ORNL Cray XT3/4 Overview

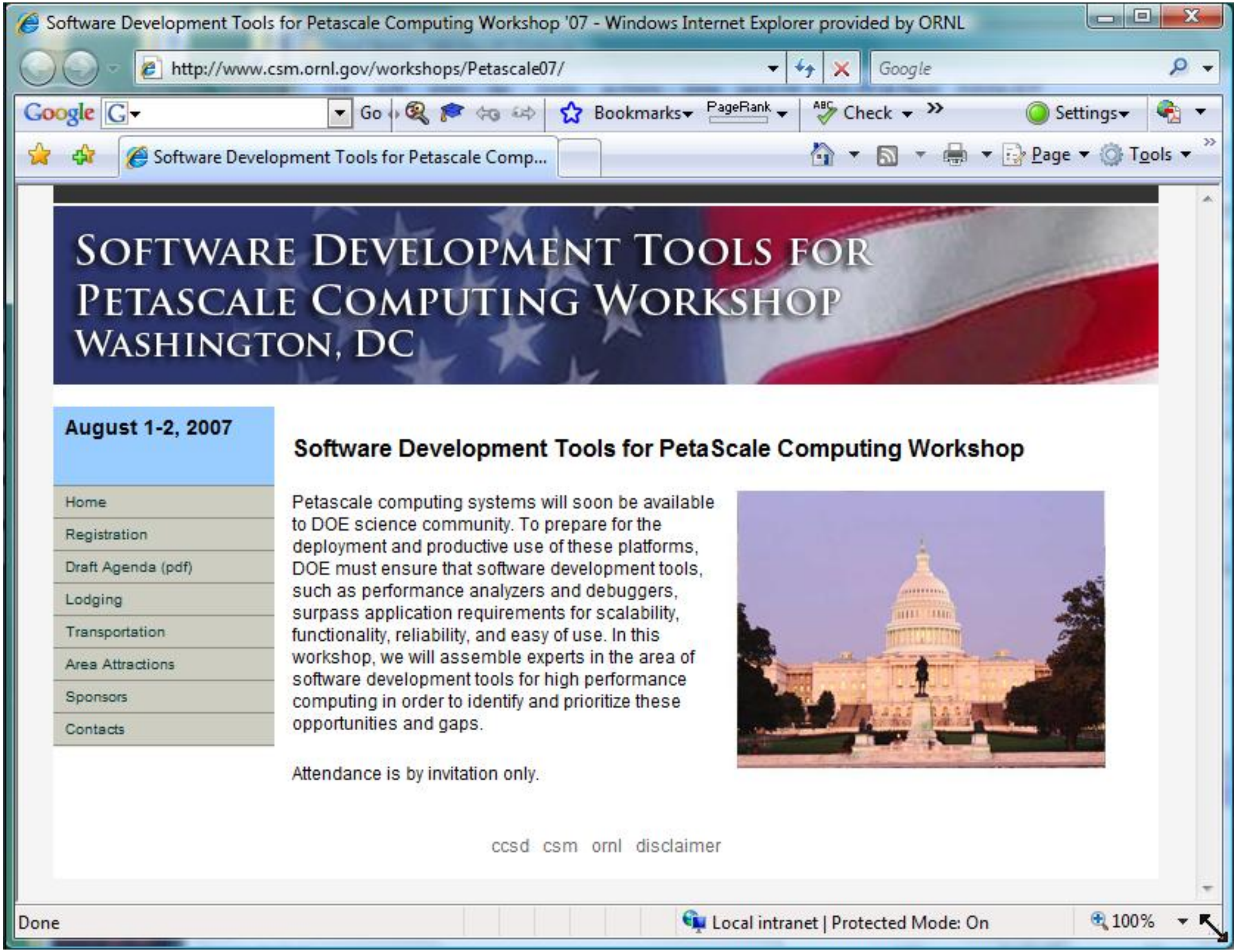
Jeffrey S. Vetter

And a cast of dozens …

Cray, NCCS, Pat Worley, Sadaf Alam, Weikuan Yu

http://ft.ornl.gov, http://nccs.gov

Software Development Tools for Petascale Computing Workshop on Aug 1-2


Software Development Tools for Petascale Computing Workshop '07 - Windows Internet Explorer provided by ORNL
http://www.csm.ornl.gov/workshops/Petascale07/
Google
Go
Bookmarks
PageRank
Check
Settings
Software Development Tools for Petascale Comp...
Page
Tools
Done
Local intranet | Protected Mode: On
100%

SOFTWARE DEVELOPMENT TOOLS FOR PETASCALE COMPUTING WORKSHOP WASHINGTON, DC

August 1-2, 2007

- Home
- Registration
- Draft Agenda (pdf)
- Lodging
- Transportation
- Area Attractions
- Sponsors
- Contacts

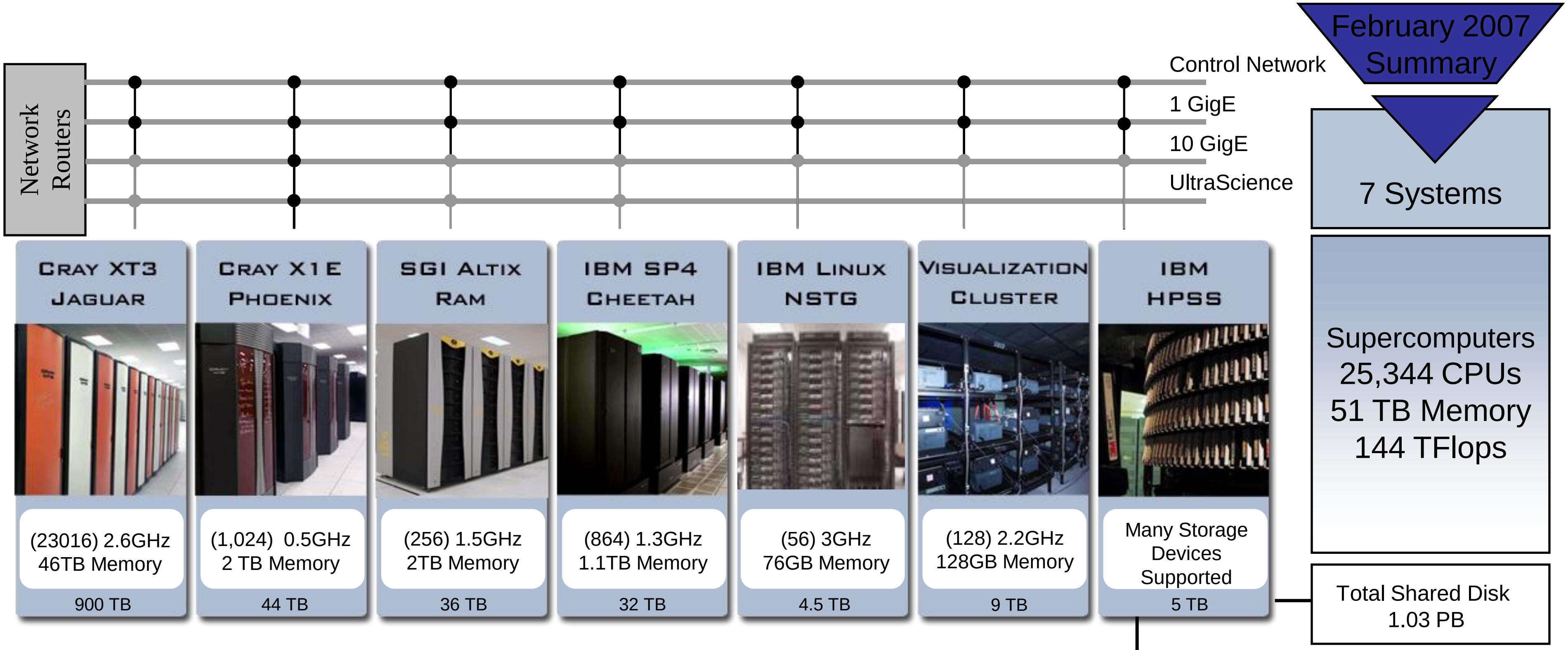
Software Development Tools for PetaScale Computing Workshop

Petascale computing systems will soon be available to DOE science community. To prepare for the deployment and productive use of these platforms, DOE must ensure that software development tools, such as performance analyzers and debuggers, surpass application requirements for scalability, functionality, reliability, and easy of use. In this workshop, we will assemble experts in the area of software development tools for high performance computing in order to identify and prioritize these opportunities and gaps.

