Future High-End and Scientific Computing:
National Leadership Computing Facility

Jack C. Wells, Computing & Computational Sciences Fellow

Jeff S. Vetter, Future Technologies Group Leader

Jeff A. Nichols, Director

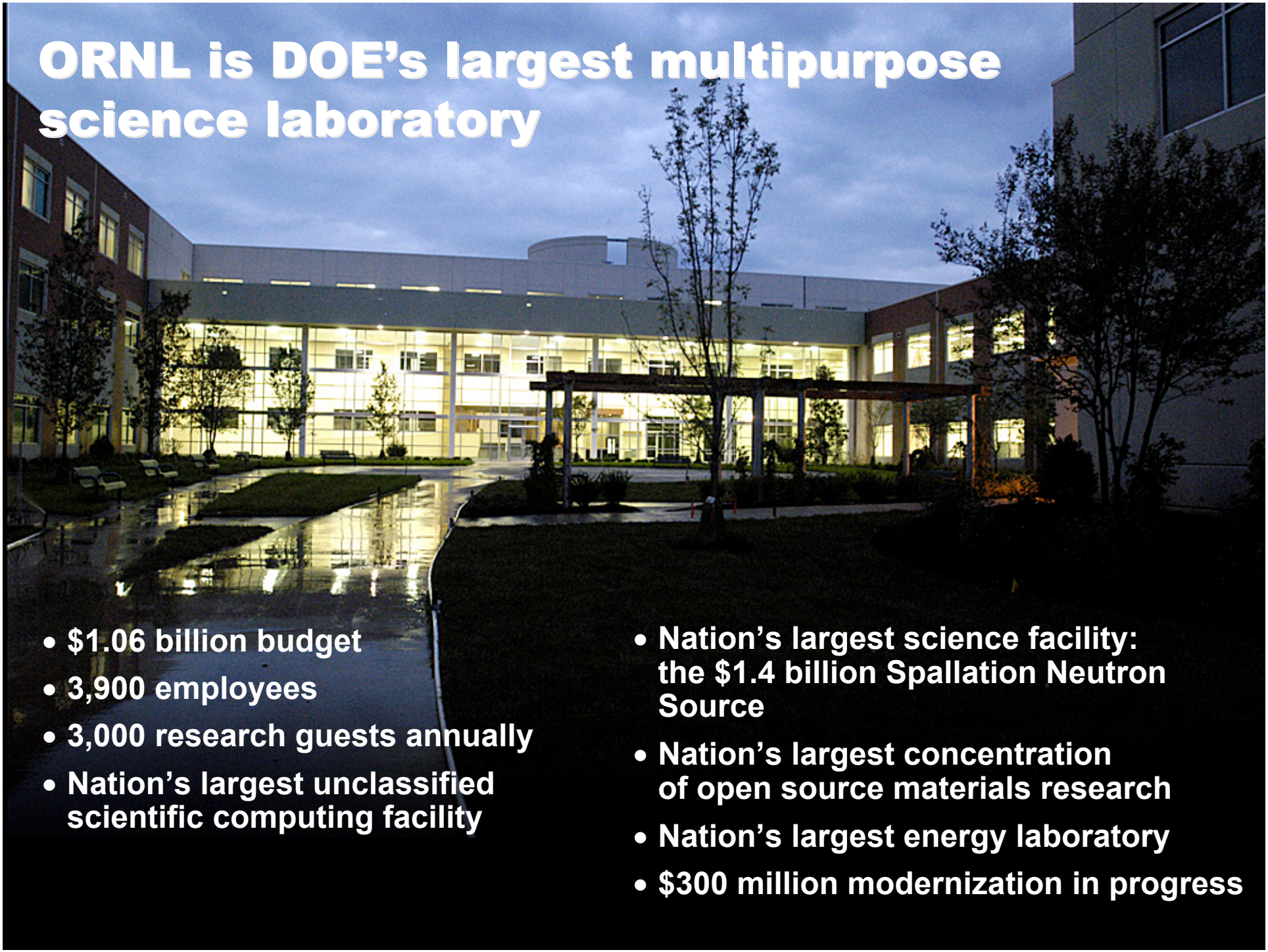
Computer Science and Mathematics Division

Oak Ridge National Laboratory

Monte Carlo 2005
March 24, 2005
Chattanooga

Research sponsored by the Office of Advanced Scientific Computing, U. S. Department of Energy, by UT-Battelle, LLC, a contractor of the U.S. Government under Contract DE-AC05-0OR22725.

ORNL is DOE's largest multipurpose science laboratory



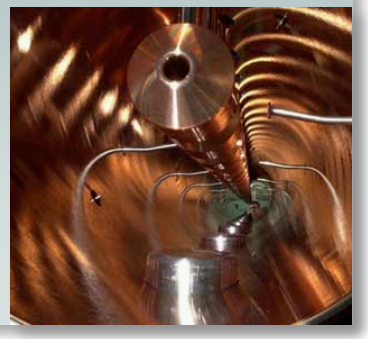
- $1.06 billion budget
- 3,900 employees
- 3,000 research guests annually
- Nation's largest unclassified scientific computing facility

- Nation's largest science facility: the $1.4 billion Spallation Neutron Source
- Nation's largest concentration of open source materials research
- Nation's largest energy laboratory
- $300 million modernization in progress

We are developing and deploying world-class tools for nanoscale R&D

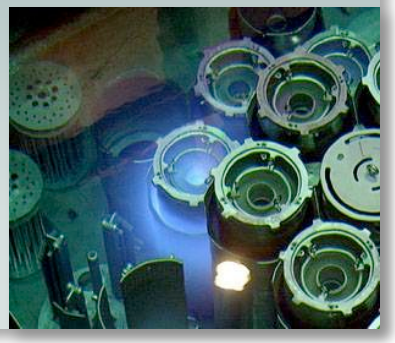
Spallation Neutron Source:

- High-intensity neutrons for materials research at the nanoscale
- 1.4 MW of beam power on target
- 16 instruments



High Flux Isotope Reactor:

- The nation's leading research reactor
- World-class instruments for neutron scattering

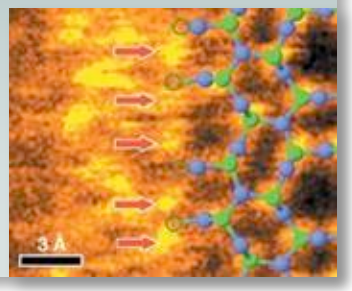


Center for Nanophase Materials Sciences

- Construction on schedule for 2006
- User program launched with 42 projects

Ultrahigh-resolution microscopy

- Advanced Microscopy Laboratory
- Aberration-corrected electron microscope
- World-record resolution: 0.6 Å



The Spallation Neutron Source

Total cost: $1.4 billion

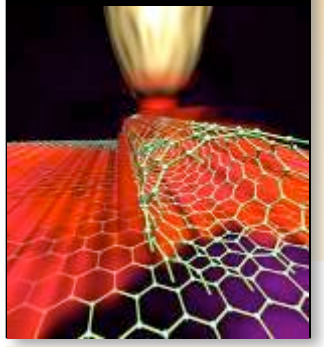
- Operational in 2006
- World's most powerful pulsed spallation source
- With complementary resources at the High Flux Isotope Reactor, Oak Ridge will lead the world in neutron scattering

Leadership Computing for Science

Critical for success in key national priorities

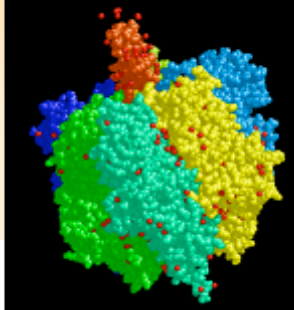
Office of Science research priorities

Manipulating the Nanoworld



Computational design of innovative nanomaterials

Taming the Microbial World

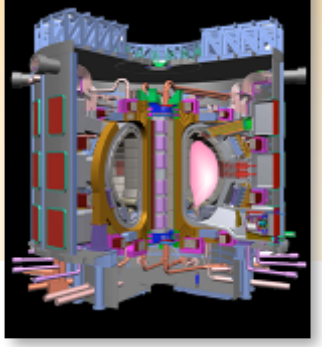


Predictive understanding of microbial molecular and cellular systems

Environment and Health

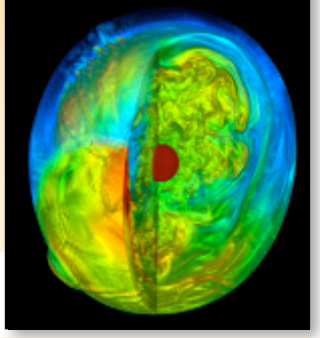
Full carbon cycle in climate prediction, IPCC

