SAND2011-1588P

Understanding the Performance of Scientific Simulation Programs on High Performance Computers using Proxies

Richard Barrett

http://www.sandia.gov/~rfbarre/

Scalable Architectures Department (1422)

Computer Science Dept Seminar

New Mexico State University

March 9, 2011

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000.

Preparing Multi-physics, Multi-scale Codes for Hybrid HPC

SIAM Computational Science and Engineering

March 3, 2011

Richard Barrett, Rich Drake, and Allen Robinson
Center for Computing Research (1400)

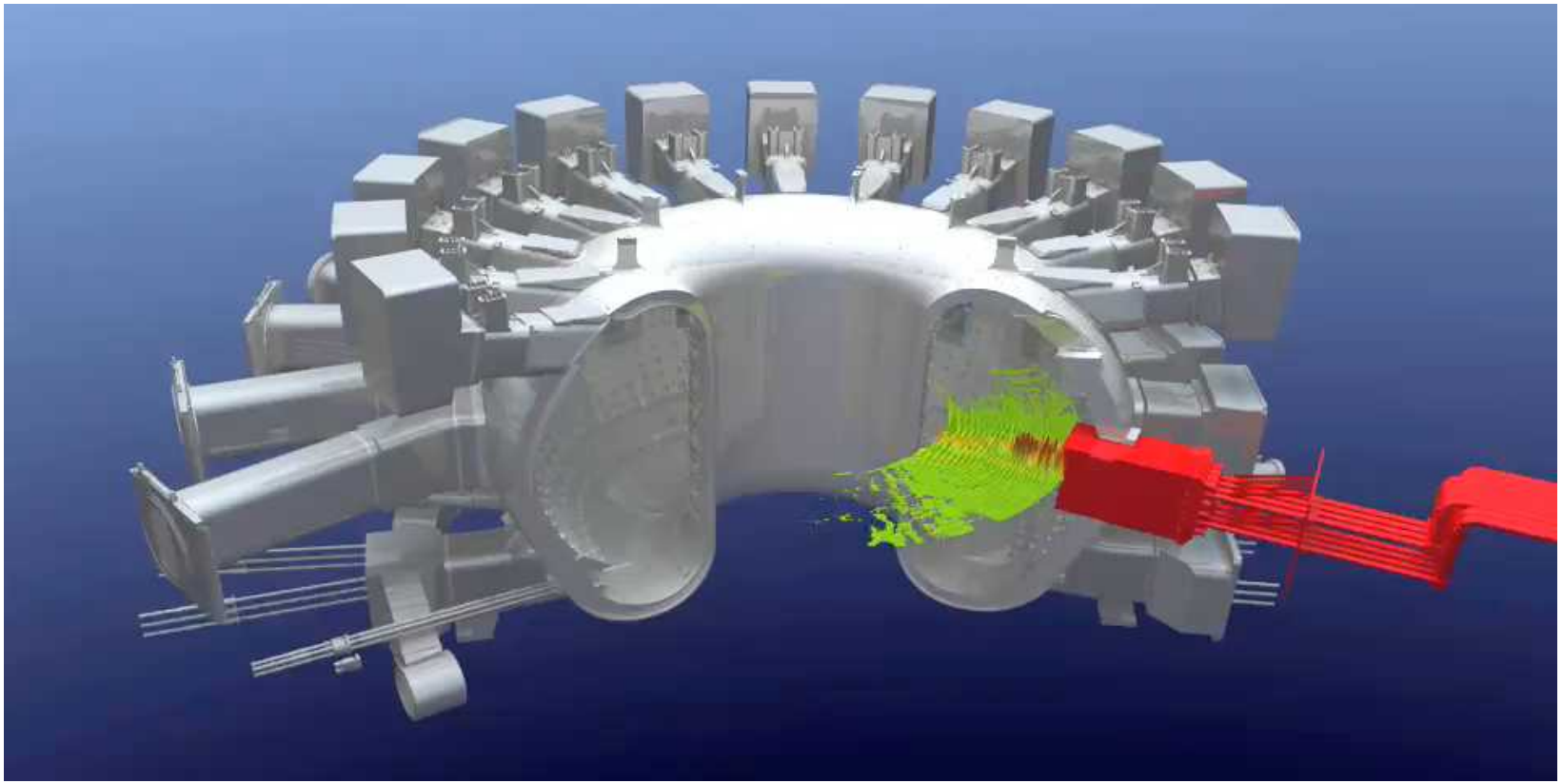
Sandia National Laboratories is a multi-program laboratory operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin company, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

National Nuclear Security Administration

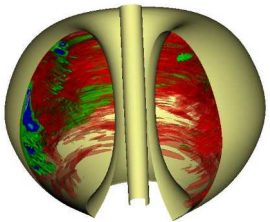
Sandia National Laboratories

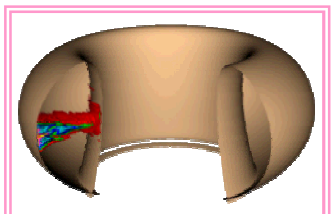
Programming model, mechanisms, etc

- How programmer views data and the computations that operate on it.
- Mechanism: MPI, OpenMP, cuda, opencl, etc
- Critical link: how machine views data and the computations that operate on it.
- Over-arching goal: science and engineering



AORSA simulation; movie by Sean Ahern@ORNL





Sandia National Laboratories

```
C     APPROXIMATE VALUES FOR SOME IMPORTANT MACHINES ARE:
C
C          IBM/195    CDC/7600  UNIVAC/1108   VAX 11/780 (UNIX)
C           (D.P.)  (S.P.,RNDG)    (D.P.)     (S.P.)     (D.P.)
C
C NSIG       16         14          18           8         17
C ENTEN    1.0D75     1.0E322     1.0D307     1.0E38    1.0D38
C ENSIG    1.0D16     1.0E14      1.0D18      1.0E8     1.0D17
C RTNSIG   1.0D-4     1.0E-4      1.0D-5      1.0E-2    1.0D-4
C ENMTEN   2.2D-78    1.0E-290    1.2D-308    1.2E-37   1.2D-37
C XLARGE   1.0D4      1.0E4       1.0D4       1.0E4     1.0D4
C EXPARG   174.0D0    740.0E0     709.0D0     88.0E0    88.0D0

c timing               on ncar"s control data 7600, besic takes about
c                    .32+.008*n milliseconds when z=(1.0,1.0).
c
c portability          ansi 1966 standard
```

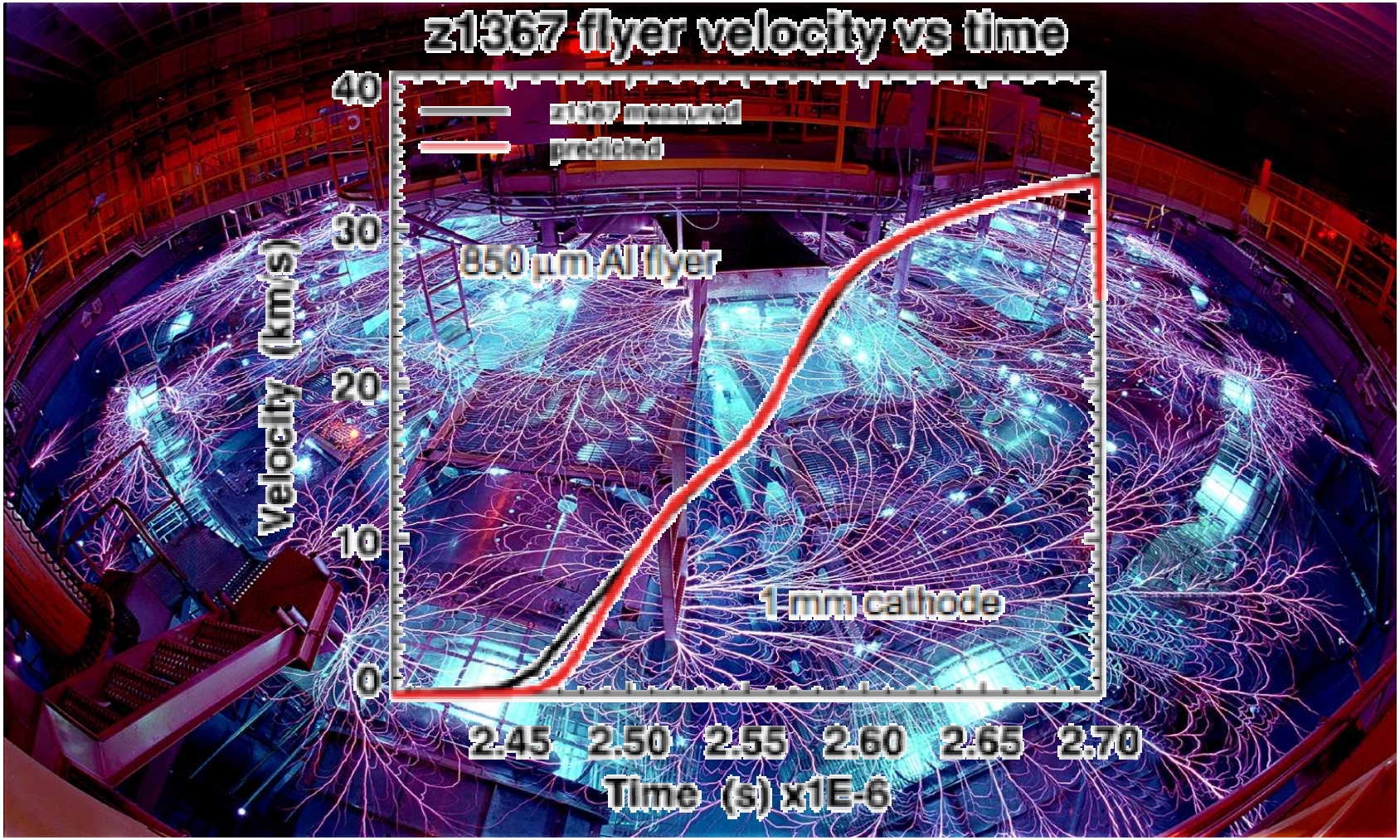
Sandia National Laboratories

ALEGRA

- Simulate large deformations and strong shock physics including solid dynamics in an Arbitrary Lagrangian-Eulerian methodology

- Also magnetics, MHD, electromechanics and a wide range of phenomena for high-energy physics applications.

