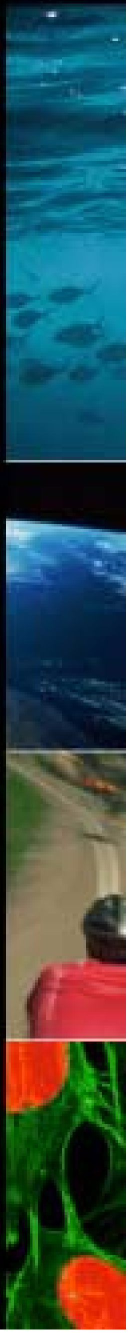
The Supercomputer Company

The Cray XT3™

MPP Supercomputer

Cray Inc.

February 2005



MPP Computing at Cray

1991

MPP Decision:

- MPP Advisory Group Formed
- 2 Year Goal to produce first machine

1993

Cray T3D:

- Unicos max
- PVM, CRAFT
- “Shmem”
- Totalview
- PATP
- F--

1996

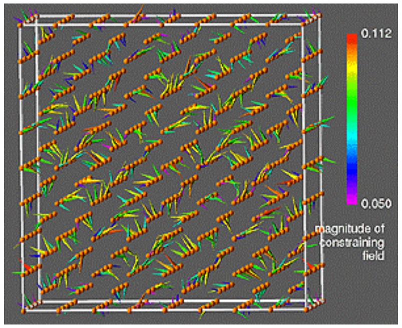
Cray T3E:

- MPI
- UNICOS mk
- Stream buffers
- Gigaring

MPP Computing at Cray

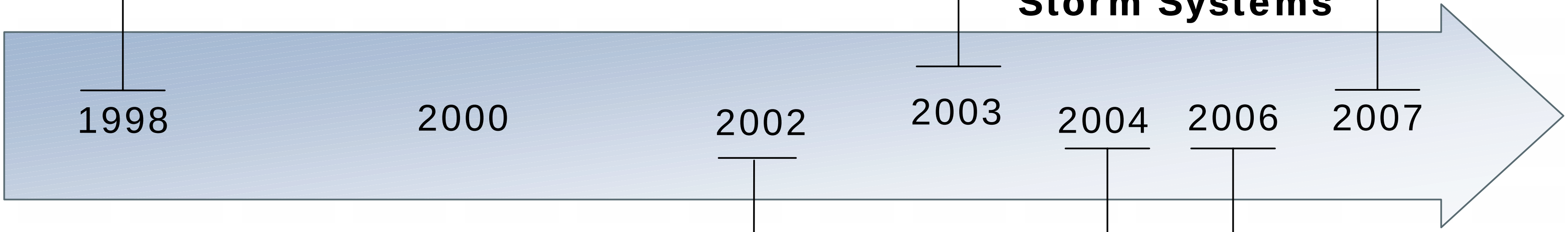
Cray T3E1200:

- Sustained Teraflop achieved on 1480 processors
- Gordon Bell Prize Winner



MPP: "Adams"

Decision to Productize Red Storm Systems



1998

2000

2002

2003

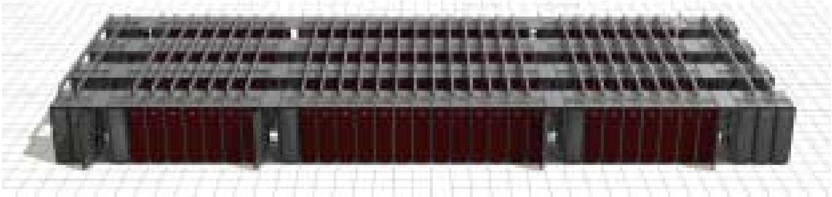
2004

2006

2007

Sandia Red Storm Contract:

- 10,000 processor machine
- Delivery in 2004
- Balanced, 40Tflops System

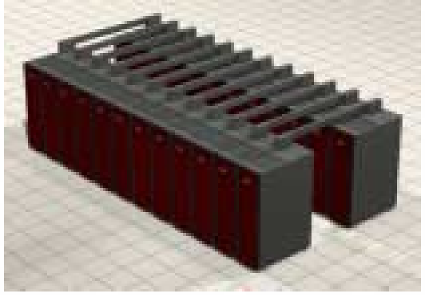


Cray XT3:

- 3rd Generation MPP
- UNICOS/lc
- First Cray XT3 Order
- First Cray XT3 Deliveries

Cray XT4:

- DDR2 Memory
- Faster Interconnect



Red Storm
Background & Status

Cray Red Storm

- Massively parallel processing supercomputer system used for analysis and stewardship of nuclear weapons at Sandia National Labs
- Key system characteristics
  - Massively parallel system – 10,000 AMD 2 GHz processors
  - High bandwidth mesh based custom interconnect
  - High performance I/O subsystem
  - Fault tolerant
- Full system delivery in 2004
- Designed to double in size—100 Tflops

