SAND2006-7241C

Carpe computem: The transformation of engineering and science through petascale computing

SAND2006-6158C

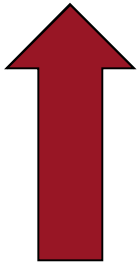
Rob Leland
Computing and Network Services
Sandia National Laboratories

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. SAND -2005-6648P

Sandia National Laboratories

“There are but four phases in any campaign: survey, plan, marshal, execute.”

-- Cicero



Part 1: The focus today

- **Scientific Transformation:** the emergence of predictive simulation as a third scientific approach on par with theory and experiment

- **Engineering Transformation:** the change from a primarily test-based approach in engineering to a primarily simulation-based approach

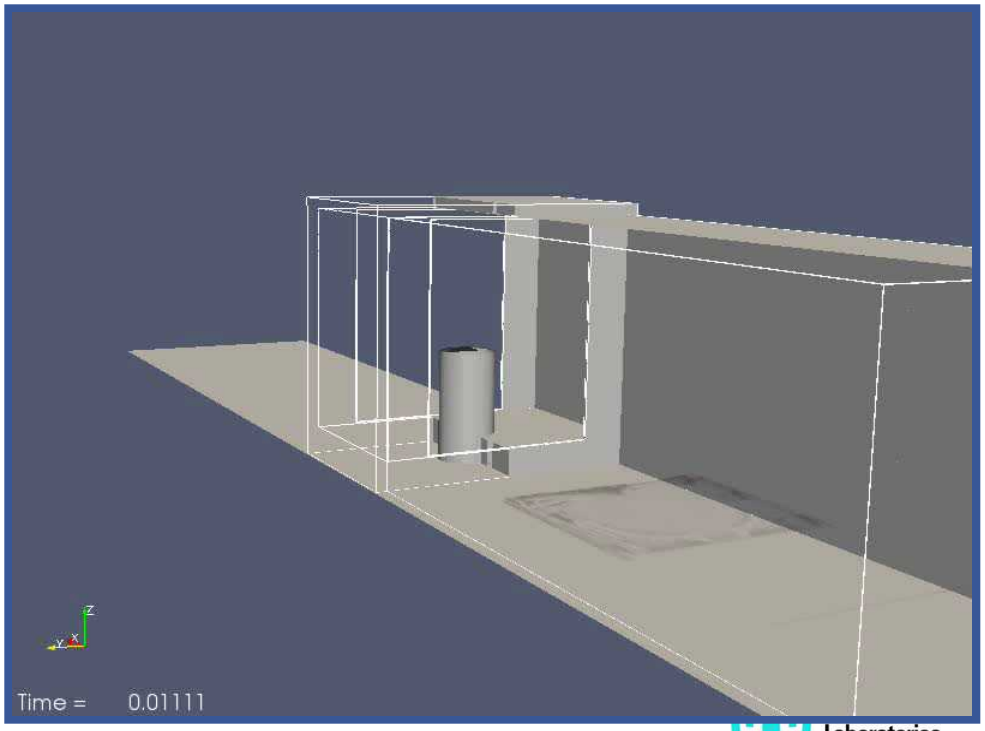
- **National drivers (one perspective)**
  A. Loss of nuclear testing capability
  B. Things we've never been able to fully test
  C. Emerging security challenges post 9/11
  D. Understanding high consequence events
  E. Sustaining economic competitiveness
  F. Accelerating scientific discovery

A. Loss of underground testing

B. Things we've never been able to fully test

Abnormal thermal environments:

- **Need to simulate**
  - **Thermal loads on objects**
  - **Heat transfer within objects**
- **Can't test full system**
- **Many possible scenarios**



Laboratories

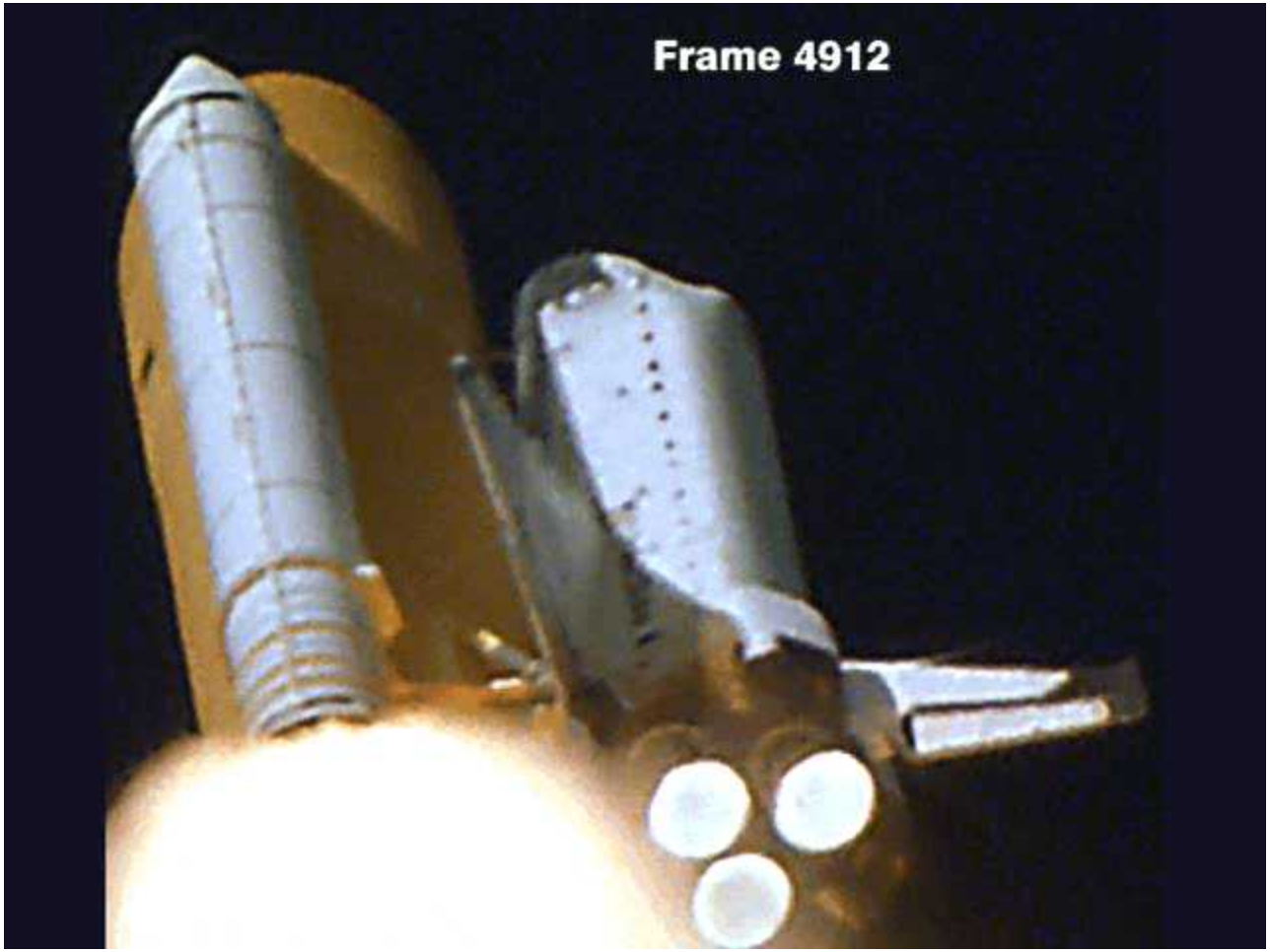
C. Emerging security challenges post 9/11

"… after operations of three years or perhaps less, the Sandia Pulsed Reactor will no longer be needed, since computer simulations will be able to assume its mission."

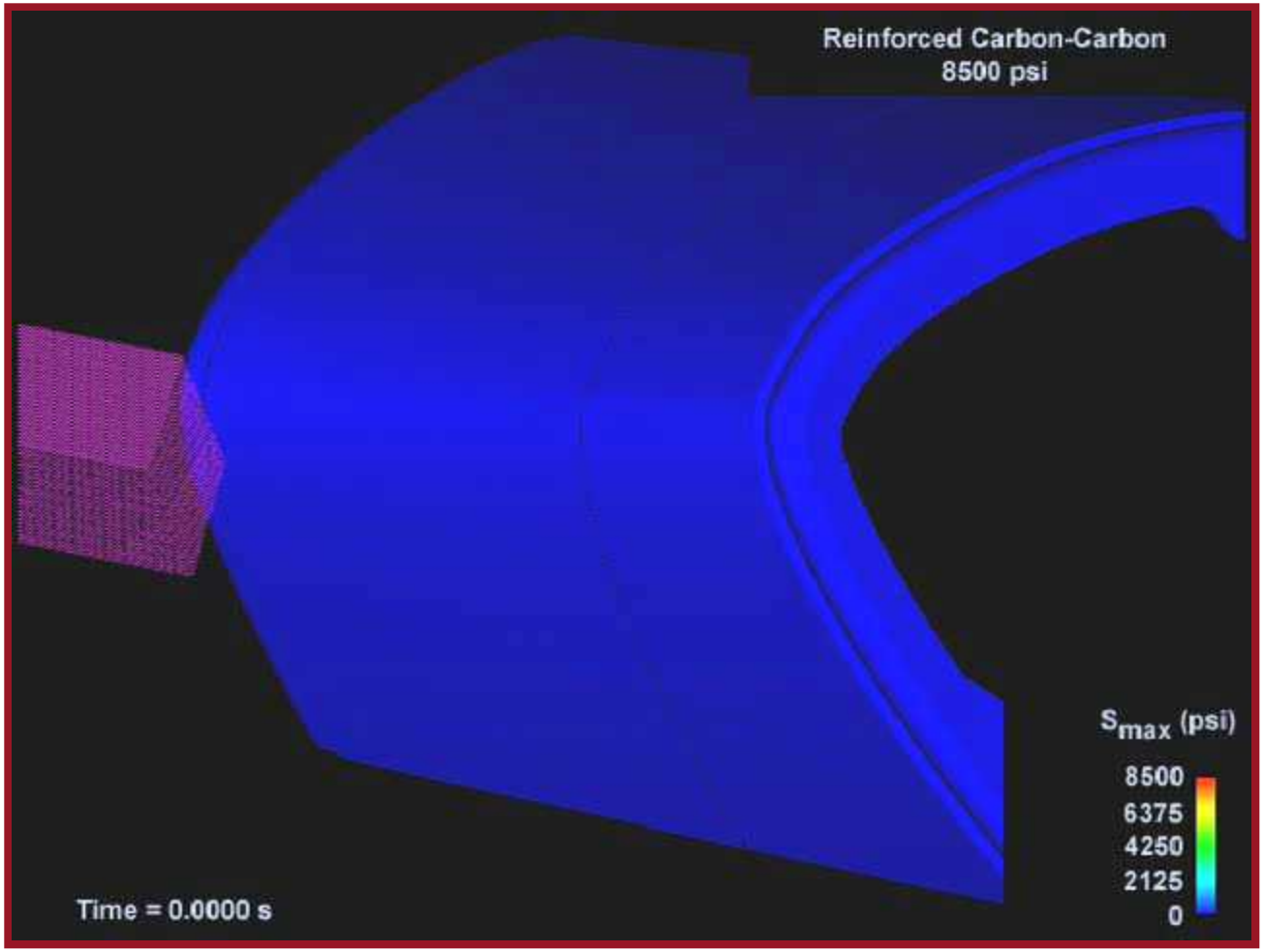
- Secretary of Energy Abraham, 2004

- **SPR historically critical to certification for hostile environments**
- **Provides unique combination of high fluence, short pulse width, mixed radiation, and high volume**
- **Being removed from service due to security concerns post 9/11**
- **Replacing SPR implies petascale computation**