ITER for Fusion Energy



Simulation of burning plasma, Fusion Simulation Project

Search for the Beginning

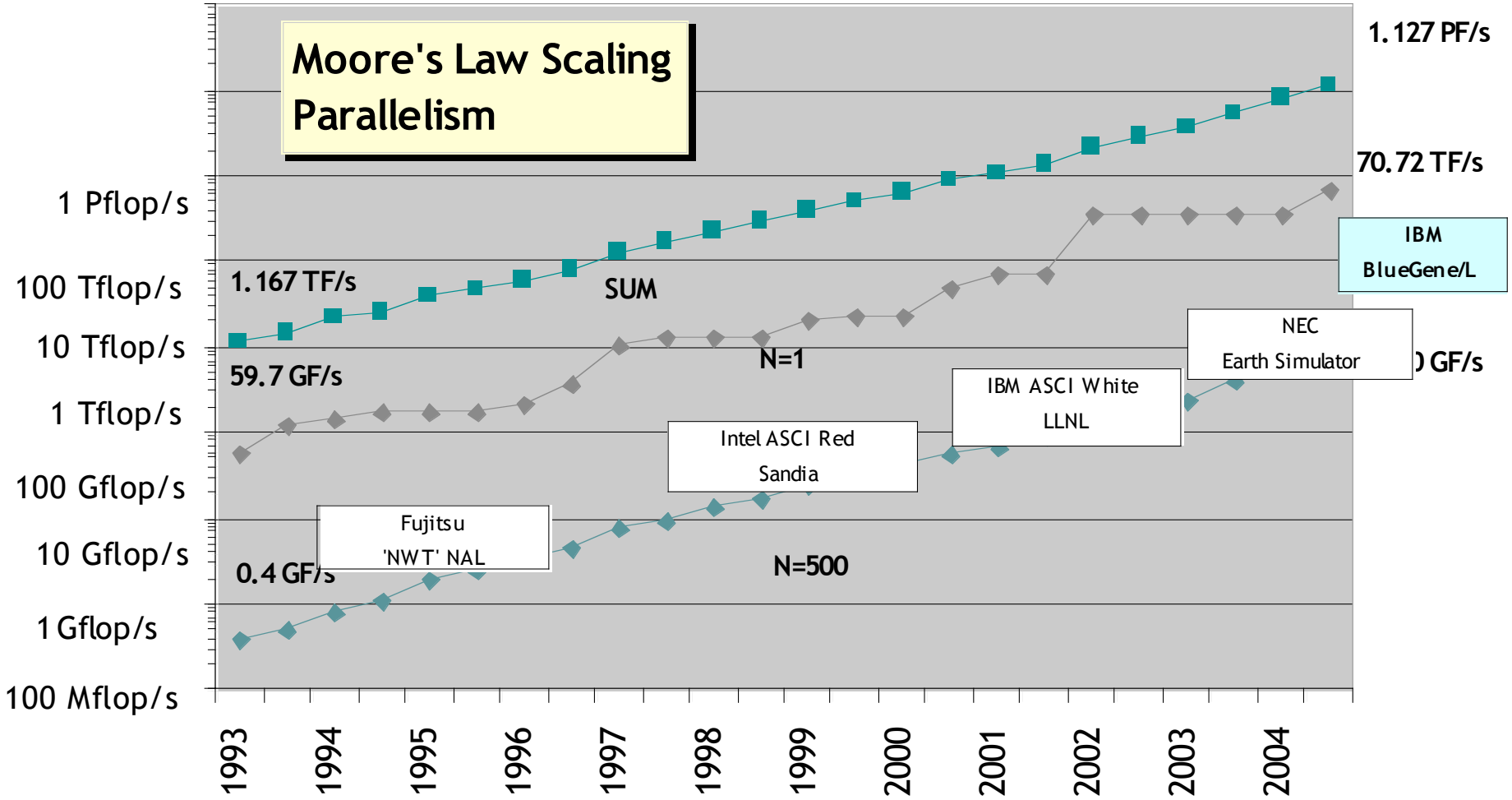


Terascale Supernovae Simulation

Theory, Mathematics, Computer Science

National Leadership-Class Computing Facility for Science

Computing has experienced exponential growth over the past five decades



Jack Dongarra, TOP500 Performance – November 2004

Examples of early evaluation over the years

Compaq AlphaServer SC 2000

SRC Prototype 1999

KSR-1 1991

GSN Switch 2000

IBM S80 1999

IBM Winterhawk And Nighthawk 1999

Intel i/PSC-2 1988

Intel I/PSC-860 1990

Intel Paragon XP/S-35 1992

Intel Paragon MP X/PS-150 1995

Cray XT3 2005

Cray XD1 2004

SGI Altix 2003

Cray X1 2003

IBM Power4 and Federation 2001-2004

New facility capable of housing leadership class computers

- Space and power:
  - 40,000 ft^2 computer center with 36-in. raised floor, 18 ft. deck-to-deck
  - 8 MW of power (expandable) @ 5c/kWhr
- Classroom and training areas for users
- High-ceiling area for visualization lab (Cave, Powerwall, Access Grid, etc.)
- Separate lab areas for computer science and network research

CCS terascale computer systems

Cray X1 – *Phoenix* (6.4 TF)
Largest X1 in the world and first in DOE
Scalable vector architecture
512 multi-streaming processors 400 MHz

IBM Power4 – *Cheetah* (4.5 TF)
First Power4 system in DOE
Cluster using IBM Federation Interconnect
864 IBM Power4 processors 1.3 GHz

SGI Altix – *Ram* (1.5 TF, 2TB)
Large globally addressable memory system
256 Intel Itanium2 processors 1.5 GHz
Linux with single operating system image

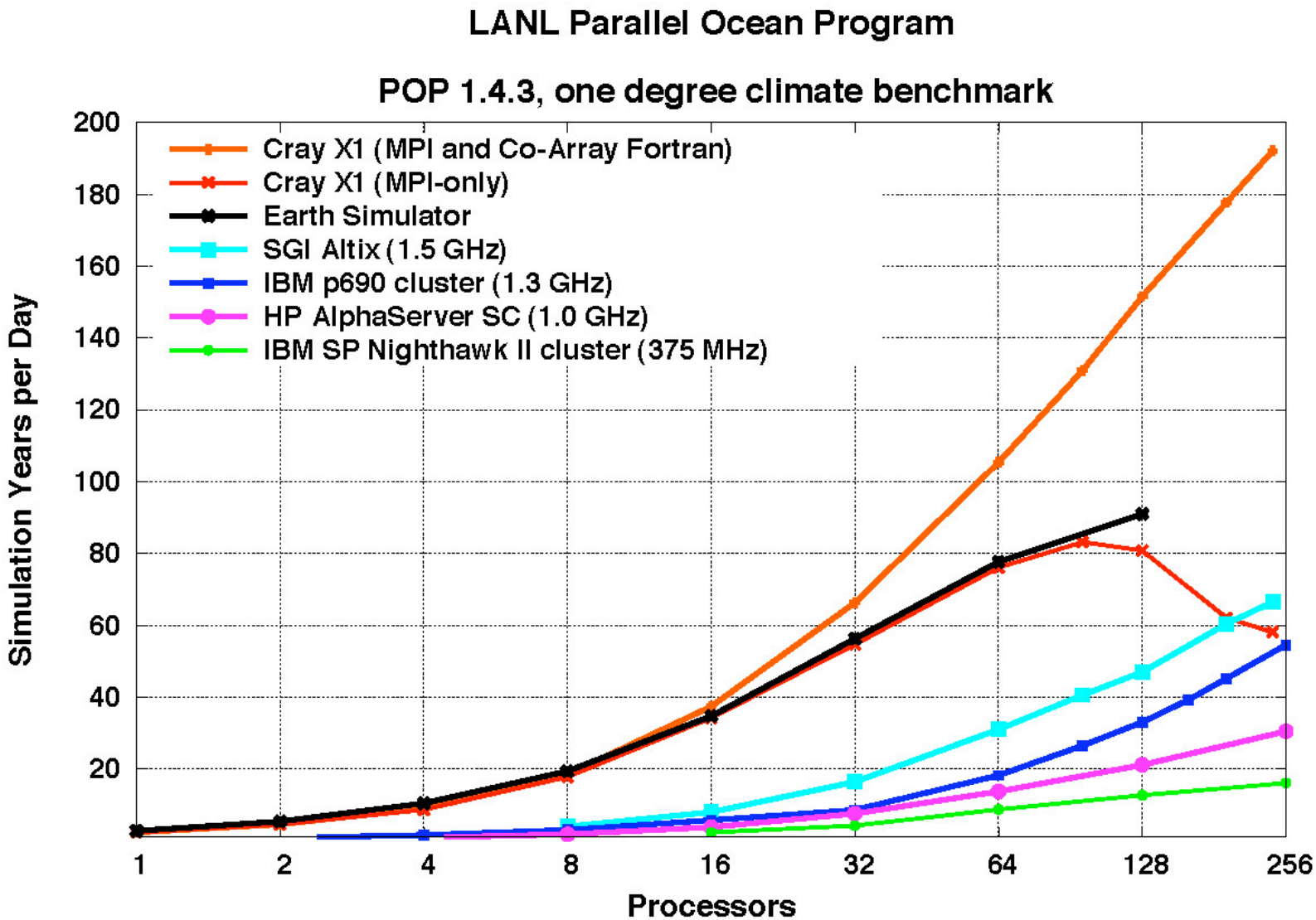
IBM Power3 – *Eagle* (1 TF)
First terascale system in DOE-SC
Cluster of SMP nodes
736 IBM Power3 processors 375 MHz

