

Critical Software Needs for Petascale Computing

Al Geist

Oak Ridge National Laboratory
Oak Ridge, TN U.S.A

SOS 10 Conference
Maui, Hawaii
March 7, 2006

Research sponsored by U.S. Department of Energy

DOE Office of Science leaps to Petascale

Ray Orbach DOE FY07 budget rollout (2-14-2006)

Increase Funding for Leadership Class Computing at Three Science labs

ORNL 250 TF Cray XT3 in FY07

1000 TF Cray XT by FY08

ANL 100 TF IBM Blue Gene in FY07

NERSC 150TF production system

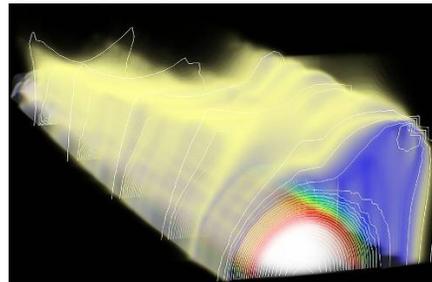
**Are we ready?
Is Science ready?**

**Ready or not
here we come!**



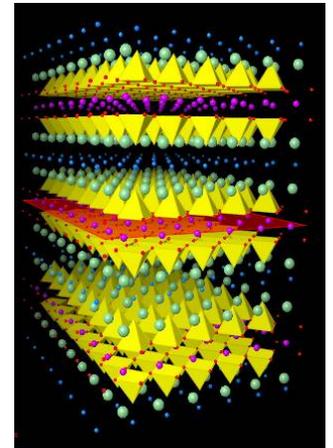
Hero efforts

Fusion



Largest simulation of plasma behavior in a tokamak

Superconductivity



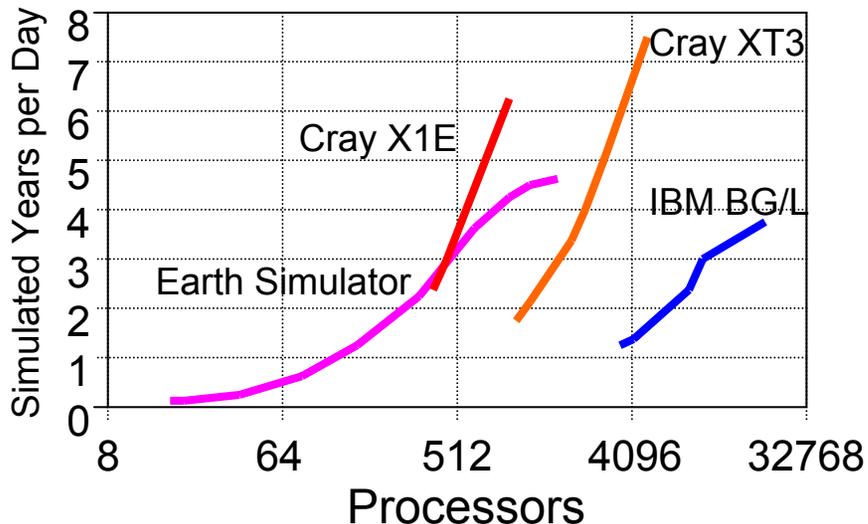
First solution of 2D Hubbard Model

Science Performance of Cray XT3

Consistently high performance across wide range of science apps and shows consistently superior scaling