D. Understanding high consequence events



Simulation of foam hitting shuttle wing

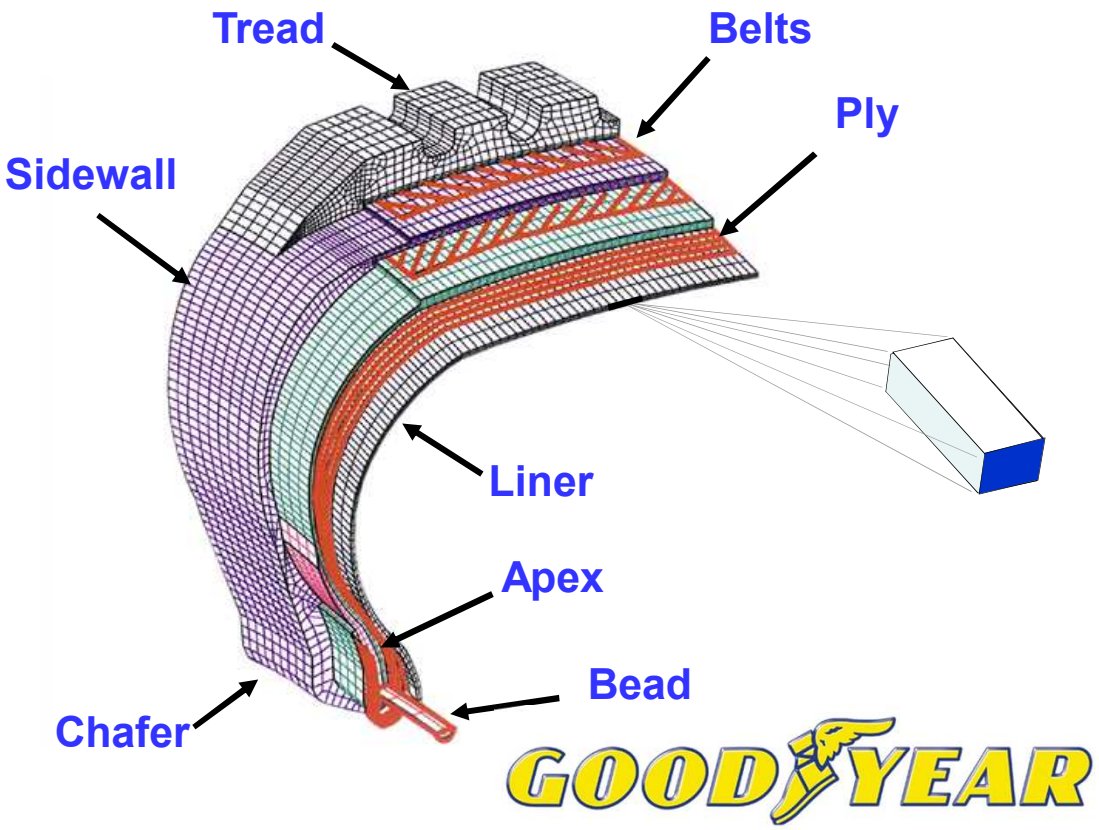


Sandia National Laboratories

Physical validation test

E. Sustaining economic competitiveness

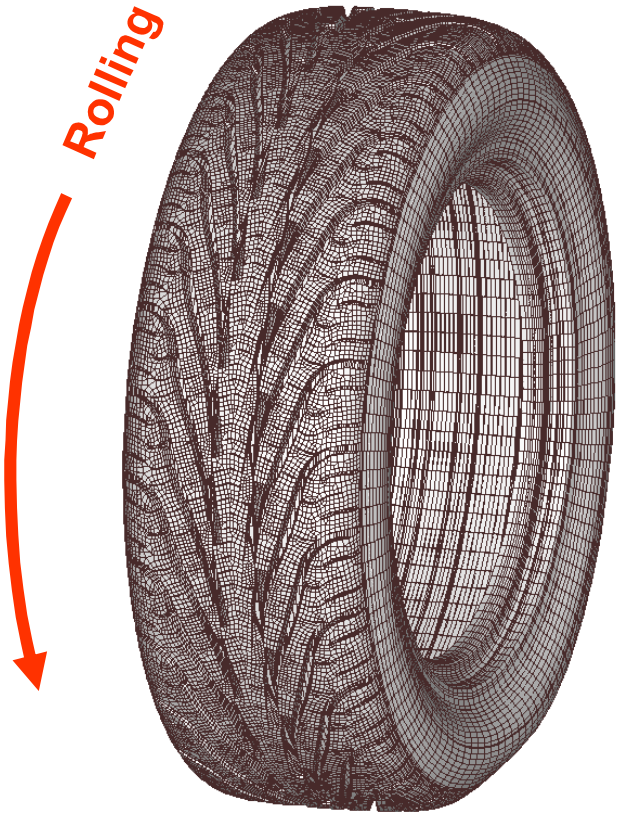
Finite Element Analysis (FEA) Computer Modeling of Tires



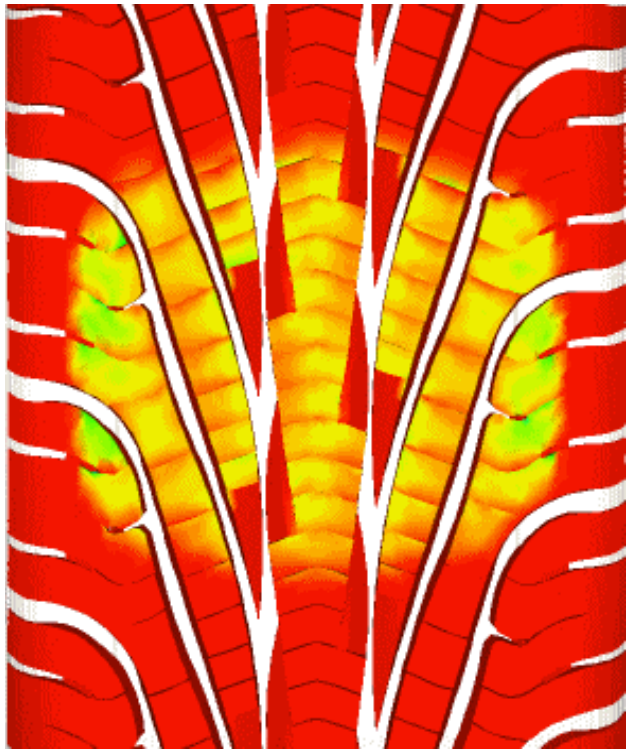
- A tire is a composite structure of at least 8 layers of different materials
- Each layer can be approximated with a large number of “finite” elements
- Most of the elements have non-linear material properties

Sandia National Laboratories

Simulation based product design



GOODYEAR



Wear simulation using Sandia derived tools

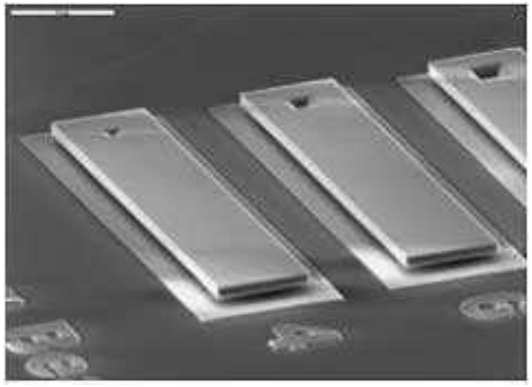
Sandia National Laboratories

How much deeper can we go?

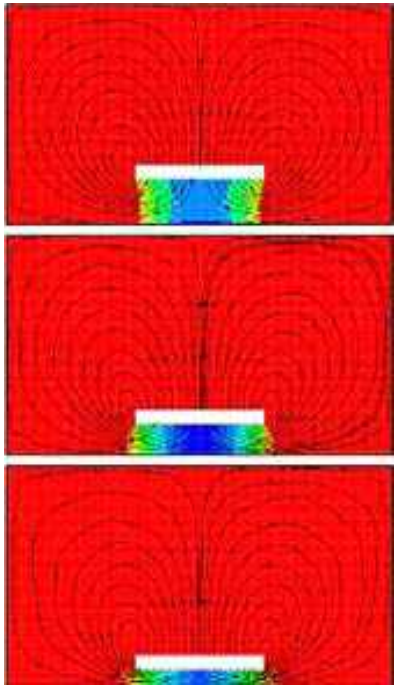
Sandia National Laboratories

A very “simple” MEMS example

- **Polycrystalline silicon beams**
  - Used in accelerometers, switches etc.
  - Specs: 100 by 20 by 2 microns; 10-100 kHz
  - “diving boards for bacteria”
  - Gas dynamics critical to performance
  - Material properties not homogenous
  - Physical intuition often wrong



Problem
- 3-d geometry
- Moving boundary
- Transient gas flow
- Full oscillation

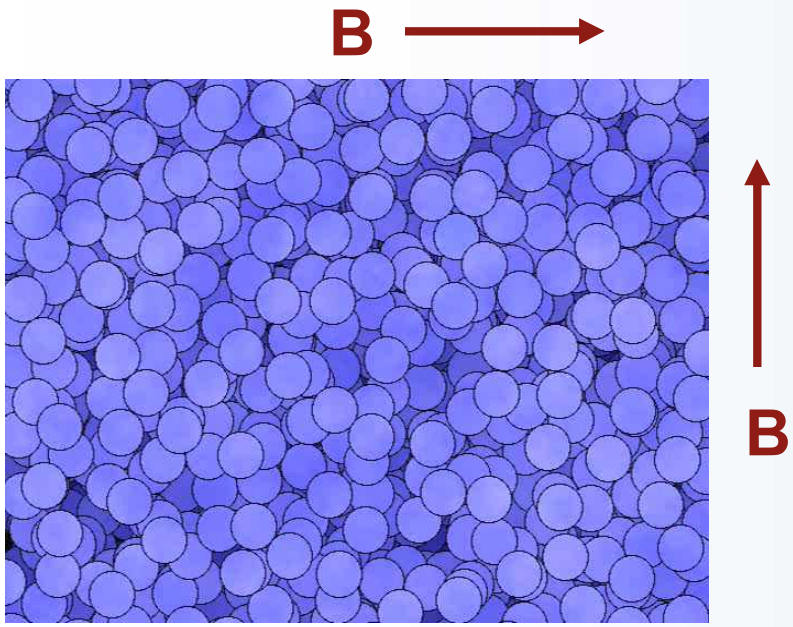


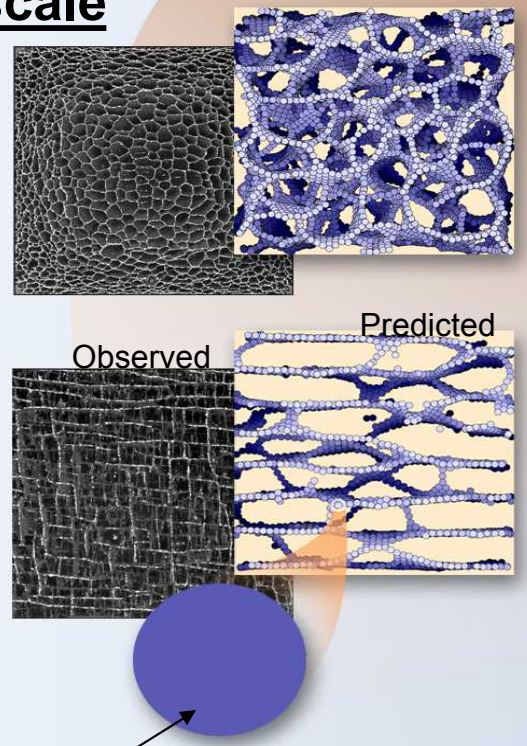
Current simulation
- 2-d geometry
- Static boundary
- Steady-state gas flow
- Distinct points in oscillation

F. Accelerating scientific discovery

Re-engineering ENGINEERING

Engineering at the nanoscale

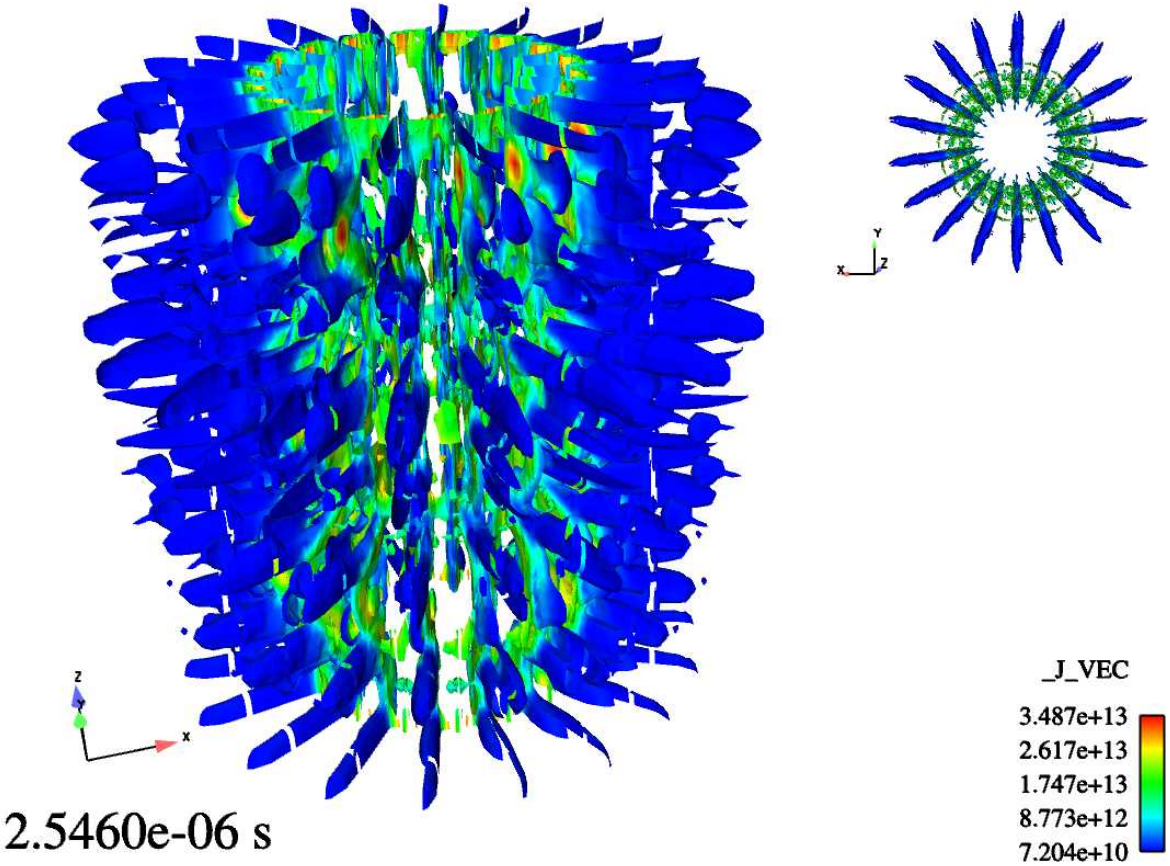




Sandia National Laboratories

LOCKHEED MARTIN

High energy density physics

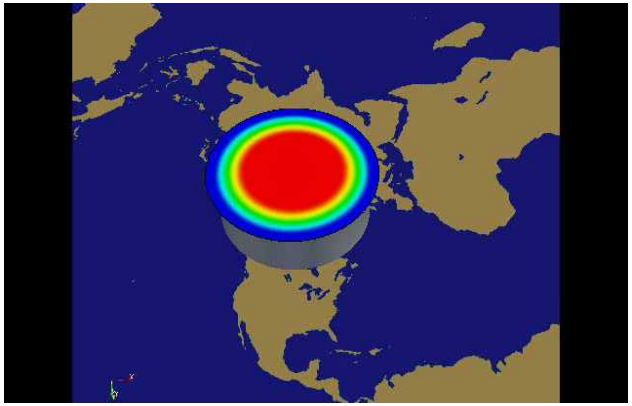


Dynamic hohlraum pinch stagnation on central foam showing 18-fold azimuthal structure

