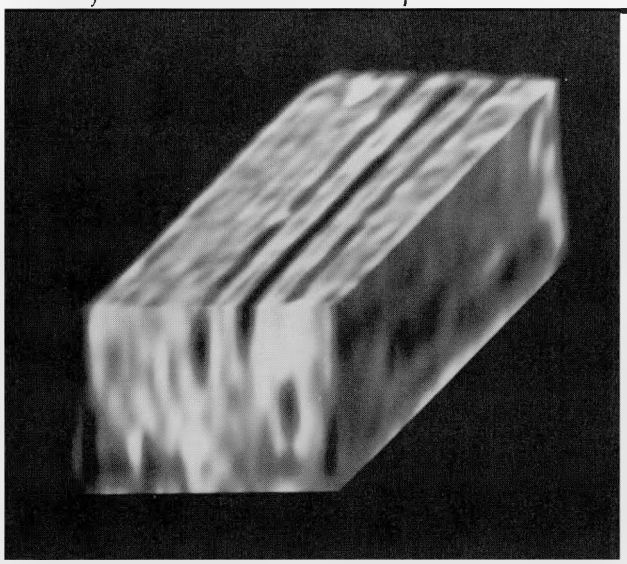


Figure 1. *Cubetool* display of microholes in a rectangular solid. Three faces of the solid are visible, simulating an isometric projection.

- Semi-transparent volumes
- Opaque objects in semi-transparent volumes with surface shading.

Useful combinations of rendering and display techniques include:

- Interactive "slicing" of opaque volumes
- Single images and film loops of three-dimensional renderings of transparent volumes
- Single images and film loops of three-dimensional renderings of semi-transparent volumes
- Single images and film loops of opaque objects in semi-transparent volumes with shading
- Interactive slicing of precomputed three-dimensional renderings of transparent and semi-transparent volumes.

In the following subsections we will describe these techniques further and discuss our progress toward making these tools available to programmatic users.

Display

Slicing is the process of displaying the data which lies along a particular plane in the volume. The particular plane is selected by the user. Two slicing tools are available—*"cubetool"* and *"slicetool."* *Cubetool* presents views of three slicing planes—one plane parallel to each of the *x-y*, *y-z*, and *x-z* planes. Which of each of these planes is displayed is interactively controlled by the user and his mouse. These three slices are then assembled into a three-dimensional view of the subvolume surface of the original data. **Figure 1** shows an example of using *cubetool* on a 3D proton energy loss tomogram. Seeing the spatial orientation of the slices together in this manner makes the data much easier to understand than viewing the slices independently. The difficulty with interactive slicing on conventional machines is the large amount of data involved. The PIC, however, can hold a sixteen million grid point volume (256 × 256 × 256) in its main memory, making this kind of interactive manipulation possible. A limitation of this type of display is that only the data along the slices is visible; i.e., data "behind" the slice is completely obscured by the data in the slice.

The other interactive slicing tool, called *slicetool*, operates in much the same way as *cubetool*, but allows slicing along any arbitrary oblique-angle plane. This tool removes the *cubetool* requirement that only orthogonal slices are visible. *Slicetool* is designed to accept user commands from the mouse to define the slice plane and to graphically display where the slice is being taken relative to the data volume. **Figure 2** is an example of the *slicetool* display of a three-dimensional x-ray tomogram of a sphere with internal flaws. On the PIC display, two views of the data volume are represented. The top view shows a wire frame representation in which a cube representing the volume is stationary and a plane (representing the slice plane) moves about the cube as the user moves the mouse. The bottom view shows the tri-linearly interpolated

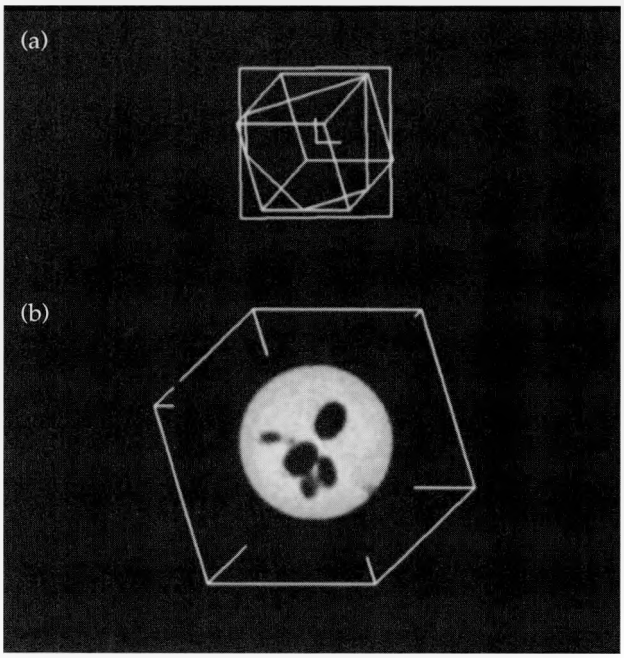


Figure 2. Display of a three-dimensional tomogram of a Delrin sphere with internal voids. (a)Wire frame representation in which a cube representing the volume is stationary and a plane moves about the cube. (b) Interpolated view of the slice plane image.

view of the slice plane image. In this view, the plane of the slice is stationary and a wire frame representation of the cube moves about the view of the slice.

Since the PIC is not fast enough to render full resolution images of the slice in the required time intervals, a simple but powerful technique known as progressive rendering was used to achieved the desired speed. With this technique, low resolution images of the slice, adequate for the user to assess the data, are computed rapidly enough to be displayed at required (i.e., interactive) rates. Higher resolution images will automatically follow if the user does not attempt to change the view. This powerful technique gives *slicetool* the ability to deal with extremely large data volumes. *Slicetool* has additional features that allow the user to remember earlier mouse motions for playback at a later time, adjust the resolution range used by the progressive renderer, and adjust the display for recording onto video tape.