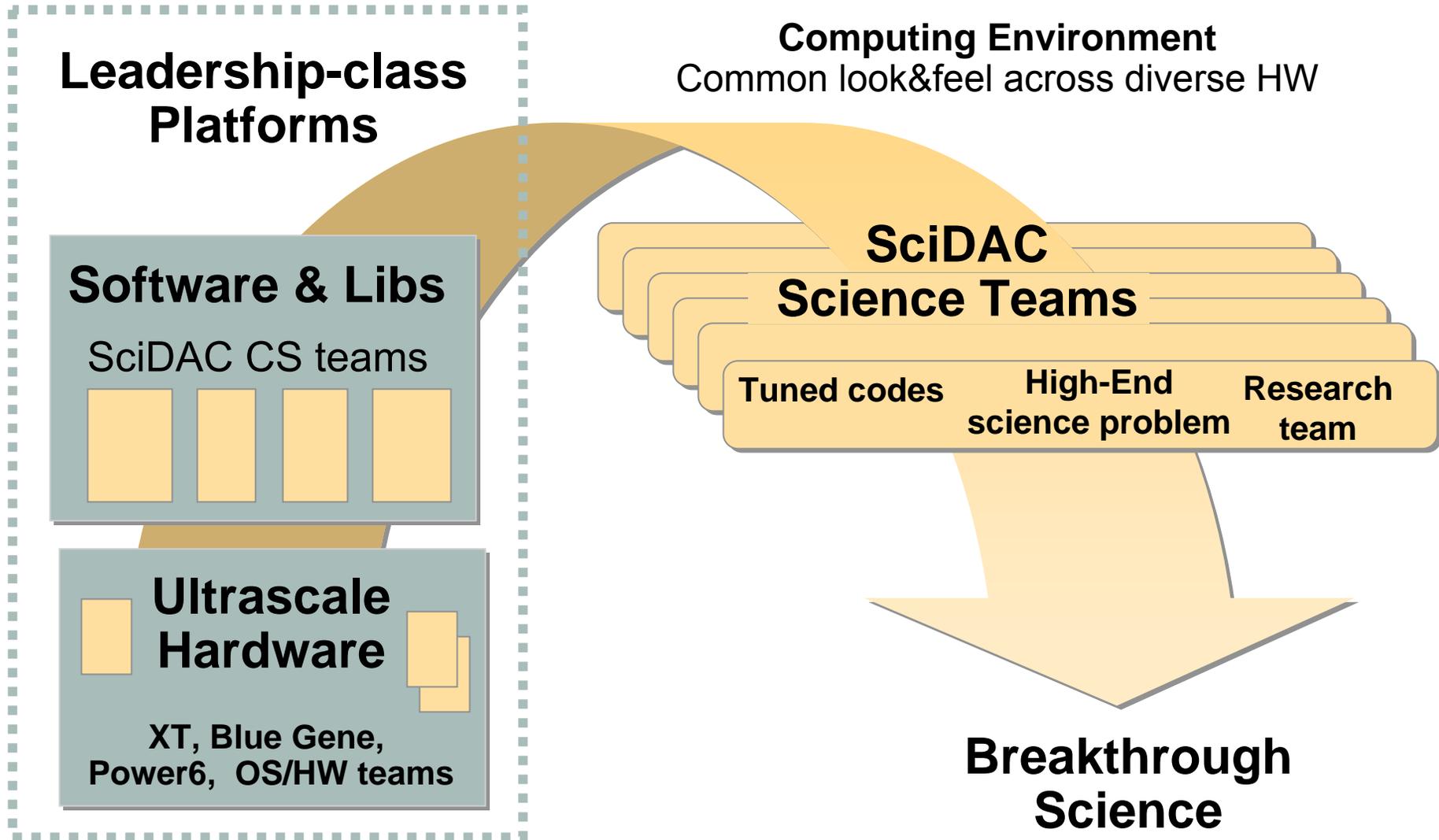
POP 1.4.3 tenth-degree benchmark



- Astrophysics
 - Better than Power 5 on VH1
- Biochemistry
 - Excellent LAMMPS scaling
- Climate
 - Tenth-degree POP (shown here)
- Combustion
 - Best S3D scaling
- Fusion
 - Beats even Cray X1E on GYRO
 - Beats all but Earth Simulator on GTC
 - Made biggest AORSA run ever
- Groundwater
 - Excellent PFLOW scaling

For why - Bill Camp's talk next

Accelerating Science Breakthroughs is the Goal – What are SW needs?



Critical SW Gaps opening up

Heterogeneous leadership-class machines

Sysadmins need scalable consistent system tools to manage machines
Science teams need to have a robust environment that presents similar programming interfaces and tools across the different machines.

Performance of File System and I/O

increasing demands of scalability, fault tolerance and data intensive computing drive I/O needs but disk speeds not keeping up