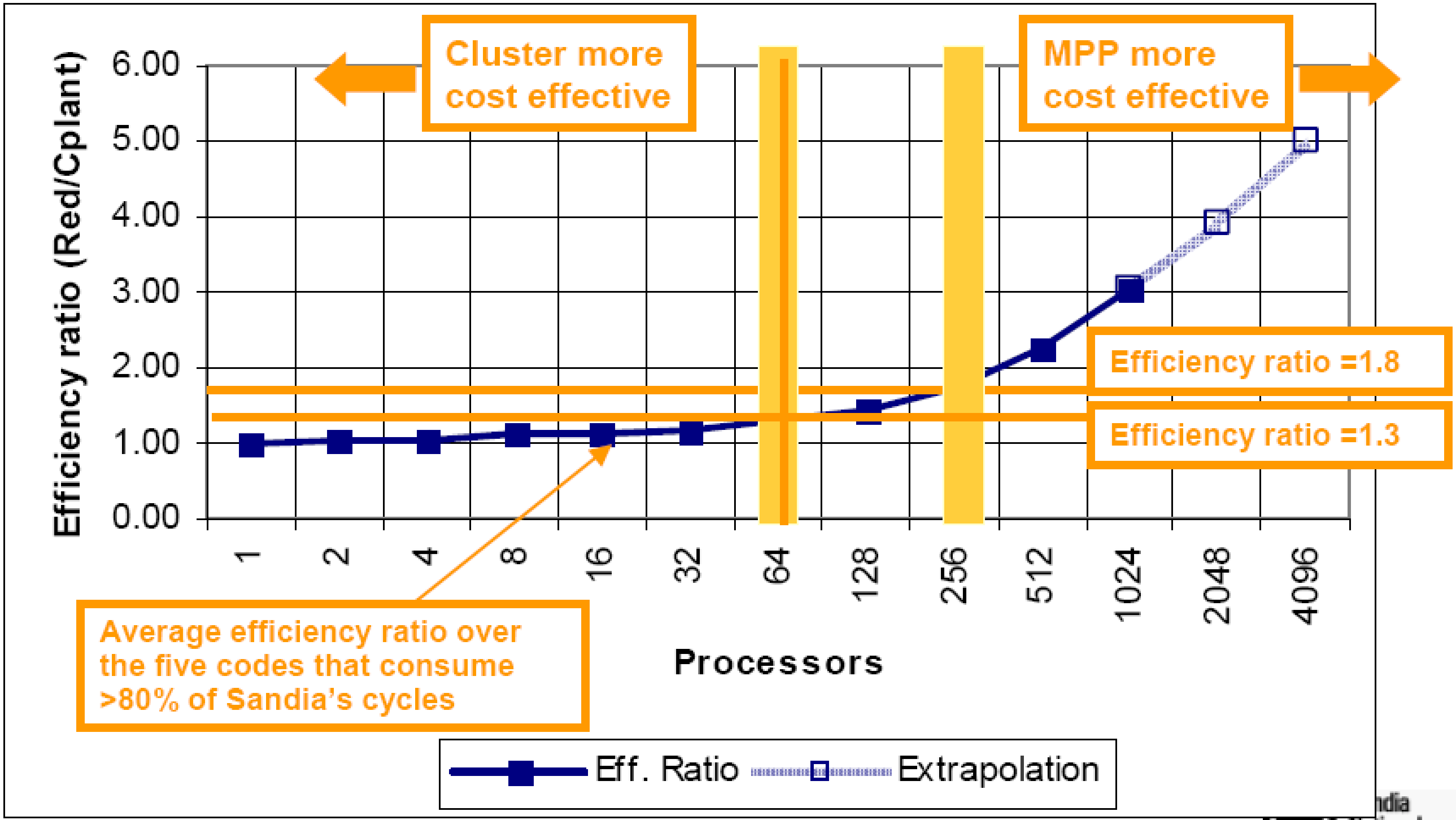
> **"We expect to get substantially more real work done, at a lower overall cost, on a highly balanced system like Red Storm than on a large-scale cluster."**
>
> **Bill Camp, Sandia Director of Computers, Computation, Information and Mathematics**

System Goals

- Balanced Performance between CPU, Memory, Interconnect, and I/O
- Highly *scalable* system hardware and software
- High speed, high *bandwidth* 3D mesh interconnect
- Run a set of applications 7 times faster than ASCI Red
- Run an ASCI Red application on *full system for 50 hours*
- Flexible partitioning for classified and non-classified computing
- High performance I/O subsystem (File system and storage)