Most recently, *slicetool* has been modified for greater user flexibility by adding three features: (1) pivot mode, (2) double buffering, and (3) improved color handling. With the addition of pivot mode, the user may select an important data point on the current slice plane image about which subsequent slice planes must pivot. This has been very useful for certain types of data sets within which important data points are lodged. Double buffering of the images improves the speed of display and creates a smoother motion over time from one slice to the next. The enormous color mapping capabilities of the PIC are now being more effectively used in our codes. In addition to these features, many minor changes have been made to the *slicetool* user interface to make the overall code run faster and friendlier.

Another display technique for understanding the data is the use of artificial film loops. Film loops are sequences of precomputed images. The images are all loaded into the main memory of the PIC. Each image is displayed briefly on the PIC's monitor before switching to the next image. The playback of film loops can be interactively sped up, slowed down, stopped, run backwards, and stepped from frame to frame. A limitation of this type of display is that only precomputed images can be displayed. Thus, during film loop display, the user cannot view between frames or modify the rendering process. **Figure 3** shows nine frames out of a film loop of six micro-pipettes reconstructed from computed tomography data.

The advantage of these three display techniques is that users can manipulate the display and immediately view the results, thereby gaining a better and quicker grasp of the three-dimensional aspect of the data. Not only is this a better process than viewing static images, but it is also better than viewing a sequence of images where the users have

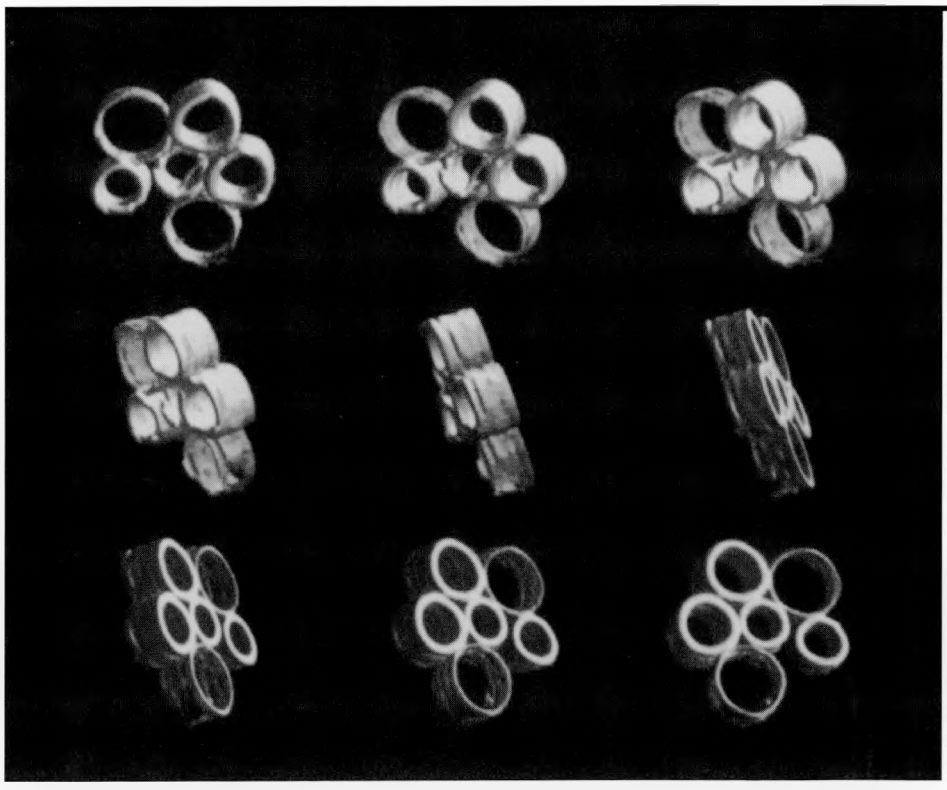


Figure 3. Film loop of six micro-pipettes reconstructed from computed tomography data.

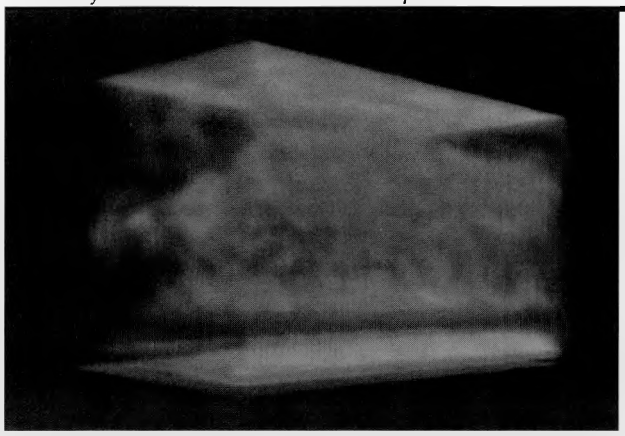


Figure 4. Example of the most complex form of rendering, in which surface shading is added to semi-transparent volumes.

no control over the sequencing. Again, the key to this improvement is the interactive control of the display process.

Rendering

The word "render" is a term from architecture which means: "to draw a realistic picture." The simplest rendering technique used with volumetric data is to draw only the data on the surface of the volume, or along a particular slice, as in the display techniques just described. This technique is very fast, so interactivity is possible, but it is also very limited. This is because any relationships between the displayed data and any other data in the volume (e.g., shadows, obscuring objects, etc.) will not be visible. The user must cut deeper into the volume if he wishes to see such data.

The next most complex form of rendering is to consider the volume to be completely transparent. As the volume is projected to two dimensions, the values of the three-dimensional voxels (volume elements resembling small cubes) are integrated over distance to give the values of the two-dimensional pixels. This type of rendering is similar to a conventional negative radiogram (i.e., medical x ray). Sections of the volume that have a higher density are brighter in the image, while sections with lower density are darker. This type of rendering is more useful than the simpler techniques, since more of the internal volume information is presented in the two-dimensional image, but it is also limited because no depth information is preserved in the final image. Combining this rendering technique with interactive display techniques can effectively provide a snapshot of three-dimensional data.