Attendance is by invitation only.

ccsd csm ornl disclaimer

Current NCCS Resources



Cray XT3 Jaguar	Cray X1E Phoenix	SGI Altix Ram	IBM SP4 Cheetah	IBM Linux NSTG	Visualization Cluster	IBM HPSS
(23016) 2.6GHz 46TB Memory	(1,024) 0.5GHz 2 TB Memory	(256) 1.5GHz 2TB Memory	(864) 1.3GHz 1.1TB Memory	(56) 3GHz 76GB Memory	(128) 2.2GHz 128GB Memory	Many Storage Devices Supported
900 TB	44 TB	36 TB	32 TB	4.5 TB	9 TB	5 TB

7 Systems

Supercomputers
25,344 CPUs
51 TB Memory
144 TFlops

Total Shared Disk
1.03 PB

Scientific Visualization Lab

27 projector,
35 megapixel,
Power Wall

Test Systems

- 96 processor Cray XT3
- 32 processor Cray X1E*
- 16 Processor SGI Altix

Evaluation Platforms

- 144 processor Cray XD1 with FPGAs
- SRC Mapstation
- Clearspeed

Backup Storage

10PB

10 PB

CSM Experimental Computing Lab houses IBM CELL, GPUs, Clearspeed, FPGAs, etc

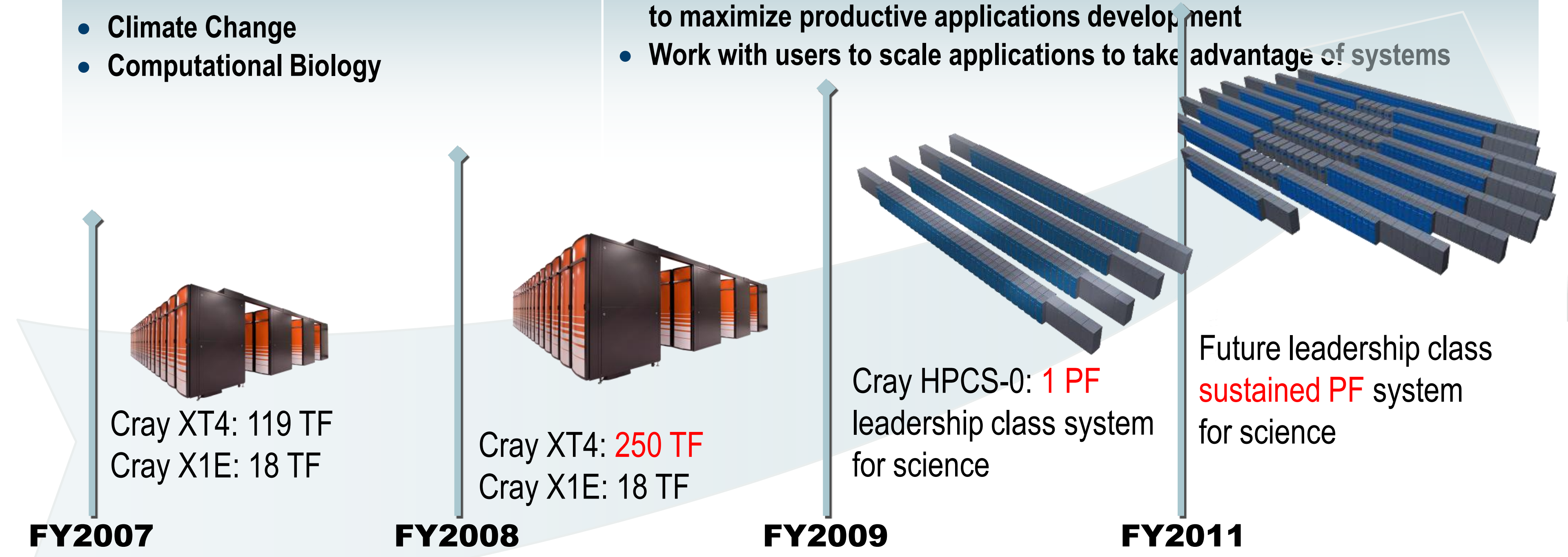
NCCS Roadmap for Leadership Computing

Mission: Deploy and operate the computational resources needed to tackle global challenges

- Future Energy
- Understanding the universe
- Nanoscale materials
- Climate Change
- Computational Biology

Vision: Maximize scientific productivity and progress on the largest scale computational problems

- Providing world class computational resources and specialized services
- Providing a stable hardware/software path of increasing scale to maximize productive applications development
- Work with users to scale applications to take advantage of systems



Cray XT4: 119 TF
Cray X1E: 18 TF

Cray XT4: 250 TF
Cray X1E: 18 TF

Cray HPCS-0: 1 PF leadership class system for science

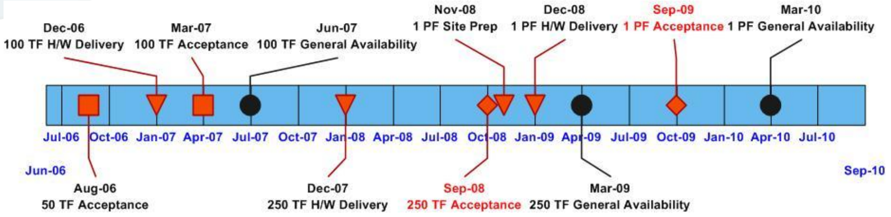
Future leadership class sustained PF system for science

FY2007

FY2008

FY2009

FY2011



Dec-06 100 TF H/W Delivery

Mar-07 100 TF Acceptance

Jun-07 100 TF General Availability

Nov-08 1 PF Site Prep

Dec-08 1 PF H/W Delivery

Sep-09 1 PF Acceptance

Mar-10 1 PF General Availability

Jun-06

Aug-06 50 TF Acceptance

Dec-07 250 TF H/W Delivery

Sep-08 250 TF Acceptance

Mar-09 250 TF General Availability

Sep-10

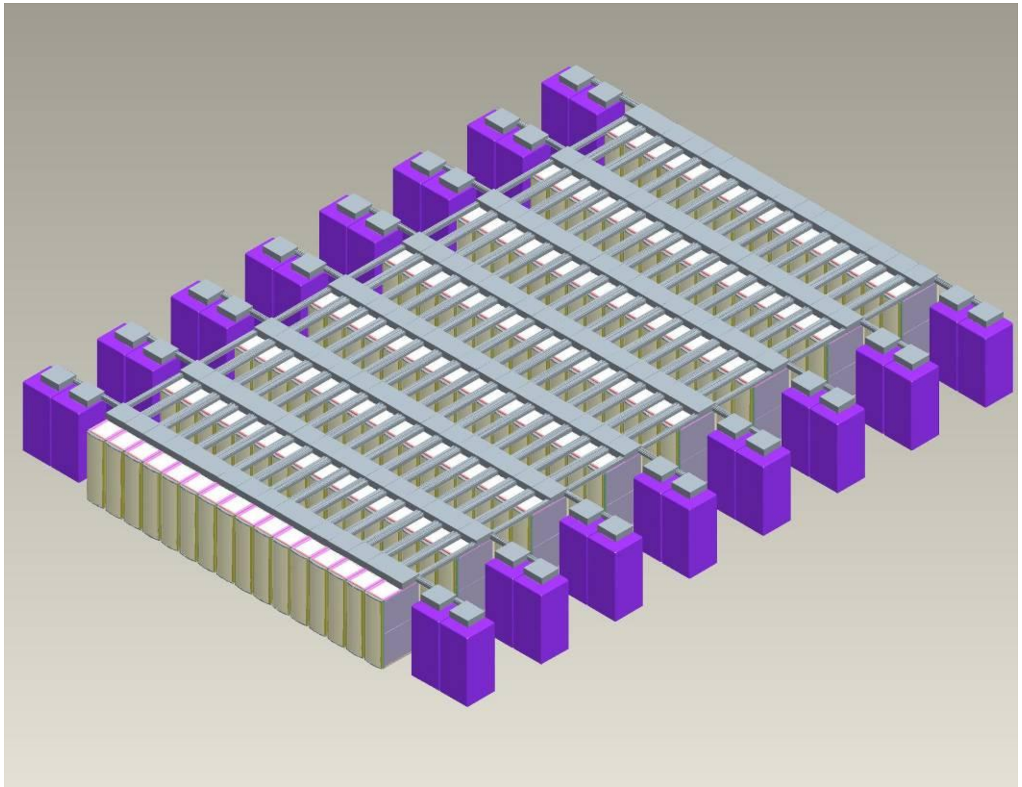
Jaguar – Cray XT4 with 11,706 Dual-Core AMD Opteron Processors