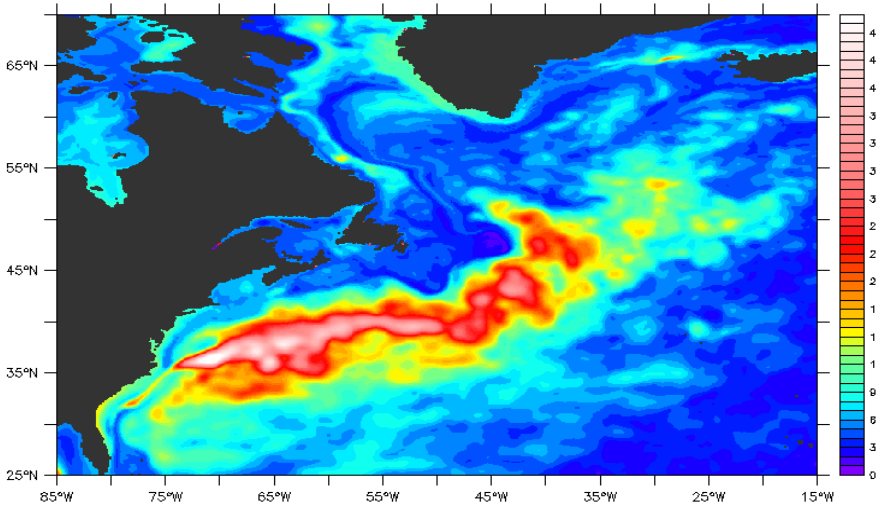
Sandia National Laboratories

High impact environmental science

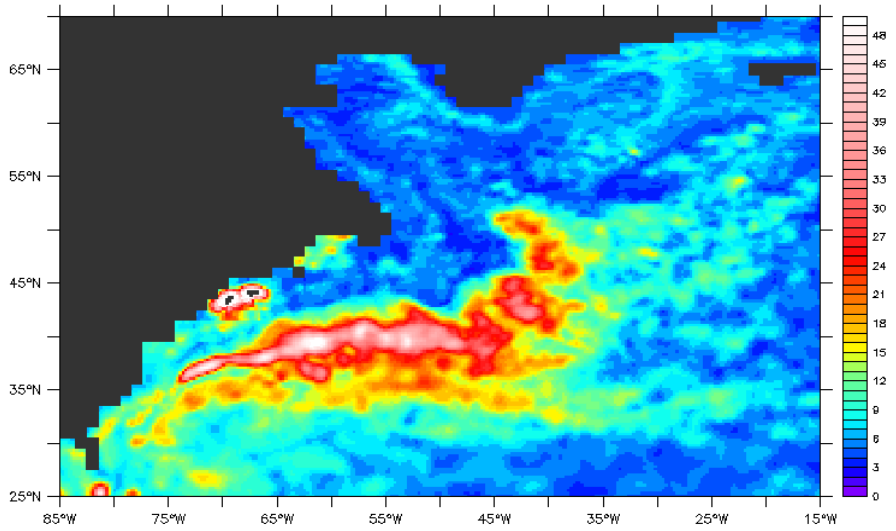
- **Polar vortex breakdown modeling**
- **Ocean simulation at 10km resolution**