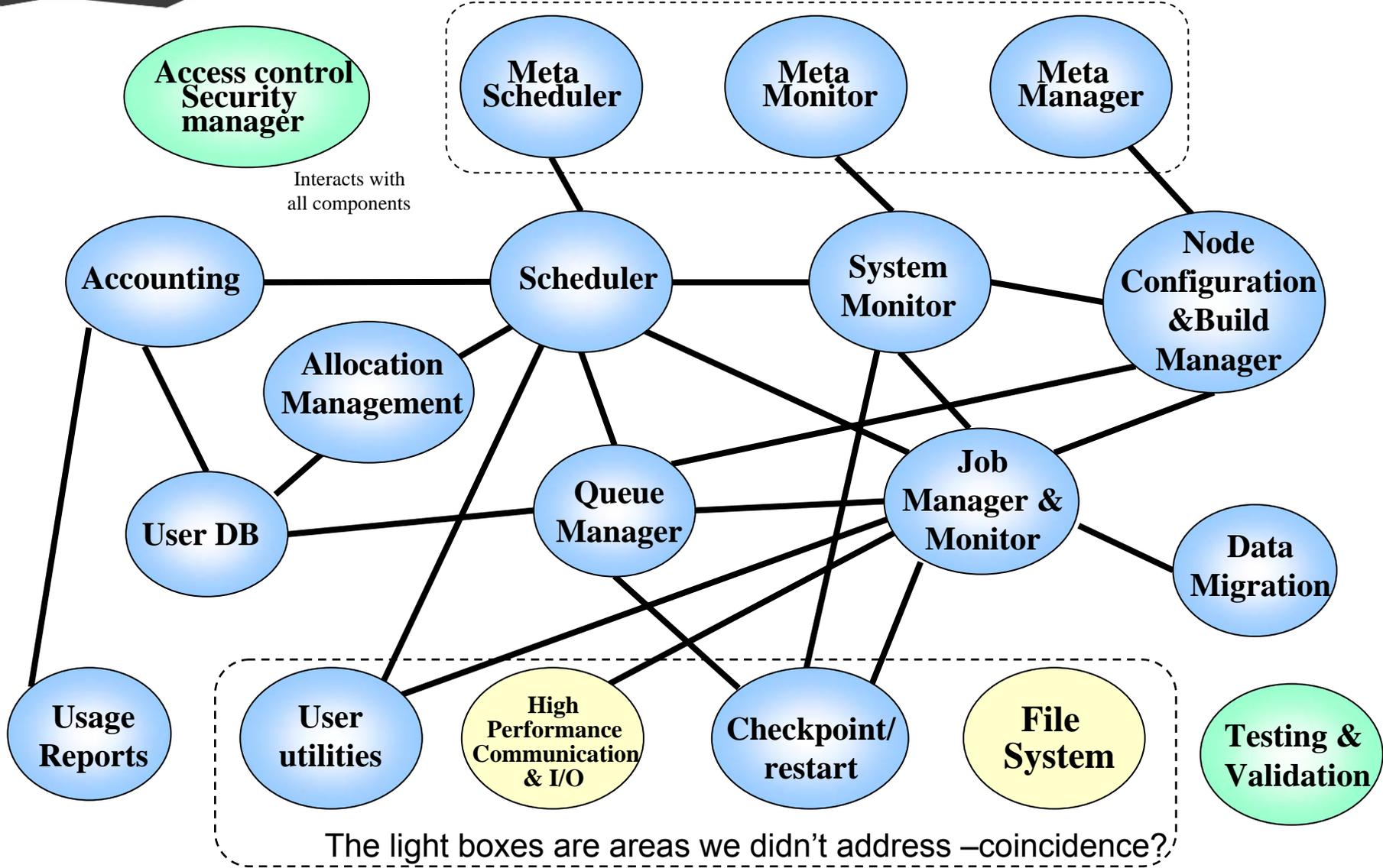
Fault tolerance requirements in apps and systems software

if the app doesn't complete there may be no data to be intensive about

Support for application users submitting interactive jobs

computational steering as a way to shift from data intensive to knowledge intensive computing

Scalable System Software Project



Data Intensive I/O for Petascale

I/O BW has lowest exponential growth rate – thus this hurdle will continue to get harder – hence the term “hard drive”

Technology	Growth e^k
Transistors per chip	$K= 0.46$
LINPACK on TOP10	0.58
Capacity hard drive	0.62
GB disk per dollar	0.83
I/Os per second	0.20
BW from hard drive	0.26

Assume a “balanced” 1 PF machine with 1 TB/s I/O.