- 119 TF peak performance
- 46 TB main memory
- 2.5 MW of power (20 KW per rack)
- 124 cabinets – 96 Opterons per cabinet
- Air cooled bottom-to-top
- Single 3000 CFM variable speed fan per rack
- 2.3 miles of interconnect cables
- Upgrade to Quad-core processors in Fall, 2007

1000 TF Cray “Baker” system in 2008

System configuration

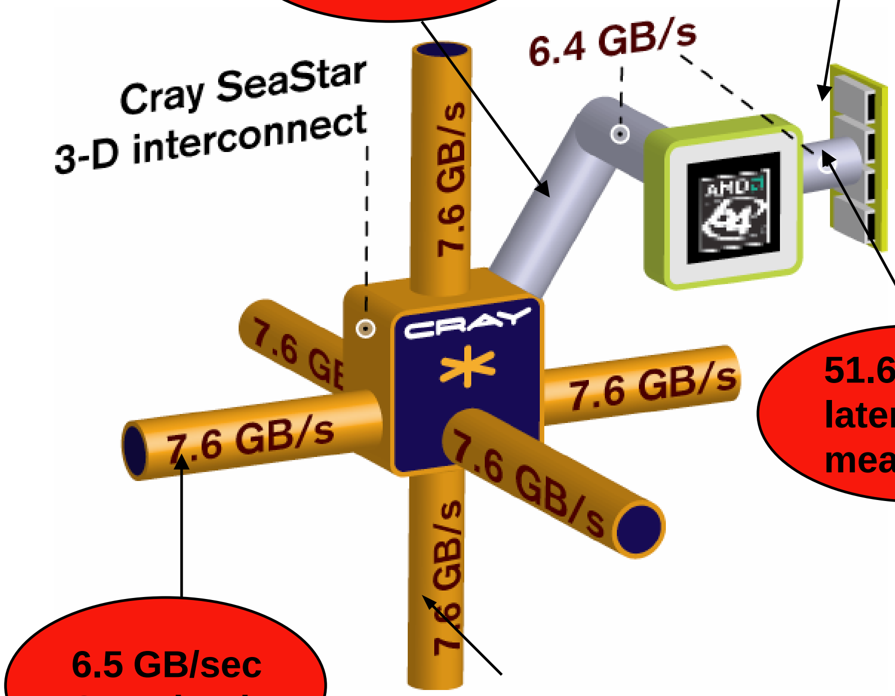
- 1 PF peak
- ~24,000 quad-core processors
- ~50 KW per cabinet
- ~7 MW power
- Over 3,000 watts/ft^2
- 40+ heat exchange units (chilled water to R-134a)



1 PF Cray system in 2008

Used by permission: Cray, Inc.

Cray XT3 Processing Element: Measured Performance



5.7 GB/sec Sustained

2.17 GB/sec Sustained

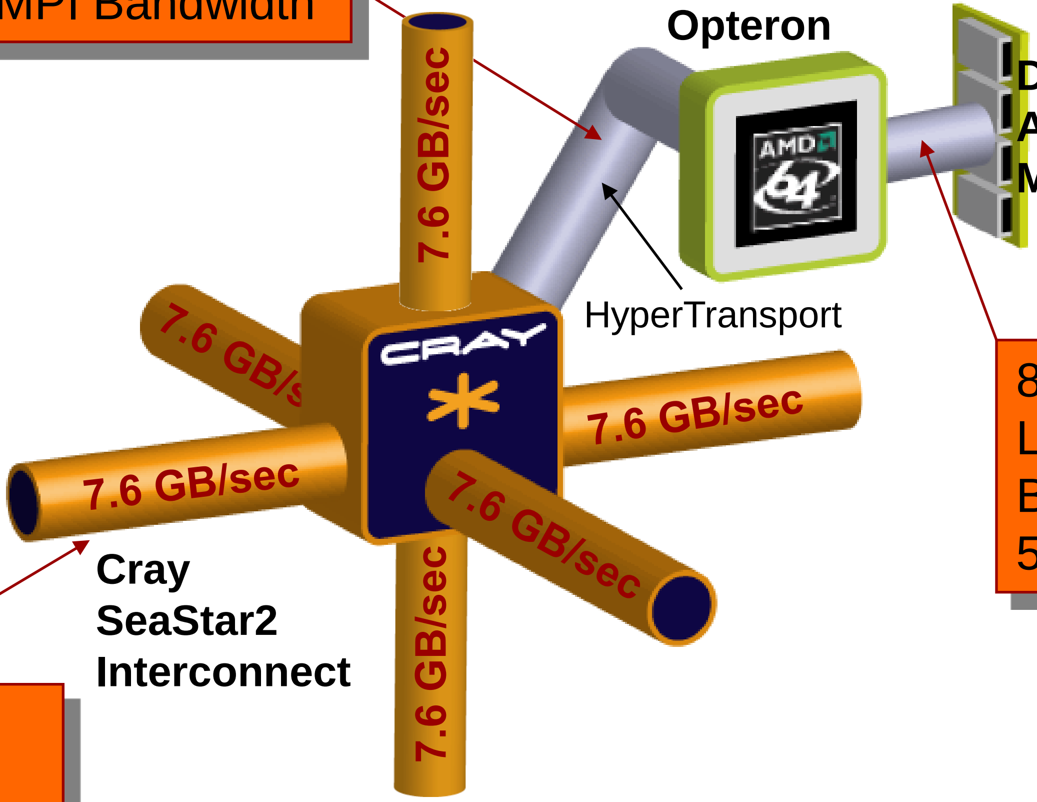
51.6 ns latency measured

6.5 GB/sec Sustained

Six Network Links
Each >3 GB/s x 2
(7.6 GB/sec Peak for each link)

- SDRAM memory controller and function of Northbridge is pulled onto the Opteron die. Memory latency reduced to <60 ns
- No Northbridge chip results in savings in heat, power, complexity and an increase in performance
- Interface off the chip is an open standard (HyperTransport)