Pulsed power: Z-machine



Sandia National Laboratories

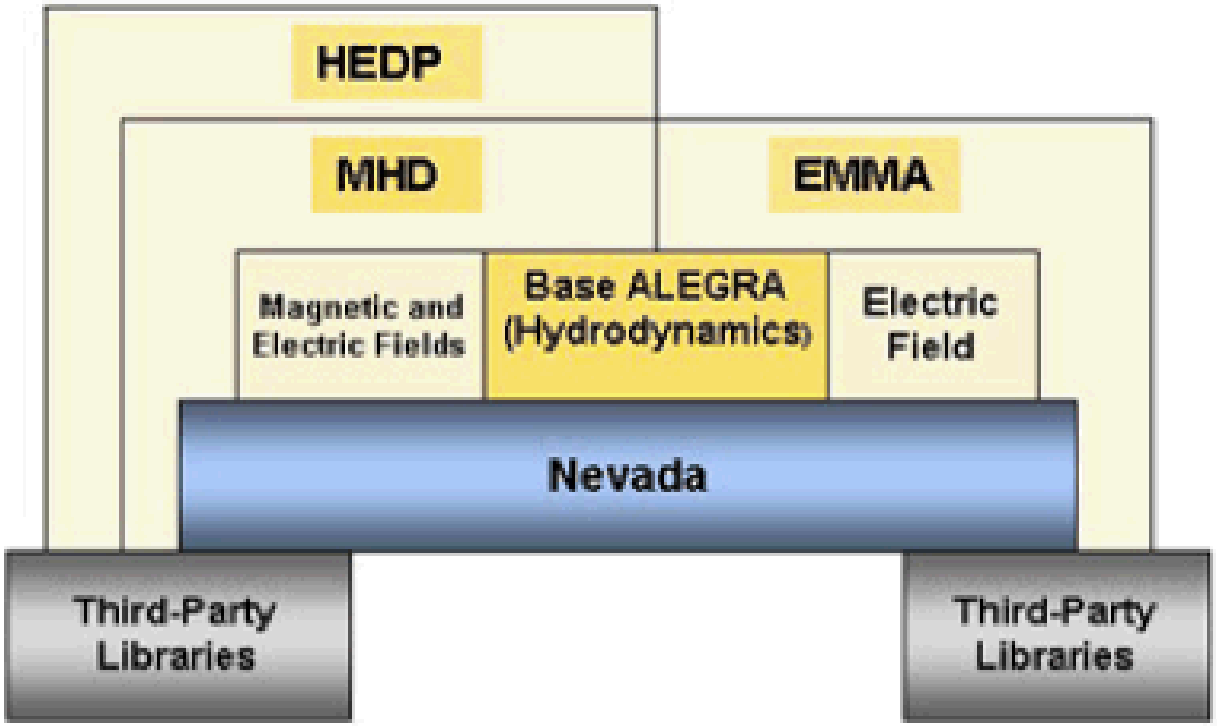
ALEGRA code base*
(project began 1990)

```
                    C/C++ SOURCE LINES OF CODE COUNTING PROGRAM
          ( c ) Copyright 1998 - 2000 University of Southern California, CodeCount (TM)

   University of Southern California retains ownership of this copy of software.  It is licensed to you.
Use, duplication, or sale of this product, except as described in the CodeCount License Agreement, is
strictly prohibited.  This License and your right to use the software automatically terminate if
you fail to comply with any provisions of the License Agreement.  Violators may be prosecuted.
This product is licensed to : USC CSE and COCOMO II Affiliates


The Totals
   Total    Blank |      Comments      |  Compiler  Data    Exec.   |  Number   |          File  SLOC
   Lines    Lines |   Whole  Embedded  |  Direct.   Decl.   Instr.  | of Files  |   SLOC   Type  Definition
-------------------------------------------------------------------------------------------------------
--------------------
  388275    62268 |   72506      8267  |    14688   64562   174252  |   1241    |  253502  CODE  Physical
  388275    62268 |   72506      8267  |    14622   32912   116441  |   1241    |  163975  CODE  Logical
    5388      778 |       0         0  |        0    4610        0  |     68    |    4610  DATA  Physical

Number of files successfully accessed.....................  1309 out of 1353

Ratio of Physical to Logical SLOC.........................     1.55

Number of files with :
        Executable Instructions        >     100          =     289
        Data Declarations              >     100          =      48
        Percentage of Comments to SLOC <      60.0 %      =     697      Ave. Percentage of Comments to Logical
 SLOC =  49.3

REVISION AG4  SOURCE PROGRAM -> C_LINES          This output produced on Wed Feb 23 10:20:26 2011
```

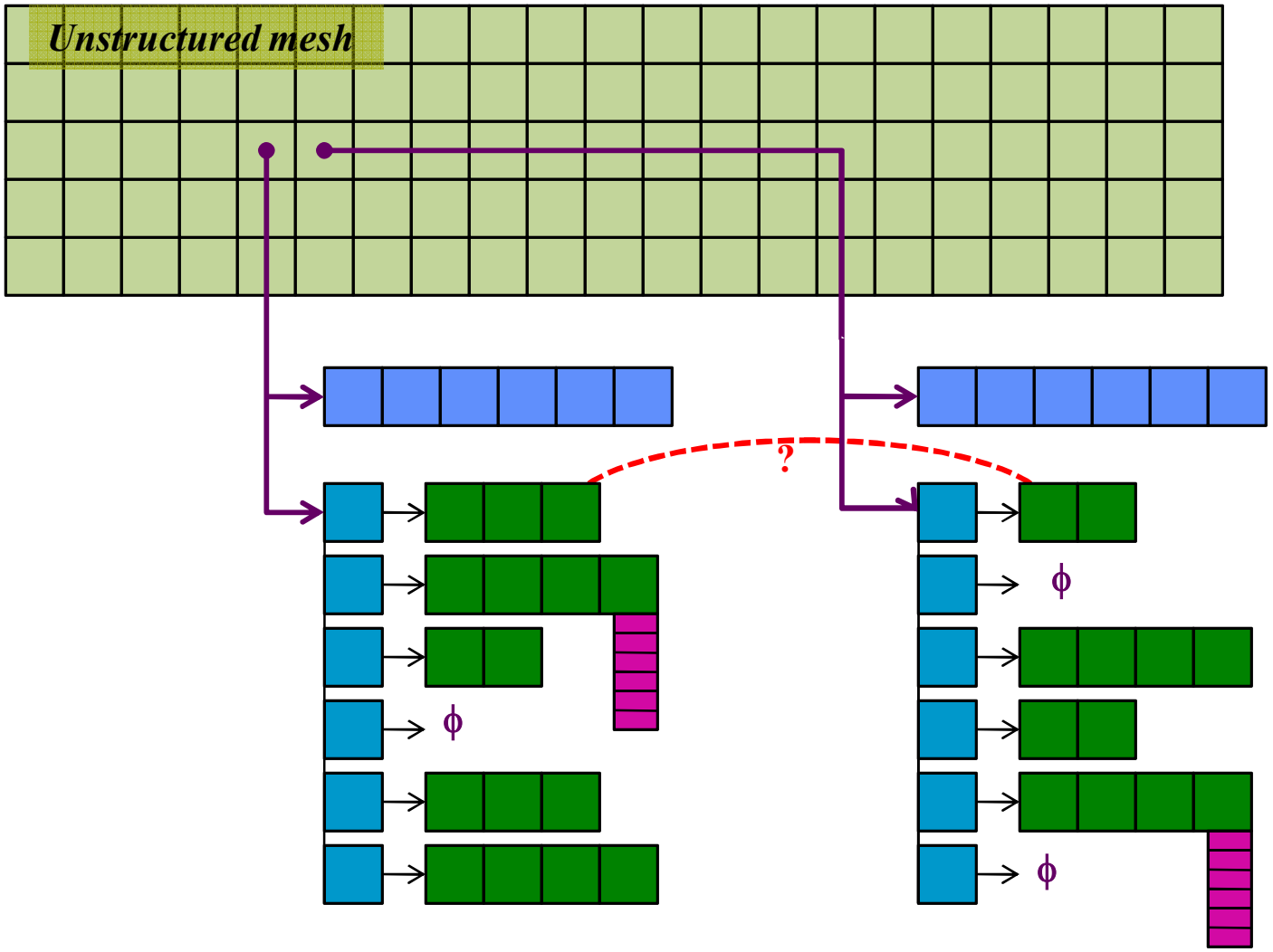
** Excluding some Fortran (58k@121f), python, xml, etc, some uncounted files, and the Nevada framework.*