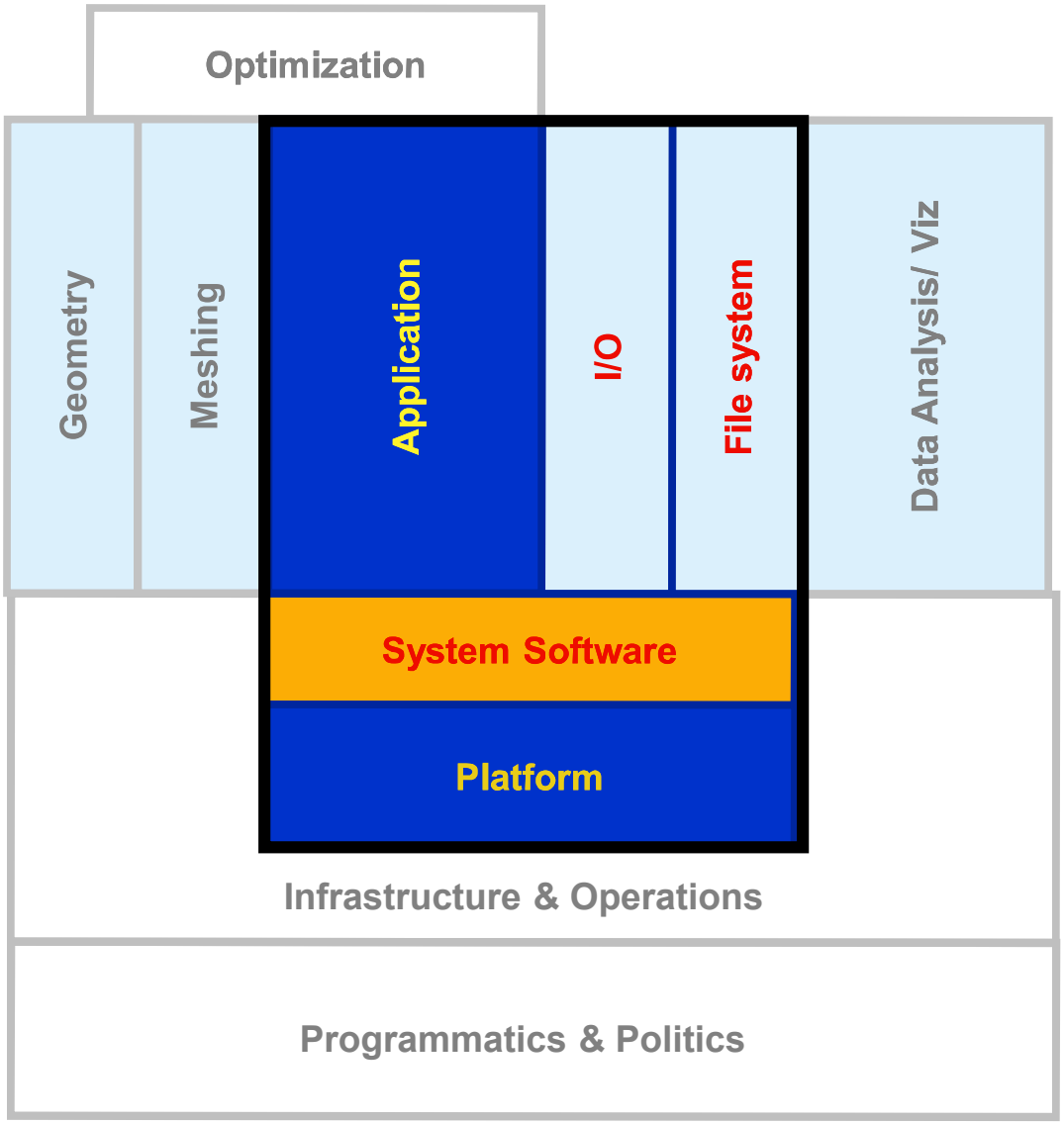
High Resolution Simulation



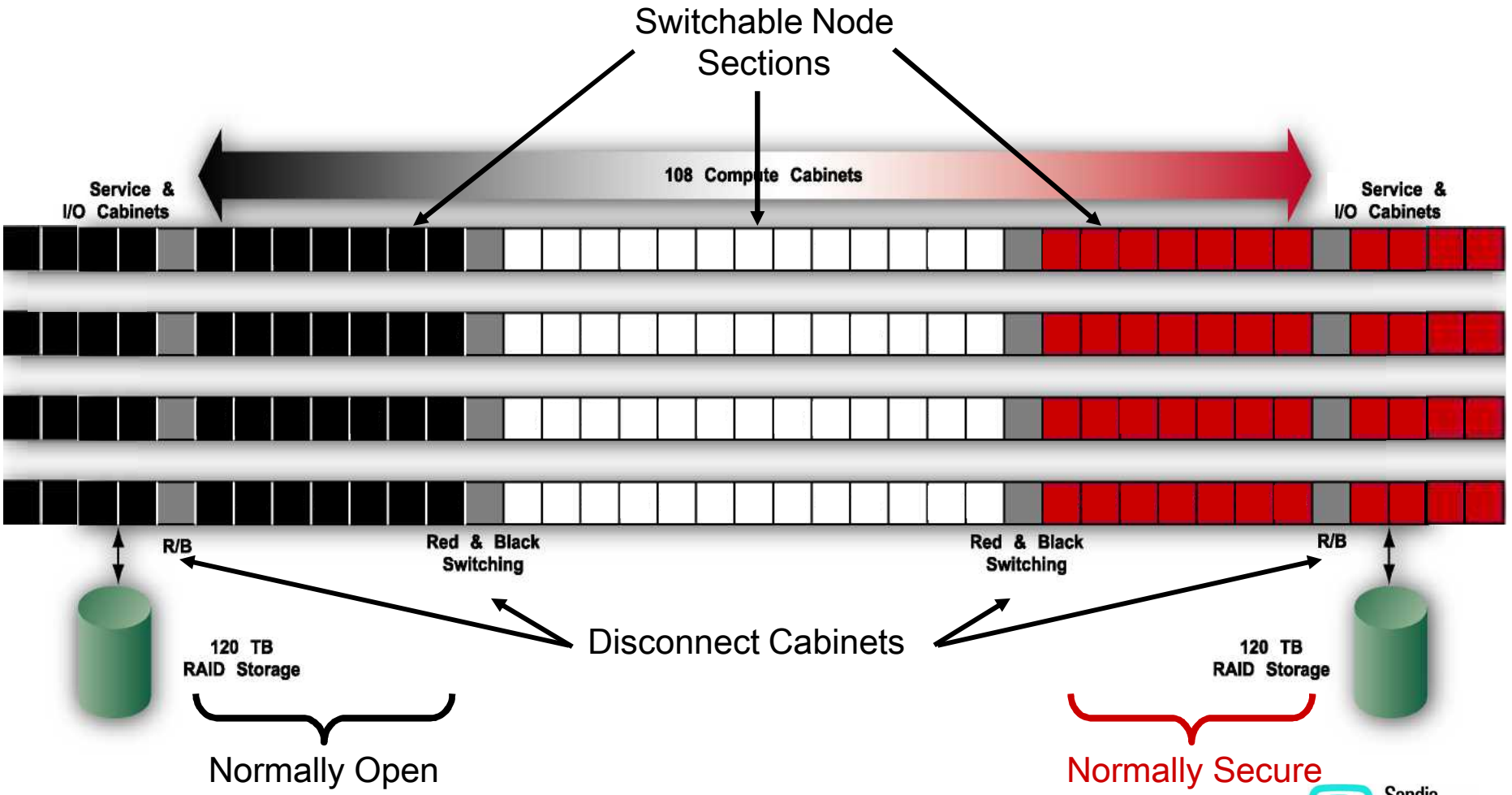
Observations



Part 2: The differentiating toolset

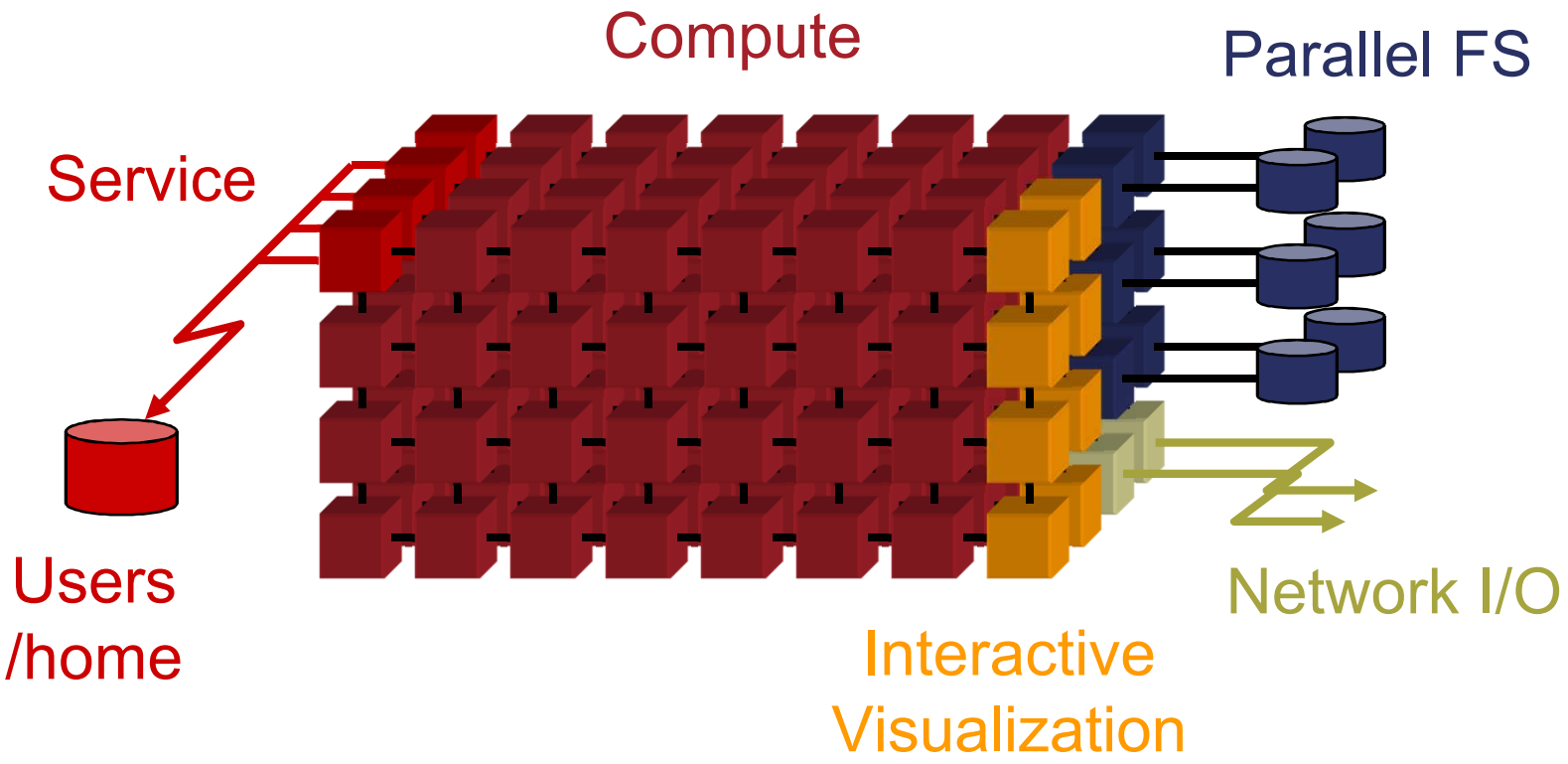


Red Storm: a state of the art capability system



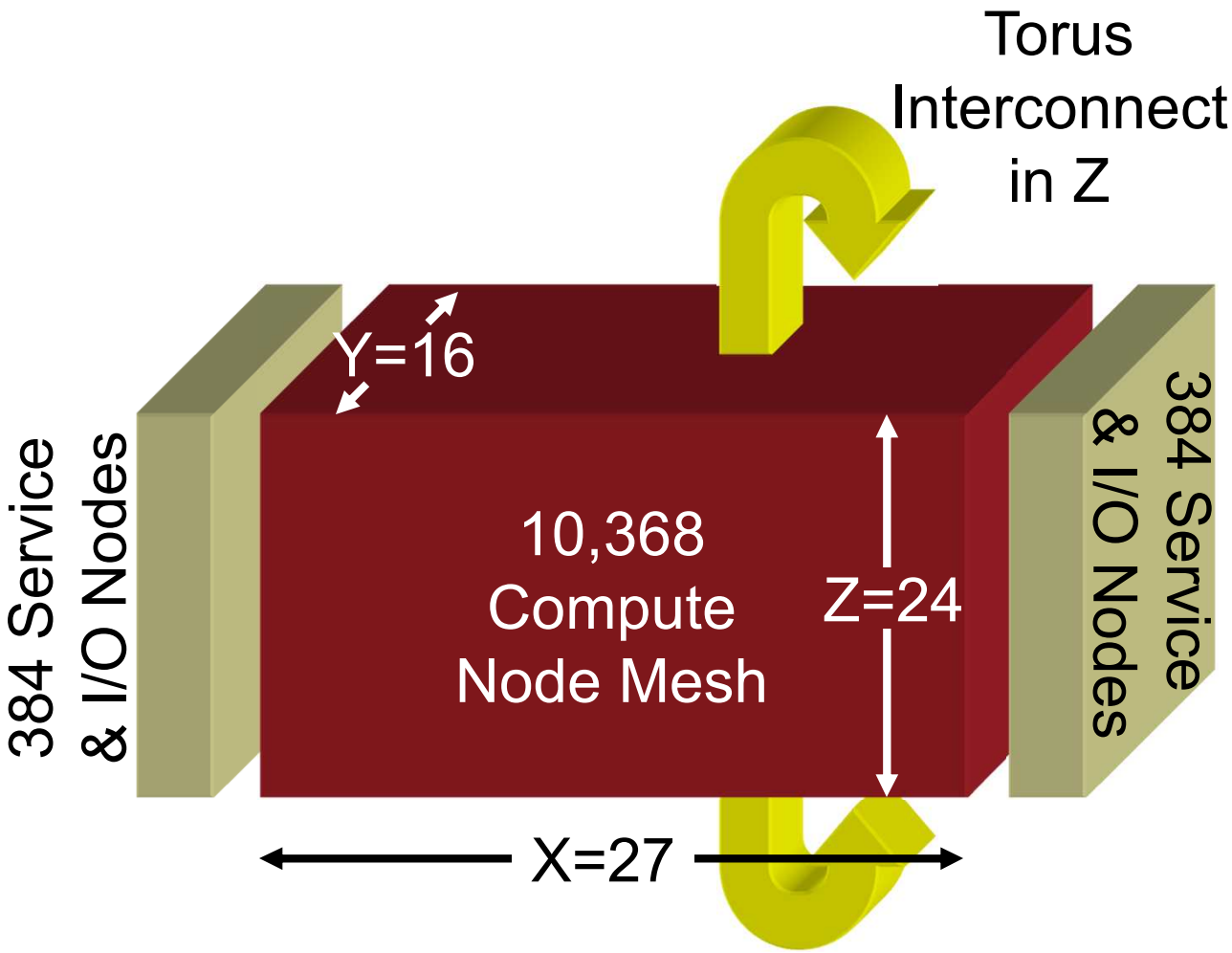
Sandia National Laboratories

Functional architecture



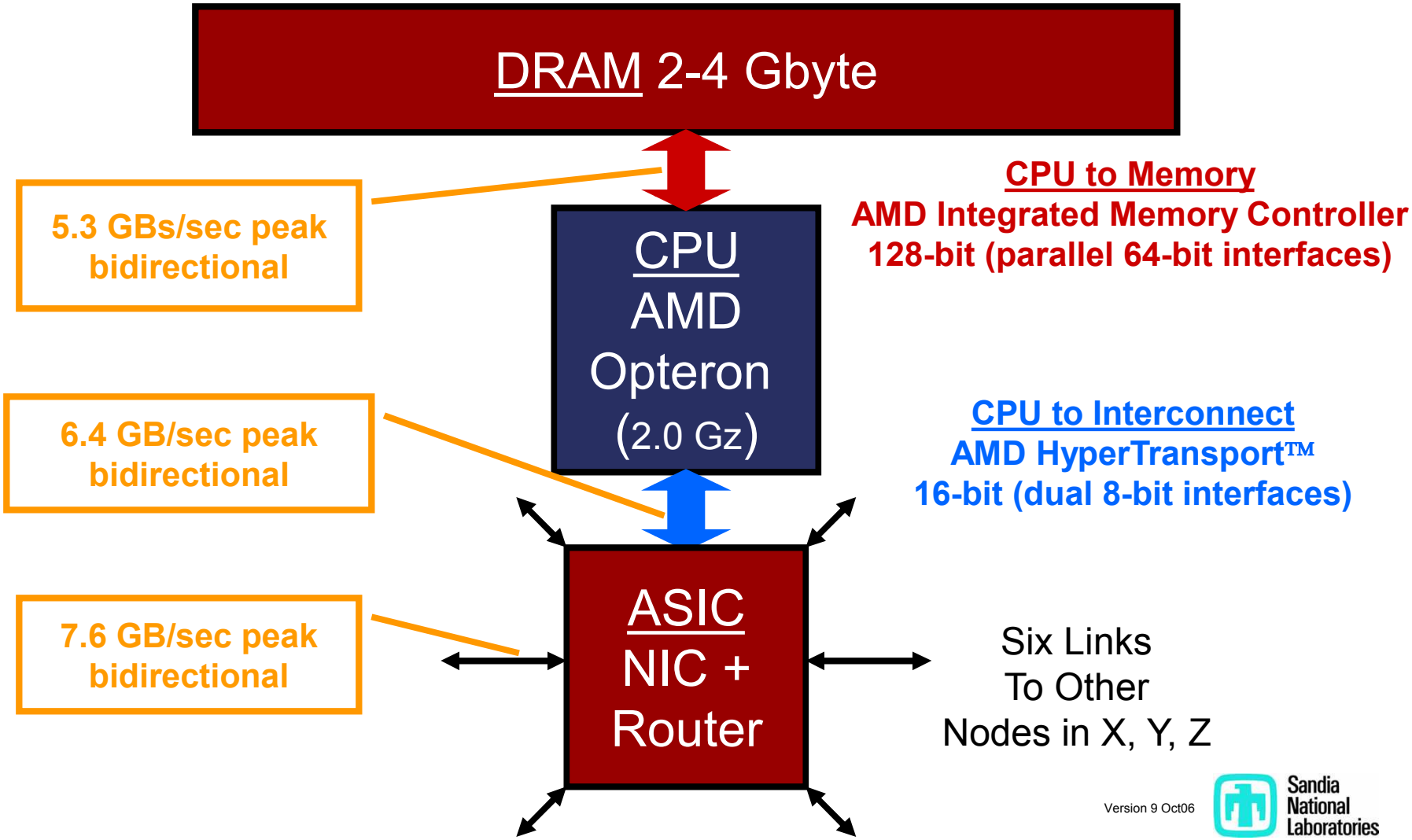
Sandia National Laboratories

Topological architecture



Sandia National Laboratories

Compute node architecture



Sandia National Laboratories

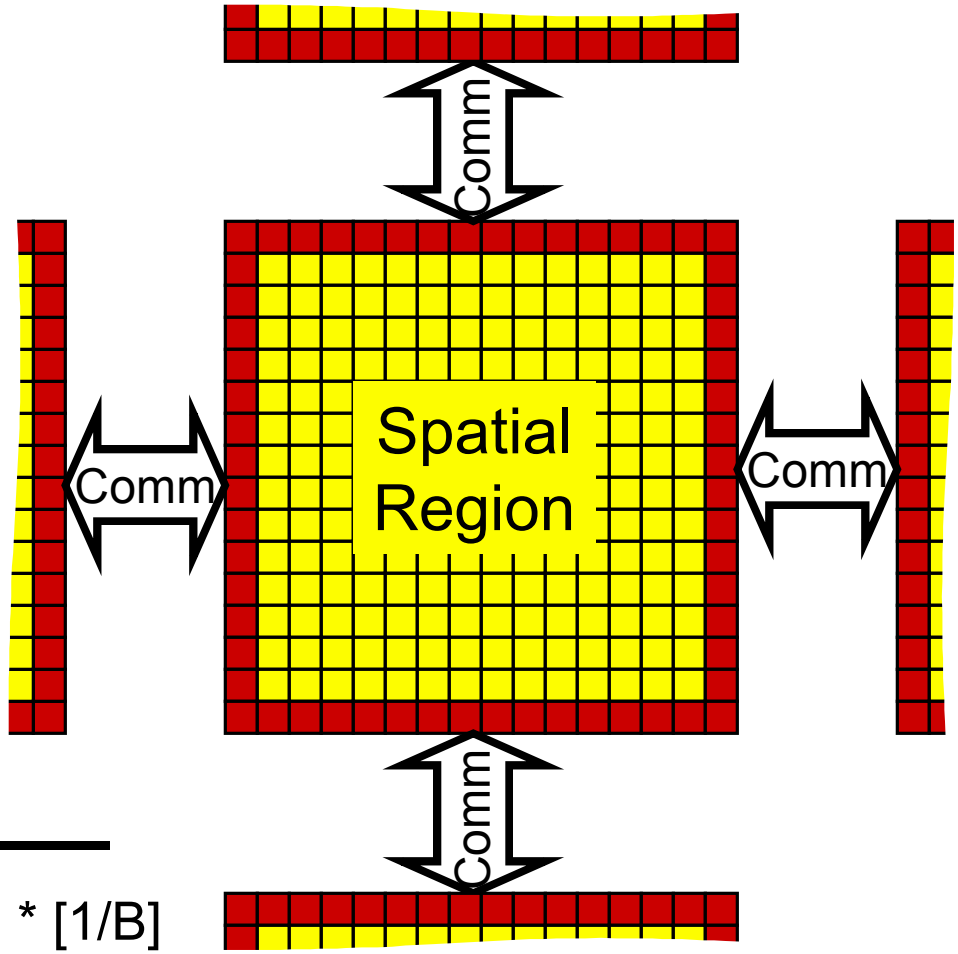
Design philosophy

- Scalability
- Usability
- Reliability
- Economy

} “SURE” architectural principles

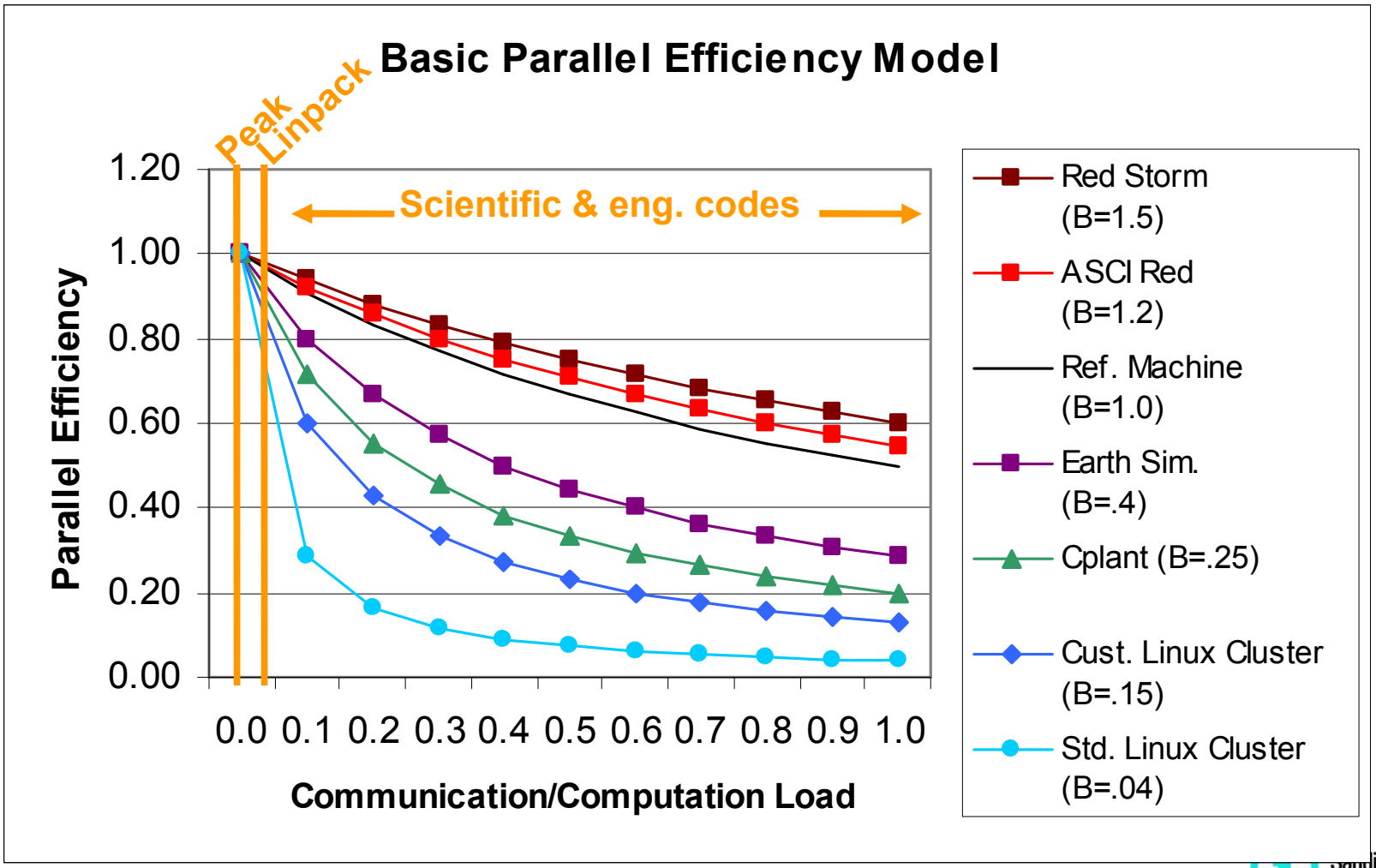
Scalability: a simple model

- E.g. nearest neighbor 2D relaxation
- Interior values updated by computation each iteration
- Boundary values must be communicated each iteration