The Cray XT4 Processing Element:
Providing a bandwidth-rich environment



4 GB/sec
MPI Bandwidth

AMD
Opteron

Direct
Attached
Memory

HyperTransport

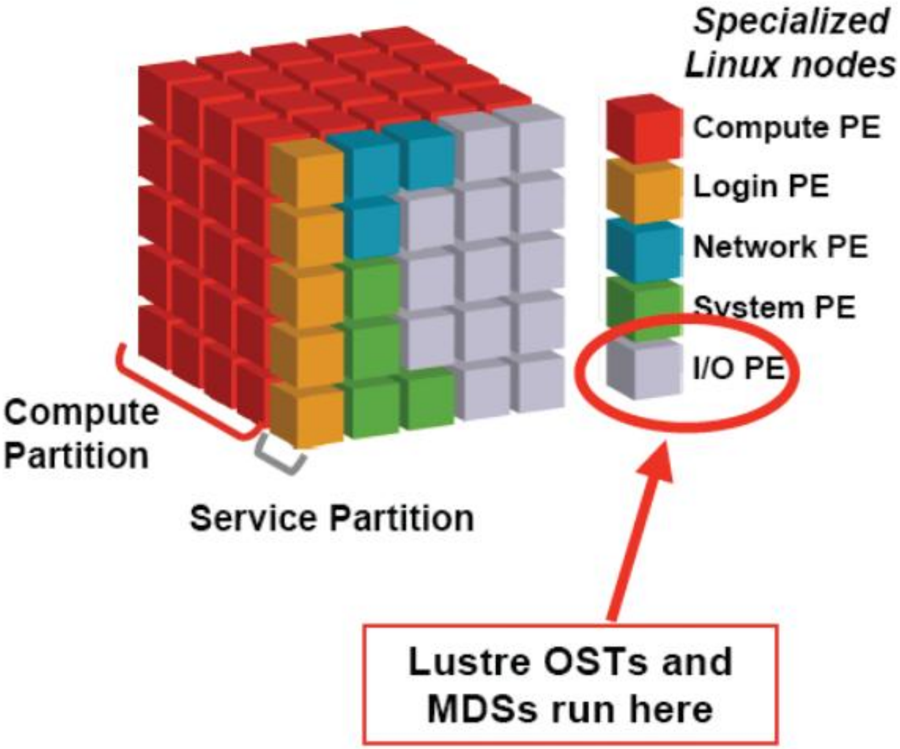
8.5 GB/sec
Local Memory
Bandwidth
50 ns latency

Cray
SeaStar2
Interconnect

6.5 GB/sec
Torus Link
Bandwidth

I/O

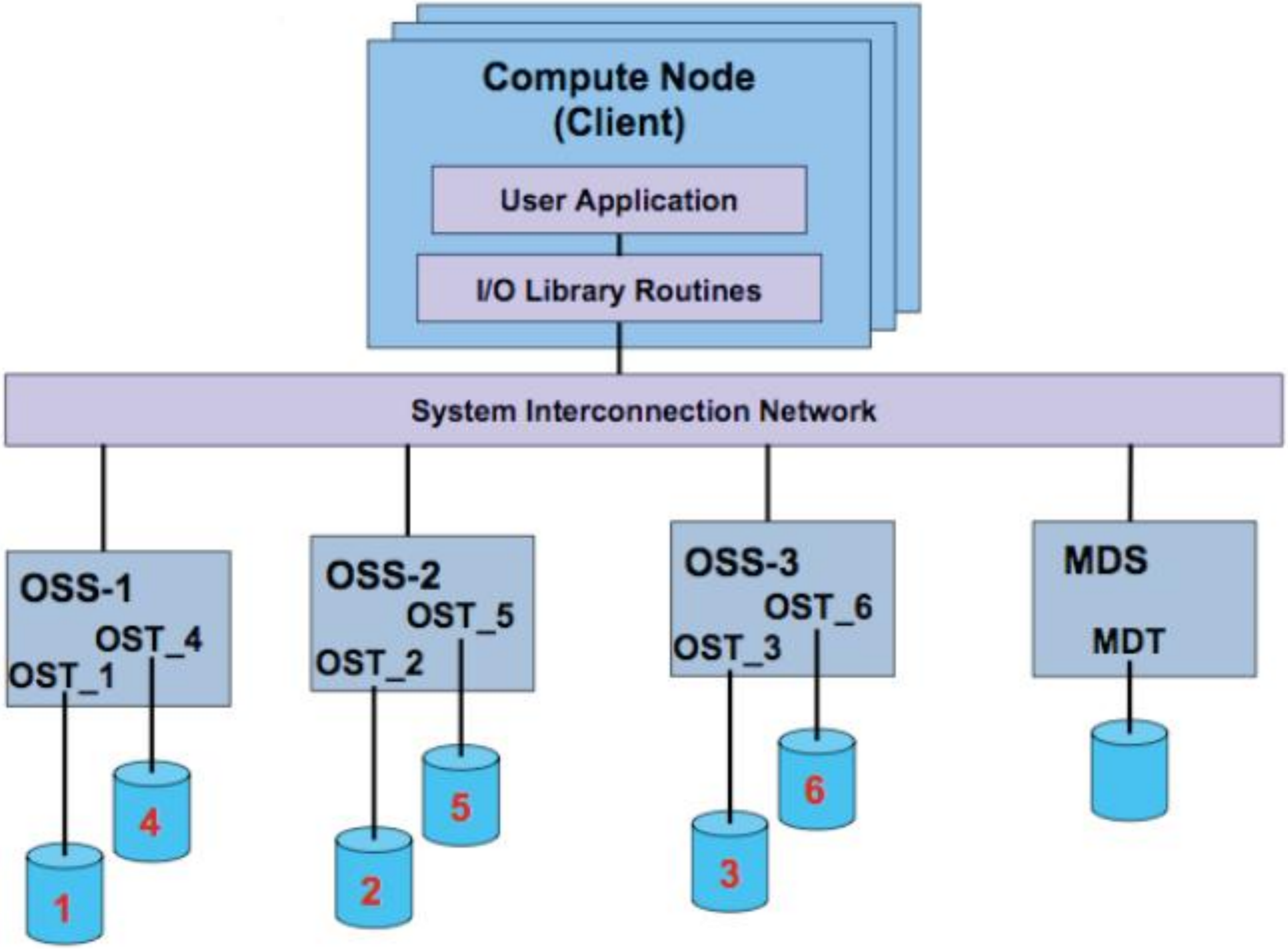
I/O Configuration



Lustre OSTs and MDSs run here

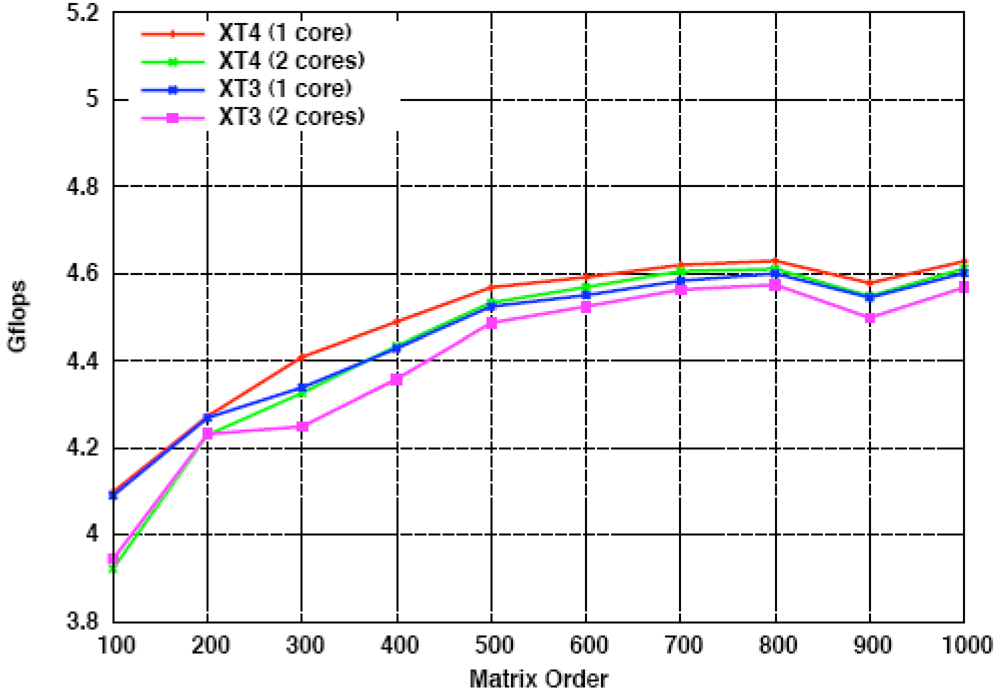
- Lustre filesystems
  - Serviced by 80 I/O nodes
  - /lustre/scr144
    - 144 OSTs
    - Peak 72 GB/s
    - Target ~48 GB/s
    - Early results
      - Read 45 GB/s
      - Write 25 GB/s
  - /lustre/scr72[a,b]
    - 72 OSTs each
    - Default scratch