Sandia National Laboratories

ALEGRA software infrastructure



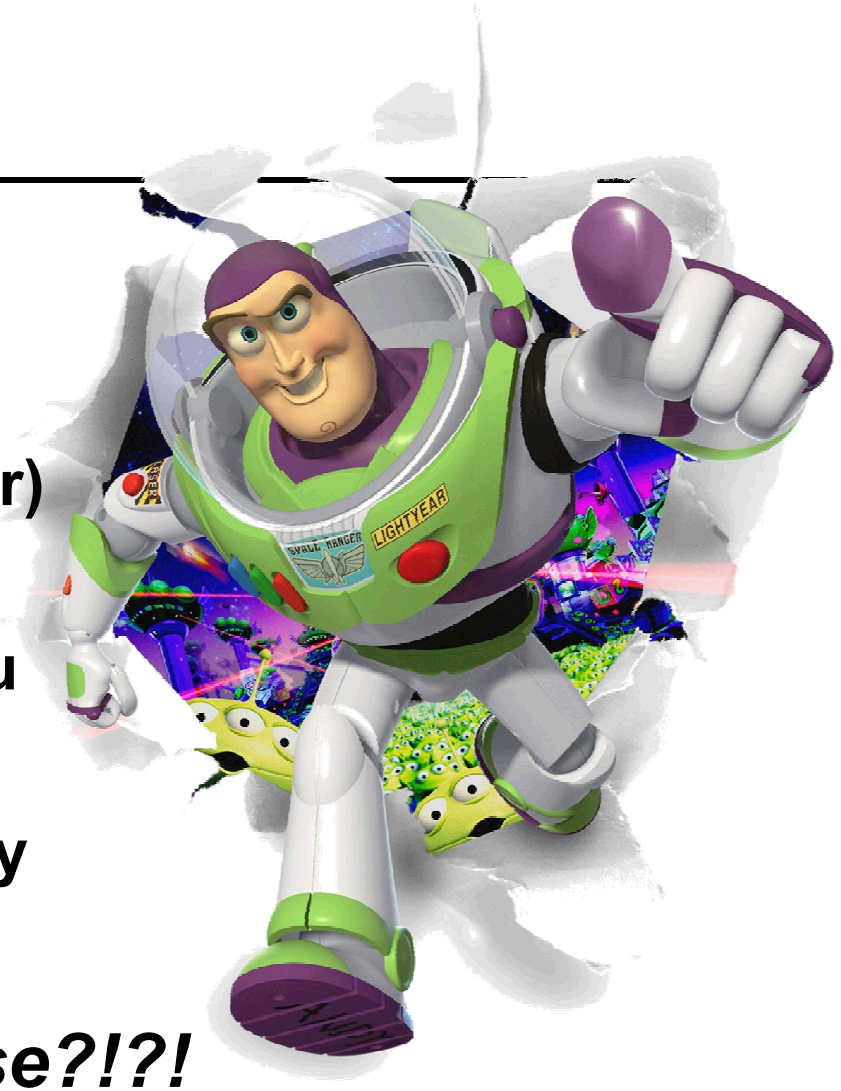
ALEGRA data structure



Target architectures

- Small clusters: linux, SunOS, IRIX, AIX
- MPP: Red Storm, Red Sky
- New ASC capability: Cielo

…and beyond



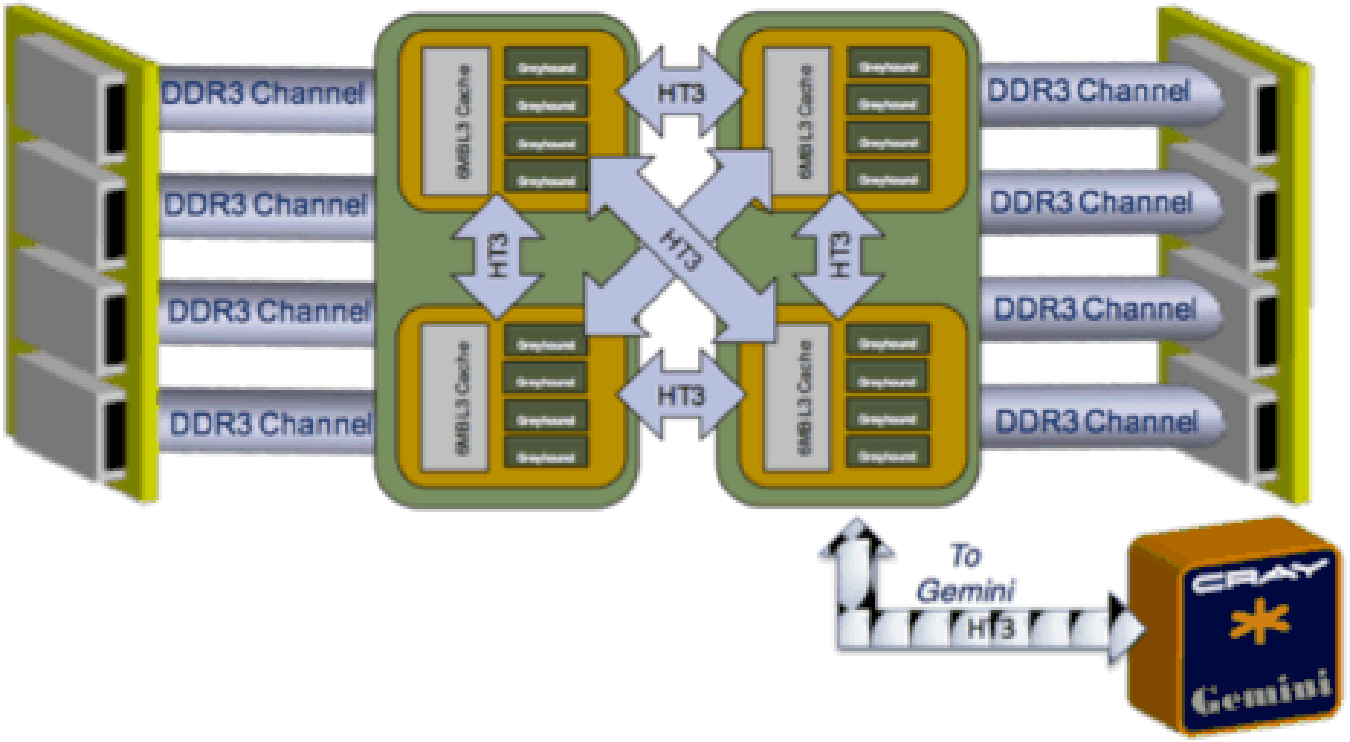
"Accelerator"-based arch

- Cell + Opterons (Roadrunner)
- gpu + x86
- LCF3@ORNL: 20PF, mc+gpu

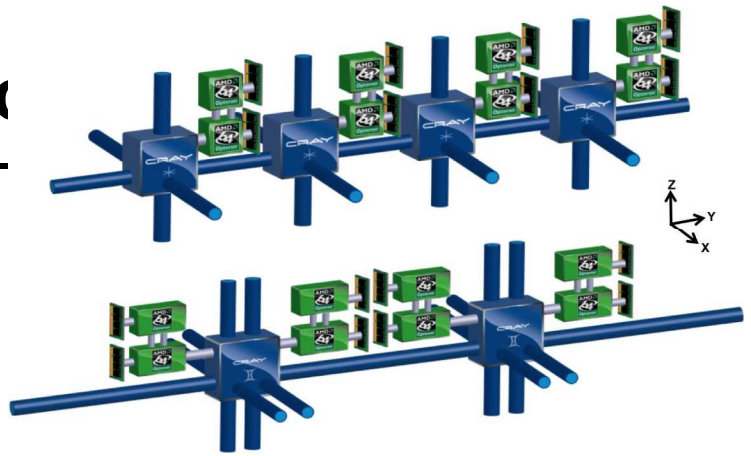
Intel many-core, eg Knights Ferry

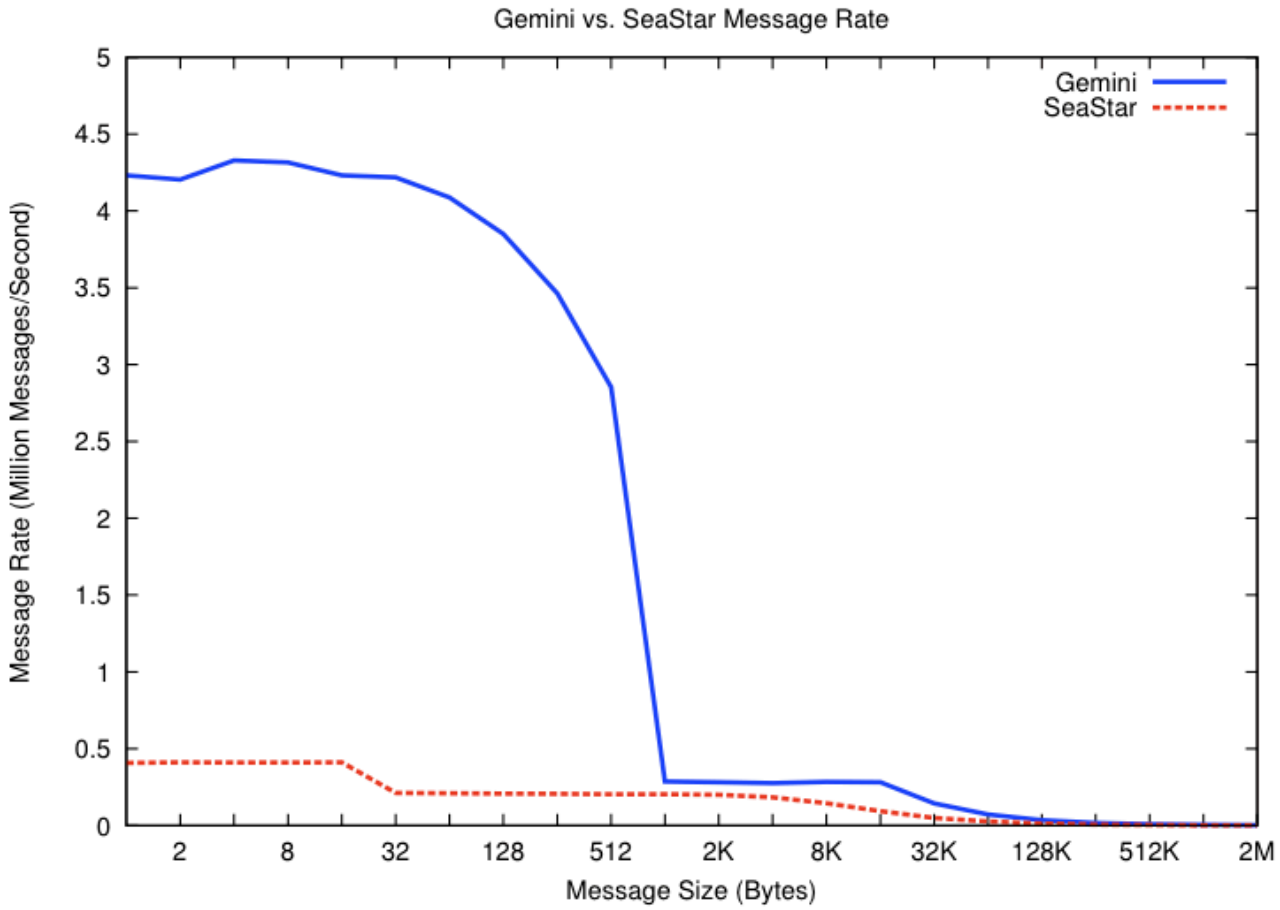
So how do we program these?!?!