500 controllers and thousands of disks

Could cost \$30M-\$60M in 2008

A 1 TB/s filesystem in 2008 requires approximately 2,500 file servers

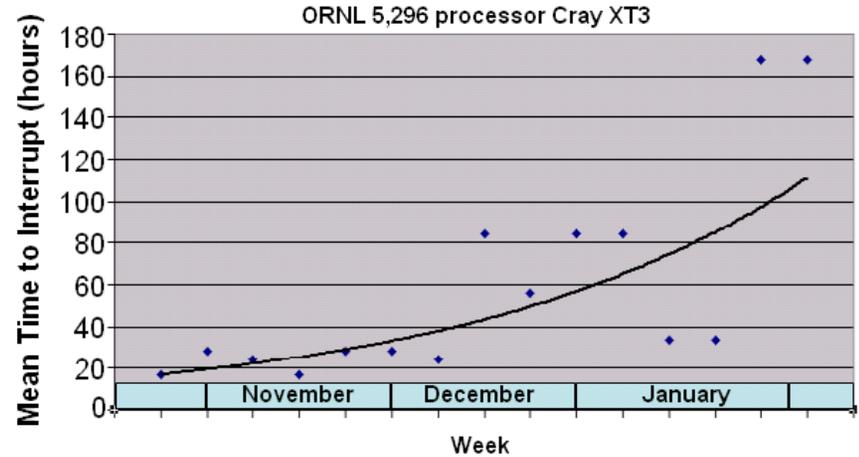
One choice is to reduce the I/O bandwidth balance



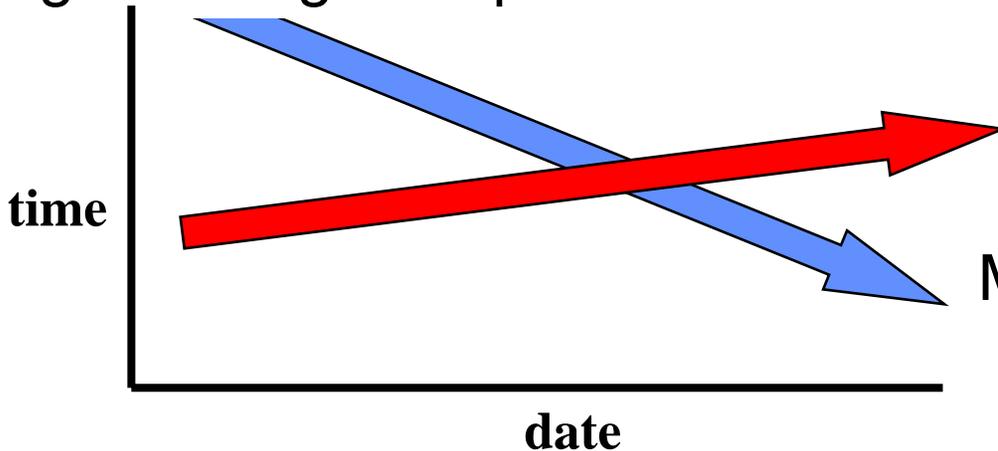
But this increases the time to checkpoint applications
A common I/O intensive job for large applications

The End of Fault Tolerance as We Know It

Petascale systems will have 100,000+ processors on multi-core chips. The time before **some** failure will be measured in minutes. Checkpointing and restarting this large a system could take longer than the time to the next failure!



Cross over point: Time to checkpoint grows larger as problem size increases



Good news is the MTBF is better than expected for BG/L and XT3 – days not minutes

MTBF grows smaller as number of parts increases

The System Can't Ignore Faults

Even if system survives, application often doesn't



Fault tolerance in systems software – the best type of fault is the one that never happens