User view of I/O

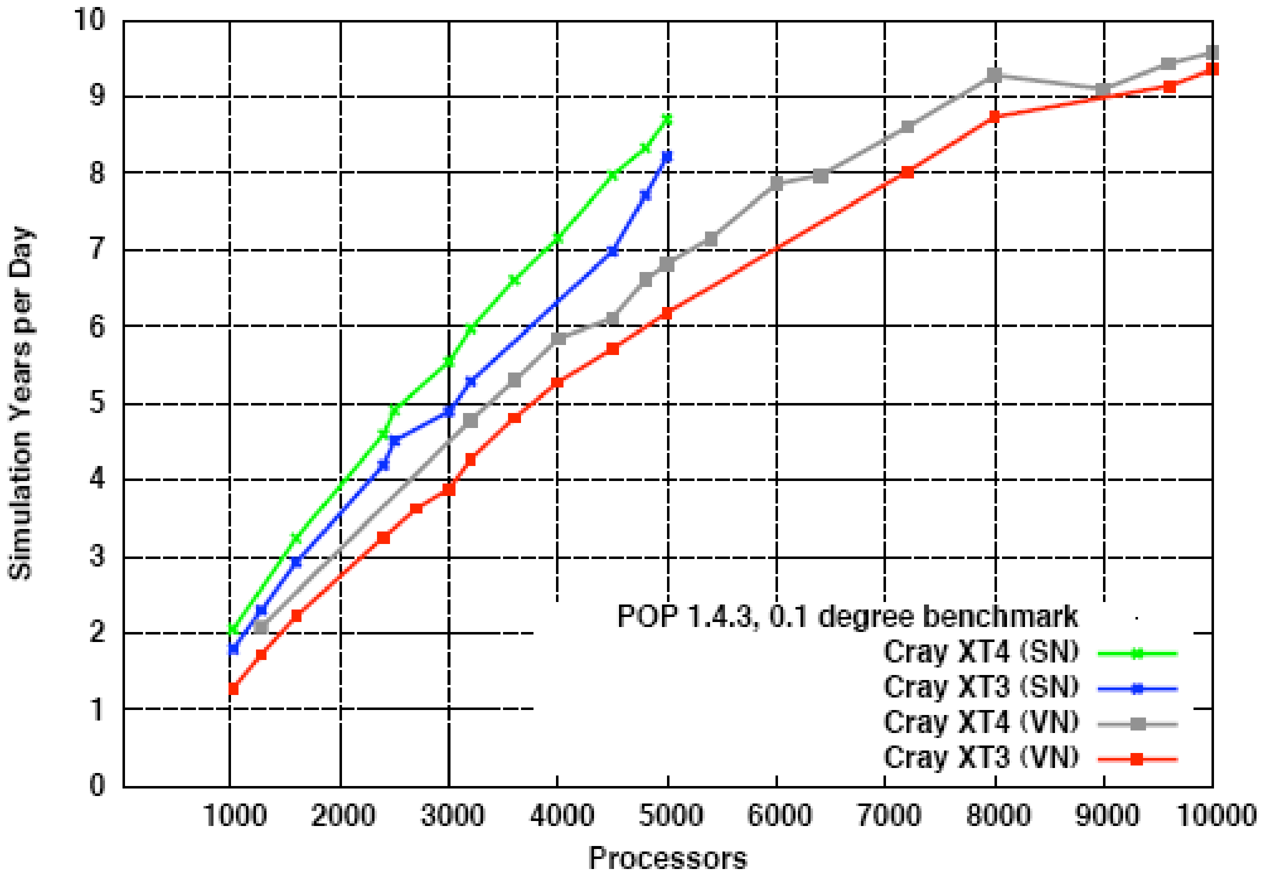


Performance

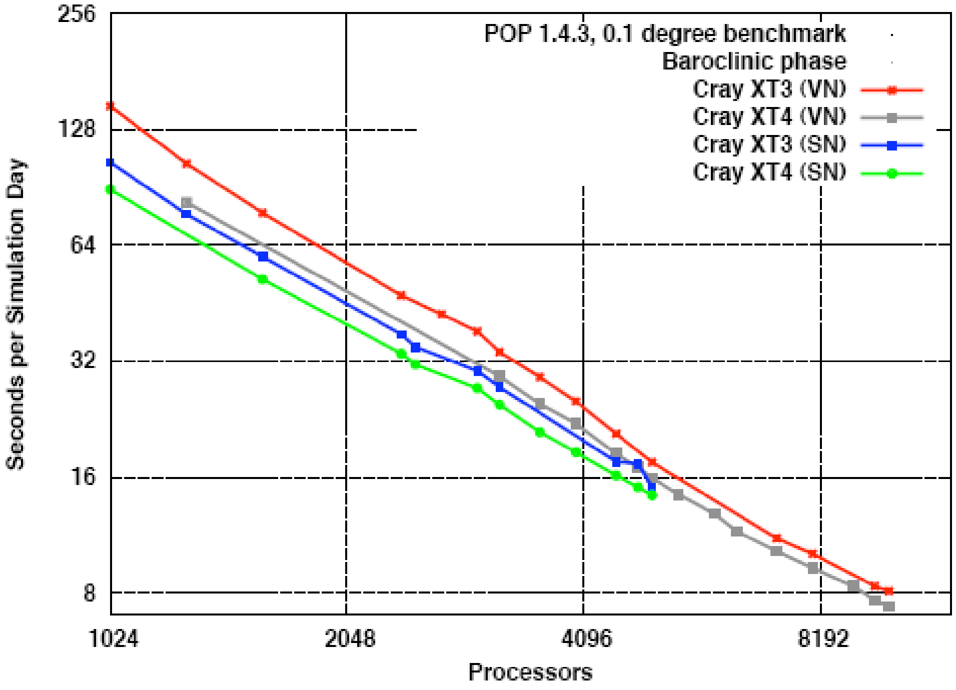
Performance > Matrix Multiply

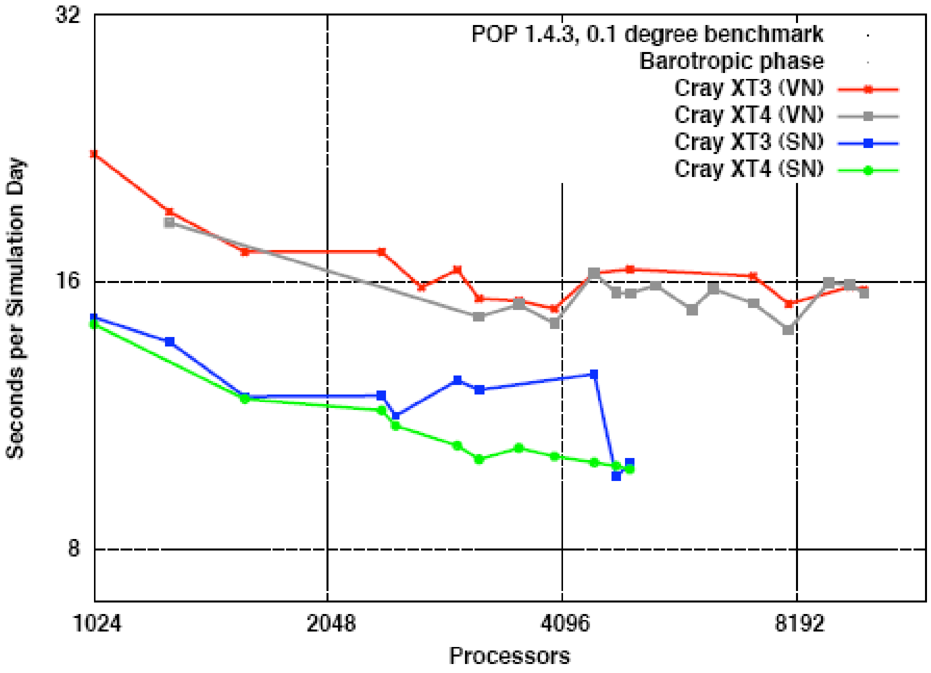


Performance > POP



Performance > POP > Phases





Performance Considerations

- Compiler options
- Page Size
  - 4KB v. 2MB (def)
- SN v. VN
- Process Mapping
  - Manage logical to physical placement of tasks
  - Cray
    - MPICH_RANK_REORDER_METHOD
    - Wrap, smp-style
- Collectives
  - MPI_COLL_OPT_ON