- Overhead is proportional to ratio of boundary to interior points
- Overhead is inversely proportional to communication/computation rate
- Hence efficiency goes like …



$$\frac{1}{1 + \text{overhead}} = \frac{1}{1 + [1/\text{sqrt}(N)] * [1/B]}$$

System balance determines scalability

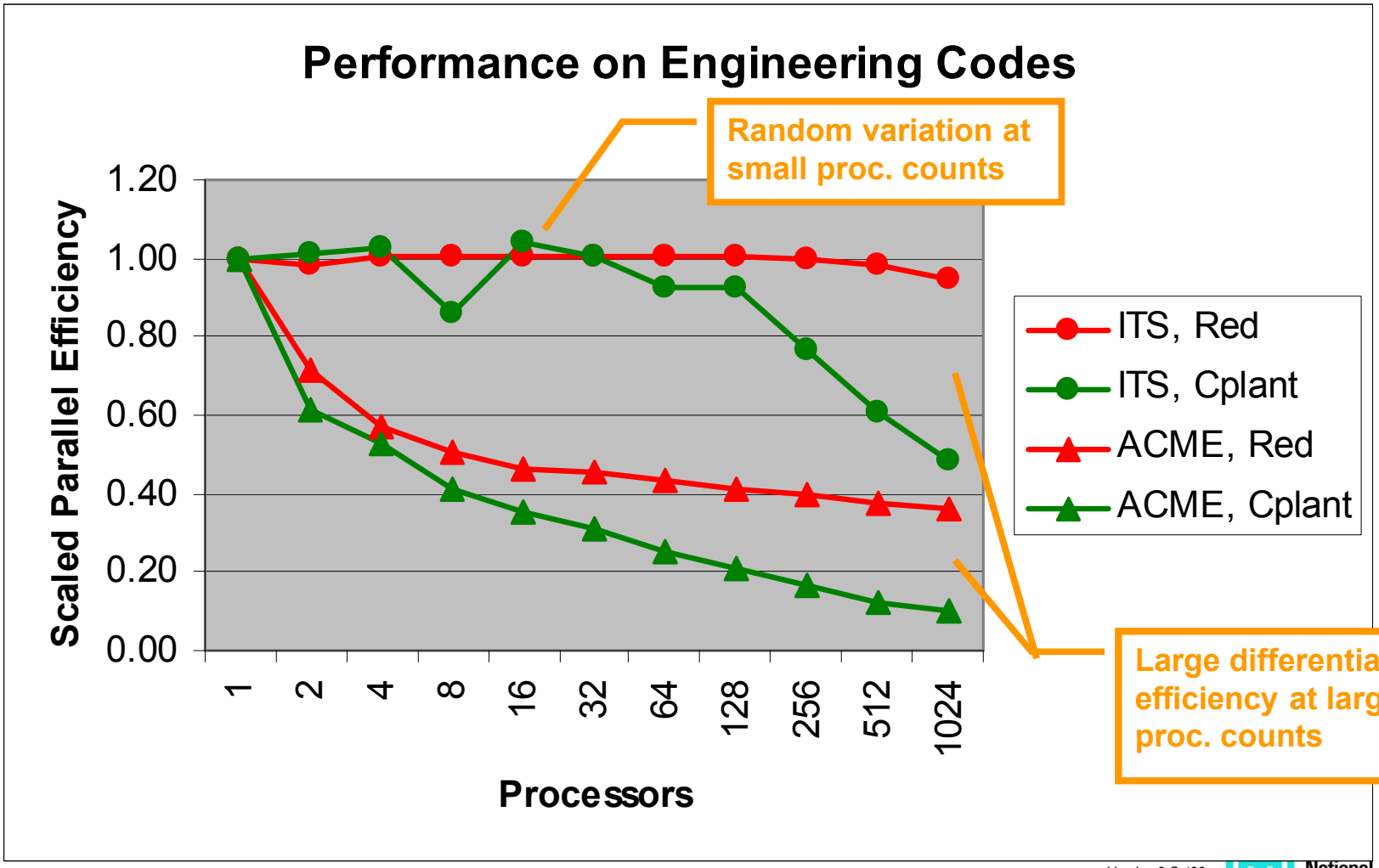


Scalability per HPCC benchmarks

System	HPL	PTRANS	RANDOM	FFT	STREAM
Red Storm (SNL)	41	100	100	51	79
Blue Gene Light (LLNL)	100	19	6.5	100	97
Purple (LLNL)	72	31	17	39	100

… percentage of best achieved so far on baseline benchmarks

Scalability of some key engineering codes



National Laboratories

Other key scalability issues

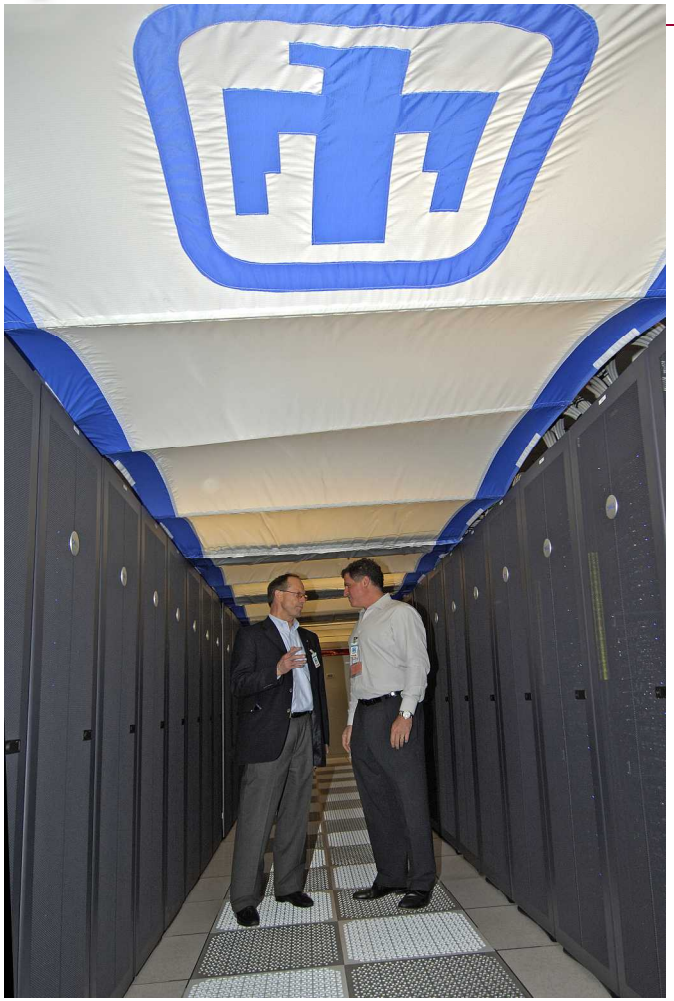
- Light weight kernal
- Scalable system services
- Light weight file system
- Physical scalability
- Scalable enabling toolset

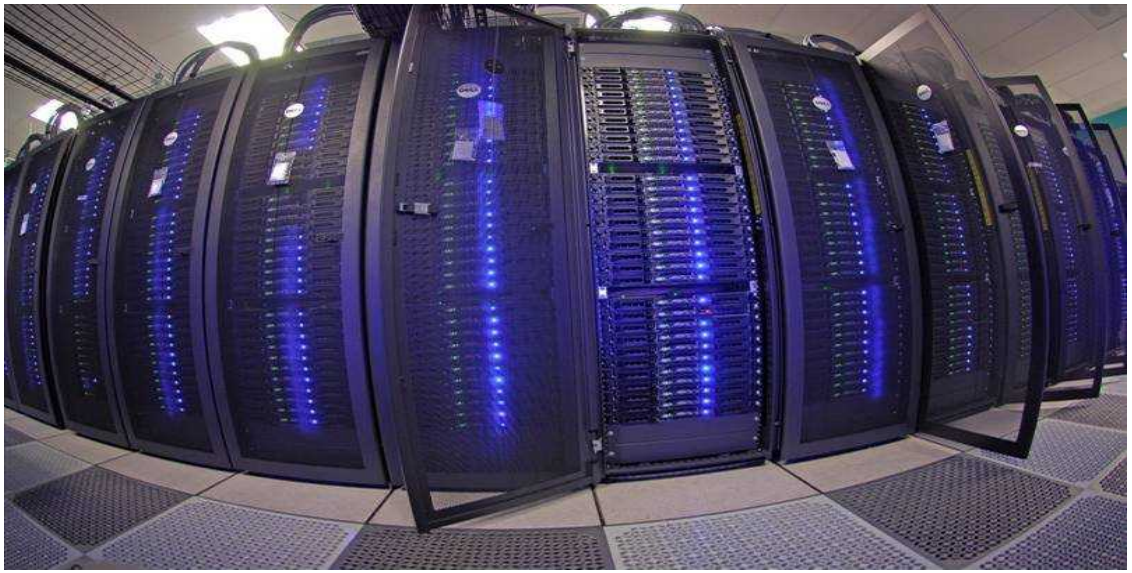
- … Scalability must permeate every aspect of the design