Cielo Cray XE6



Cielo Gemini Interc





Sandia National Laboratories

Computer Science Goals

- Port to new platforms
- Improve performance
- Investigate new algorithms, programming models, etc
- Machine procurement
- Manage power requirements
- Simulators
- Engage external research communities, vendors, 3rd party code developers, etc.

The problem

- Codes are million lines,
- under constant development,
- controlled distribution,
- limited access to computing environments,
- etc.

Goal :
At most, one and a half code re-writes

1: Revolutionary: programming model

½ : Evolutionary: programming mechanism

The Mantevo Project

- focused on developing tools to accelerate and improve the design of high performance computers and applications by providing application and library proxies to the high performance computing community.
- Proxies written by application code developers
- http://software.sandia.gov/mantevo

Mantevo* Project

* Greek: augur, guess, predict, presage

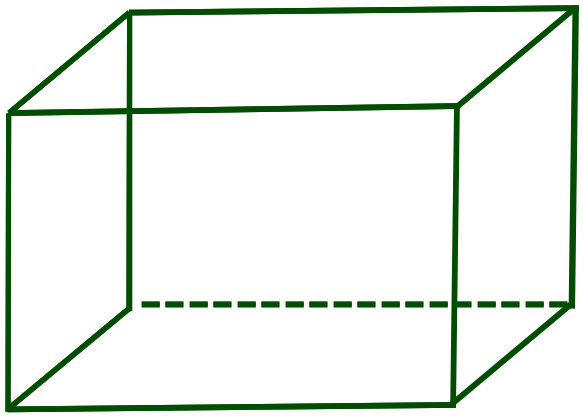
- Multi-faceted application performance project.
- Three types of packages:
  - **Miniapps**: Small, self-contained programs.
    - **MiniFE/HPCCG**: unstructured implicit FEM/FVM.
    - **phdMesh**: explicit FEM, contact detection.
    - **MiniMD**: MD Force computations.
    - **MiniXyce**: Circuit RC ladder.
  - **Minidrivers**: Wrappers around Trilinos packages.
    - **Beam**: Intrepid+FEI+Trilinos solvers.
    - **Epetra Benchmark Tests**: Core Epetra kernels.
  - **Motif framework**: Collection of “dwarves”.
    - **Prolego**: Parameterized, composable fragment collection to mimic real apps.
- Open Source (LGPL)
- Staffing: Application & Library developers.

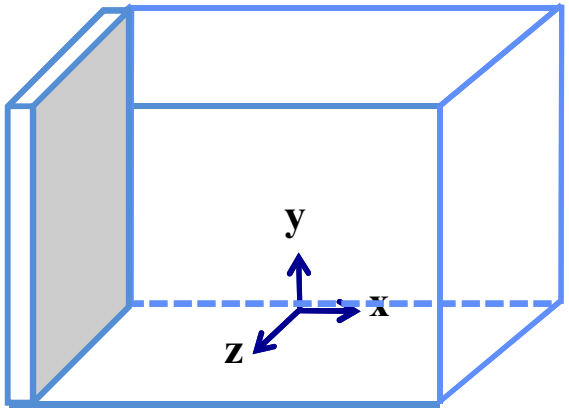
Mantevo Characterization

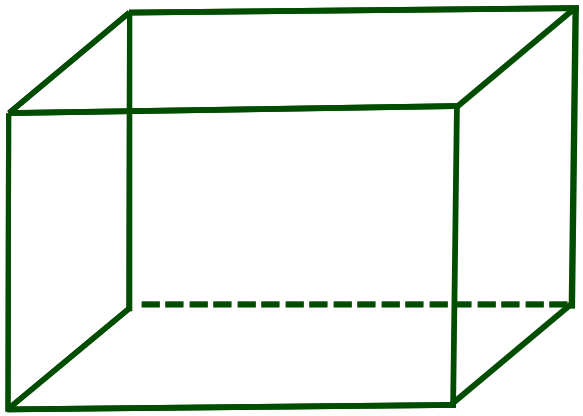
- Development of “co-design vehicles”, i.e. miniapps.
- Roles:
  - App developer: Developer & owner of miniapp (key).
  - Algorithms expert: Knowledge of algorithm options.
  - Runtime/OS expert: Knowledge of system SW.
  - HW expert: Component selection, arch trends.
  - Benchmark expert: Focused performance studies.
- Goal:
  - Concrete foundation for design studies.
    - Dwarves: “Even as cartoon characters they are sketchy.” J. Lewis.
  - Starting point for:
    - Performance studies (many kinds).
    - Algorithm replacement studies.
    - New programming models (even total rewrites).
  - Elevated conversation between all interested parties.

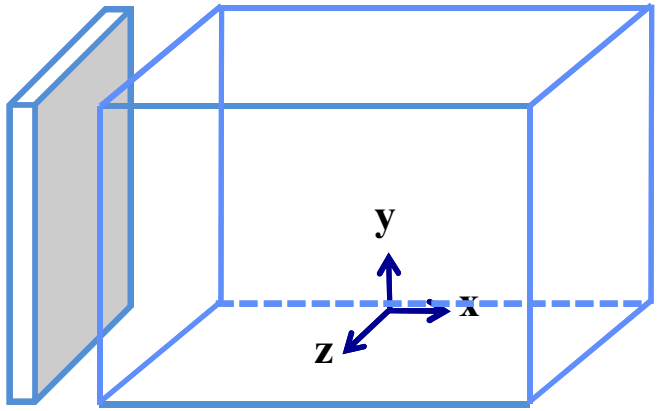
Two developing mini-apps

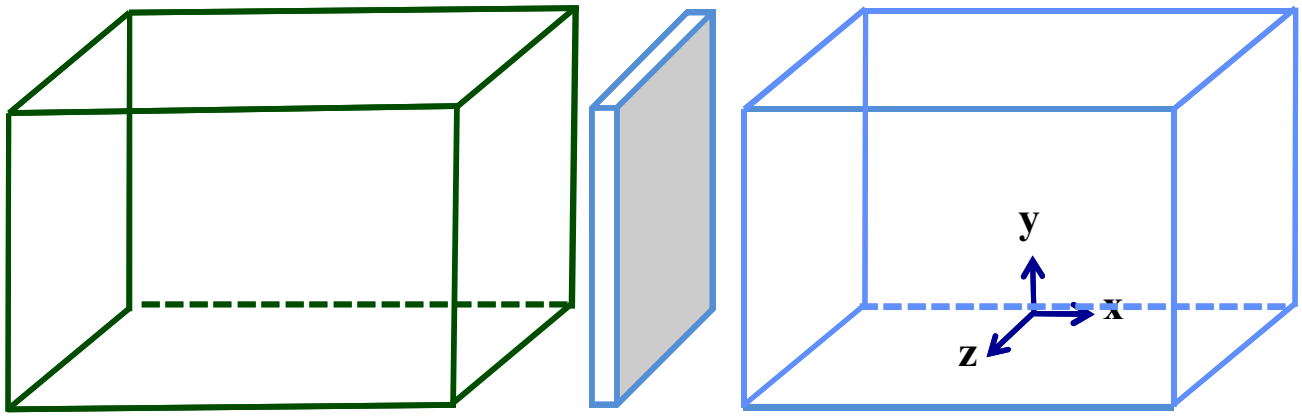
- ALE re-map
- Eulerian using BSP with message aggregation