Software

Cray XT3/4 Software

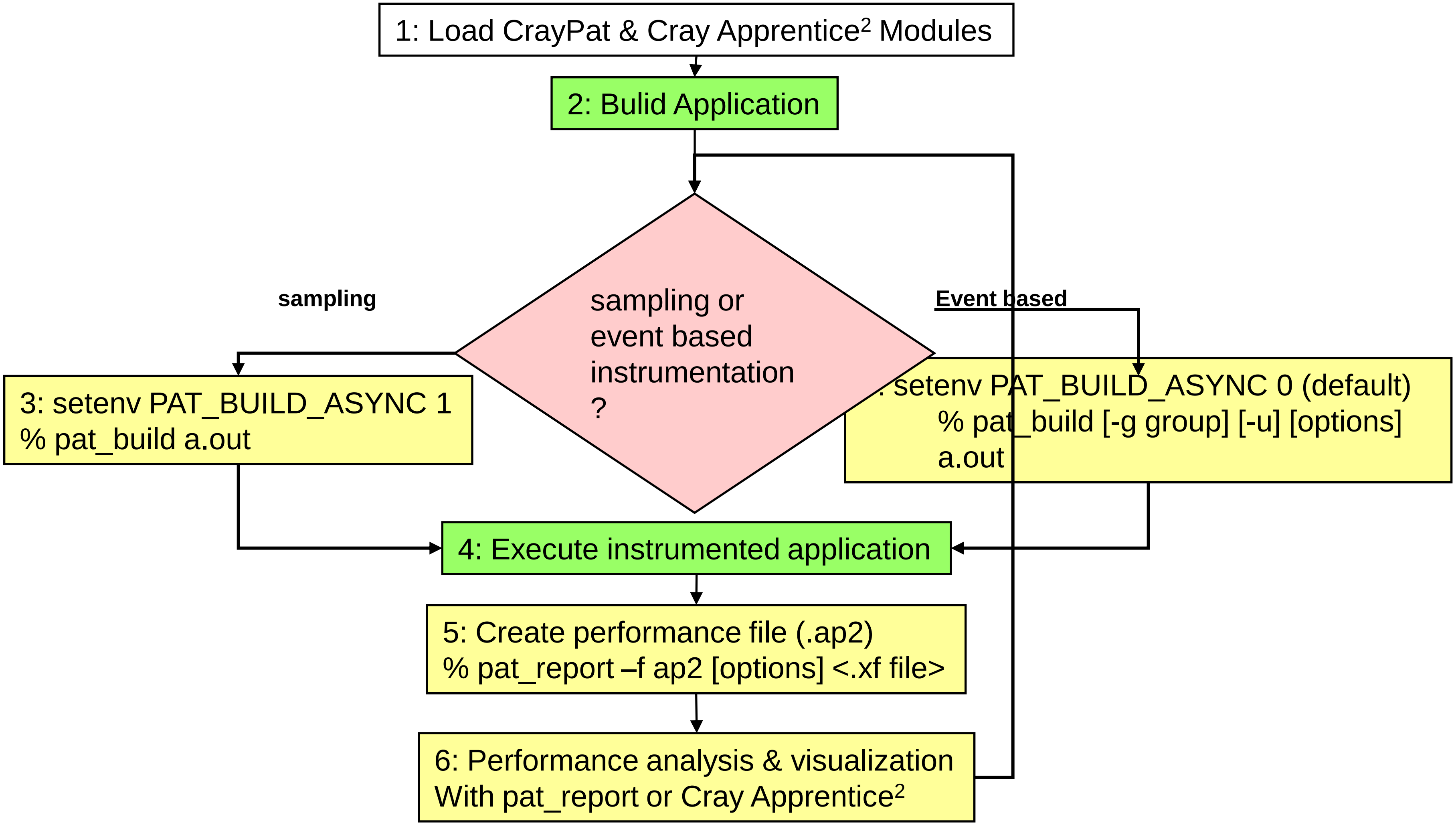
- Operating system
  - Catamount
    - Lightweight OS
  - Compute Node Linux (in testing)
    - Derived from Linux
    - Targeting quad core release
    - More functionality
      - User threads
- Filesystem
  - Lustre
- Tools
  - Performance, debugging
    - Cray PAT and Apprentice
    - Tau
    - PAPI
    - MPIP
    - Totalview
- Compilers
  - PGI, Pathscale
- PBS/Moab batch scheduler

Cray Performance Analysis Infrastructure

- ➡ **CrayPat**
  - pat_build: Utility for automatic application instrumentation
    - No source code modification required
  - run-time library for measurements
    - transparent to the user
  - pat_report:
    - Performance reports
    - Generation of file for performance visualization
  - pat_help: Runtime help utility

- ➡ **Cray Apprentice^2**
  - Graphical performance analysis and visualization tool
    - Can be used off-line on Linux system

Performance Analysis with CrayPat & Cray Apprentice^2



1: Load CrayPat & Cray Apprentice^2 Modules

2: Bulid Application

sampling or event based instrumentation ?

sampling

Event based

3: setenv PAT_BUILD_ASYNC 1
% pat_build a.out

: setenv PAT_BUILD_ASYNC 0 (default)
% pat_build [-g group] [-u] [options] a.out

4: Execute instrumented application

5: Create performance file (.ap2)
% pat_report –f ap2 [options] <.xf file>

6: Performance analysis & visualization
With pat_report or Cray Apprentice^2

Performance Metrics Available in pat_report

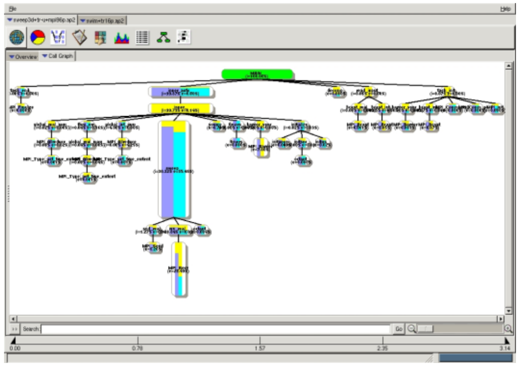
- Profile by groups
  - Threshold
  - Load imbalance information
- Function Profile
  - Flat profile
  - Call Tree view
  - Callers view
  - Hardware counters information
- MPI Profiler
  - MPI Load balance
  - MPI Stats by bin
- I/O Statistics
  - Read and Write Statistics
- Heap Statistics
  - High water mark
  - Memory leaks

CrayPat API

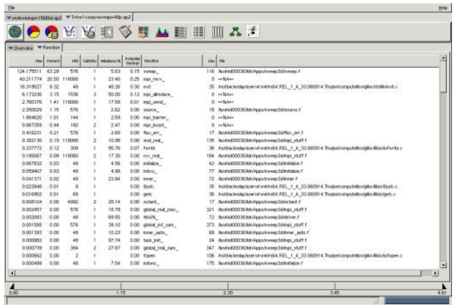
- CrayPat performs automatic instrumentation at function level
- The CrayPat API can be used for fine grain instrumentation
  - Fortran
    - call PAT_region_begin(id, “label”, ierr)
    - DO Work
    - call PAT_region_end(id, ierr)
  - C
    - include <pat_api.h>
    - …
    - ierr = PAT_region_begin(id, “label”);
    - DO_Work();
    - ierr = PAT_region_end(id);