Phoenix – The CCS Cray X1

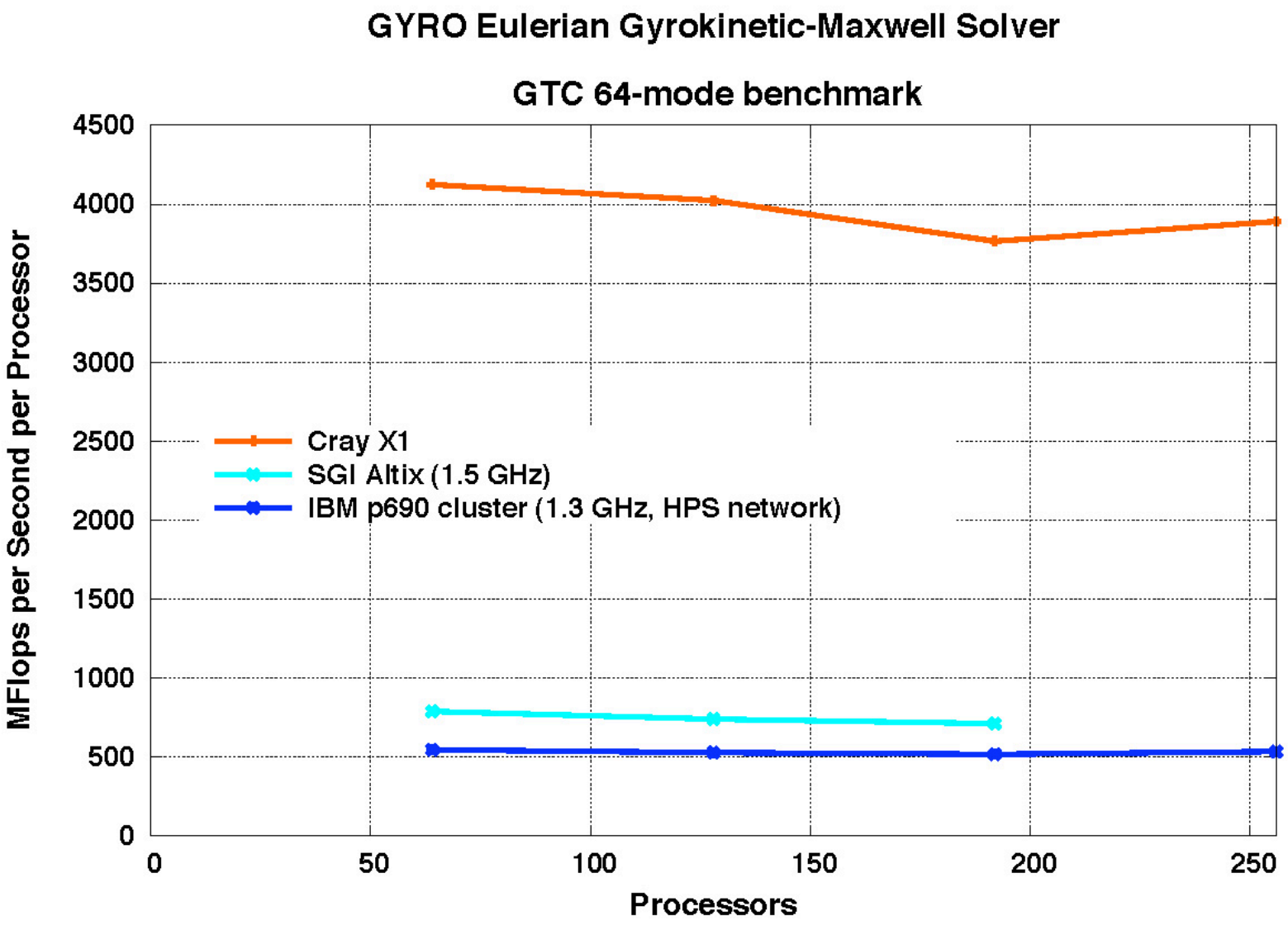
- **Largest Cray X1 in the world – 6.4TF**
- **2 TB globally addressable memory**
- **512 processors**
  - 400 MHz, 800 MHz vector units
- **32 TB of disk**
- **Most powerful processing node**
  - 12.8 GF CPU, 2-5x commodity processors
- **Highest bandwidth communication with main memory**
  - 34.1 GB/sec

- Highly scalable hardware and software
- High sustained performance on real applications

Apps: CCSM Parallel Ocean Program



Apps: GYRO Fusion Simulation



Office of Science notice to SC Laboratories

Leadership-Class Computing Capability for Science

"The focus of the proposed effort should be on capability computing in support of high-end science – rather than on enhanced computing capacity for general science users . . . "

"The proposed effort must be a user facility providing leadership class computing capability to scientists and engineers nationwide independent of their institutional affiliation or source of funding."

National Leadership Computing Facility

May 12, 2004

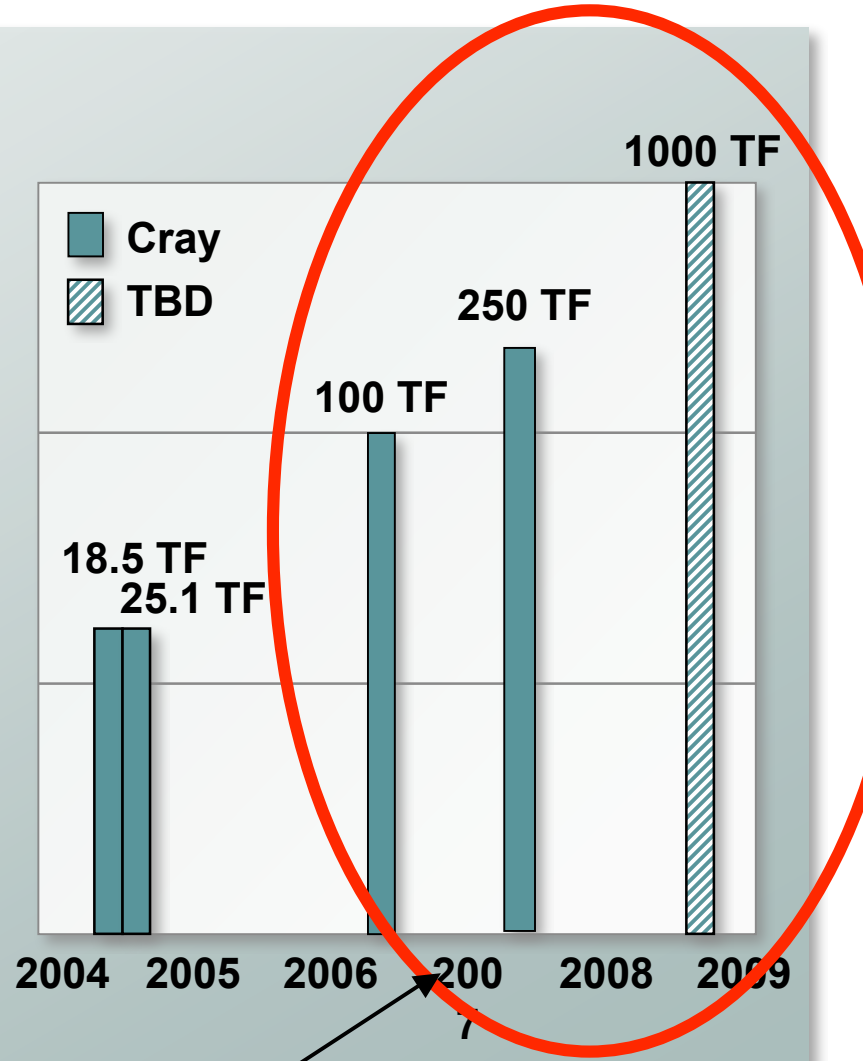
NLCF hardware roadmap

In 2005,

- Deploy 18.5TF Cray X1E and 25.1TF Cray XT3 systems
- Cray forms and supports "Supercomputing *Center of Excellence*"
- Develop/deploy complementary software environment
- Full operational support of NLCF as a capability computing center
- Deliver computationally intensive projects at scale and high impact