Relating Scalability and Cost Effectiveness of Red Storm Architecture

Source: Sandia National Labs



Cluster more cost effective

MPP more cost effective

Efficiency ratio =1.8

Efficiency ratio =1.3

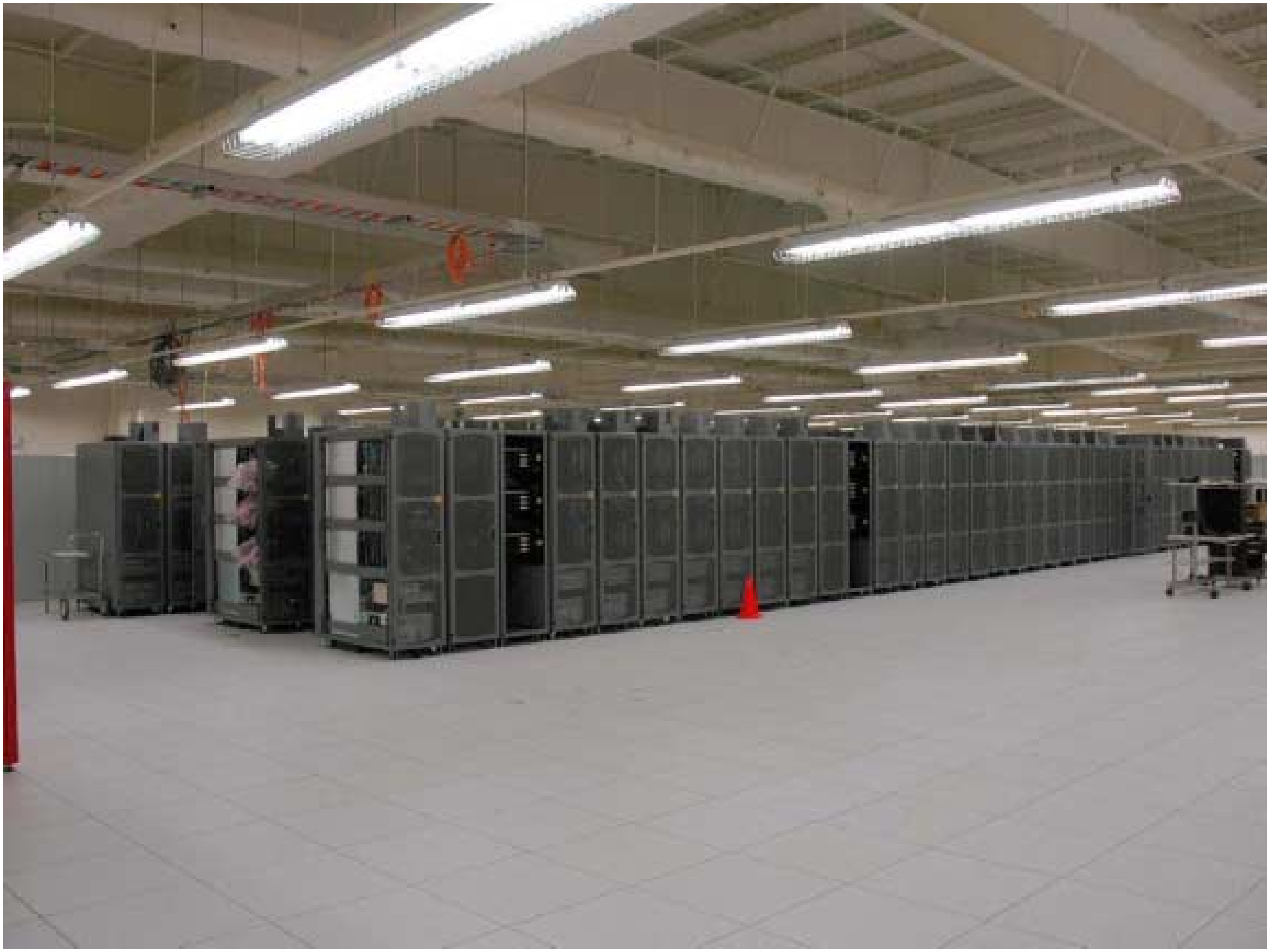
Average efficiency ratio over the five codes that consume >80% of Sandia's cycles

We believe the Cray XT3 will have the same characteristics; More cost effective than clusters somewhere between 64 and 256 MPI tasks

SeaStar ASIC

- SeaStar was checked out in September
- We started assembling and testing individual cabinets in September
- First shipment to Sandia was October 8th
- First row of Red Storm was shipped at the end of October
- 100% of the system now installed