The third most complex rendering technique is to consider the volume to be semi-transparent. Here each voxel consists not only of a degree of brightness or a color which that voxel contributes to the picture, but also an opacity which specifies how much this voxel covers the voxels behind it. A completely transparent voxel would have an opacity of zero while a totally opaque voxel would have an opacity of one. This rendering technique preserves some of the depth-ordering information of the original volume of data, giving a more informative image. The image is still not very realistic, since the model for visualization is basically semi-transparent luminous clouds.

The most complex form of rendering we have used adds surface shading to the semi-transparent volumes. The transition between a transparent region and an opaque region can be detected and the position and orientation of a surface representing the transition can be calculated. This surface can then be "shaded" as if illuminated by a light source from a particular direction. The surface itself can then be rendered as transparent, semi-transparent, or opaque. The resulting picture is closer to our real world experiences and thus easier to understand. **Figure 4** shows an example of this technique in which R Program material is rendered from CT data.

In this form of rendering we have full volumetric information (every voxel is accounted for); therefore, a part of the data can be "cut away" to expose internal components and re-rendered. (And we can do this for all rendering methods: opaque, transparent, semi-transparent, opaque objects in semi-transparent volumes with surface shading.) As an example, **Fig. 5** shows two views of a teacup which was scanned by an x-ray computed tomography system and then rendered with surface shading. The cutaway picture reveals the plastic spoon which was placed in the cup before the scan.

Other Tools

Wherever possible, we chose to use industry standard tools for this project. This included VIEW, X-windows, and the Sun workstation. With this base, the interactivity was fast and software development was made easy. Standard Pixar tools were modified so that they could be accessed directly by VIEW users and so that they could run concurrently with other codes in the system. The VIEW utilities and Pixar codes complement each other in that the former is primarily intended for

image processing and the latter is intended primarily for rendering and displaying multidimensional data.

Some of the data that the Scientific Visualization Project is asked to process does not fit neatly into our volumetric category; that is, the voxels are not equally spaced in a hypothetical volume. An example would be the three-dimensional display of seismic motion over time, a project worked on for Treaty Verification. Under such circumstances, we use a fast-line or polygon-drawing device to display the data.

One problem with the numerous machines and software tools we have available is that of data compatability; often vastly different formats of raw data are in need of display. We have taken steps to make that conversion as painless for the user as possible. For example, the process of converting data to the Pixar format and creating a Pixar volume or film loop has been automated. A program called *makevol* will read VIEW files, Rasfiles (see below), or raw data files and create a Pixar volume from them automatically. In addition, it will

- Set the desired size of the volume
- Rotate the resulting volume to a desired orientation
- Set the boundaries of the volume to eliminate "empty" space.

These programs allow users to convert data from their own format to the Pixar format so that they can use the Pixar tools mentioned above.

Another file format we developed, called Rasfiles,[2] will store images, time sequences of images, volumes, and time sequences of volumes. We have written a code called *rastool* to manipulate the data in the Rasfiles format. We have also written a code called *xtras* that displays Rasfile images, slices of volumes, and sequences of images or slices on the X window system. Code features include slow or fast playback, single frame stepping, and display of frame number. It accepts sequences of files or a file with a sequence of images, and uses the X-windows system to display data on remote systems. The *xtras* code can read data that are in the following formats:

- Black and white
- Pseudo-color
- Images in 24-bit full color.

We have written a library of volume manipulation routines similar in function to the routines executed by the Chap processor of the PIC. The routines make it possible to port volume rendering code from the PIC to general purpose machines, so if the PIC is not available to users or users want to work on their own systems, they can. However, the code runs slower and with less memory than it does on the PIC. Therefore, we wrote a fast program called *vrender* for visualizing volumetric data on general purpose hardware. This program uses an algorithm that is 30 times faster than the ray tracing algorithm commonly used for this type of data. Currently, the following options are available using this program:

- Partially transparent densities
- Surface enhancement/extraction
- Directional light source
- Specular reflections
- Arbitrary orientation of data
- Rotation of the data about arbitrary axis for a film loop
- Shadows
- Multiple materials
- Variable density and color transfer functions.

The code is portable, extensible, and uses Rasfiles for input/output.

Each of these programs or modules was imple-

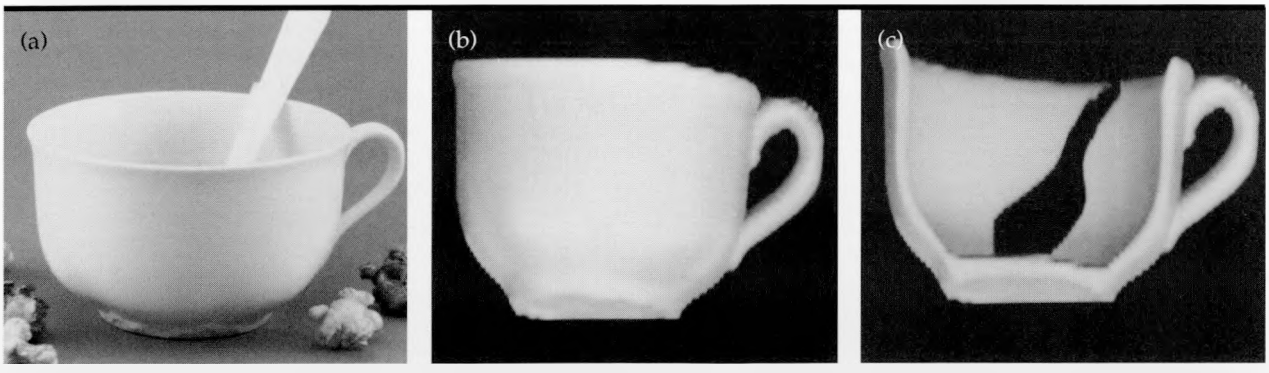


Figure 5. Reconstruction of images from three-dimensional data sets using the Pixar computer. (a) Photograph of a ceramic teacup. (b, c) Graphical renderings of three-dimensional reconstructions. The full volume of data was read into a Pixar imaging computer, and a simulated light source was used to render the data visible and generate the image. For (c), the reconstruction parameters were selected to show "internal structure" (here, a spoon).

mented, tested, and used on three different types of computers: DEC station 3100, Stellar GS-100, and Sun 3 workstations.