Cray Apprentice^2

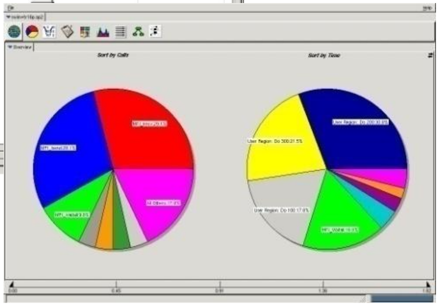
Call Graph Profile



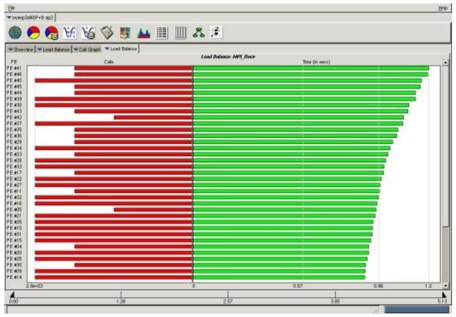
Function Profile



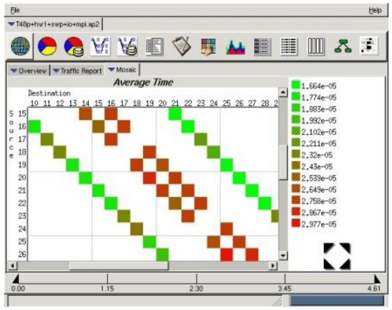
Function Overview



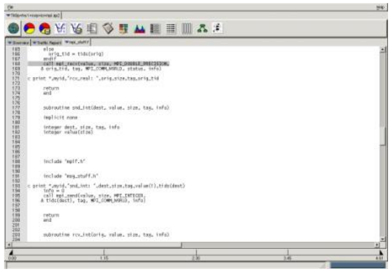
Load balance views



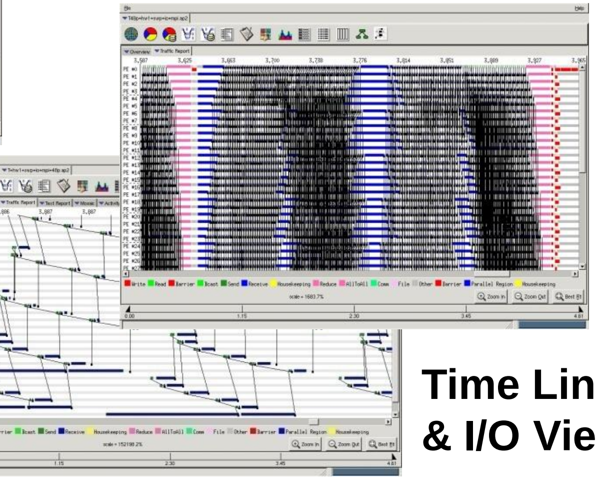
Pair-wise Communication View



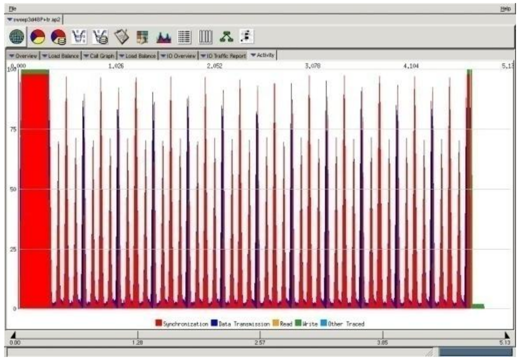
Source code mapping



Time Line & I/O Views



Communication & I/O Activity View

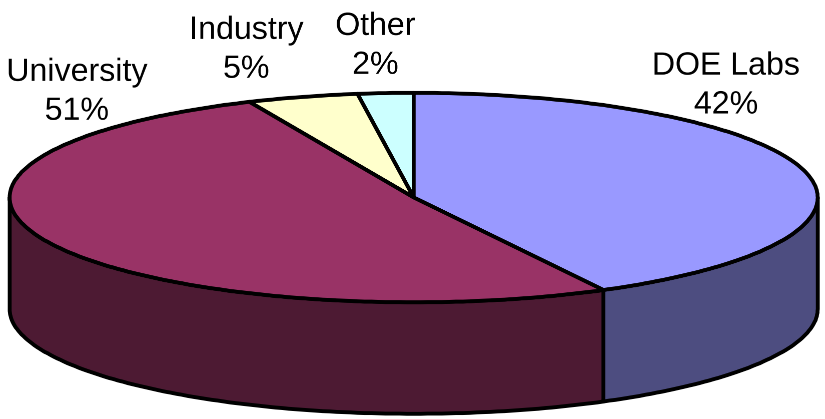


Usage

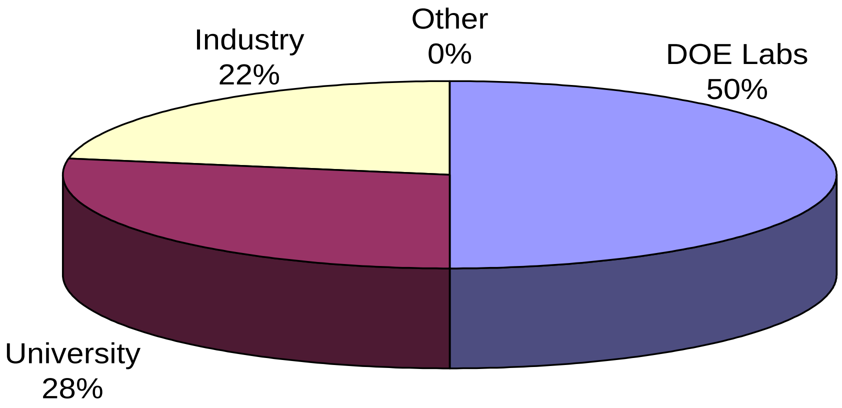
NCCS Demographics in 2006

2006 Projects			
Accelerator physics	1	Engineering	1
Astrophysics	3	Fusion	4
Chemistry	1	High energy physics	1
Climate change	3	Biology	2
Combustion	1	Materials science	2
Computer science	2	Nuclear physics	1