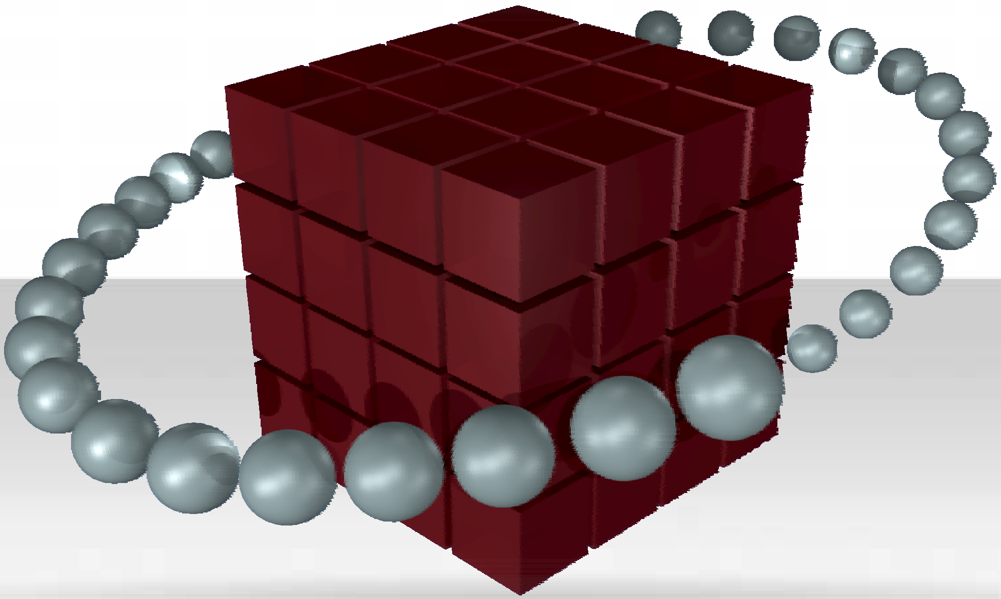
Three Rows at Sandia

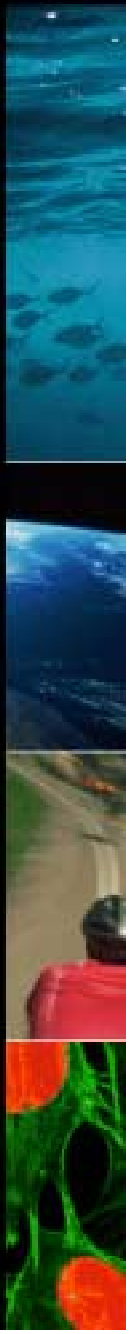


The 4 x 9 System

- We assembled the last row of Red Storm as a 4 rows by 9 cabinet configuration in Chippewa Falls
- All connections were tested and verified
- This was torn down and shipped on 1/17 (4 semi trucks)