Future Work

The Scientific Visualization Project will continue to research and develop interactive visualization algorithms for programmatic use. In the next fiscal year we will reduce turn-around time, make the software more accessible to users, and refine the algorithms.

Acknowledgments

Most of the computer operations were performed by C. Grant, M. Miller, and M. Allison. Key elements of the Scientific Visualization Laboratory have been assembled by B. Cabral, J. Dias, and R. Shectman. The data for **Figs. 1, 3,** and **4** are courtesy of A. Pontau, SNLL; the data for **Figs. 2** and **5** are courtesy of the NDE group.

1. J. M. Brase, V. J. Miller, and M. G. Wieting, "The VIEW Signal and Image Processing System", LLNL Report UCID-21368, March 1988.
2. B. Cabral and C. L. Hunter, Visualization Tools at LLNL, IEEE Computer **22**, No 8, pp 77–84 (Aug. 1989).

Two-Dimensional Visualization System

Brian K. Cabral, Michael J. Allison, Bryan S. Lawver, and Robert M. Shectman
Engineering Research Division
Electronics Engineering

During FY 89, we developed SURFACE, an integrated visualization computer program for displaying two-dimensional (2-D) surfaces embedded in three-dimensional (3-D) Euclidean space. This program enables engineers throughout the Laboratory to display many types of surface data, including mechanical engineering TAURUS data files, 3-D contours of volume data, and traditional XY versus Z plots. SURFACE is the first such tool at LLNL which runs on many different workstations, supports a variety of different data types, allows for animation, and is highly interactive. As a result, SURFACE has been successfully exported to a number of different programmatic efforts, which are now using SURFACE as a part of their daily research activities.

Introduction

Many Laboratory programs have an increasing need for a better way of visualizing complex data sets, which require some sort of surface rendering in order to make the science they represent more comprehensible. SURFACE is a tool which uses several different surface display techniques to render data. SURFACE enables the user to rotate, zoom, animate, and control the display of many different data types using a single integrated user interface. This level of integration has the added advantage of exploiting the software commonalities shared by all surface-rendering algorithms. Most notably, the graphics transformations and polygon rendering which must occur for any surface display are carefully factored out to get the maximum possible reuse of software. The resulting surface polygons are then fed to GLIB, a portable high performance graphics package. By using GLIB, SURFACE minimizes the machine dependencies introduced into the end product. The net result is a tool which can run on many different workstation platforms while exploiting the enhanced performance of certain high-end workstations.

Progress

The current version of SURFACE was derived from a prototype version developed in FY 88. The prototype did not use a portable graphics package, did not support multiple data types, and did not exploit high end workstations. The current version of SURFACE, developed during FY 89, rectified these inadequacies and added some significant enhancements.

The salient features and enhancements to SURFACE are quite diverse. The progress made in the 2-D visualization effort will therefore be discussed in terms of a number of subtopics.

The GAB User Interface

The acronym GAB stands for Graphics ABstraction. This interface abstracts higher level graphics functionality into more usable modules. The primary set of modules supports an user interface. SURFACE uses a combination of GAB buttons and a slide bar to control the display, and a number of buttons to select display or rendering options. Some buttons connect the slide bar to a data variable such as scaling, rotation about an axis, or translation along an axis. The user interface provides a great deal of freedom for the user to browse through the data.

Iso-Contouring

Iso-contouring is a data rendering option which creates contours of equal potential through uniform 3-D gridding of data. One of many dependent data values available at the nodes is used to interpolate points, which form polygons. The polygons are plotted as iso-surfaces. Additional dependent values can be interpolated to produce shading of the surfaces.

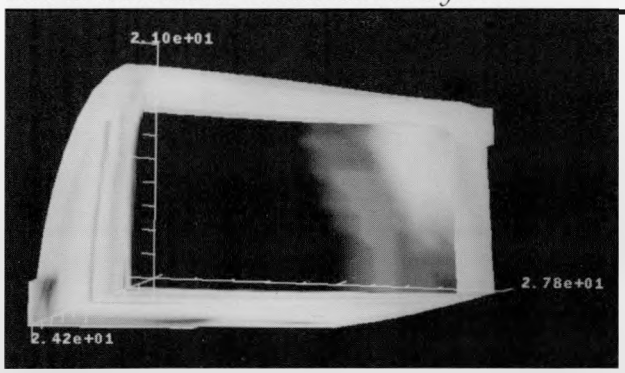


Figure 1. Three-dimensional display of a Taurus database. The object shown represents one quarter of a high-explosive cylindrical pellet. The three orthogonal scales are a three-dimensional coordinate system. Differences in pellet density or composition appear as shades of gray in the displayed shape.

TAURUS Displays

TAURUS is a plot data format of 3-D finite element codes. The data consist of "bricks" and the velocities at the corners of the bricks. From time step to time step, the bricks deform and move. Each brick is associated with a material type. SURFACE renders materials by finding exterior faces of all bricks of a particular material type **(Fig. 1).** Bricks which are completely surrounded by bricks of the same type are discarded, as are faces which touch faces of the same material type. Either the material type or the velocities at the brick corners can shade the face.

Surface Plots

Any 2-D gridded data can be used to produce 3-D perspective surface plots. The dependent data at grid points becomes the third dimension. A second dependent data value can shade the surface to produce a view of the data which displays two dependent data values simultaneously.