Usability, reliability and economy

- Usability
  - Linux front end
  - MPI programming model
  - Scalable system response
- Reliability
  - NOT high availability
  - 100 hrs. MTBI for hardware & software … 500T working parts
  - 50 hrs. MTBI for apps … 10^19 calculations w/o failure
  - Independent RAS “immune” systems
  - Highly redundant power supplies & memories
- Economy
  - Leverage commodity markets for components
  - Design for lifecycle costs
  - Encourage a wider market for system

Thunderbird: state of the art capacity is essential





Sandia National Laboratories

Part 3: Looking forward five years

- "Moonshot" goals
  - Predictive simulation across full system lifecycle
  - Component design cycle times reduced from 7 to 2 years
  - Rapid design of custom microsystems without prototyping
  - Completely virtual test environments and facilities
  - Scientific breakthroughs, e.g.
    - target design for break-even fusion on Z
    - 10km resolution climate simulation
    - virtual cell modeling
- All require (at least) petascale computation

A fearless forecast

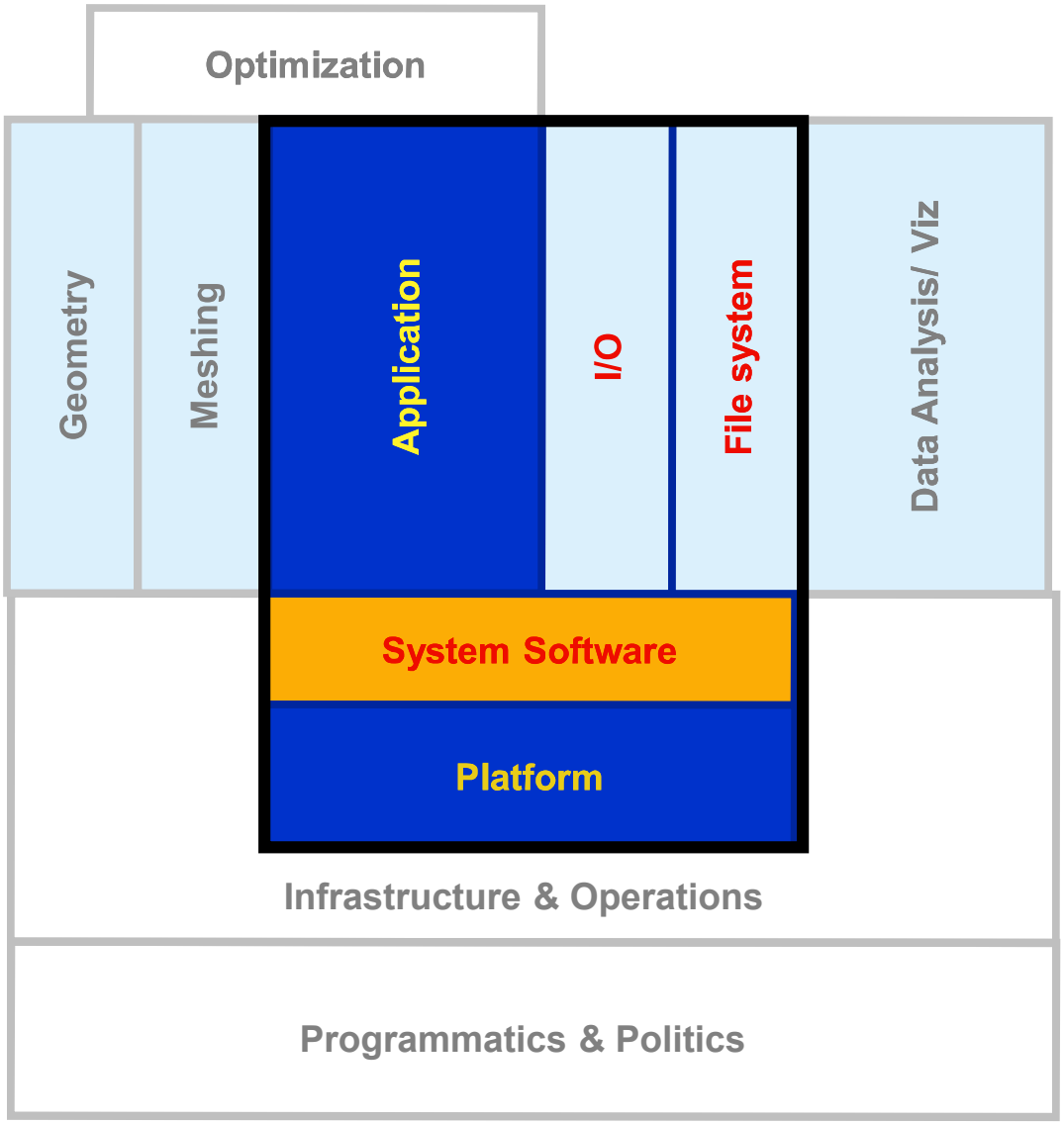
By 2011:

- **Topology: mesh**
- **Processors: 10k**
- **Cores: 1/3 million**
- **Peak: 10 TF**
- **Links: 10 GB/s**
- **Memory: 1/3 PB**

Part 4: Key challenges

- **Not comprehensive**
- **Under appreciated (my view)**
- **Biased by my experience**
- **Some interesting highlights**

Application and system challenges



System challenges

- Scalability of computer hardware and system software
  - Maintaining balance in the architecture
  - Coping with the drive to multi-core
  - Power consumption
  - Optical interconnect crossover point
  - Resiliency despite growing system complexity
  - Quantitative understanding of latency impact
- Processor design
  - Much higher fraction of chip area devoted to processing
  - Memory wall
- File system performance
  - Viable LWFS

Applications challenges

- Maintaining parallel efficiency with a million threads
  - New programming models
    - Fractal approach?
    - Implicit within socket, explicit without?
- Compelling non technical applications with similar needs
- Developing (accepting) better benchmarks
- Embracing V&V
  - Tools, methodology, computing power and incentives

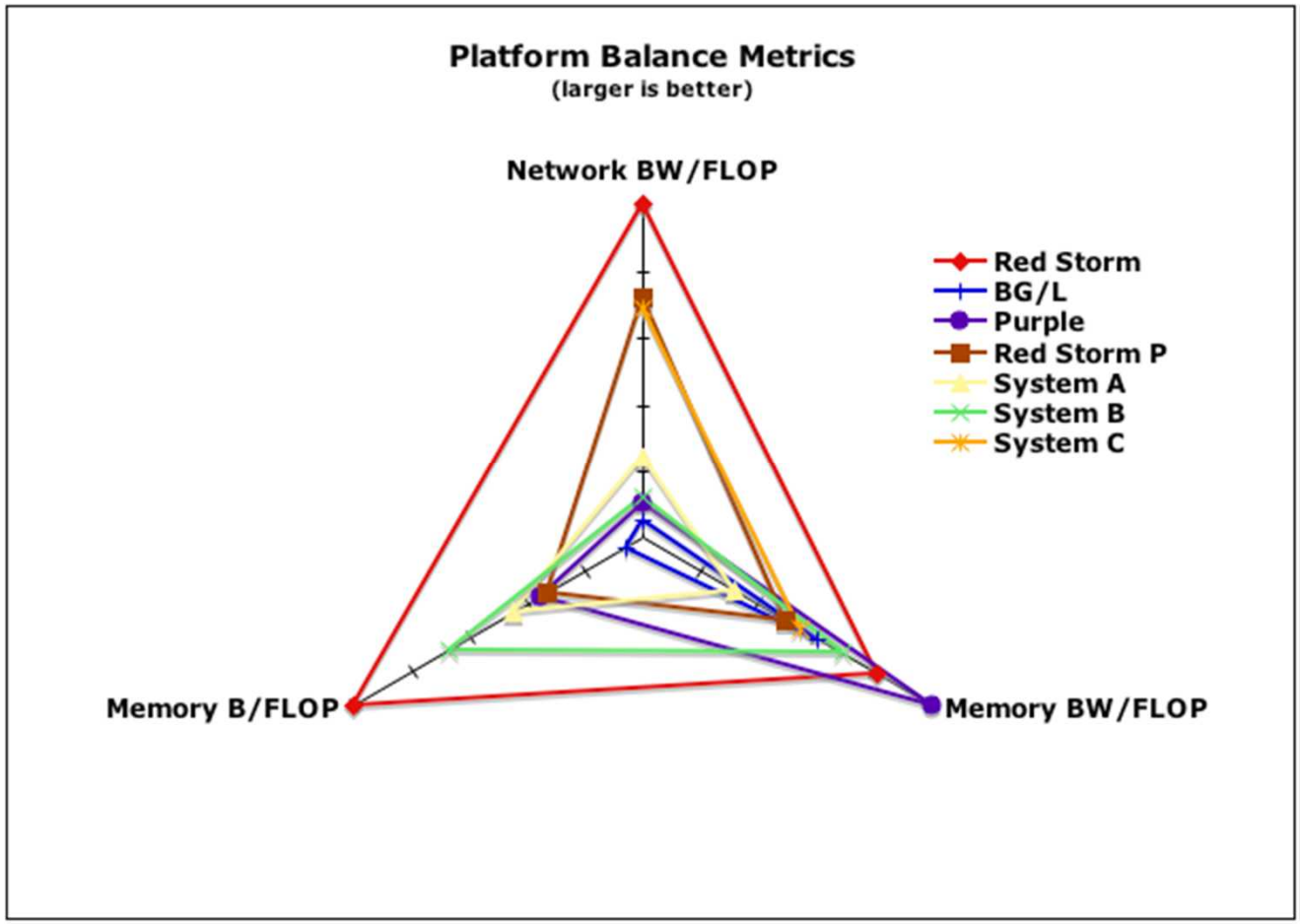
Extreme parallel programming

- **"I think that HPC probably won't drive the fundamental advancements in parallel programming. I think it had that opportunity, but that window of leadership is rapidly closing …. If these new [multicore] architectures are going to be successful, a lot of people are going to have to program them and they're not going to be satisfied with the kinds of tools available in HPC today."**

Justin Rattner, Intel CTO

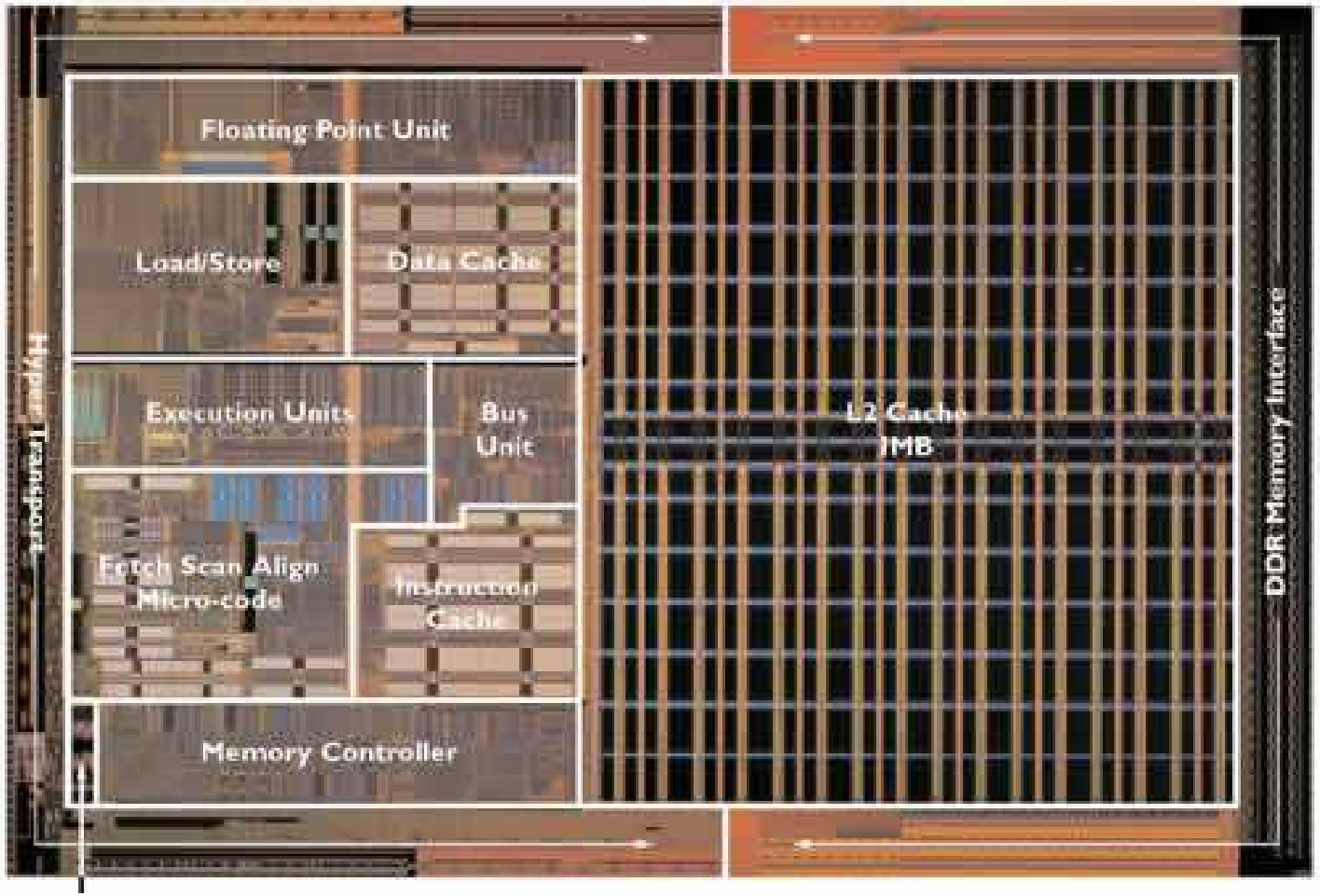
Impact of a system balance

- Balance factors predict HPCC results pretty well
- Architectural balance with low system noise is key to a scalable performance