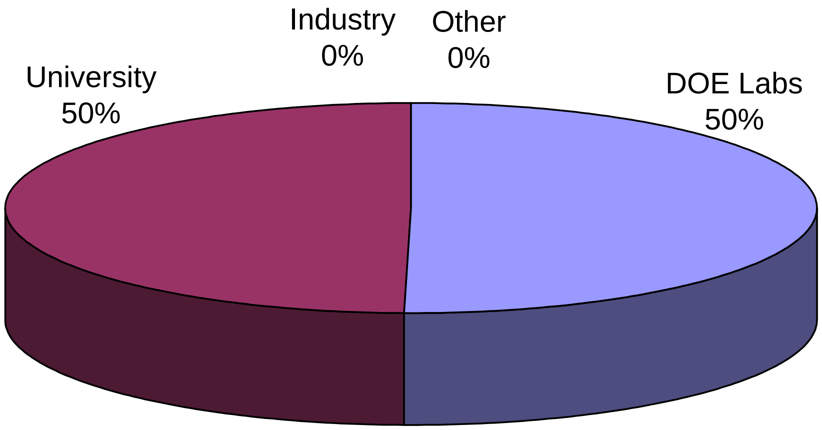
LCF Users by Affiliation



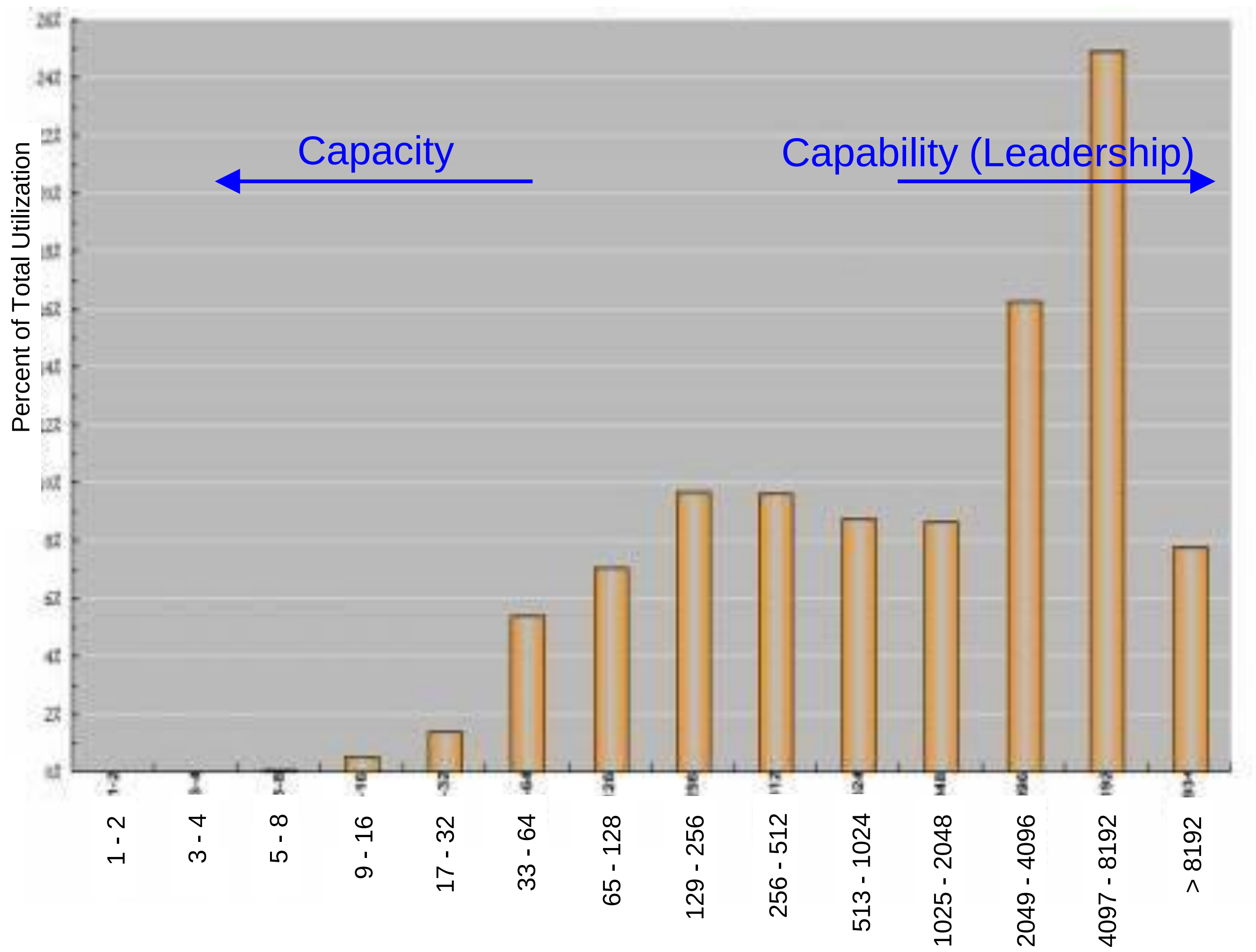
FY 2006 Phoenix Usage by Affiliation



FY 2006 Jaguar Usage by Affiliation

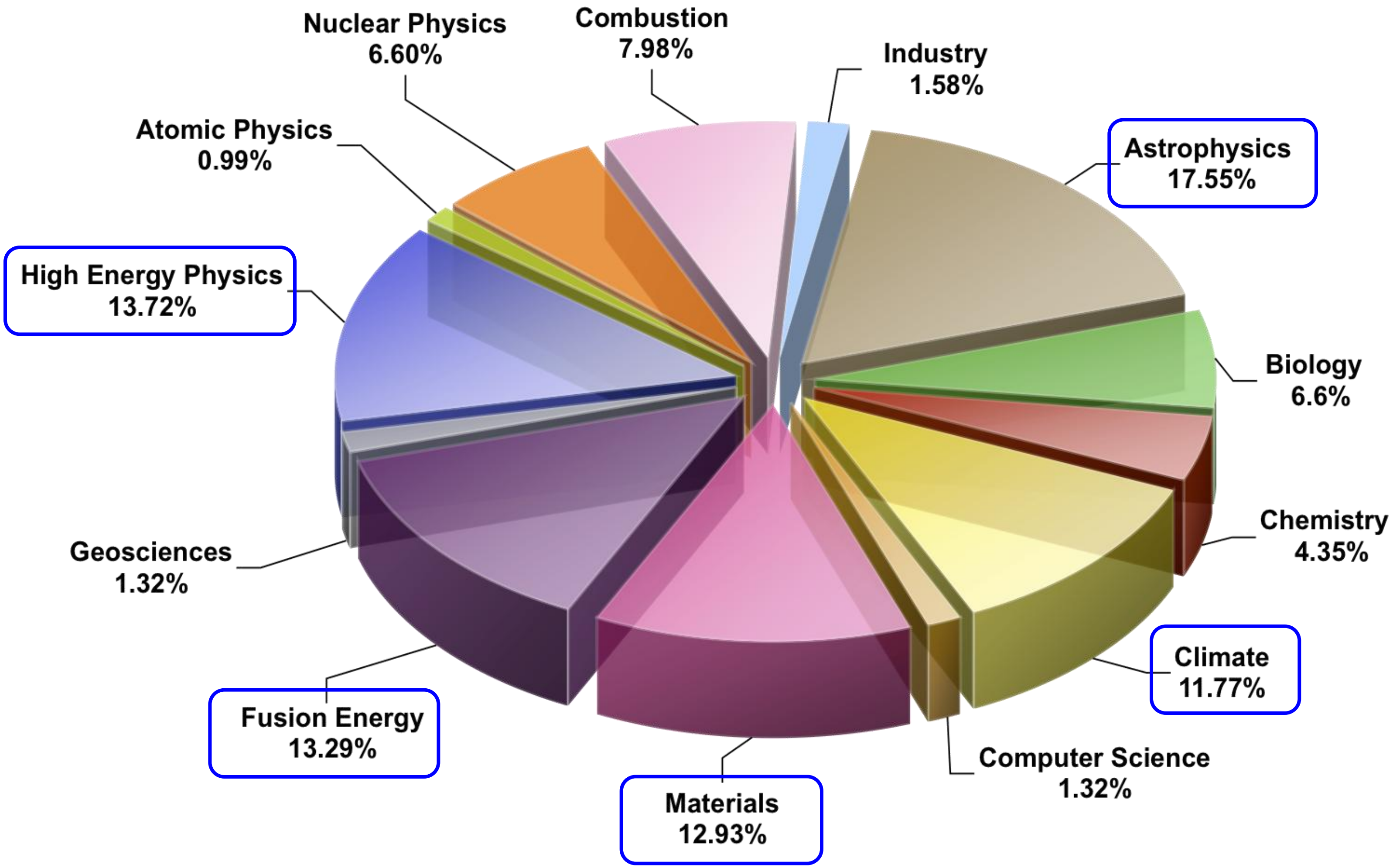


Simulation Job Size Distribution for Science Applications on the ORNL Cray XT3 in 2006



2007 INCITE Allocations at the ORNL LCF

Breakdown by Discipline



Circled domains make up 70% of total allocation

NERSC Cray XT4

- Very similar to ORNL platform
- Main differences
  - XT4 nodes (homogeneous system)
  - GPFS filesystem