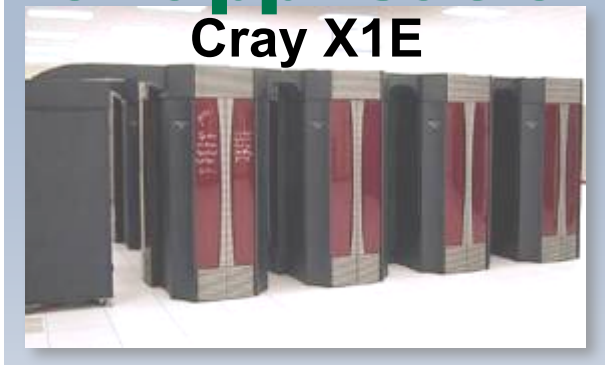
In 2006, deploy 100TF Cray XT3

In 2007-8, deploy 250TF Cray Rainier

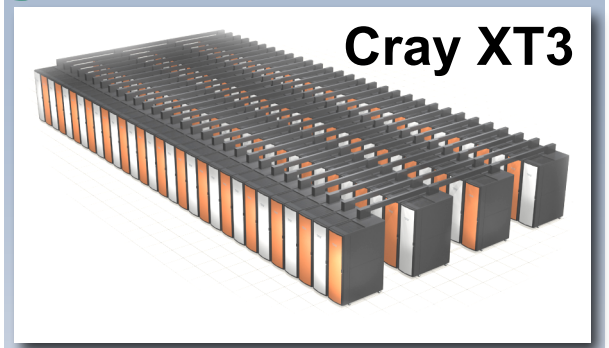


Future Technology

Three purpose-built architectures optimized for applications



- Proven architecture for performance and reliability
- Most powerful processors and interconnect
- Scalable, globally addressable memory and bandwidth
- Offers capability computing for key applications



- Extremely low latency, high bandwidth interconnect
- Efficient scalar processors, balanced interconnect
- Known system architecture – based on ASCI Red
- Linux operating system on service processors with microkernel on compute processors

- Multi-paradigm computing including vector, scalar, and potentially reconfigurable and multi-threaded processors
- Single solution for diverse workloads
- Single user interface and environment
- Improved performance by matching processor to job
- Leverage all engineering resources

Jaguar – The CCS XT3 system

Cabinets	Performance	Processors	Memory	Disk Space	I/O Bandwidth
56	25 TF	5,304	10.5 TB	120 TB	15 GB/s
120	54.6 TF	11,374	23 TB	240 TB	30 GB/s
120	109 TF	22,748	46 TB	480 TB	60 GB/s

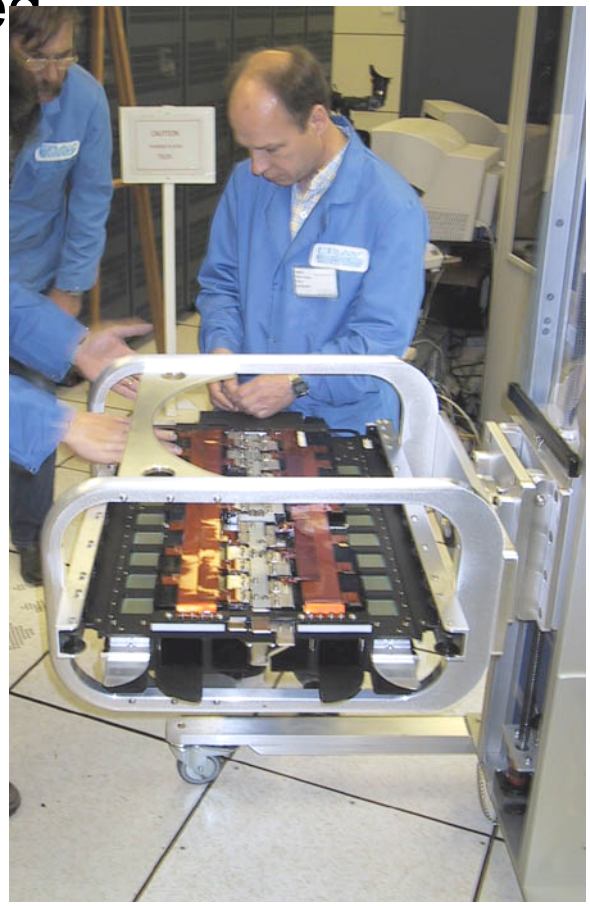
"Jaguar" first 11 cabinets

Phoenix – Cray X1 to X1E upgrade

- One-fourth of machine will be upgraded at a time
- X1 and X1E nodes will be partitioned into separate systems until upgrade completed

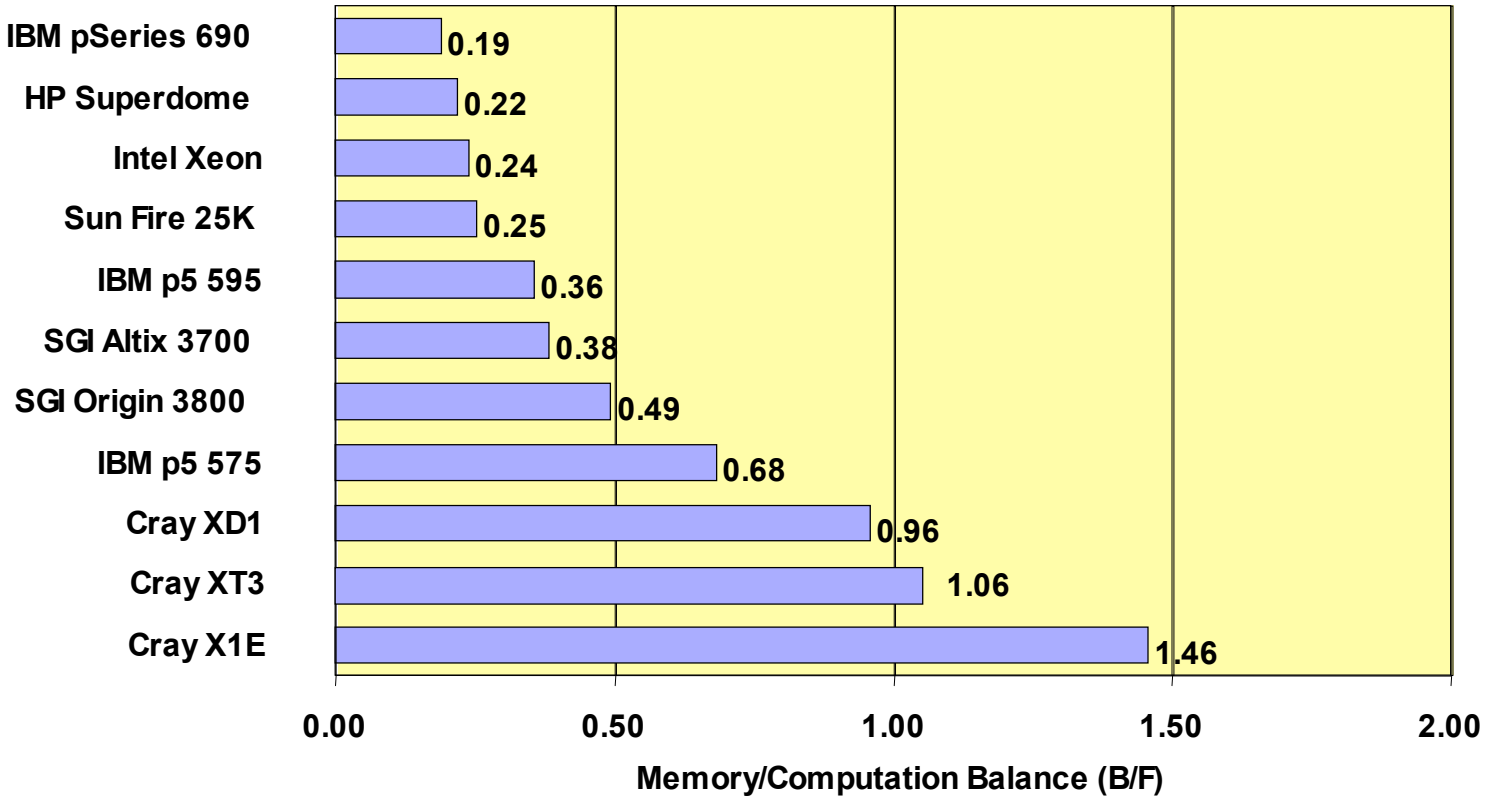
Upgrades begin in June and continue through summer