CRAY XT3
Balanced Architecture

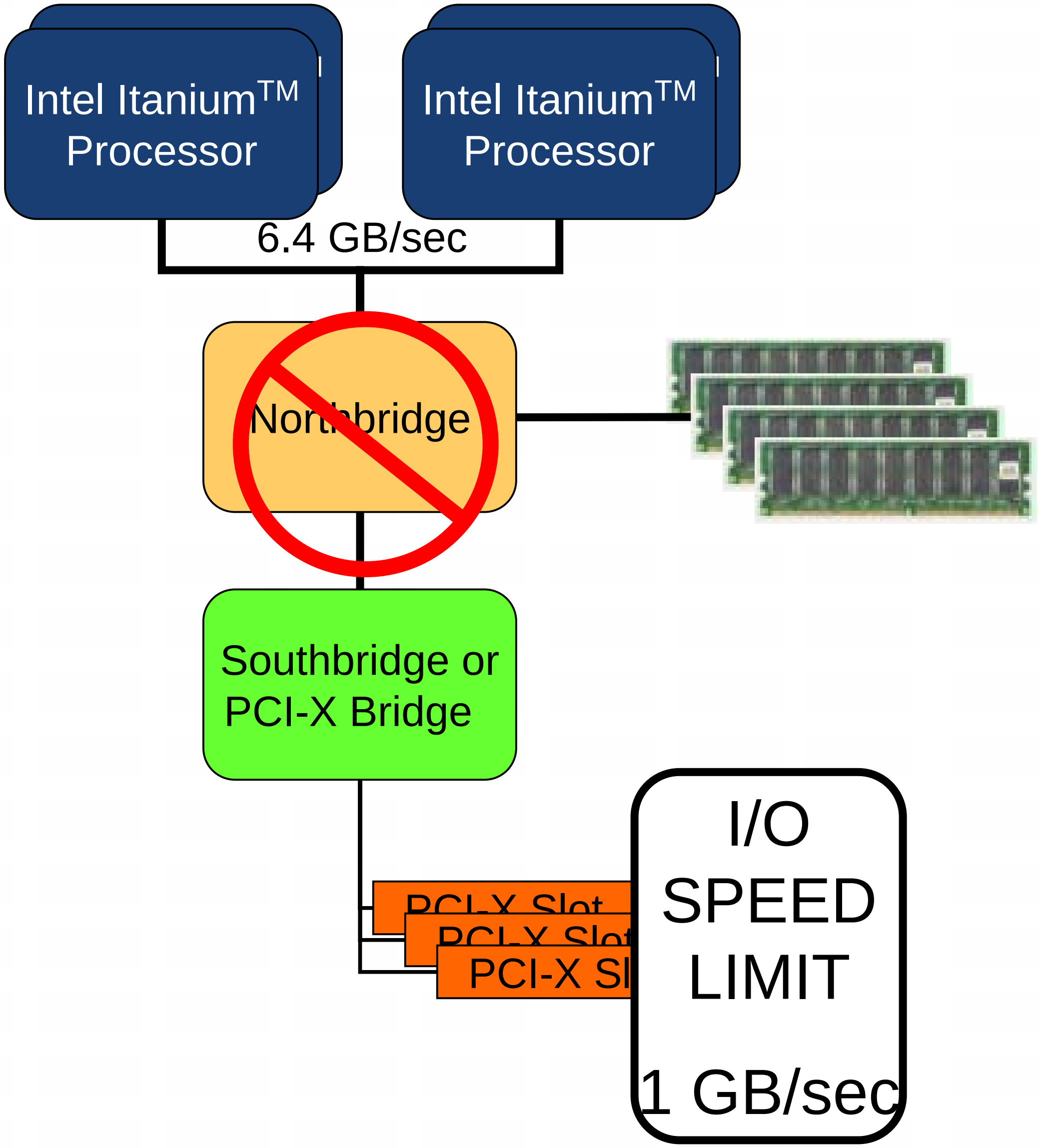




Recipe for a good MPP

1. Select Best Microprocessor
2. Surround it with a balanced or "bandwidth rich" environment
3. Eliminate "barriers" to scalability
   - SMPs don't help here
   - Eliminate Operating System Interference (OS Jitter)
   - Reliability must be designed in
   - Resiliency is key
   - System Management
   - I/O
   - System Service Life

Picking the best Processor: Why not Intel?



Intel Itanium™ Processor

Intel Itanium™ Processor

6.4 GB/sec

Northbridge

Southbridge or PCI-X Bridge

PCI-X Slot

PCI-X Slot

PCI-X Sl

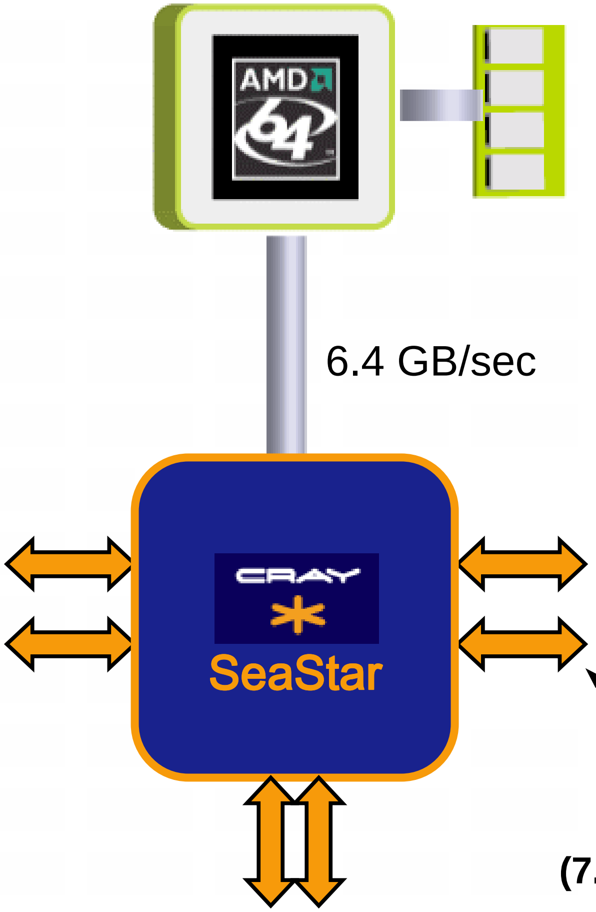
I/O SPEED LIMIT

1 GB/sec

- **Memory latency ~ 160 ns and *bandwidth is shared* between mutliple processors**
- **Northbridge chip is 2nd most complex chip on the board. Typical chip uses about 11 Watts**
- **Any interconnect limited by speed of PCI-X since it's the fastest place to "plug in"**
- **Best place to tie in a high performance interconnect would be through the Northbridge, but this is difficult to do legally without an Intel bus license**