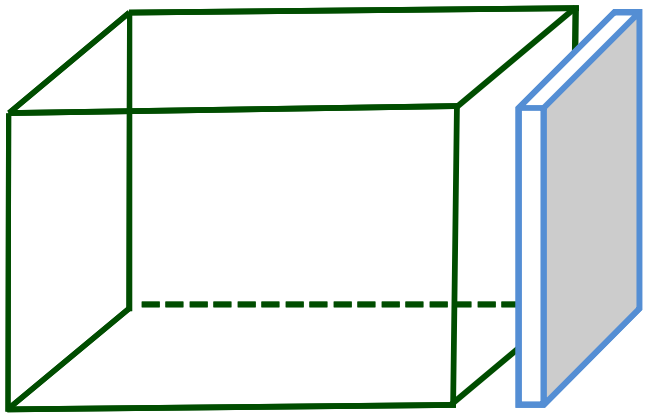


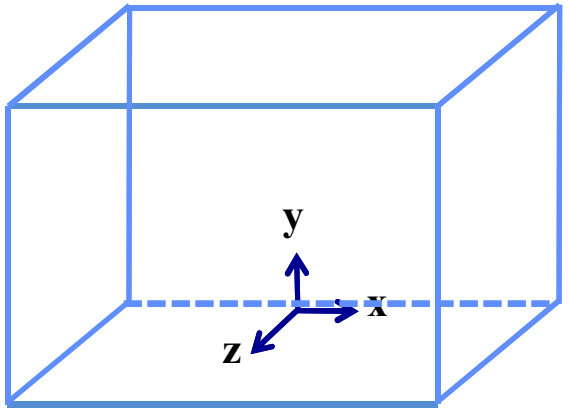


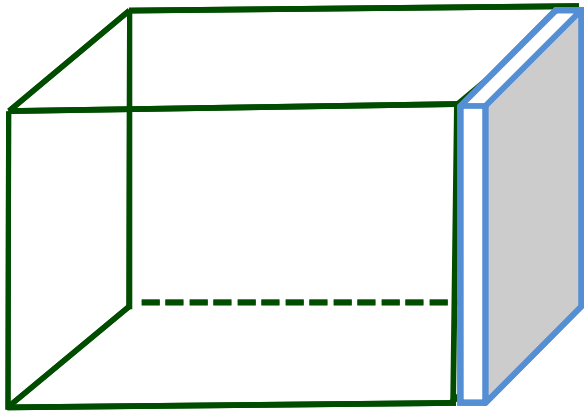


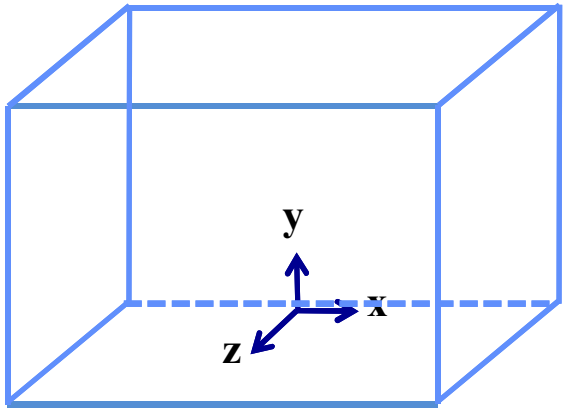








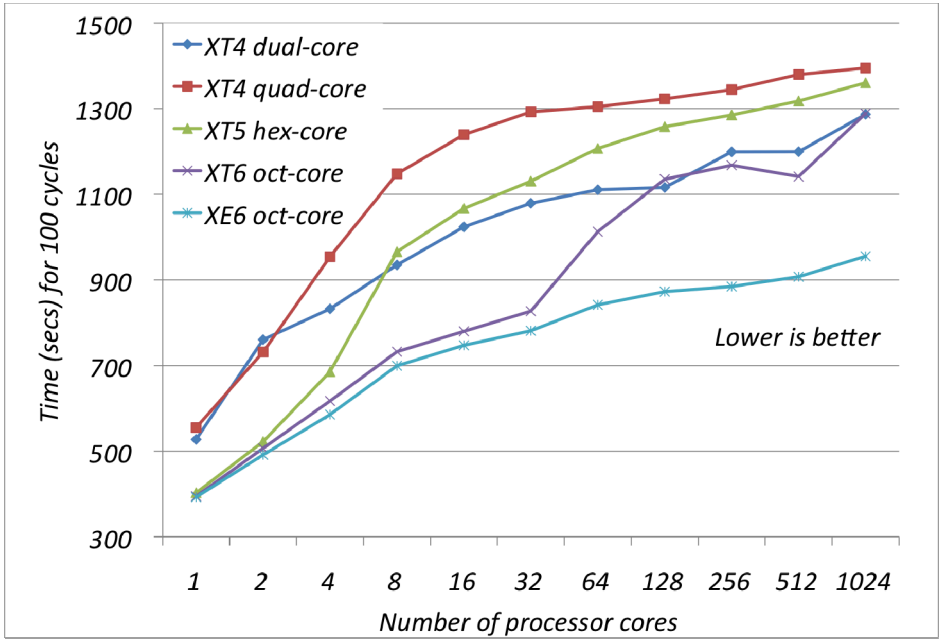




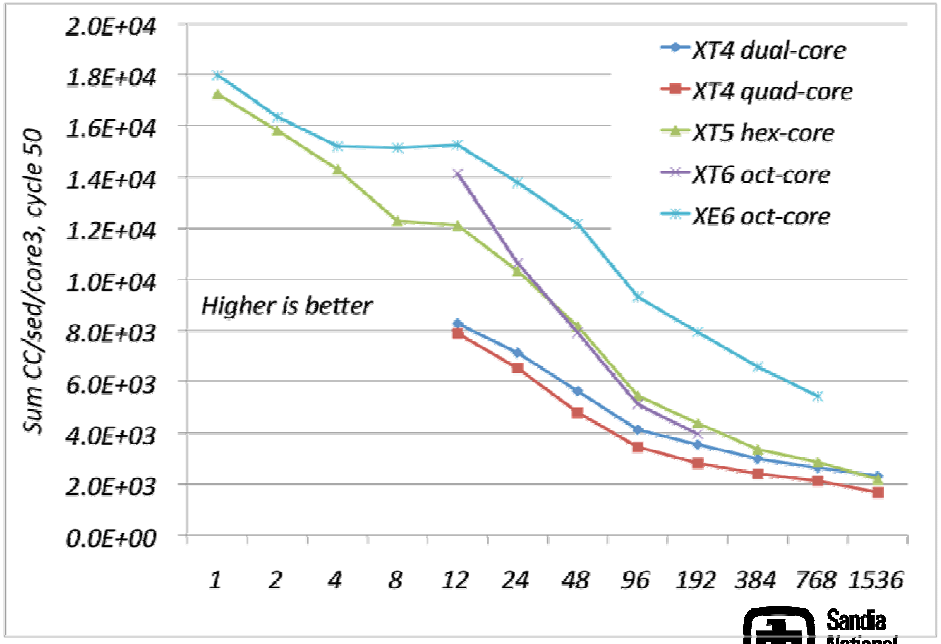
Each time step:

- Boundaries (2d faces) of up to 40 variables exchanged 19 times
  - up to 6 nearest neighbors
  - ~3 MByte messages
- 90 reductions (mostly MPI_Allreduce (8 bytes))
- Stencil sweeps over variables

CTH

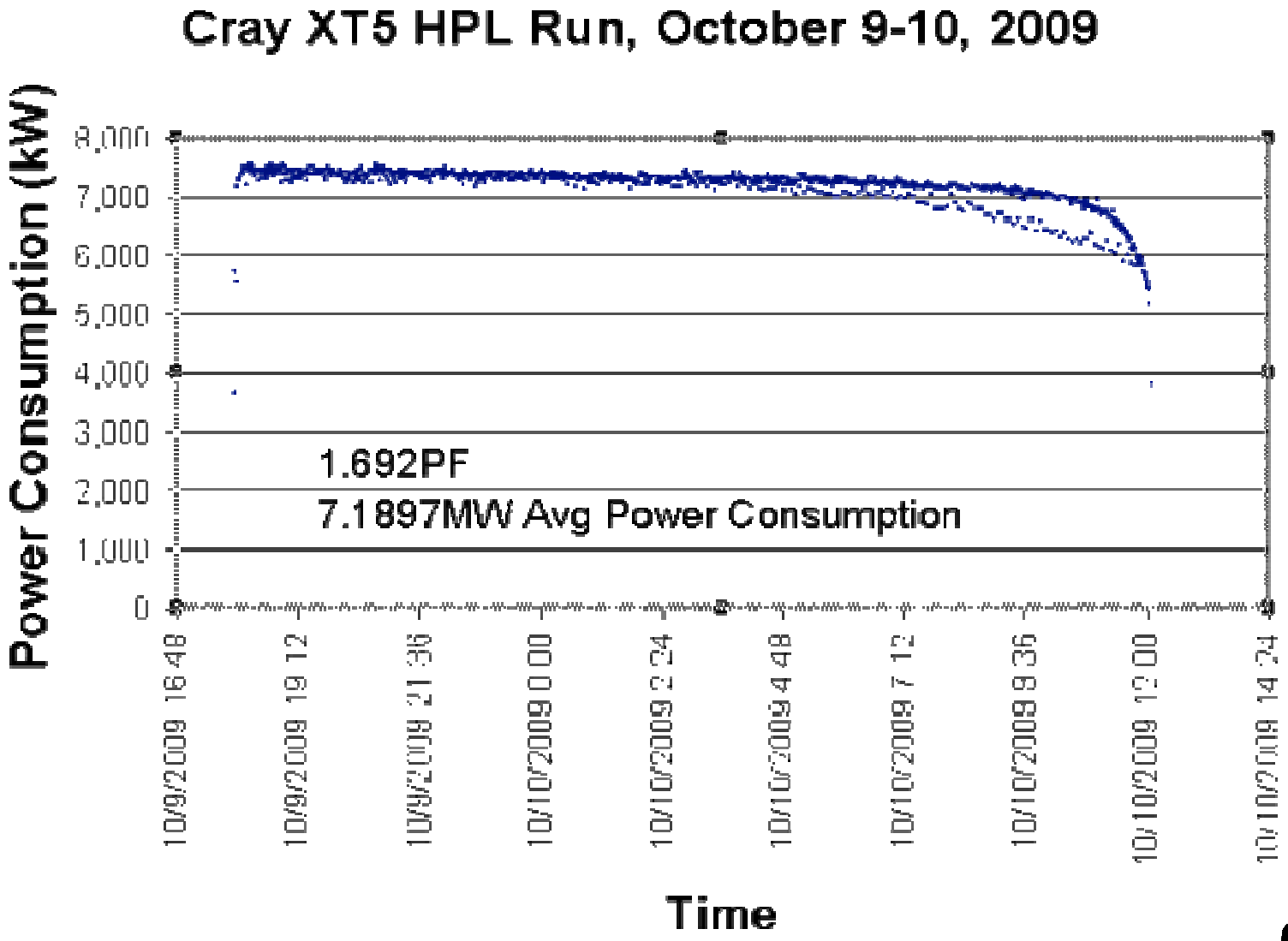


xNobel



Sandia National Laboratories

It's the power…



Sandia National Laboratories

AMG2006*

Platform: Jaguar

Architecture: XT4

CPU: AMD Quad

P-states (Frequency States)

- **P0: 2.1 GHz , 1.25V**
- **P1: 2.1 GHz , 1.25V**
- **P2: 1.7 GHz, 1.1625V**
- **P3: 1.4 GHz, 1.125V**
- **P4: 1.1 GHz, 1.1V**