End result is 1024 processor, 18.5 TF X1E

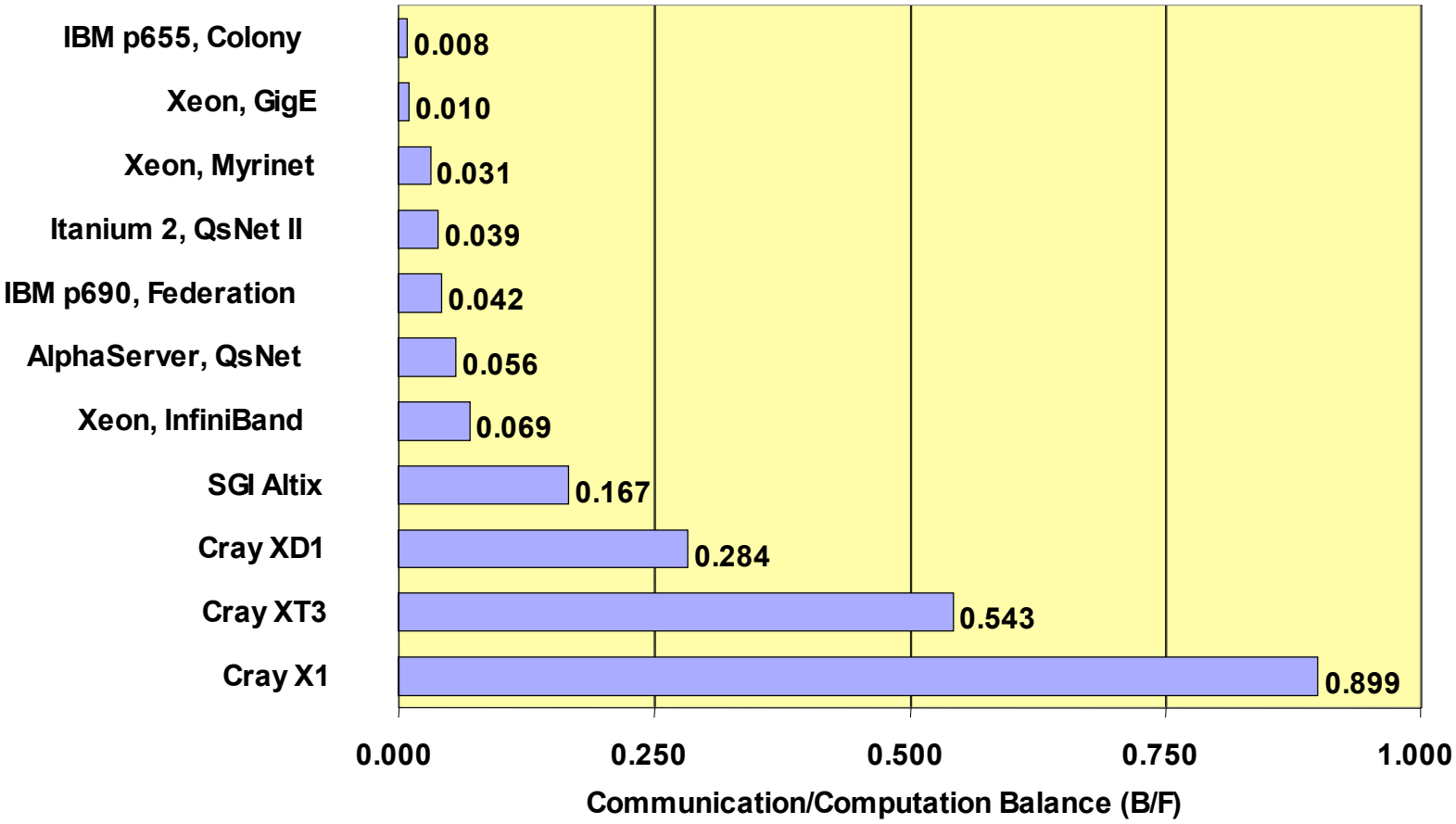


System offers memory bandwidth

Memory bandwidth in bytes per second/FLOPS calculated from memory bandwidth measured via STREAM Triad benchmark