AMD Opteron Generic System

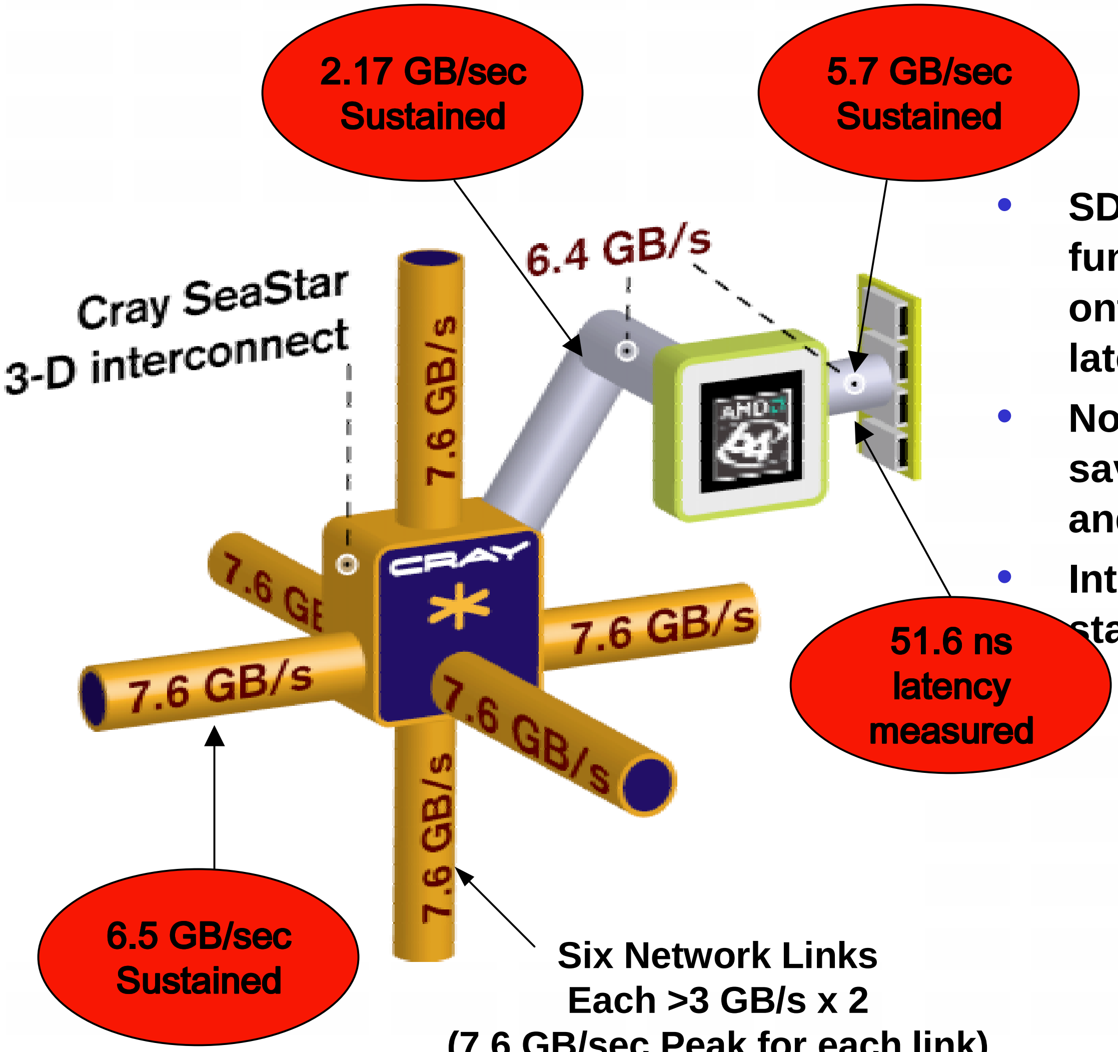
CRAY XT3 PE



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 60-90 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)**