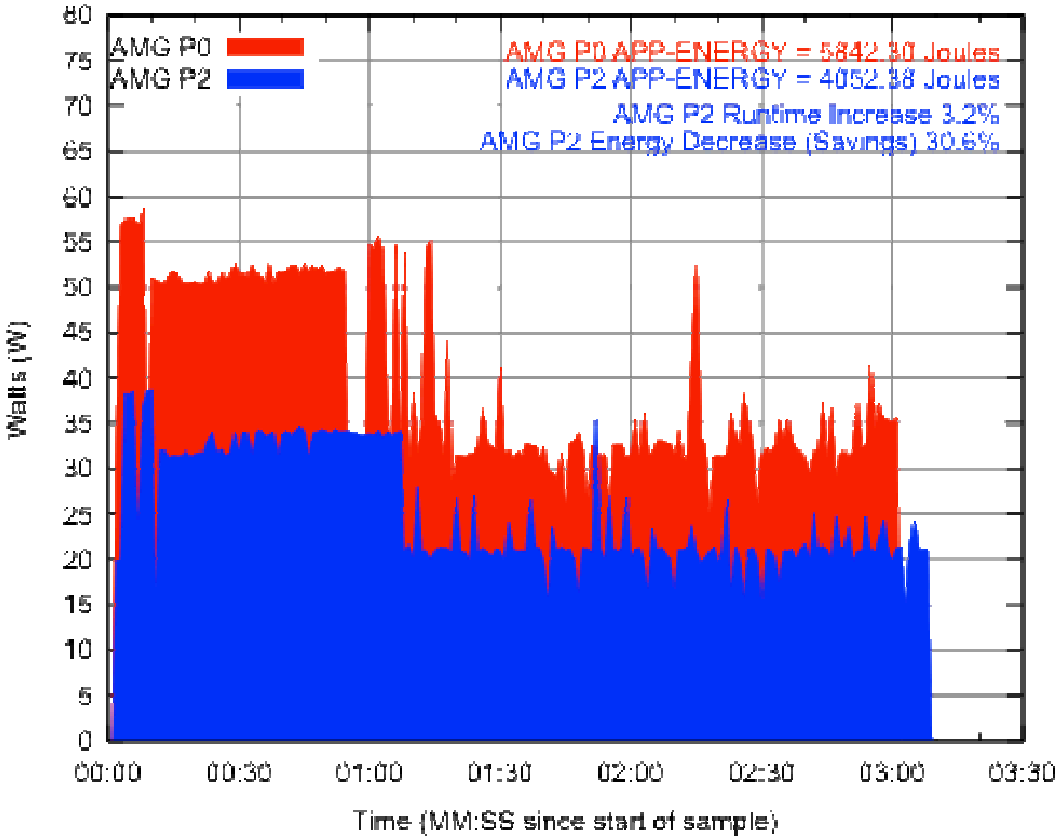
Nodes: 6144

Runtime Increase: 3.2%

Energy Decrease (Savings): 30.6%

<u>Order of magnitude energy savings vs. performance impact!</u>

Two application runs, same physical nodes, statically altering CPU frequency (P-state) allows lowering input voltage to chip resulting in larger energy savings.



Single node capture of watts over time for each run of AMG2006 varying P-states

** Work of Jim Laros@Sandia*

Sandia National Laboratories

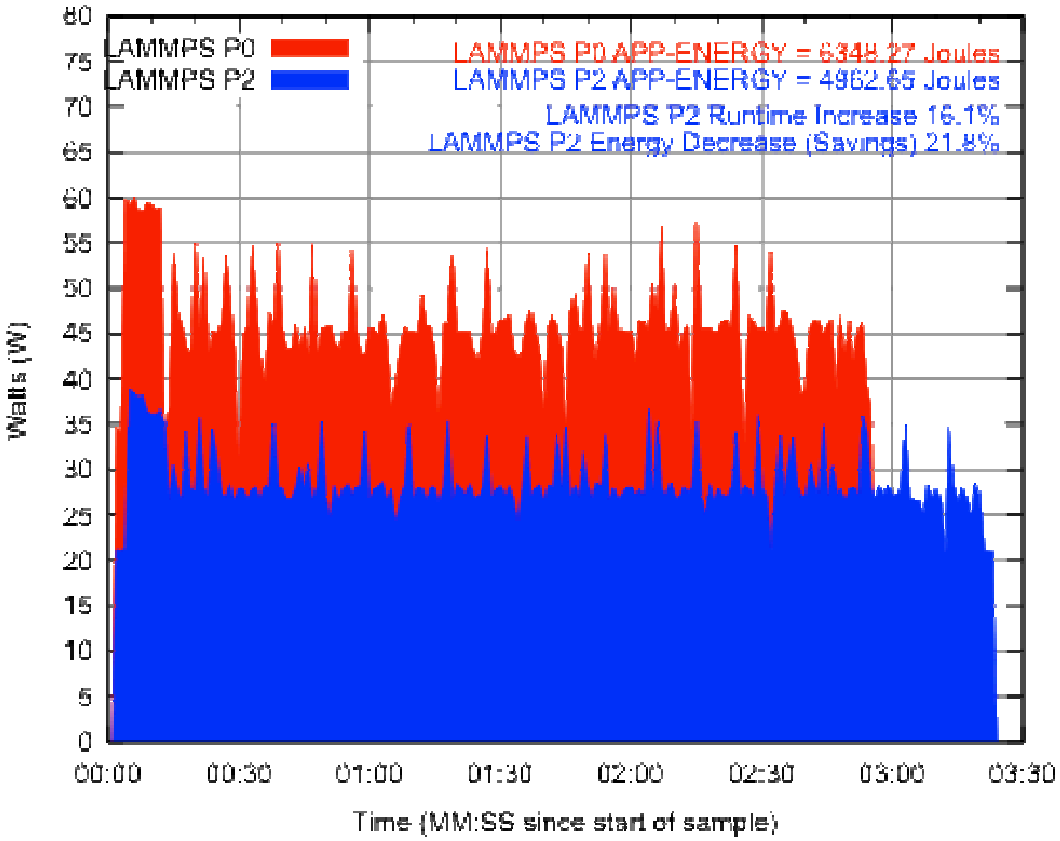
LAMMPS*

Platform: Jaguar
Architecture: XT4
CPU: AMD Quad
P-states (Frequency States)
- **P0: 2.1 GHz , 1.25V**
- **P1: 2.1 GHz , 1.25V**
- **P2: 1.7 GHz, 1.1625V**
- **P3: 1.4 GHz, 1.125V**
- **P4: 1.1 GHz, 1.1V**

Nodes: 4096
Runtime Increase: 16.1%
Energy Decrease (Savings): 21.8%

Compute intensive application, still observe significant energy savings. Illustrates which applications can expect most benefit.

Two application runs, same physical nodes, statically altering CPU frequency (P-state) allows lowering input voltage to chip resulting in larger energy savings.

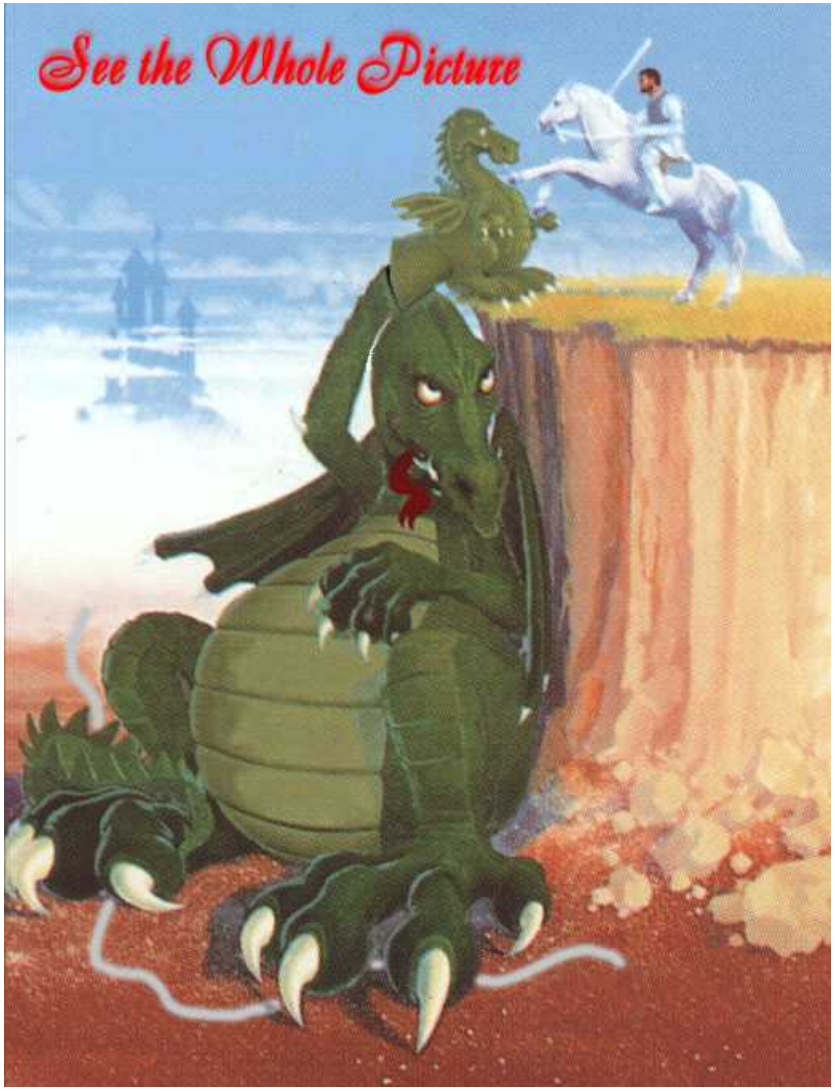


Single node capture of watts over time for each run of LAMMPS, varying P-states

** Work of Jim Laros@Sandia*

Sandia National Laboratories

What are we evaluating/measuring?