GLIB

The name GLIB stands for Graphics LIBrary, and is yet another type of graphics interface. GLIB is meant to be a very thin layer between higher abstraction such as GAB and the multitude of vendor-specific graphics libraries such as X-windows and the IRIS library. GLIB provides us with a buffering layer to mitigate these violations so that applications can be platform-independent. SURFACE is the first tool of its power anywhere in the industry to be platform-independent. Here at the Laboratory diversity of workstations is the rule, and GLIB allows SURFACE to move into many different programs.

Advanced Display Features

Where the display platform supports advanced capabilities, SURFACE has been extended to incorporate those features. For example, one workstation can generate Gourard-shaded polygons that produce a smooth transition of color across objects. Also supported is the capability to plot node points as shaded and lit spheres with specular lighting. Another feature is transparent surfaces on a workstation which supports alpha blending. Each feature extends the ability of the users to visualize their data, but SURFACE is still a powerful tool on simpler workstations which do not support these advanced features.

Future Work

During FY 90 we expect to complete the development of SURFACE into a finished product which will include a complete users manual, users quick reference, internal documentation, and a migration of the code to a support group. Furthermore, enhancements to the animation system which began in FY 89 will continue in FY 90, and in particular, utilizing quaternion algebra as a means to unify interactive and animation-scripting rotation schemes. During FY 90, we plan to acquire some form of scripting language based on an existing interpretive language, such as LISP, and integrate it into SURFACE and our other rendering software systems.

X-Ray Imaging System

H. Joseph Weaver and Gary E. Sommargren (consultant)
Engineering Sciences Division
Mechanical Engineering

We are developing a new type of x-ray imaging system in which an opaque sphere is used to image x rays by diffraction. We have successfully described the diffraction and imaging processes analytically and have shown excellent agreement with experimental results taken in the visible region of the spectrum. (We are presently preparing the x-ray experiments.) Because of the simplicity of the system, the potential for high spatial resolution, and the extremely large depth of field, this technique could have a significant impact on x-ray lithography (a field involved with the printing of high-density integrated circuits), the imaging of living cells with the aid of high-resolution microscopes, material studies, and diagnostics in the Weapons Programs.

Introduction

The diffraction of light (or any part of the electromagnetic spectrum including x rays) by an opaque sphere such that a point source of illumination is focused to a bright spot behind the sphere - at the center of the shadow cast by the sphere - has been a curiosity ever since Arago demonstrated the phenomenon experimentally in 1818. **Figure 1** shows the basic arrangement of the source, opaque sphere, and observation plane. The distribution of light in the observation plane was derived by Mie in 1908 using rigorous electromagnetic theory. The resultant formulas, however, are quite involved and cannot easily be used to study the characteristics of the diffraction pattern. Some authors have described the pattern using scalar diffraction theory. Most have concentrated on the axial intensity of the central bright spot, with Osterberg and Smith solving the Rayleigh diffraction integral exactly. Their results showed that the bright spot exists at all axial points behind the sphere, but that the intensity is not increased beyond that which would be measured if the sphere were removed. Others have derived the radial intensity of the central bright spot using various levels of approximation.

In 1914 Hufford realized that this diffraction phenomenon could also be used to form an image. He demonstrated this using a carbon arc lamp, but never pursued the problem analytically or with invisible radiation. Since then, this phenomenon has received little attention, particularly for visible imaging applications where high quality lenses and mirrors are readily available.

However, for short, invisible wavelengths such as x rays, there are no transmissive materials with refractive indices high enough to be used to fabricate lenses. Thus, x rays are typically focused and imaged using grazing incidence mirrors, multilayer mirrors, or zone plates, each of which is formidibly difficult to fabricate. It is, however, relatively easy to fabricate a sphere, and if it could be shown that such a sphere could be used to image x rays, it would prove to be a useful addition to x-ray technology.

Our Approach

The goal of this project is to describe analytically and verify experimentally that an opaque sphere

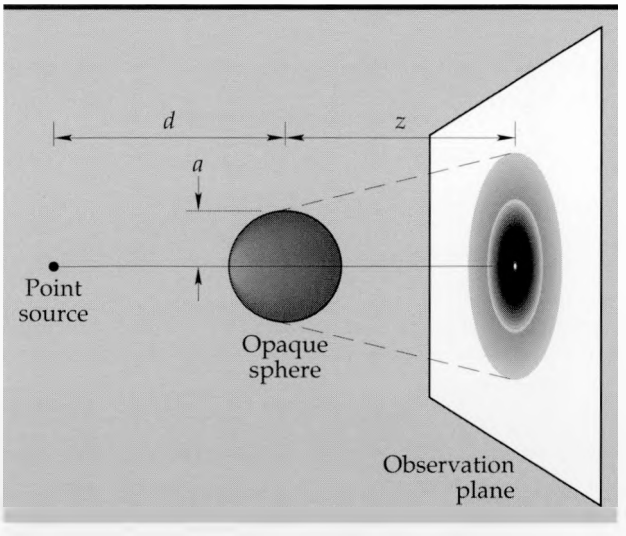


Figure 1. Schematic diagram of the diffraction by an opaque sphere of light from a point source to form a focused bright spot at the center of the shadow cast by the sphere.

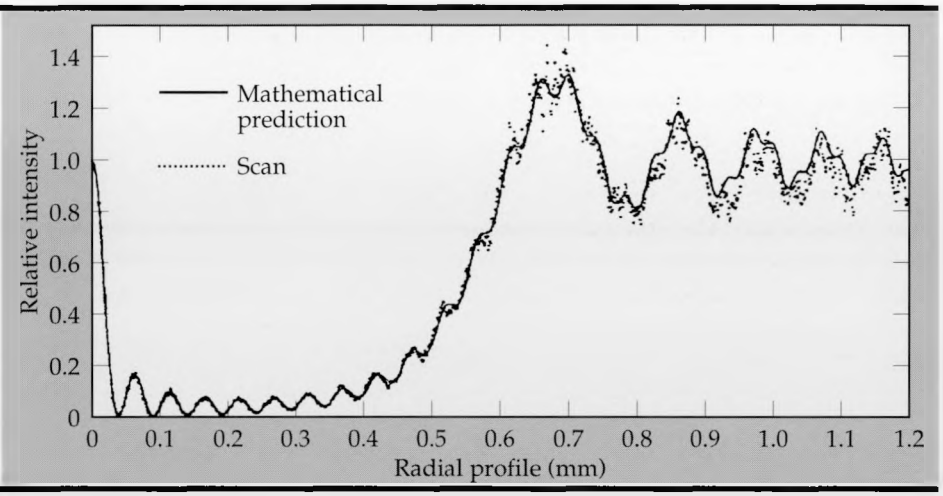


Figure 2. Comparison of the diffraction pattern of a point source with a microdensitometer scan of an actual diffraction pattern.