Cray XT3 Processing Element: Measured Performance



2.17 GB/sec Sustained

5.7 GB/sec Sustained

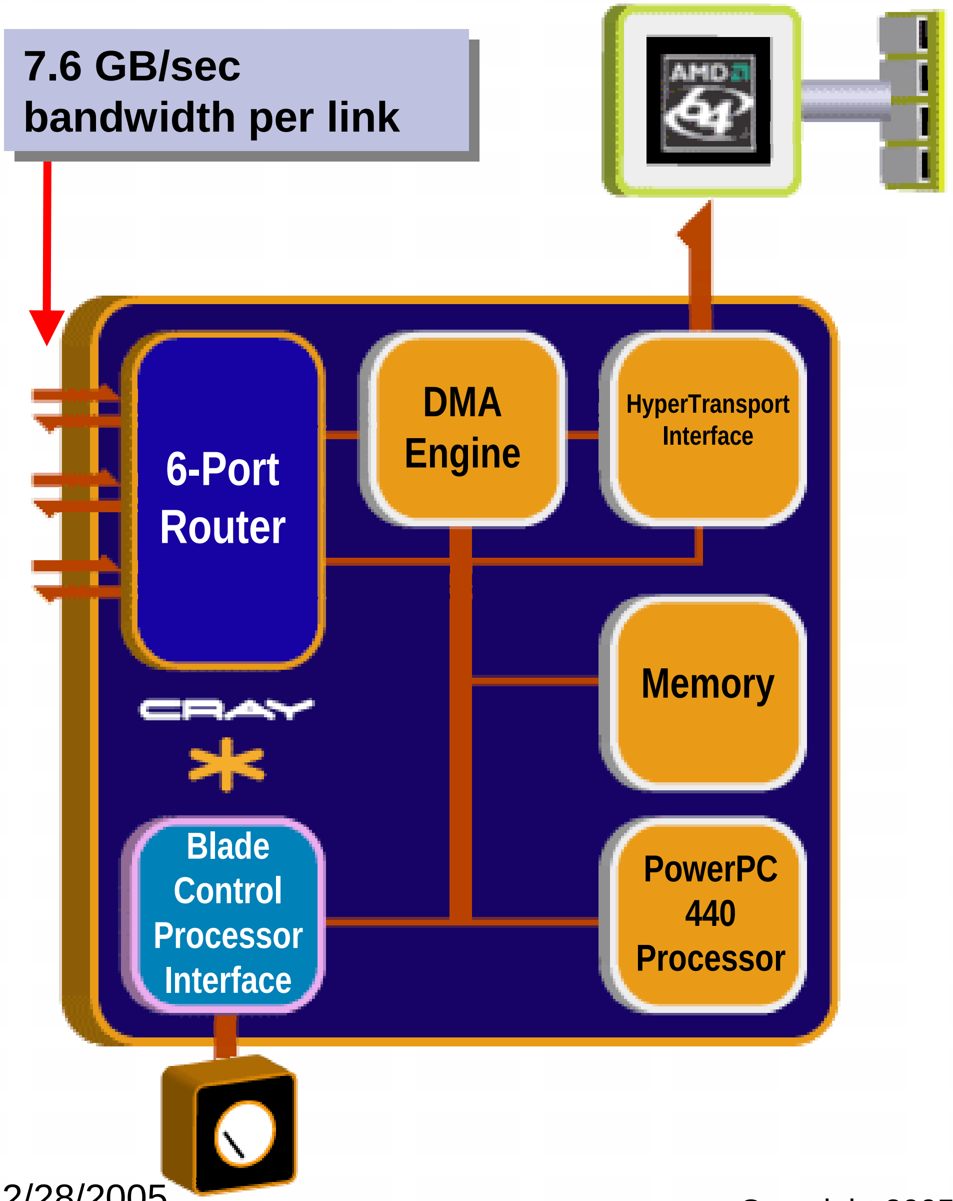
51.6 ns latency measured

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)