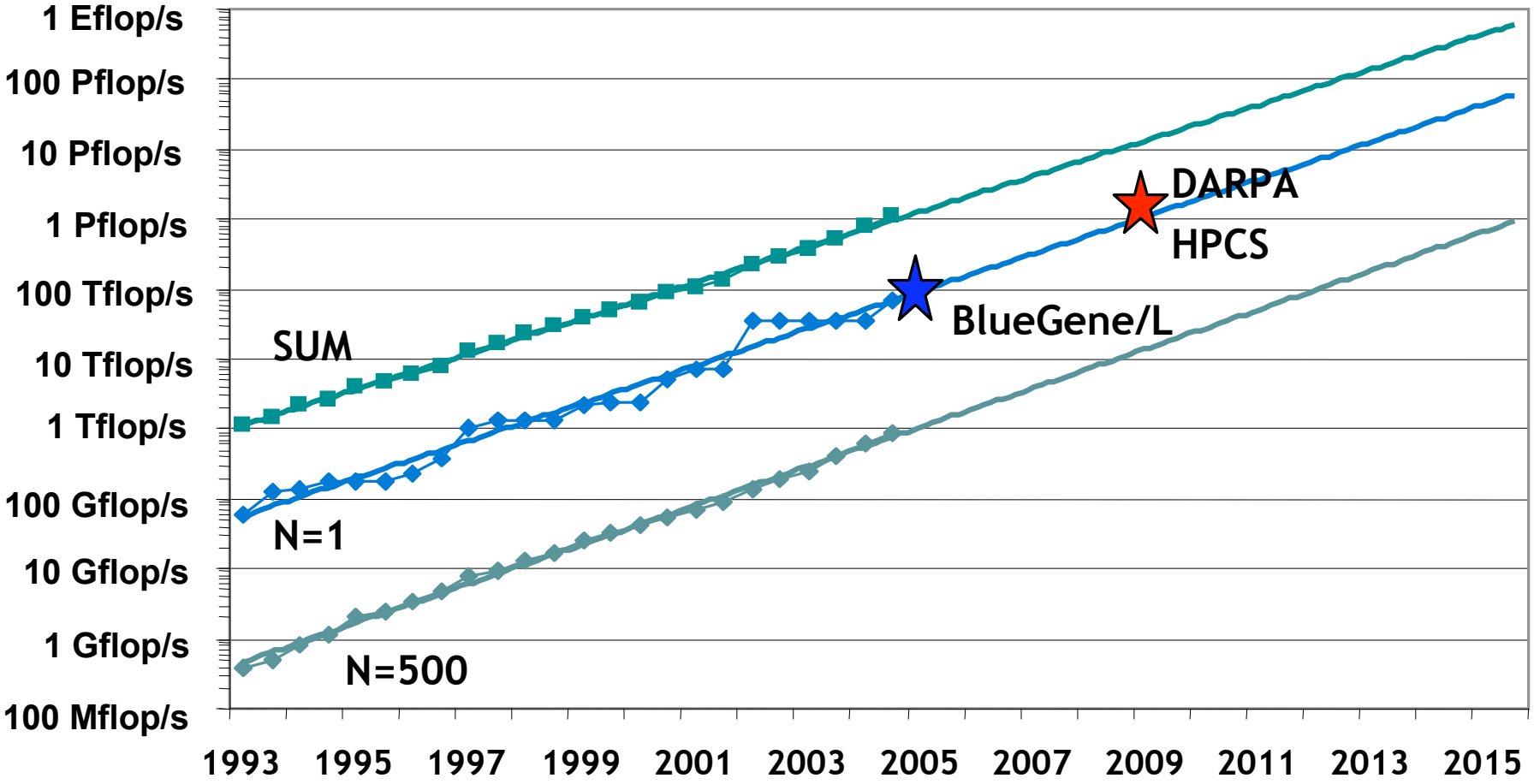


... and interconnect balance



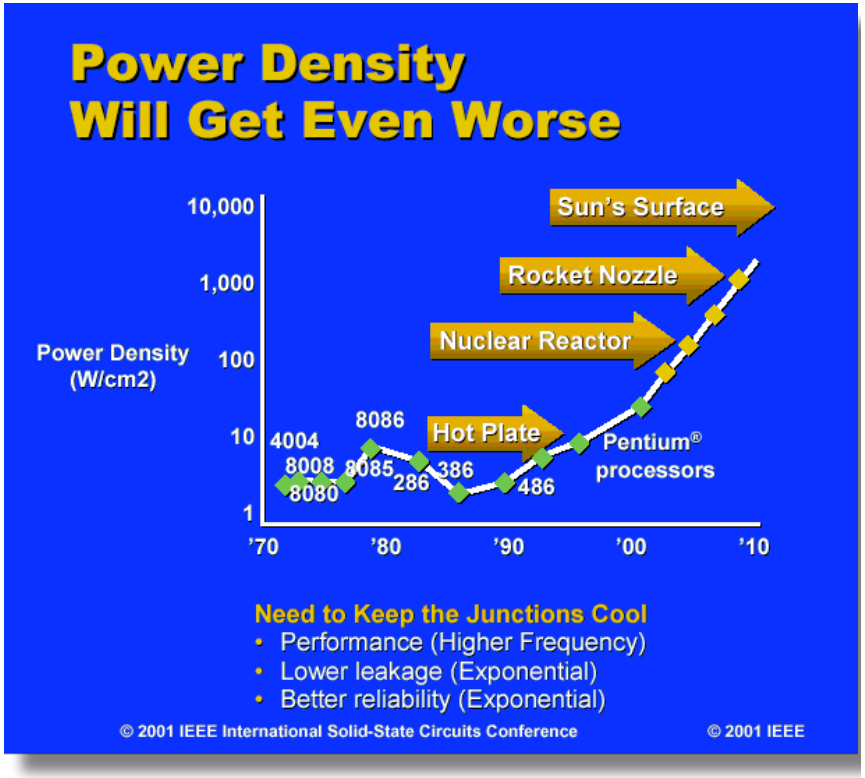
Network bandwidth is maximum bidirectional data exchange rate between two nodes using MPI shown in bytes per second/FLOPS

Growth rate is expected to continue
Requires $$ (Moore's 2nd Law) and new solutions

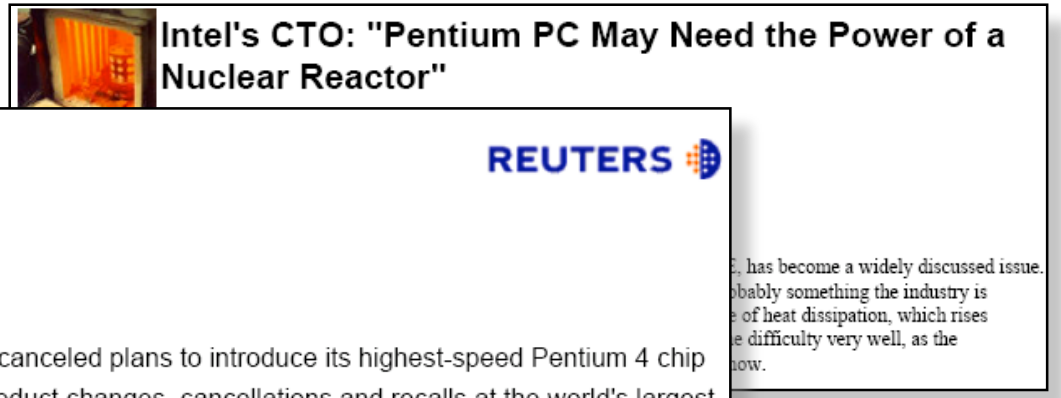


Observations

- Memory-processor performance gap leading to low sustained performance on important applications
- ITRS predicts that Moore's Law will be coming to an end in 2012-2015
- Practical considerations (power, cooling, floor space) are forcing architects to change their designs now



Source: Pat Gelsinger, CTO, Intel



Intel's CTO: "Pentium PC May Need the Power of a Nuclear Reactor"

, has become a widely discussed issue.
obably something the industry is
e of heat dissipation, which rises
e difficulty very well, as the
now.

Intel Cancels Top-Speed Pentium 4 Chip

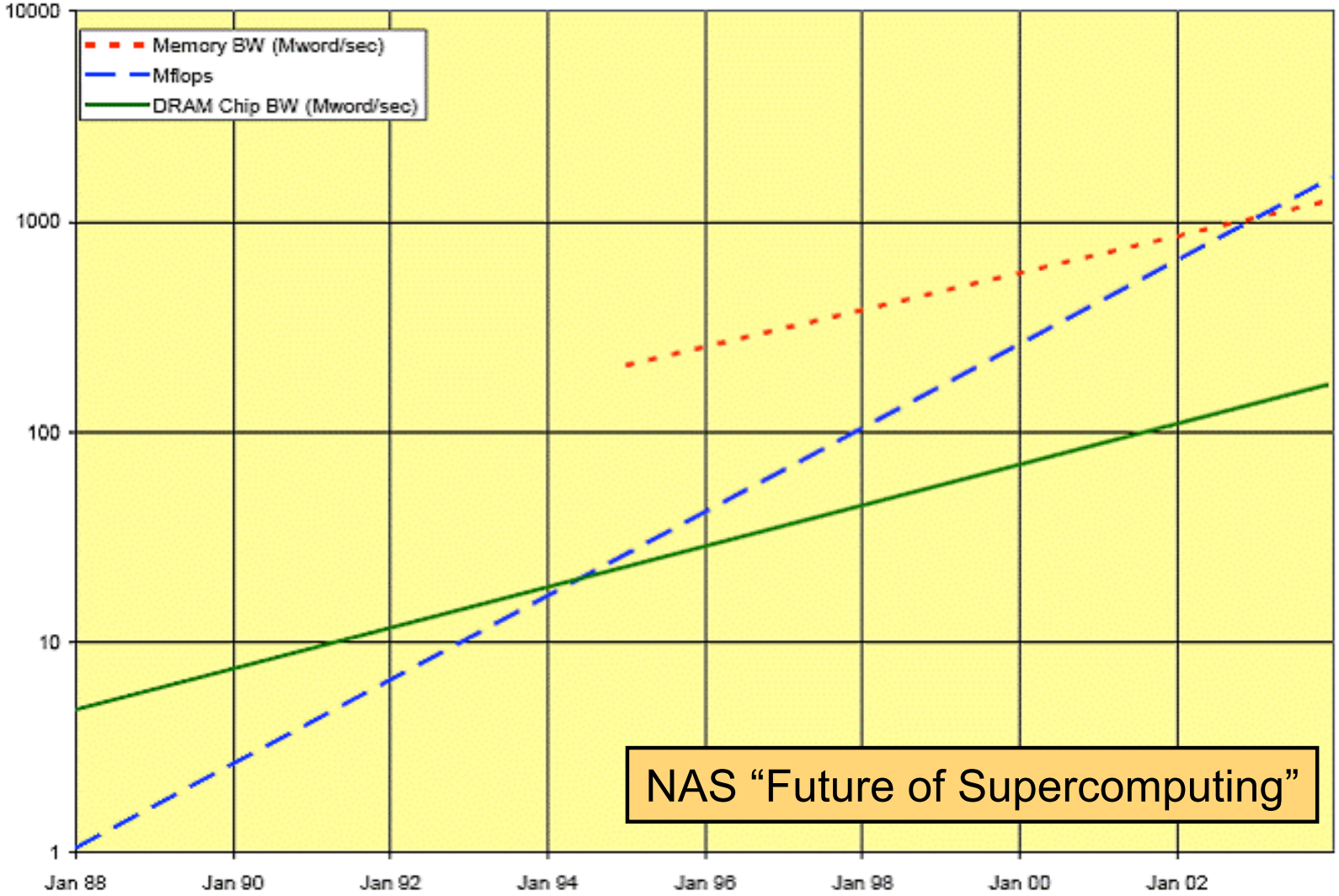
REUTERS

1 hour, 42 minutes ago
October 15, 2004

By Daniel Sorid

SAN FRANCISCO (Reuters) - Intel Corp. (Nasdaq:INTC - news) on Thursday canceled plans to introduce its highest-speed Pentium 4 chip for desktop computers, marking another in a string of unexpected product changes, cancellations and recalls at the world's largest computer chip maker.

Disparity between memory B/W, processor and memory performance



Trends in Design of Mainstream Systems

- **Processors**
  - More cache per chip
  - SIMD units (e.g. SS3)
  - Multiple cores per cpu
  - Multithreading
  - Power efficient execution
    - variable clock speeds

- **It is NOT clear how these trends impact our mission applications**
  - →Many of these systems require us to rethink algorithms and programming systems.