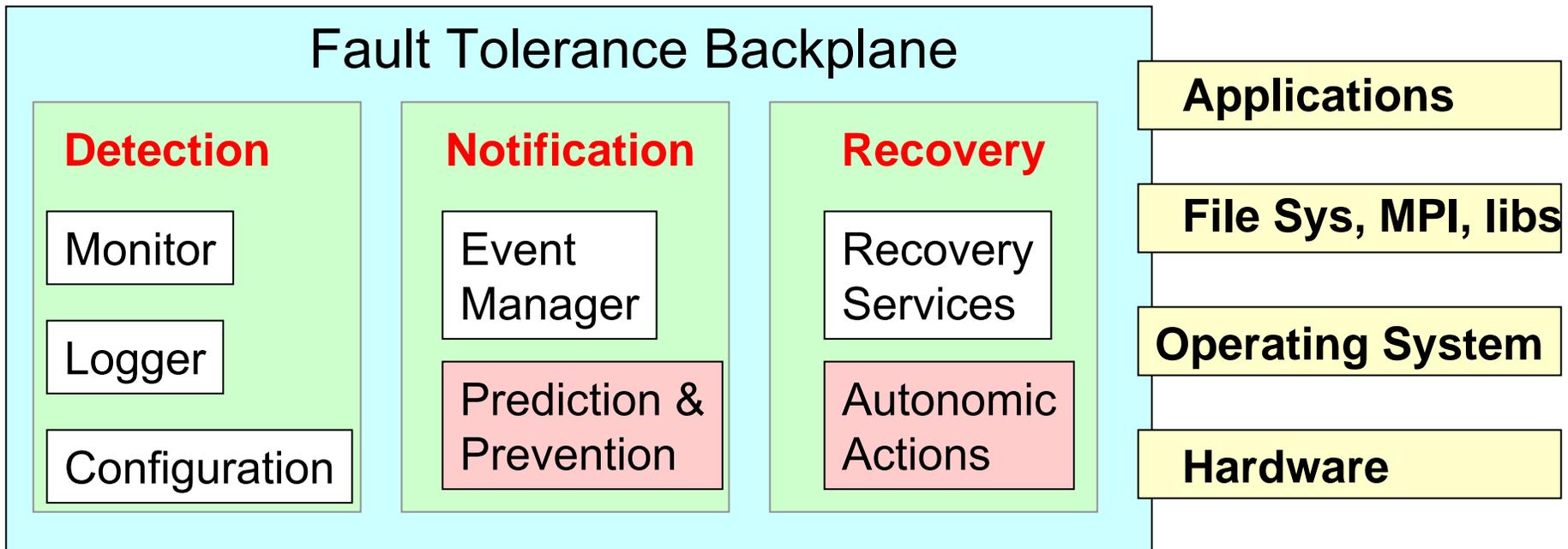
- Ideal case is prediction and prevention
- Survivability and resiliency when faults can not be avoided i.e. unpredictable events

Holistic Fault Tolerance

- Research into active fault management that takes into account the full impact of faults across application, middleware, OS, and hardware (heterogeneous nodes – compute, I/O, service)

Holistic Fault Tolerance

Backplane to provide Fault awareness, prediction and recovery across the entire HPC system from the application to the hardware



Lots to be learned about how middleware should survive

Enhancing the Application Software Ecosystem

With petaflop systems only 2-3 years away, took a critical look at the potential software gaps that could affect the ability of DOE's Leadership Computing Facilities from delivering immediate successes in scientific discovery

Divided into two tiers:

1. Critical - software required for "Hero" programmers to make science breakthroughs on petascale system
2. Important - software required for the typical LCF user to make productive, efficient use of the petascale system.

Critical Software for Heroic App Scientists

These scientists are used to many hardships associated with running on the biggest computers in the world, but even they need a working operating system, a file system, a tuned MPI, fast numerical libraries, and a computer that stays up long enough for them to get their science done.

For Cray XT series this means:

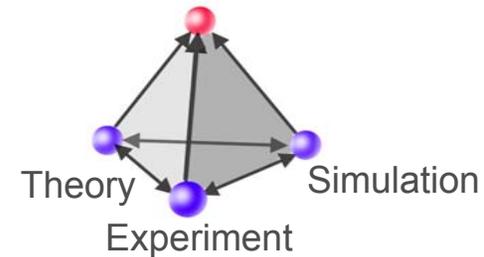
- Highly tuned MPI and Math Libraries
- Light Weight Kernel – Linux version tuned and stable
- Petascale I/O and file system
- Networking the resources to the world
- Reliability and Fault Tolerance for
 - science applications, middleware, and system software

Plug and Play Petascale

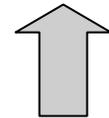
The long term productivity of the Leadership computers will require a lot of work in creating a programming environment and runtime interface that allows top scientists, who may not be experts on the Leadership systems, to produce efficient, scalable applications.

- **Common Programming Environment:** Application scientists expend considerable time and effort dealing with development, deployment, and run time interfacing activities that are significantly different on each high-end platform.
- **Advanced Debugging:** Petascale systems will present significant challenges for debugging.
- **Automated Performance Tuning:** users will need performance tools that are easy to use and able to automate the performance tuning process.