can be used as an imaging element for x rays. To do this we started with scalar diffraction theory, using minimal approximations, and derived analytic expressions that describe the diffraction of light from a point source by an opaque sphere. Using the geometry in **Fig. 1,** we found that the intensity, $I(s)$, of the diffraction pattern[1] is given by

$$I(s) = \frac{1}{(z+d)^2}\left[V_0(u, v)^2 + V_1(u, v)^2\right]$$

where

s = radial distance from center of spot in observation plane

d = distance from source to center of sphere

z = distance from the center of sphere to observation plane

$V_0(u, v)$ and $V_1(u,v)$ are Lommel functions

$$u = ka^2\left(\frac{z+d}{zd}\right)$$

$$v = \left(\frac{kas}{z}\right)$$

a = radius of sphere

$k = 2\pi, \lambda$

λ = wavelength of the source radiation.

This equation can be applied to both visible light and x rays, which means that we can verify it by performing visible light experiments (a much easier set of experiments to do). Using the above equation, we calculated the diffraction pattern of a point source and compared the results to a microdensitometer scan of an actual diffraction pattern. **Figure 2** shows excellent agreement between calculation and experiment. As predicted, the image of the point source appeared as the central peak of the diffraction pattern, indicating that the opaque sphere did act as an imaging element.[2]

Although the above experiment shows that an opaque sphere can be used an an imaging element, it does not indicate its image quality. One way to measure how well an opaque sphere images is to determine the minimum separation of two point sources, the images of which appear distinct and separate. (This is called spatial resolution.) Theoretically, we found the resolution of the opaque sphere to be given by

$$\text{Resolution} = \frac{0.38\lambda z}{a}.$$

This was verified by imaging two incoherent point sources that were separated by a distance equal to the resolution limit given by the above equation. **Figure 3** shows both the calculated and experimental results. As can be seen, the model is in good agreement with the experiment.

We then looked at how accurately the sphere could image extended objects (as opposed to points). Extended objects may be considered an aggregate of point sources, each one producing a corresponding diffraction pattern in the observation plane. The sum of these diffraction patterns is the resultant image. **Figure 4** shows the image of a standard three-bar target using visible illumination and the corresponding calculated image. Again, there is good agreement between theory and experiment; however, there is a loss of contrast in the image. This is related to the efficiency by which the imaging system passes different spatial frequencies. It is described mathematically by the modulation transfer function of the opaque sphere, which we are presently studying.

Discussion

Although imaging with an opaque sphere is not a new idea, it has never been fully developed, either analytically or experimentally. This is because it is not a practical alternative to imaging elements,

such as lenses and mirrors, in the visible part of the spectrum. However, at x-ray wavelengths, where grazing incidence mirrors, multilayer mirrors, and zone plates are difficult to fabricate to tolerances required to achieve their theoretical resolution, an imaging opaque sphere may be a viable addition to x-ray technology. The art of manufacturing extremely round, smooth spheres (attributes that enhance their imaging performance) has improved considerably over the last several years, primarily due to the need for accurate spheres for general relativity experiments.

The imaging system described here has several unique characteristics that set it apart from conventional imaging systems. First, it has an almost infinite depth of field; i.e., the image exists at all distances, z, behind the sphere. This is advantageous for x-ray lithography because we are working to very close tolerances. In conventional systems the image is localized within a small interval. The dimension of this interval is related to the spatial resolution. Increasing the resolution by a given factor decreases the depth of field by that same factor squared. For high resolution conventional systems the depth of field is on the order of the wavelength of the illumination. For x rays this presents significant alignment and stability problems.

Secondly, imaging with an opaque sphere is independent of the wavelength. The source of illumination can be monochromatic, broad band or can be varied during its use without affecting the image or the alignment. Conventional systems using multilayer mirrors and zone plates must be designed for a specific wavelength for optimum performance.

Finally, the opaque sphere images at different magnifications by simply changing the ratio of z to d (see **Fig. 1**). When $z > d$, the system gives a magnified image. This is the geometry used when imaging living cells with a microscope. When $z < d$, the image is reduced, a geometry useful in x-ray lithography where the printing of integrated circuits in highly dense arrays calls for the exact

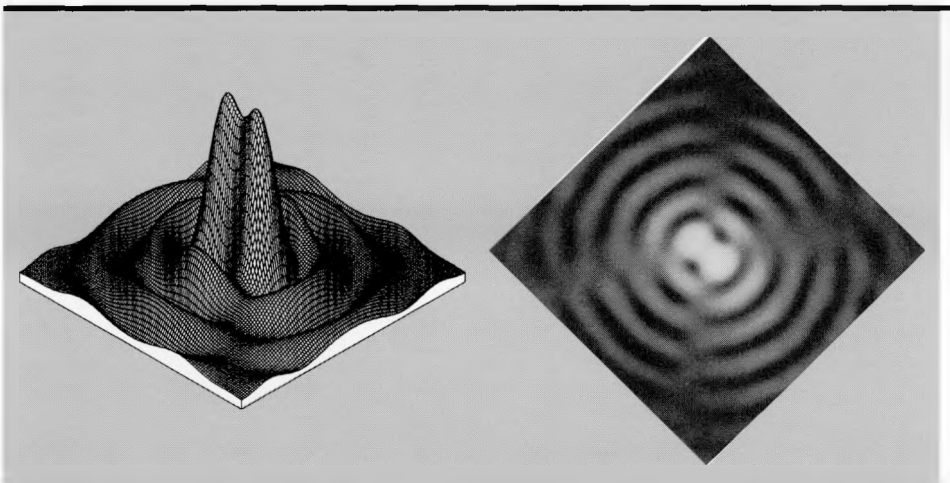


Figure 3. Comparison of a model and experiment of two incoherent point sources where the points are separated by a distance equal to the minimum predicted resolution.