Motivation and Opportunity for the Exploration of Alternative Technologies

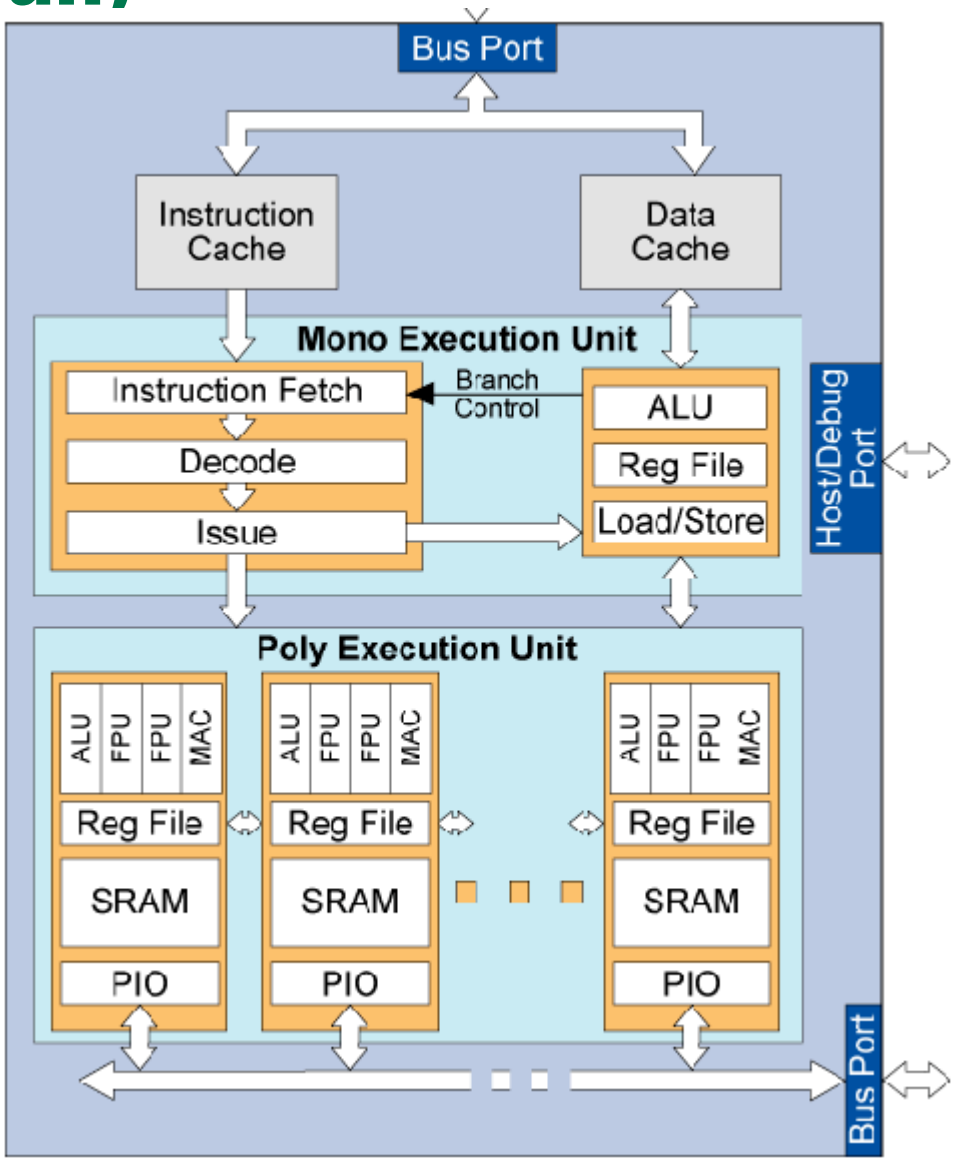
- A balanced Petaflop architecture will most likely require major changes in the hardware and software
- In fact, vendors and other agencies are already beginning to explore this space
  - Cray, SGI, and others are exploring reconfigurable computing
  - Cray, HP, and others are exploring PIM, co-processor architectures in the next-generation architectures
- Multi-paradigm computing
  - Scalar procs, vector procs, and other technologies FPGA, PIM all on the same scalable interconnect network

Alternative, Potentially Disruptive Technologies Allow Leaps in Performance

- **ClearSpeed Multithreaded Array Processor**
  - 96-way parallel @ 250Mhz, 5 watts
  - Co-processor capable of 50GF/card, 3TF/rack
- **SRC Computers 6E MAP Processor**
  - Reconfigures hardware for each application
  - Improves efficiency
- **Enlight256® Optical Digital Signal Processor**
  - Offers over 3 orders of magnitude computational capability over competitors
  - 8Tops in fixed point macs per second

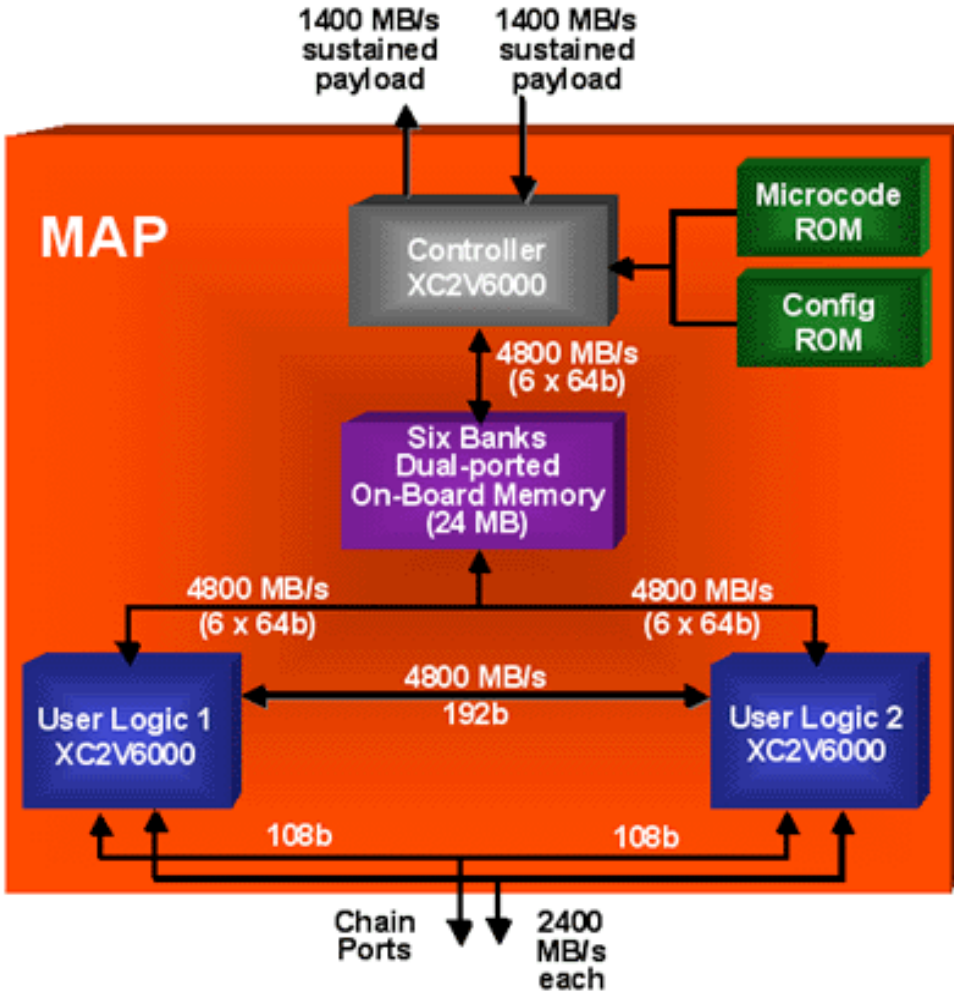
ClearSpeed CSX600 (Jun)

- **Multithreaded Array Processor**
  - 100GF/s per card
- **Each PE**
  - Is a VLIW core
  - Has a address generator
  - Has both integer, 64b FP ALU, div and sqrt
  - 128B register file
  - 6KB SRAM
  - 64b virtual addressing