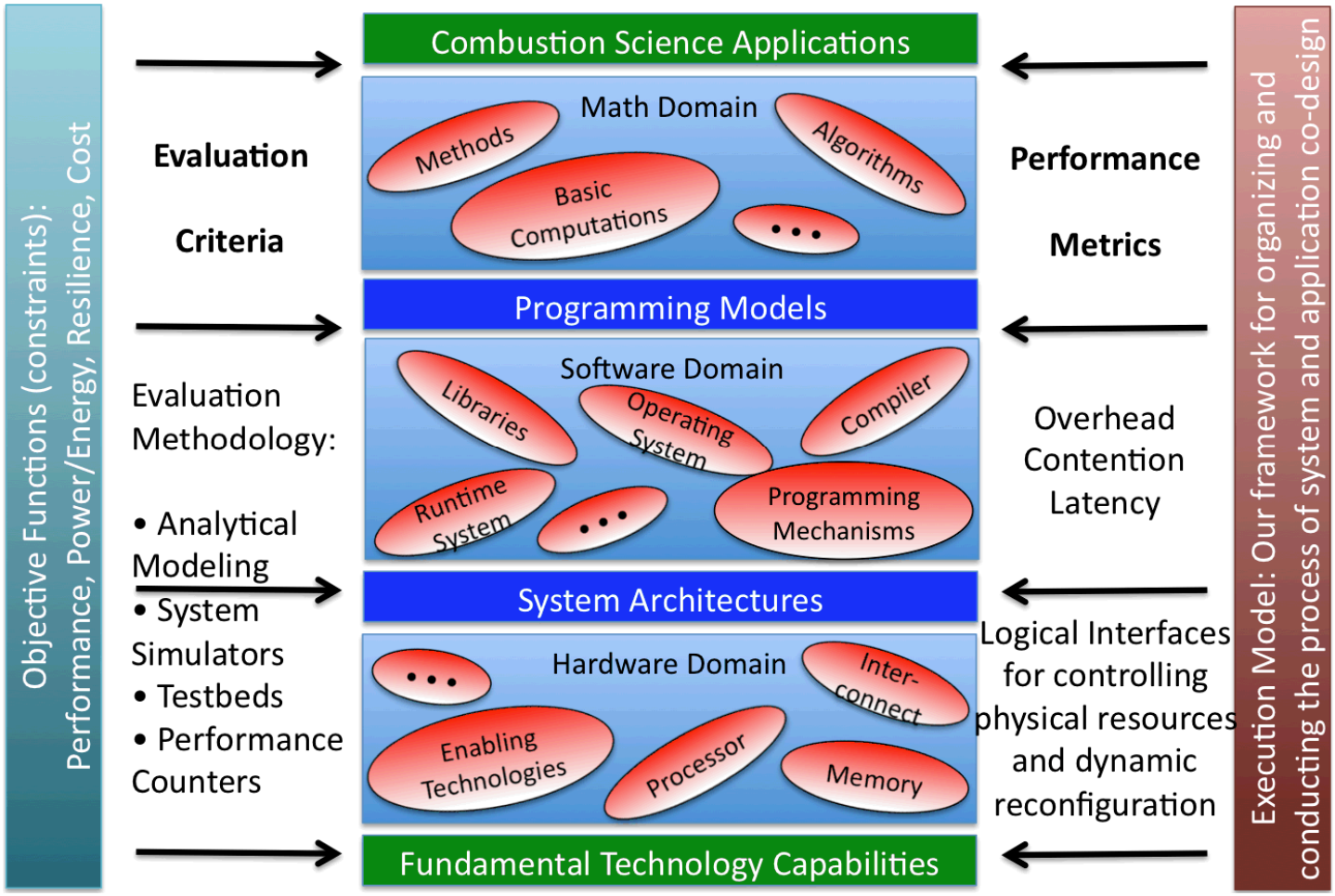


Sandia National Laboratories

Co-Design

Application
Methods
Algorithms
Basic Computations
Programming Model
Programming Mechanisms
Libraries
Compiler
Runtime System
Operating System
System Architecture
Node: Processors, Memory
Inter-node: Interconnect
Enabling Technologies
Fundamental Technology Capabilities

Co-Design


Objective Functions (constraints):
Performance, Power/Energy, Resilience, Cost
Evaluation
Criteria
Evaluation
Methodology:
• Analytical
Modeling
• System
Simulators
• Testbeds
• Performance
Counters
Combustion Science Applications
Math Domain
Methods
Algorithms
Basic
Computations
• • •
Programming Models
Software Domain
Libraries
Operating
System
Compiler
Runtime
System
• • •
Programming
Mechanisms
System Architectures
Hardware Domain
• • •
Inter-
connect
Enabling
Technologies
Processor
Memory
Fundamental Technology Capabilities
Performance
Metrics
Overhead
Contention
Latency
Logical Interfaces
for controlling
physical resources
and dynamic
reconfiguration
Execution Model: Our framework for organizing and
conducting the process of system and application co-design

DOE Exascale Co-design Centers

Exascale Co-Design Consortium (ECDC):

- **Exascale Co-Design Center for Materials in Extreme Environments**
- **Co-design for Exascale Research in Fusion(CERF).**
- **Chemistry Exascale Co-design Center (CECC)**
- **High Energy Density Physics**
- **Center for Exascale Simulation of Advanced Reactors**

Charon

- Analysis of electric potential and electron and hole concentrations in semiconductors.
- Solves stabilized weak form of drift diffusion equation

$$F_\psi = \int_\Omega R_\psi \, \phi \, d\Omega = 0,$$

$$F_n = \int_\Omega R_n \, \phi \, d\Omega \; - \; \sum_e \int_{\Omega_e} \tau_n \left[\mu_n \mathbf{E} \cdot \nabla \phi \right] R_n \, d\Omega = 0,$$

$$F_p = \int_\Omega R_p \, \phi \, d\Omega \; + \; \sum_e \int_{\Omega_e} \tau_p \left[\mu_p \mathbf{E} \cdot \nabla \phi \right] R_p \, d\Omega = 0,$$

where

$$\mathbf{E} = -\nabla \psi.$$

- Discretization yields sparse strongly coupled nonlinear system,
- Solved using Newton-Krylov
  - TFQMR or GMRes, various preconditioners

Case study: MiniFE

Solves the element diffusion matrix for the steady conduction equation[1]

$$(K_{12}^e)_{xy} = \int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1} k_{xy}\left(J_{11}^*\frac{\partial\psi_1}{\partial\xi} + J_{12}^*\frac{\partial\psi_1}{\partial\nu} + J_{13}^*\frac{\partial\psi_1}{\partial\zeta}\right) \cdot \left(J_{21}^*\frac{\partial\psi_2}{\partial\xi} + J_{22}^*\frac{\partial\psi_2}{\partial\nu} + J_{23}^*\frac{\partial\psi_1}{\partial\zeta}\right)|J|d\xi d\nu d\zeta$$

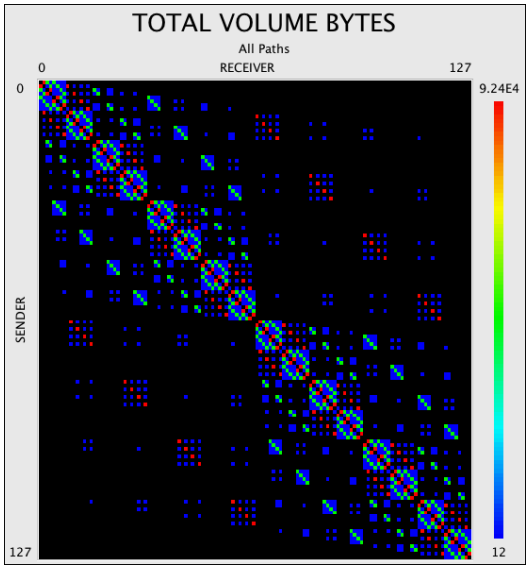
$$\int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1} F(\xi,\nu,\zeta)d\xi d\nu d\zeta \approx \Sigma_{I=1}^{M}\Sigma_{J=1}^{N}\Sigma_{K=1}^{P} F(\xi_I,\nu_J,\zeta_K)W_I,W_J,W_K$$

[1] *"The Finite Element Method in Heat Transfer and Fluid Dynamics, 2nd Edition", Reddy and Gartling, CRC Press, 2001.*

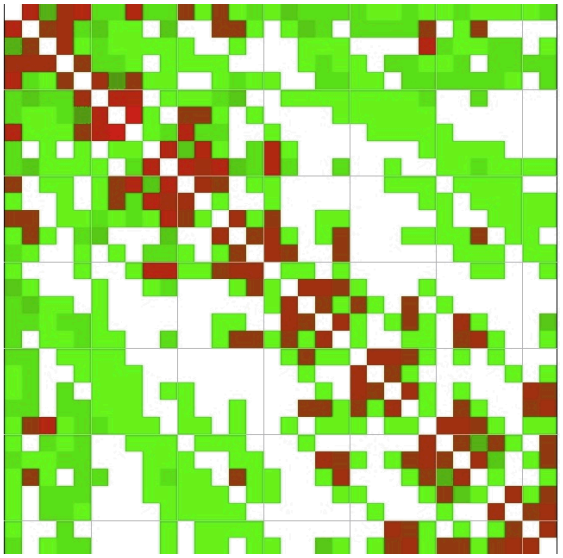
Sandia National Laboratories

Charon	miniFE
Unstructured	Structured *(but not exploited)*
Implicit	Implicit
Newton-Krylov TFQMR/GMRes; ML precon	CG, no precond
mpi	mpi, omp, p/qthreads, cuda, etc
SLOC	5,836 SLOC (2,159 exec)

Communication Patterns



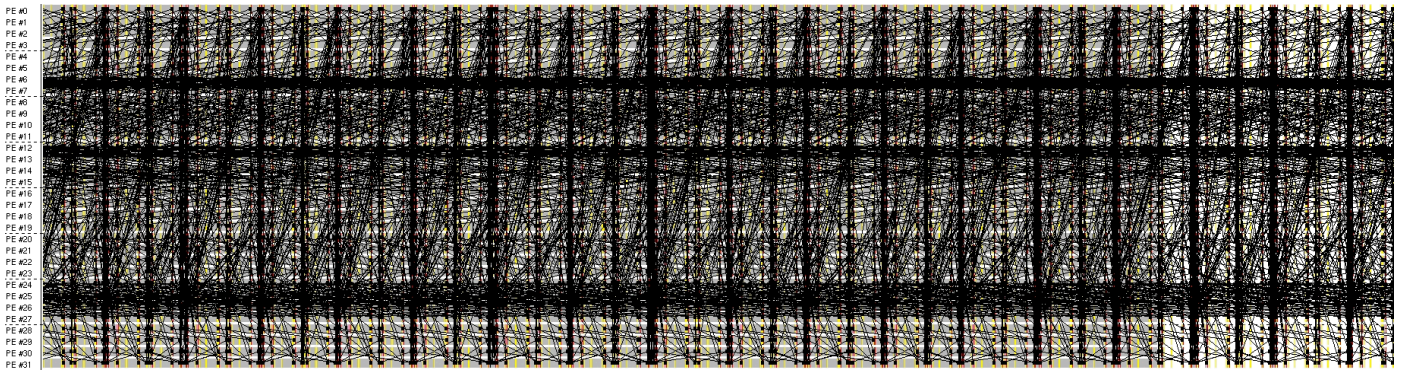
miniFE (tau)