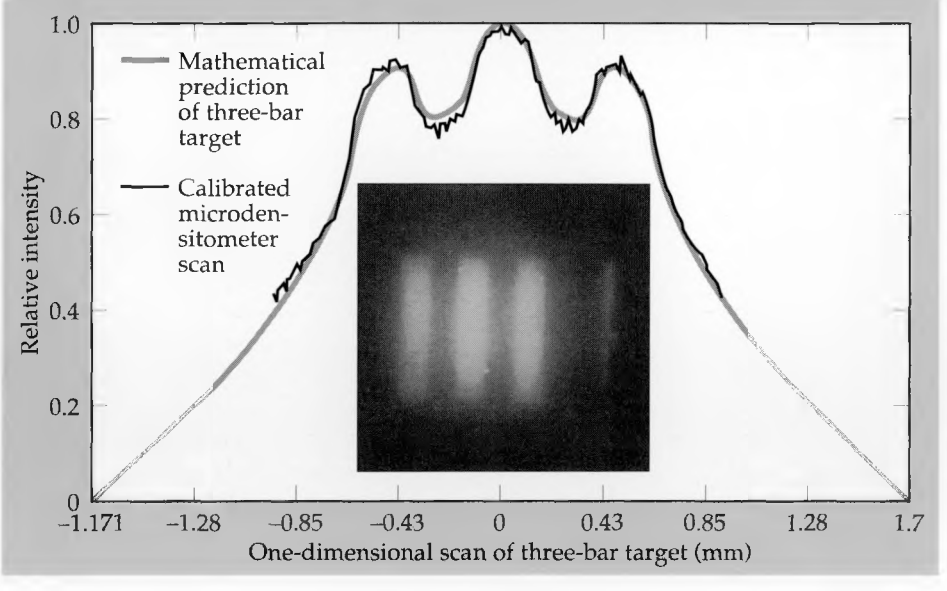


Figure 4. Comparison of the model and experiment of a three-bar target.

Table 1. Our approach to the problem of developing an x-ray imaging scheme based on the diffractive focusing of radiation by an opaque sphere.

- ■ Mathematical description of the diffraction problem
- ■ Experimental confirmation of mathematical analysis
- ■ Two-point resolution of imaging system
  - ■ Analytical studies
  - ■ Experimental studies
- ■ Mathematical description of sphere as imaging device
- ◪ Experimental demonstration of imaging
  - ■ Optical domain
  - □ X-ray domain
- ◪ Transfer function resolution of imaging system
  - ■ Analytical studies
  - □ Experimental studies
- □ Possible applications
  - □ X-ray lithography
  - □ Biological/medical research
  - □ LLNL Weapons Program

■ Completed work
◪ Started work
□ Work planned but not yet started

placement of extremely small lines. Conventional systems, on the other hand, must be designed for a particular magnification and application to keep aberrations to a minimum.

Future Work

Up to now, all our experiments have been done with visible illumination, primarily because it was a simple way of verifying our calculations. Now that we believe that our analysis is sound, we have started setting up an imaging system with an x-ray source. We expect to have preliminary results within the next few months. **Table 1** summarizes what has been accomplished and what future work needs to be done.

1. G. E. Sommargren and H. J. Weaver, "Diffraction of Light by a Circular Obstruction, Part 1: Description and Properties of the Diffraction Pattern," *Appl. Opt.* (to be published).
2. G. E. Sommargren and H. J. Weaver, "Diffraction of Light by a Circular Obstruction, Part II: Image Formation and Resolution Considerations," *Appl. Opt.* (submitted for publication).

A
Allison, Michael J. 9-31
Alvarez Jr., Raymond A. 1-37
Andreason, Sterling J. 3-18
Armatis, Philip M. 3-18
Azevedo, Stephen G. 7-15, 9-25

B
Barth, Marvin J. 1-25
Belak, James F. 4-8
Berryman, James G. 8-1
Blaedel, Kenneth L. 4-1
Boercker, David B. 4-17
Brand, Hal R. 9-13
Brase, James M. 8-6, 8-16, 9-25
Brentjes, Jay 5-5

C
Cabral, Brian K. 9-22, 9-31
Camacho, J. Frank 6-4
Candy, James V. 1-1
Ciarlo, Dino R. 1-9, 3-4
Clark, Gregory A. 8-16
Cooper, Gregory A. 3-8, 3-30
Craig, George D. 1-47

D
Davidson, J. Courtney 3-30
Davis, Pete J. 4-1
Dease, Clyde G. 1-13
Deford, John F. 1-18, 1-47
DeGroot, Anthony J. 4-8, 9-1
Dini, Jack W. 4-29
Druce, Robert L. 6-21

E
Engelmann, Bruce 2-1

F
Falabella, Steven 4-17
Feng, William W. 5-1
Fitch, J. Patrick 8-6, 8-9, 8-16
Flower, Elane C. 5-23
Folta, James A. 3-4
Fuchs, Baruch A. 4-21

G
Goodman, Dennis M. 8-9, 8-12
Gourdin, William H. 4-29
Grant, Chuck W. 9-22, 9-25
Groves, Scott E. 5-14

H
Hagans, Karla G. 3-26
Hee, Ed W. 1-9
Hofer, W. Wayne 6-21
Hoover, Carol G. 4-8
Hoover, William G. 4-8
Hui, Wing C. 3-4