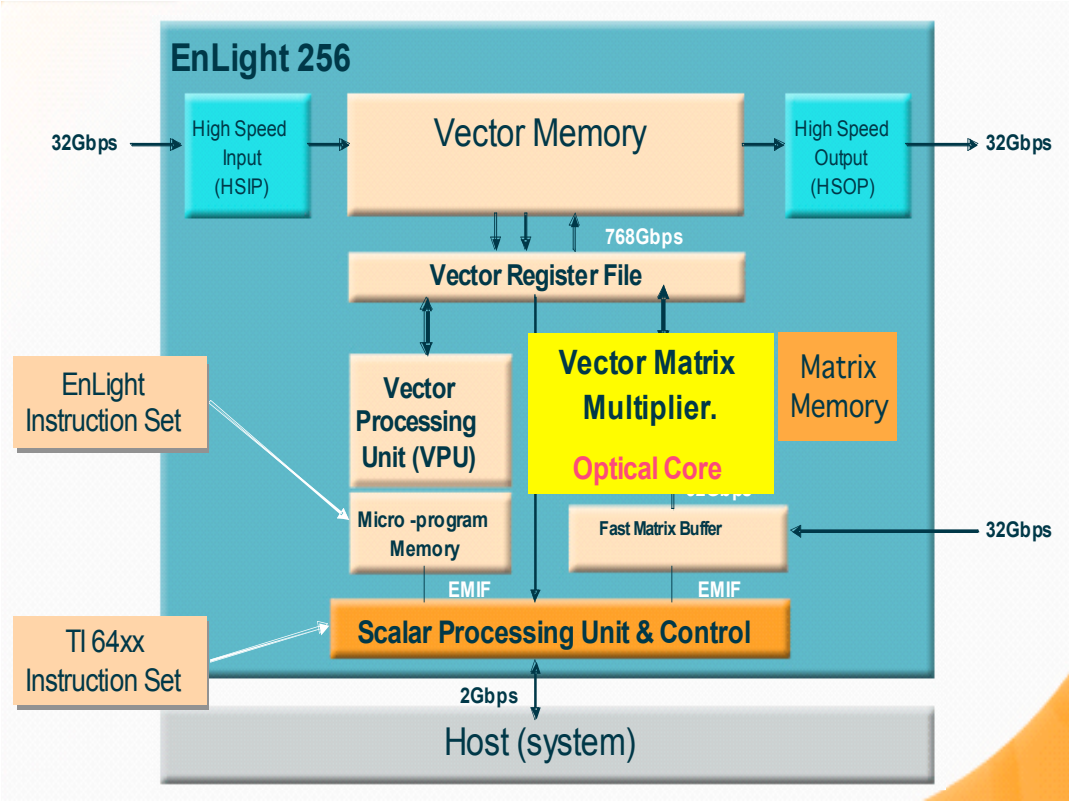
SRC Computers' MAP Processor

- Augment cluster hardware with MAP processors directly connected in memory slots
- Two User Programmable Logic devices available
  - Finally getting to regime of 64b FP math
- Multiple banks of On-Board memory maximizes local memory bandwidth
- Programming environment targeting scientific applications
  - Fortran support!



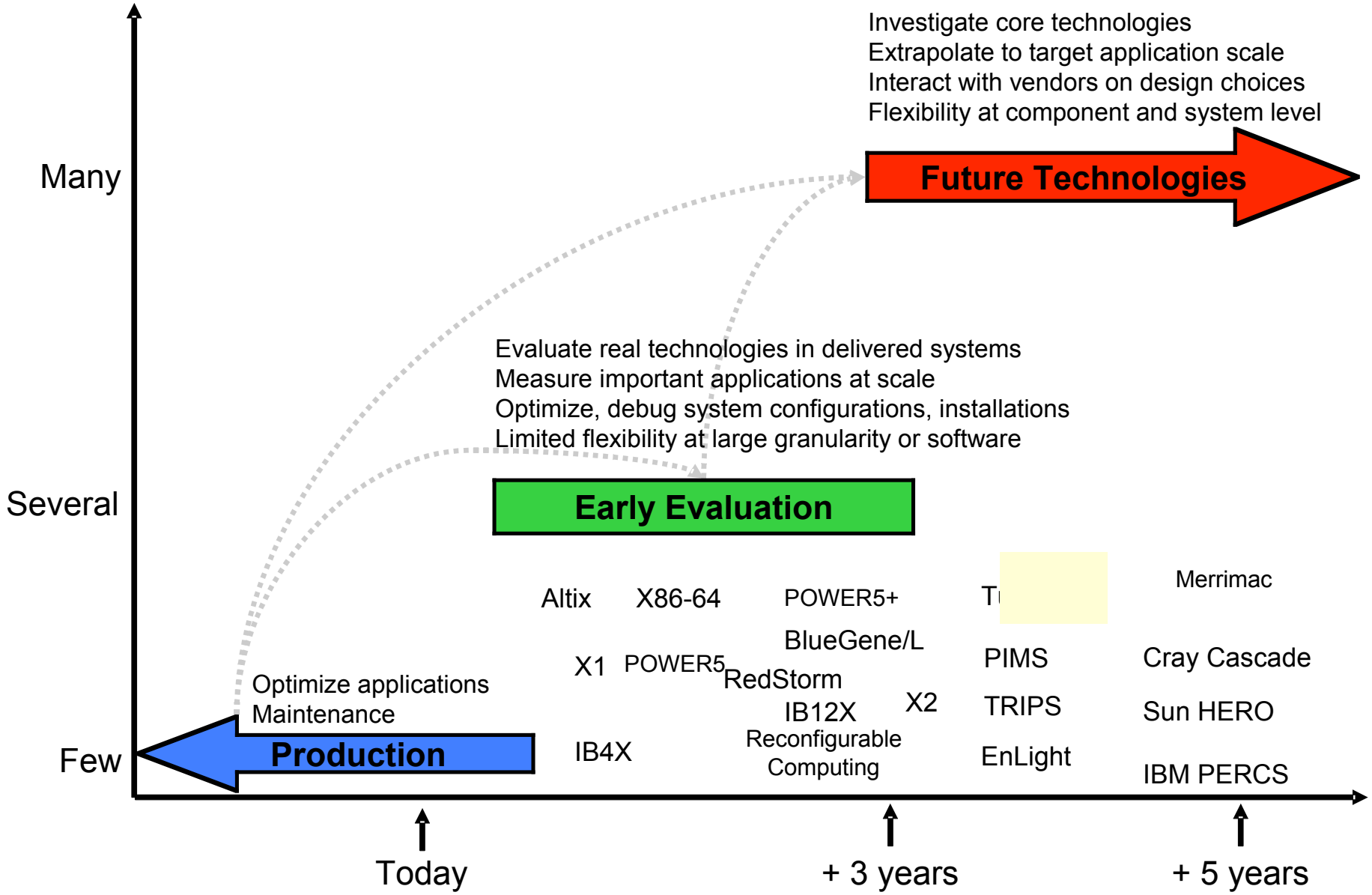
Enlight256® Optical Digital Signal Processor

- First product in a line of ultra-fast Digital Signal Processors, specified to 8-Tera (10^{12}) Multiply Accumulate operations per second
- This performance mark is about 3 orders of magnitude better than current state of the art DSP's
- The device is based on a 3-core engine: An optical Vector-Matrix Multiplier (VMM), a Vector Digital Processing Unit (VPU), and an off-the-shelf scalar DSP



Limited Precision: 8-bit, 12-bit

Future Technologies R&D complements Production Computing and Early Evaluation



Integrate core capabilities to deliver computing for the frontiers of science

Develop and evaluate next-generation architectures with industry

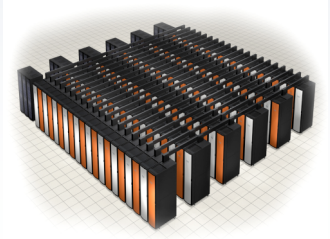
IBM

SRC

CRAY

sgi

Provide leadership-class computing resources for the Nation



Create math and CS methods to enable use of resources

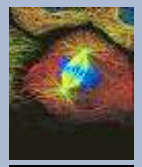
SciDAC ISICs

Scientific Applications Partnerships

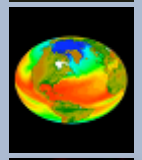
Modeling and simulation expertise

Transform scientific discovery through advanced computing

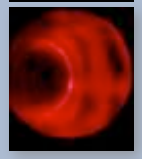
Computational End Stations



Biology



Climate



Fusion

Materials

Industry/ other agencies

National Leadership Computing Facility
2005 Call for Proposals

Eligibility

- Principal investigators engaged in scientific research with intent to publish results in open peer-reviewed literature are eligible
- Program specifically encourages proposals from universities, other research institutions and industry

Applications will be accepted electronically

- Call for proposals to be issued soon.
- Awards are expected to be announced by late July, 2005, and access to NLCF facilities for awardees established on October 1, 2005, and remain in effect until September 30, 2006

Thank You