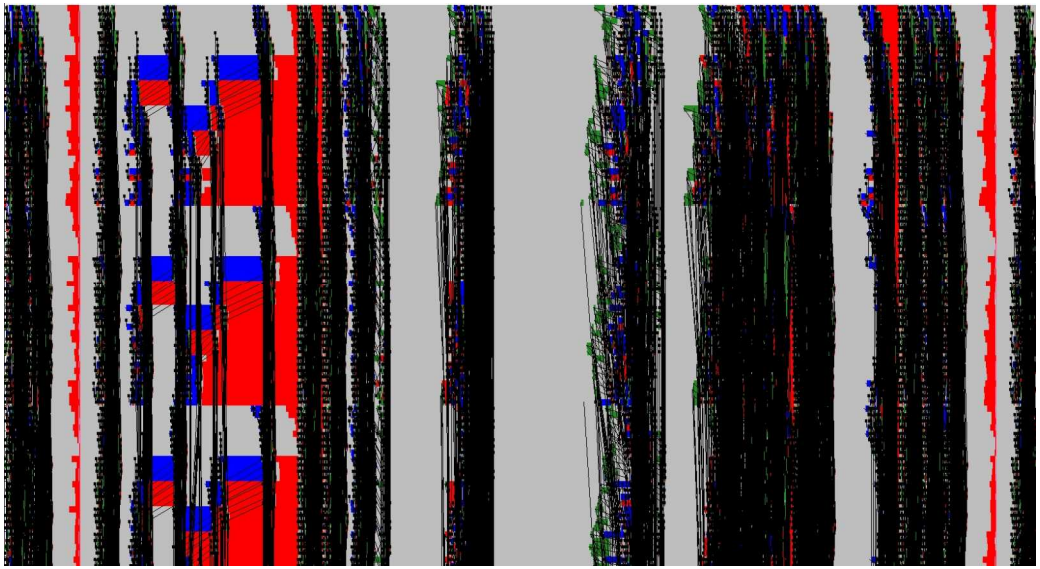
Charon (CrayPAT)

Runtime profiles

Charon



CTH



Sandia National Laboratories

Does MiniFE Predict Charon Behavior?
Processor Ranking: 8 MPI tasks; 31k DOF/core

- **Charon steady-state drift-diffusion BJT; GMRES linear solve; ML precond**
- **Nehalem: Intel 11.0.081 –O2 –xsse4.2; dual-socket QC)**
- **12-core Magny-Cours (Intel 11.0.081 –O2; one socket, 4 MPI tasks/die)**
- **Barcelona (Intel 11.1.064 –O2; two sockets out of the quad)**

MiniFE		
1	CG	FE assem+BC
2	Nehalem	Nehalem
3	MC(1.7)	MC(1.7)
	Barc(2.7)	Barc(1.8)

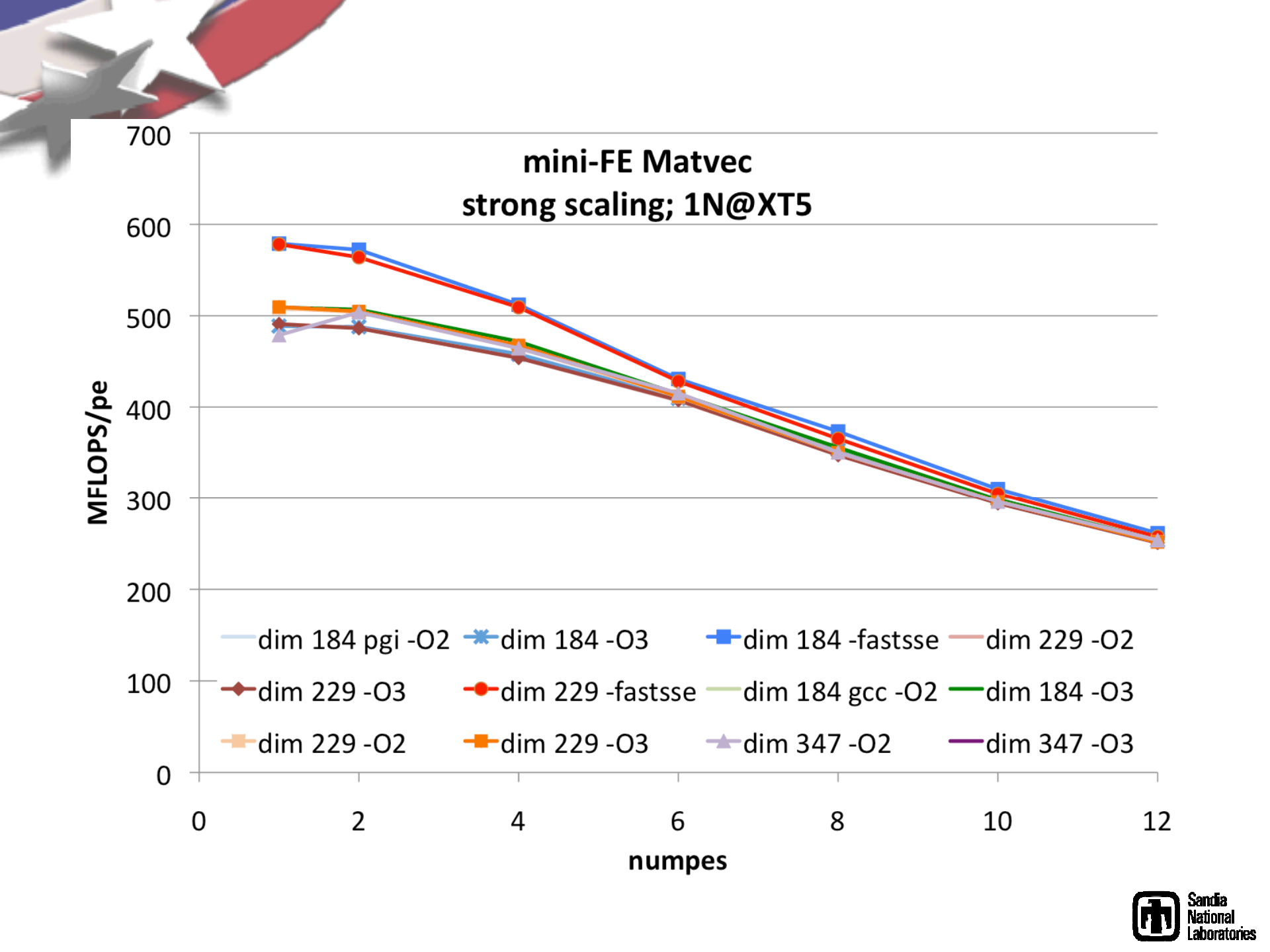
Charon			
	LS w/o ps	LS w/ ps	Mat+RHS
1	Nehalem	Nehalem	Nehalem
2	MC(1.7)	MC(1.8)	MC(1.46)
3	Barc(2.8)	Barc(2.5)	Barc(1.52)

Number in parenthesis is factor greater than #1 time

Terminal — ssh — 103×61

```
Mini-Application Name: miniFE
Mini-Application Version: 0.7
Global Run Parameters:
  dimensions:
    nx: 605
    ny: 605
    nz: 605
  load_imbalance: 0
  mv_overlap_comm_comp: 0 (no)
  number of processors: 128
  ScalarType: double
  GlobalOrdinalType: int
  LocalOrdinalType: int
Platform:
  hostname: mzlogin01e
  kernel name: Linux
  kernel release: 2.6.27.45-0.1-default
  processor: x86_64
Build:
  CXX: /home/ssshend/projects/tau_latest/craycnl/bin/tau_cxx.sh
  compiler version: /opt/cray/xt-asyncpe/4.7/bin/CC: INFO: Compiling for compute nodes running CLE.
  CXXFLAGS: -O3
  using MPI: yes
  Threading: none
Run Date/Time: 2011-02-03, 16:48:53
Rows-per-proc Load Imbalance:
  Largest (from avg, %): 1.49499
  Std Dev (%): 0.760431
Matrix structure generation:
  Time: 9.57472
FE assembly:
  Time: 17.5709
Matrix attributes:
  Global Nrows: 222545016
  Global NNZ: 5988906496
  Global Memory (GB): 68.5893
  Pll Memory Overhead (MB): 110.714
  Rows per proc MIN: 1710075
  Rows per proc MAX: 1755904
  Rows per proc AVG: 1.73863e+06
  NNZ per proc MIN: 46172025
  NNZ per proc MAX: 47201700
  NNZ per proc AVG: 4.67883e+07
CG solve:
  Iterations: 50
  Final Resid Norm: 3.03155e-14
  WAXPY Time: 1.85761
  WAXPY Flops: 1.00979e+11
  WAXPY Mflops: 54359.6
  DOT Time: 1.12053
  DOT Flops: 4.4289e+10
  DOT Mflops: 39525.1
  MATVEC Time: 11.8695
  MATVEC Flops: 6.10868e+11
  MATVEC Mflops: 51465.2
  Total:
    Total CG Time: 14.9933
    Total CG Flops: 7.56136e+11
    Total CG Mflops: 50431.5
  Time per iteration: 0.299866
"miniFE.605x605x605.P128.PGI.2011:02:03-16:49:40.yaml" 61L, 1548C                61,3          Bot
```

Sandia National Laboratories



Summary

- Architectures in flux (but converging?)
- Programming mechanisms in flux
- Revolutionary code re-write a huge undertaking
    - TPL dependencies same
- Not a computer science exercise

  (but publications are to be had)
- Science and engineering trust must be maintained throughout

Acknowledgements

- Sandia CSRF
- NNSA ASC CSSE

Sandia National Laboratories

Thanks

Sandia National Laboratories