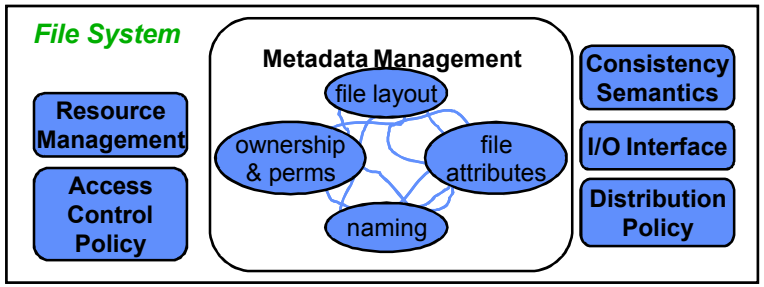
Sandia National Laboratories

Modern architecture example: The AMD Opteron



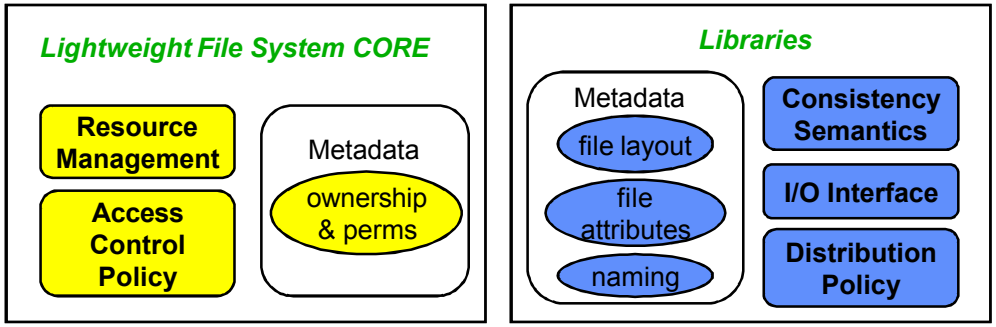
Lightweight file systems

- Current PFS solutions do not provide scalable I/O
- General-purpose file systems have too much functionality
- Lustre, PVFS2, Panasas and most others use this model



Traditional FS

- Lightweight approach has promise
  - only do what you have to do
  - put the burden on the library, not the file system



LWFS-core Provides

Direct Access to Storage

Scalable Security Model

Efficient Data Movement

Libraries Provide

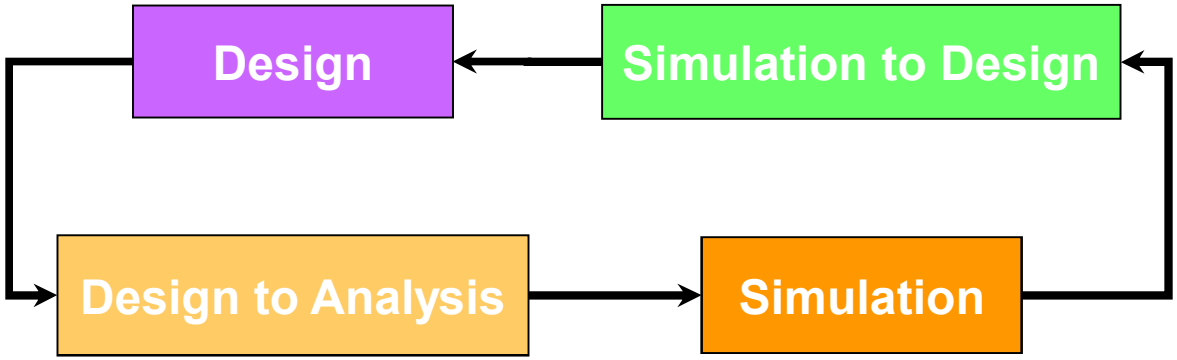
Everything else

Sandia National Laboratories

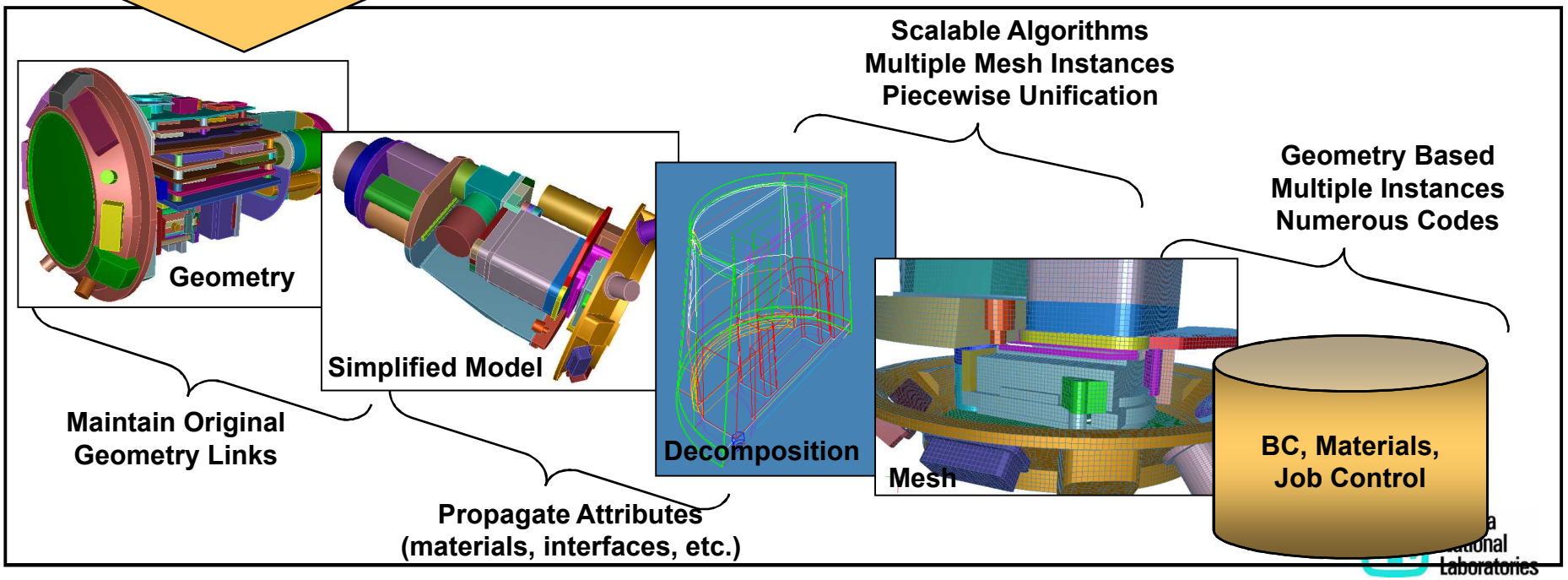
Enabling technology challenges

- Cracking “human in the loop” problems
  - Geometry, meshing
  - Visualization & interpretation (automating attention focus)
  - Design through analysis
- Exquisite load balancing technology
- Solver scalability
- Hierarchical approach across data infrastructure

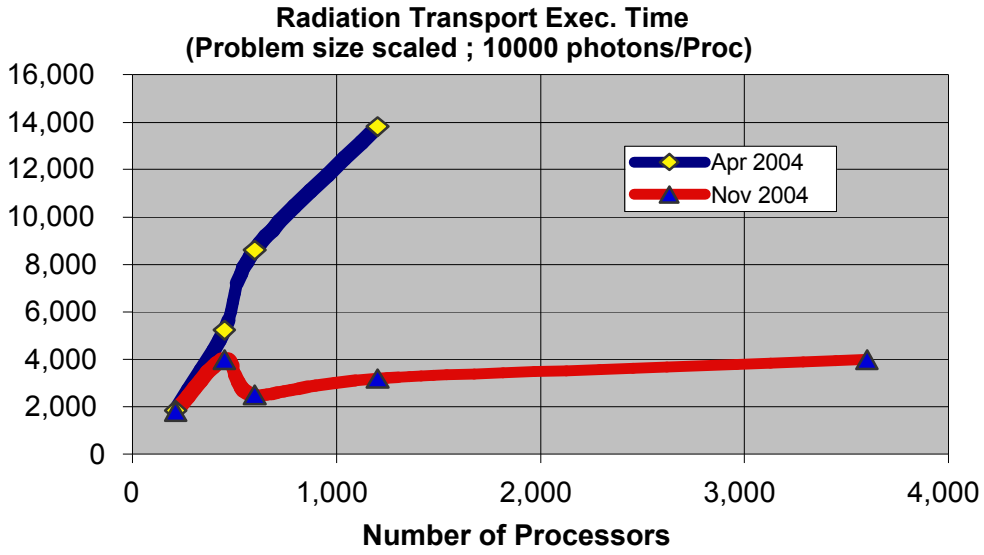
Design to analysis



- D2A dominates simulation cycle time, typically by an order of magnitude
- Majority of the time is spent in geometric manipulation
- Years are spent in some cases; 6 mos. Is not unusual
- The difficulty scales with resolution