I

J
Johansson, Erik M. 9-1

K
Khanaka, Gizzing H. 1-13, 1-33
Kihara, Ron 6-1, 6-9
Krammen, Jerome E. 9-25
Kulke, Bernhard 1-47

L
Lager, Darrel L. 9-13
Lawrence, Taylor W. 8-9
Lawver, Bryan S. 9-1, 9-31
Lehman, Sean K. 8-16
Lewis, D. Kent 1-21, 7-3

M
Madsen, Niel K. 1-29
Magnuson Jr., Waldo G. 1-6
Mariella Jr., Raymond P. 3-8, 3-10
Martz, Harry E. 7-15
Maurer, William J. 9-13
Mayhall, David J. 1-37
McConaghy, Charles F. 1-9, 3-12
McLeod, Robert R. 1-18, 1-47
McWright, Glen M. 3-1
Miller, Dale E. 3-21

N
Nikkel Jr., Daniel J. 4-1, 5-10

O
Orvis, William J. 1-9, 1-33

P
Pennock, Steven T. 1-18
Pocha, Michael D. 6-21
Poole, Brian R. 6-4

Q

R
Rhodes, Mark 6-4
Riddle, Robert A. 5-10
Roberson, George P. 7-15
Rosenbury, E. Tom 6-4
Ryon, Richard W. 7-1

S
Sanders, David M. 4-17
Schmucker, Ronald F. 1-47
Schneberk, Daniel J. 7-15
Shapiro, Arthur B. 2-10
Shectman, Robert M. 9-31
Skeate, Michael F. 7-15
Sommargren, Gary E. 9-33
Stein, Werner 2-7
Stowers, Irving F. 4-8
Streit, Ronald D. 5-10
Swierkowski, Stefan P. 3-30
Syn, Chol K. 4-21, 7-3, 7-7

T
Taylor, John S. 4-21

U

V Velsko, Stephan P. 4-21

W Wahlstrand, Karna J. 1-37
Weaver, H. Joseph 9-33
Wheelock, Everett W. 9-9
White III, W. Travis 1-13, 1-33, 9-1
Whirley, Robert G. 2-12
Wieting, Melvin G. 9-25

X

Y Yee, Jick H. 1-9, 1-33, 1-37
Yorkey, Thomas J. 8-16

Z Zentler, Jan-Mark 6-14
Ziolkowski, Richard W. 1-1, 1-21,1-25, 1-29, 1-37

A
Accelerator 1-18
Acoustic tomography 8-1
Advanced test accelerator 9-13
AMOS code 1-47
Artificial intelligence 9-13
Atmospheric compensated imaging 8-9

B
Backward-wave oscillator 6-4

C
Calibrating optical transient detectors 3-18
Cathodic arc 4-17
CELL-TRACER code 9-22
Charge-coupled device 3-30
Circuit modeling 1-6
Computer simulation of electronic circuits 1-6
Computerized tomography 7-15

D
Deconvolution algorithms 8-12
Diamond turning 4-21
Ductile grinding 4-1
Ductile metal, strain of 2-1
DYNA3D code 2-12

E
Elastoplastic shells 2-12
Electro-optic transient recording systems 3-26, 3-30
Electron-gun modulator 6-1
Experimental test accelerator 9-13
Expert system 9-13

F
Failure modeling 5-10, 5-14
FAR code 1-25
Fiber-reinforced composite 5-1, 5-5, 5-14
Finite difference time domain 1-25, 1-47
Finite element time domain 1-25
Fluid flow problems 2-7
Free-electron laser 1-18

G

H
Heat transfer problems 2-7, 2-10
High-power microwave sources 1-25, 6-4
Hologram 8-16

I
Identification algorithms 8-12
Ignitron 6-9
Image processing 8-1, 8-6, 8-9
Ion-assisted etching 3-21
Induction accelerator 1-47
Integrated optics 3-1

J

K
Kirkpatrick-Baez mirror optics 7-1
KIVA-II code 2-7

L
Localized wave transmission 1-21

M
Mach-Zehnder 3-1, 3-26
Magnetic switches 6-14
Maxwell's curl equations 1-29
Microfabrication 3-4
Miniature vacuum tubes 1-9
Modified power spectrum pulses 1-1
Molecular beam epitaxy 3-8, 3-10
Multidimensional data, display of 9-25

N
NET-2 code 1-6
NIKE2D code 2-1, 5-10
Non-destructive evaluation 7-7, 7-15
Nonequilibrium molecular dynamics 4-8

O
Optical recording 3-12
Orthogonal cutting 4-8

P
Parallel processing 9-1
Photoconductor 1-13, 6-21
Photon backscatter 7-7
Physical vapor deposition 4-17
Precision machining 4-8
Pulsed fields 1-25

Q
Quantum well 1-33, 3-10

R
Radar ocean imaging 8-6
Rayleigh waves 7-3
Reactive-ion etching 3-21
Relational database 9-9
Relativistic electron beam 6-4
Residual stress in composites 5-5, 5-23

S
Seismic tomography 8-1
Shaped charge 4-29
Signal processing ..... 7-15, 8-1, -6, -9, -12, -16
Single-crystal optics 4-21
Solid-state superlattice 1-33
SPRINT 2 code 9-1
SURFACE code 9-31
Surface breakdown 1-37
Surface indentation 4-8
Surface rendering 9-31
Switches, photoconductive 1-13

T
Thermo-Fluid Mechanics Group 2-7
Thin-film coatings 4-17
TOPAZ2D code 2-10
Total internal reflection microscopy 7-7
Transient waves 1-1
Traveltime tomography 8-1
TSAR code 1-18

U Ultrasonic detection 7-3

V Vacuum tubes, miniature 1-6
Volume-visualization code 9-22
VRENDER code 9-22

W Wave propagation 1-1
Wavelength deflection 3-12

X X-ray imaging 9-33
X-ray laser holography 8-16
X-ray optics 7-1

Y

Z

U.S. Government Printing Office: 1990-788-027