Knowledge



Information



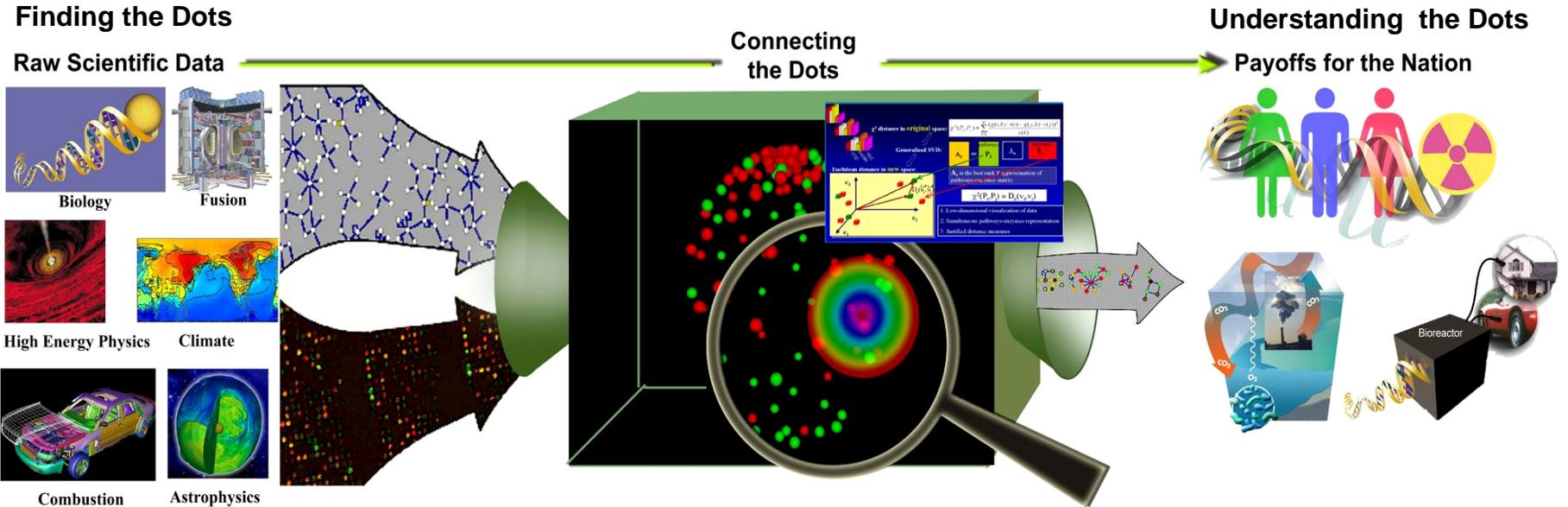
Petabytes Data



Connecting the Dots in Science

From Data Intensive to Knowledge Intensive Computing

Interactive applications benefit all three stages of discovery



Sheer Volume of Data (from Experiment and Simulation)

Climate

Now: 20-40 Terabytes/year
5 years: 5-10 Petabytes/year

Fusion

Now: 100 Megabytes/15 min
5 years: 1000 Megabytes/2 min

Advanced Mathematics and Algorithms

- Huge dimensional space
- Combinatorial challenge
- Complicated by noisy data
- Interactive steering of simulations

Providing Predictive Understanding

- New energy economies H_2 , nuclear
- Weather and climate
- Biology and health care

Questions?

Backup Slides



Advanced Scientific Computing Research (ASCR)

Research in applied mathematics and computer science to be funded at FY2006 levels. Funding is requested to:

- Advance the underlying mathematical understanding of physical, chemical and biological systems of interest to the DOE.
- Underpin further the development of advanced algorithms to describe, model and simulate complex systems.
- Ensure the effective utilization of high-performance computers to advance science in areas important to the DOE mission.

Increase in Computational Partnerships, fostered by

- The re-competition of SciDAC activities and
- The initiation of university based SciDAC Institutes in FY2006.

High-performance computing and network facilities and testbeds

- Increase funding for Leadership Class Computing
 - Upgrade capability to 250 Teraflops at ORNL, Upgrade to 1,000 TF by the end of FY 2008
 - Acquire up to 100 Teraflops of high-performance computing capability with low electrical power requirements at ANL.
- Increase funding for high performance production computing
 - Increase capacity at NERSC to 100-150 teraflops.
- Increase funding for ESnet to realize the promise of optical networks for DOE's science research missions.