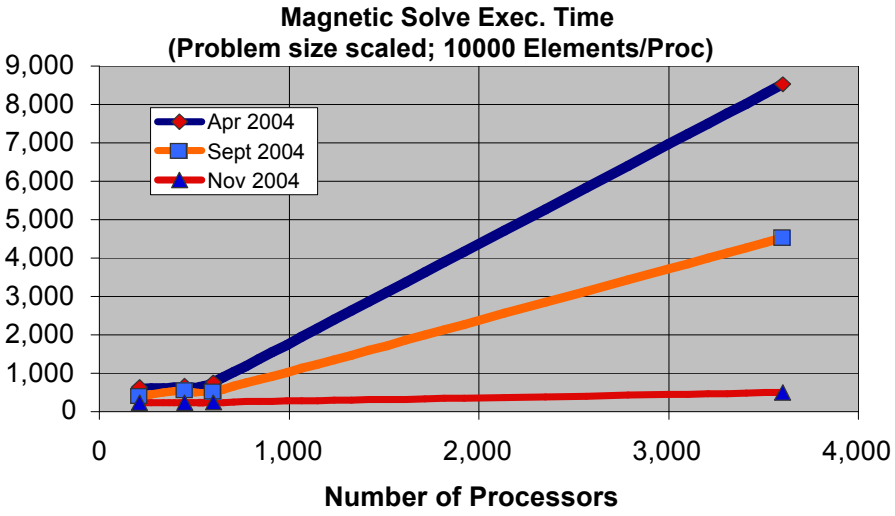


Solvers still need a lot of tuning

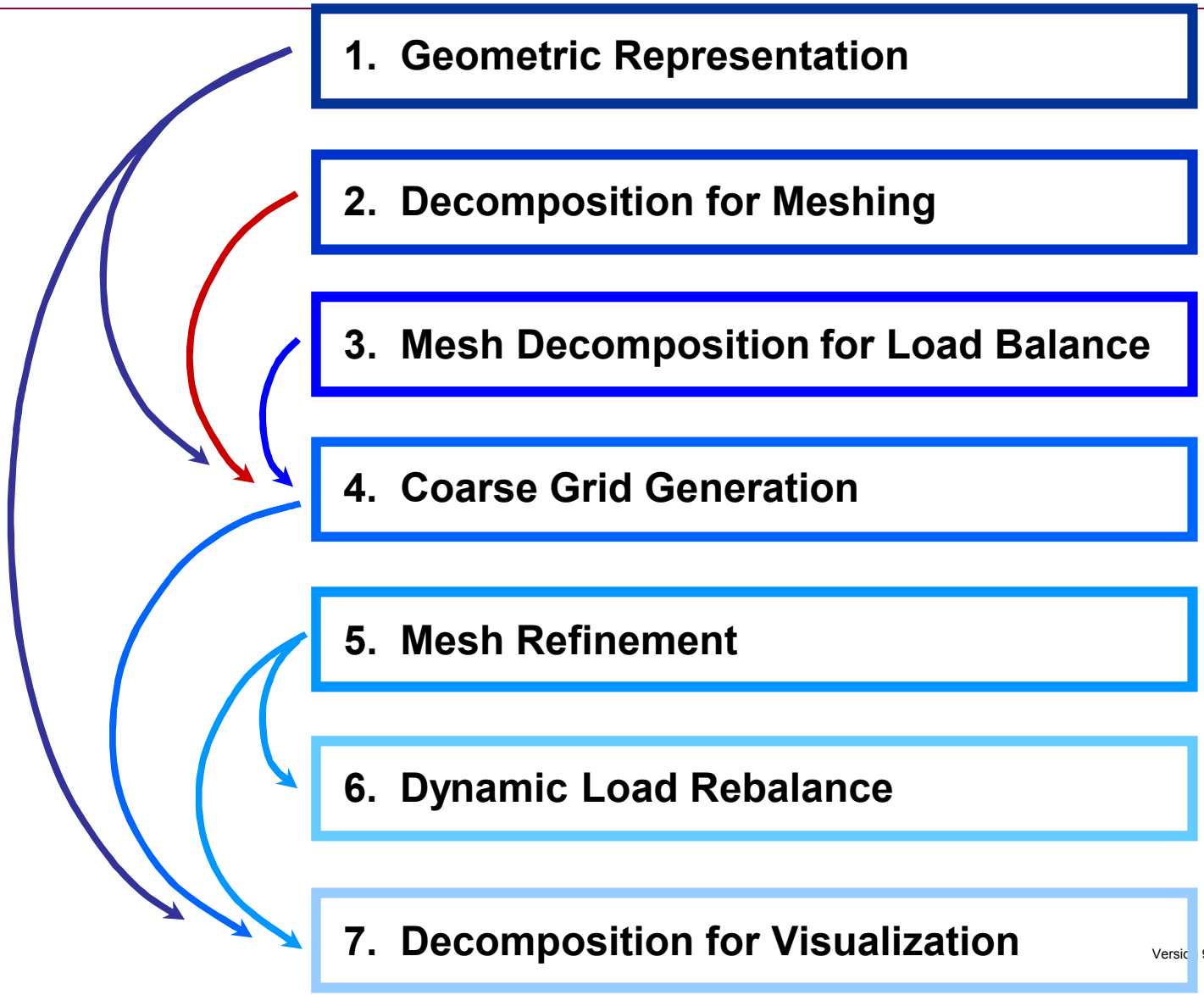


- **Improvements in memory usage and load balance (via coarse grid repartition) of the multi-level solver resulted in 17x reduction in execution time at scale**

- **Improvements in IMC radiation algorithm using non-blocking exchange resulted in 4x reduction in execution time on 1,200 processors**



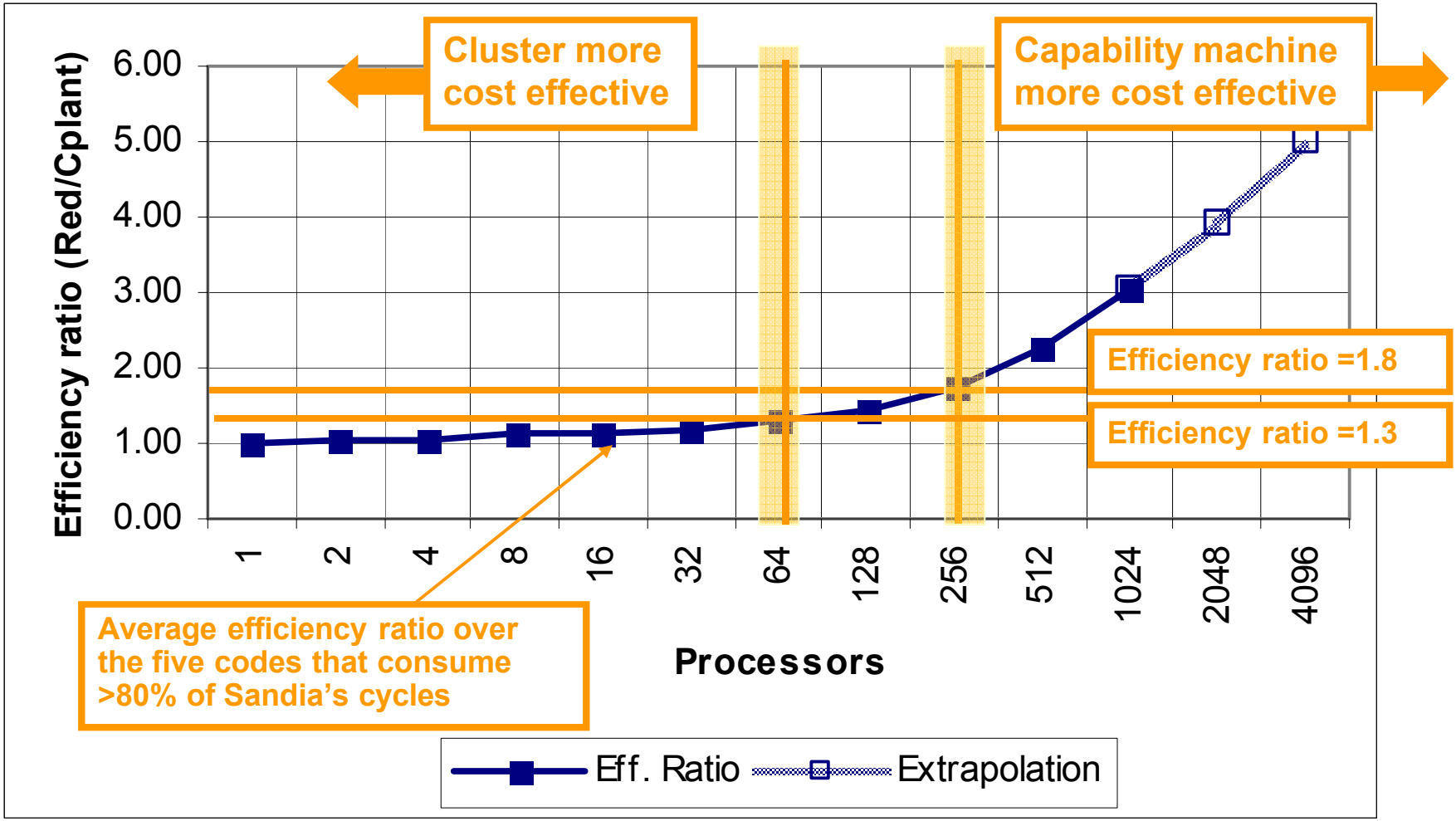
Hierarchical data architecture



Operational challenges

- Power load (really function of scale as well)
  - 20MW is pushing practical limit
  - ~$12M/yr alone for power
- Petascale networking infrastructure
  - 25X – 250X current network capability
  - An MP system in the infrastructure as well
- Getting the portfolio right
  - Capability vs. capacity mix
- Rapid integration
- Scalable user support model (for 10X more users)
- Merging with enterprise computing

Setting a cost effective machine portfolio



Sandia National Laboratories

Leadership challenges

- Delivering systems that really work
- Honoring the value proposition (impact v. viewgraphs)
- Examining computing and architectural model
  - engineering v. science v. informatics
- Combating the misuse of benchmarks
- A scalable user support model
- Projecting the future
- Maintaining political and public support

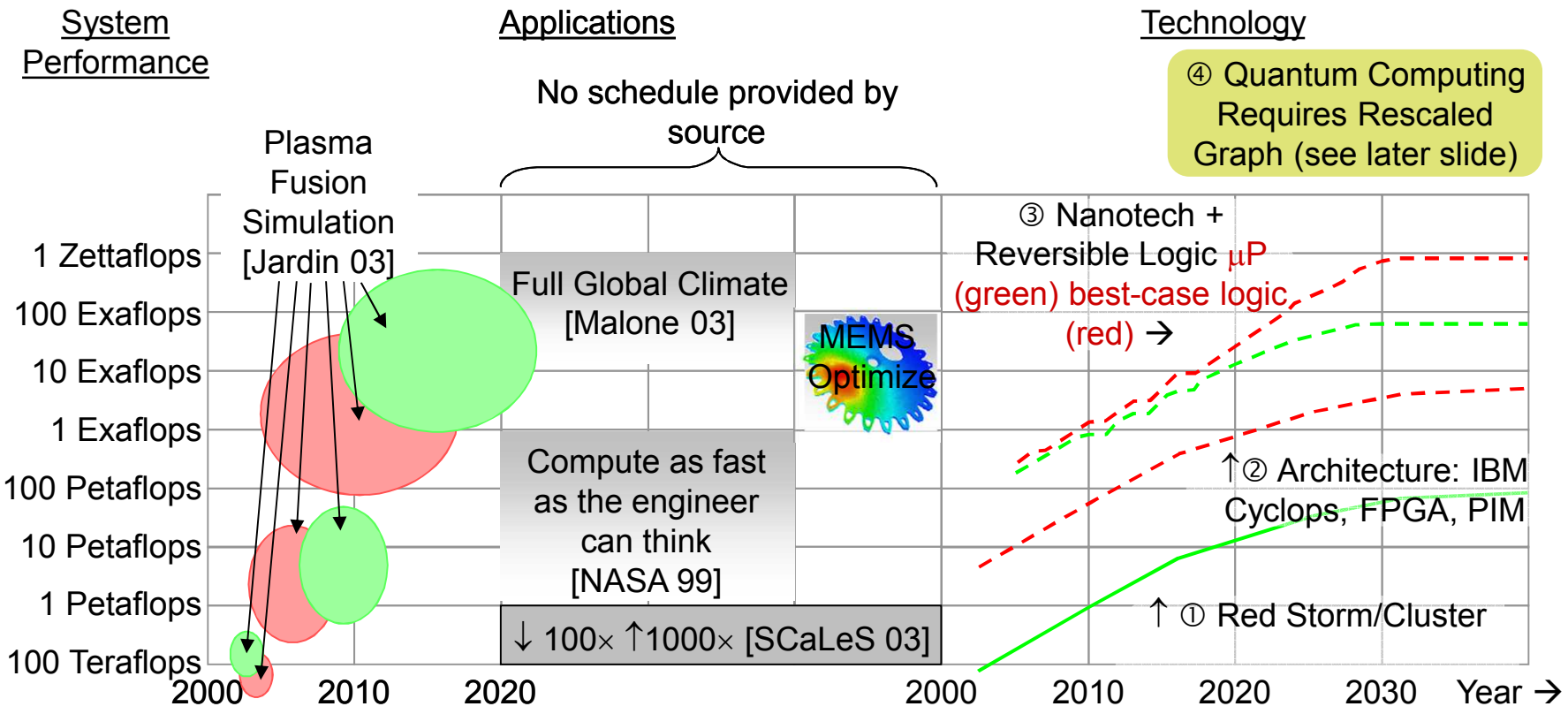
A scalable user support model

- Rollout Common Engineering Environment (CEE)
  - Study and segment the market
  - Partner with engineering communities to select best tools
  - Design mission specific operating environments
  - Integrate with Design through Analysis environment (DART)
  - Deploy via Software Acquisition/Management System (SAMS)
  - Architectural changes to desktop

- Extend CSU business model to cover CEE in depth
  - Tier 1: CCHD
  - Tier 2: In office support from simulation technicians
  - Tier 3: Route to 4300 applications specialists
  - Tier 4: Route to code development team, e.g. 1500

- Lower barriers … scalable architecture, corporately branded

The future: Sic itur ad astra



[Jardin 03] S.C. Jardin, "Plasma Science Contribution to the SCaLeS Report," Princeton Plasma Physics Laboratory, PPPL-3879 UC-70, available on Internet.
[Malone 03] Robert C. Malone, John B. Drake, Philip W. Jones, Douglas A. Rotman, "High-End Computing in Climate Modeling," contribution to SCaLeS report.
[NASA 99] R. T. Biedron, P. Mehrotra, M. L. Nelson, F. S. Preston, J. J. Rehder, J. L. Rogers, D. H. Rudy, J. Sobieski, and O. O. Storaasli, "Compute as Fast as the Engineers Can Think!" NASA/TM-1999-209715, available on Internet.
[SCaLeS 03] Workshop on the Science Case for Large-scale Simulation, June 24-25, proceedings on Internet a http://www.pnl.gov/scales/.
[DeBenedictis 04], Erik P. DeBenedictis, "Matching Supercomputing to Progress in Science," July 2004. Presentation at Lawrence Berkeley National Laboratory, also published as Sandia National Laboratories SAND report SAND2004-3333P. Sandia technical reports are available by going to http://www.sandia.gov and accessing the technical library.

Acknowledgements

- Tom Hunter
- Bill Camp & 1400
  - Jim Tomkins
  - Sue Kelly
  - Erik DeBenedictis
  - Steve Plimpton
  - Ted Blacker
  - Tom Brunner
  - Mark Taylor
  - Richard Murphy
  - Jim Ang
  - Doug Doerfler
  - Keith Underwood
  - Ron Oldfield
  - Sheldon
  - Ken Moreland
- 4300 colleagues
  - John Zepper
  - Mahesh Rajan
  - John Noe
  - David White
  - Bob Balance
  - Debra Buttry
- Art Ratzel & 1500
  - Hal Morgan
  - Sheldon Tiezen
  - Martin Heinstein
  - Sam Key
  - Ken Alvin
  - Mark Blanford
  - Ken Gwinn
  - Steve Kempka
  - Dan Rader
  - John Torczynski
  - Michael Gallis
  - Anton Sumali
  - David Epp
  - Phil Reu
  - Stefan Domino
  - Tolulope Okusanya
- Ed Barsis & Peter Mattern
- Regina Valenzuela
- Many, many others others