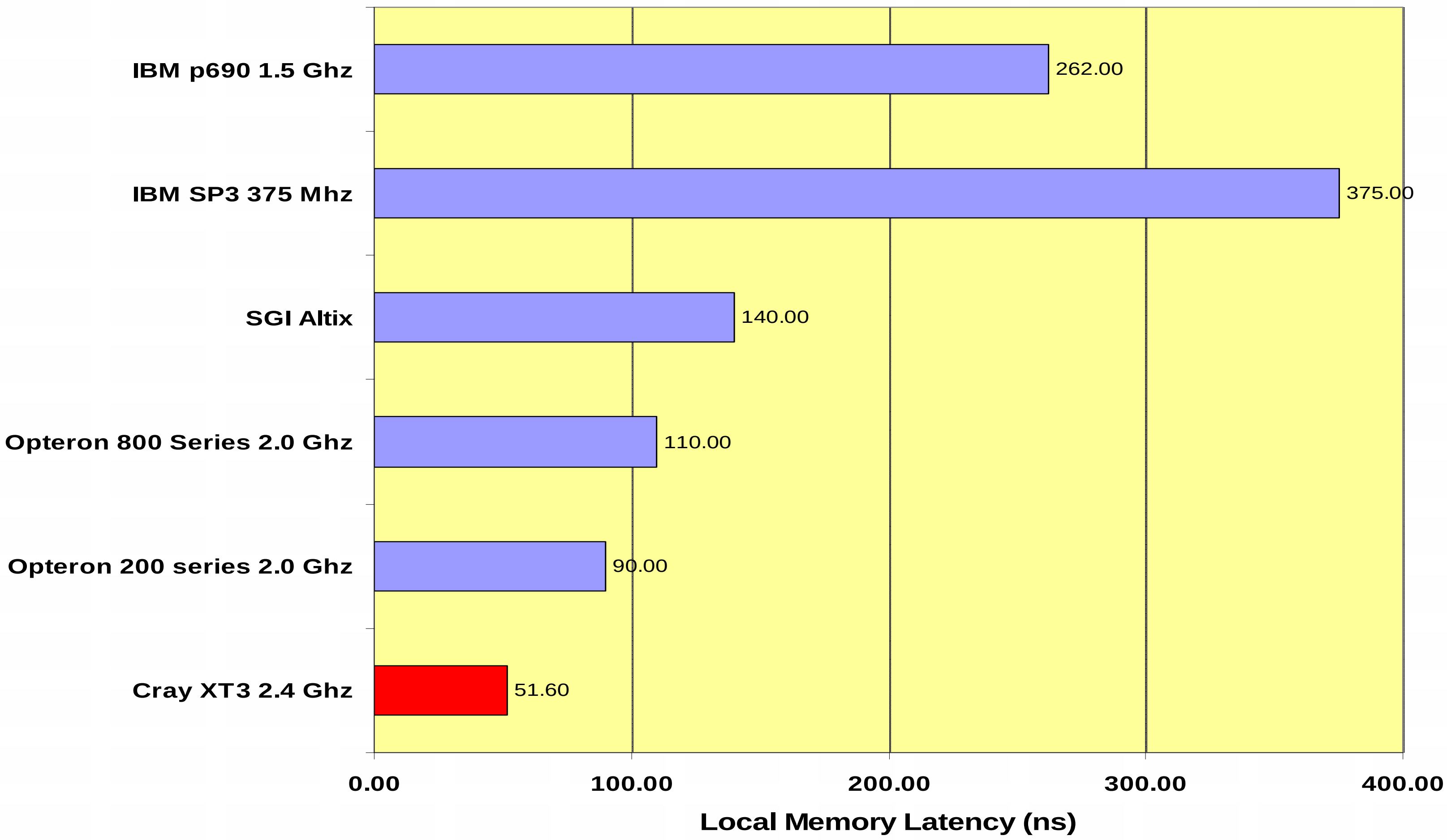
Cray SeaStar Internals



7.6 GB/sec
bandwidth per link

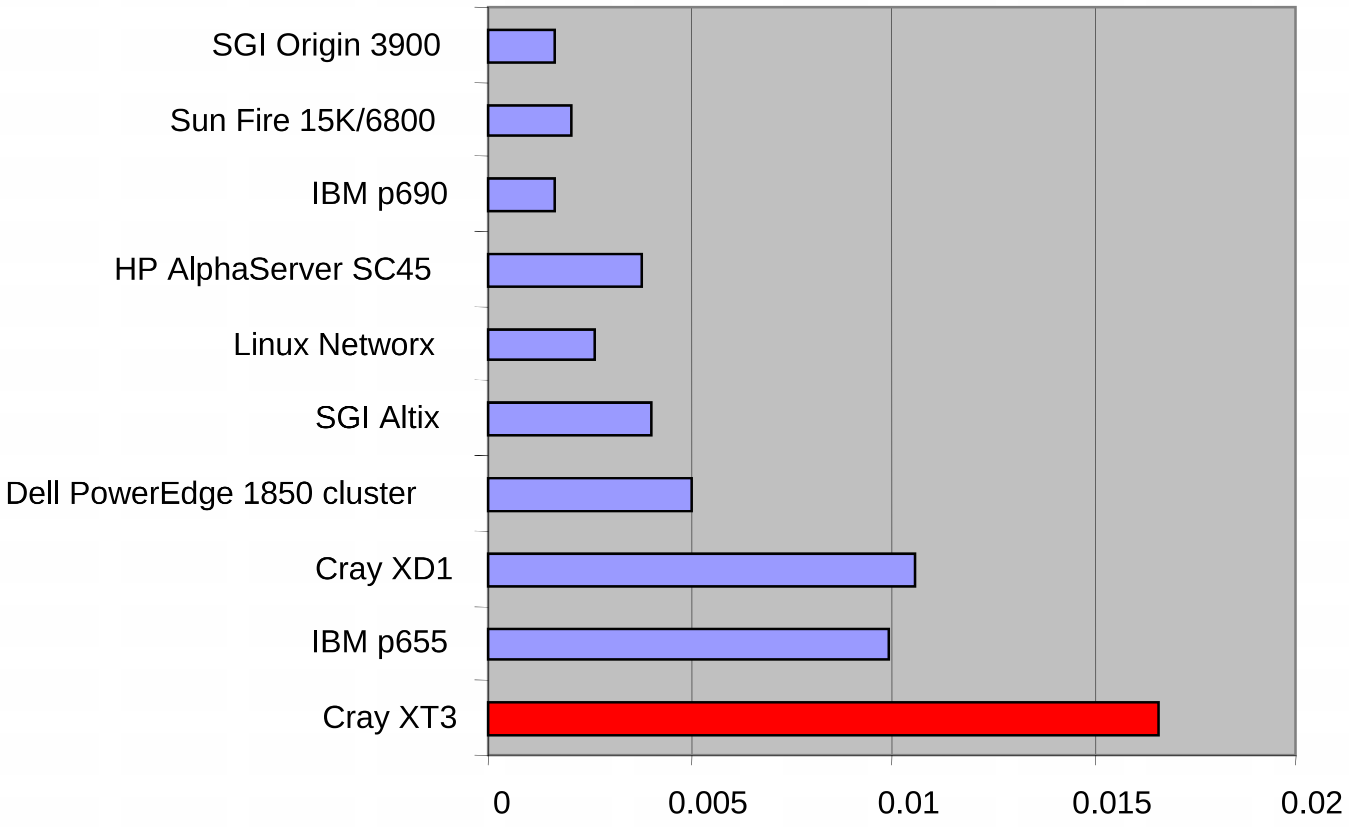
- Each Processor is directly connected to a dedicated SeaStar
- Each SeaStar contains a 6-Port router *and* communications engine
- Provides serial connection to the Cray RAS and Management System

Memory Latency

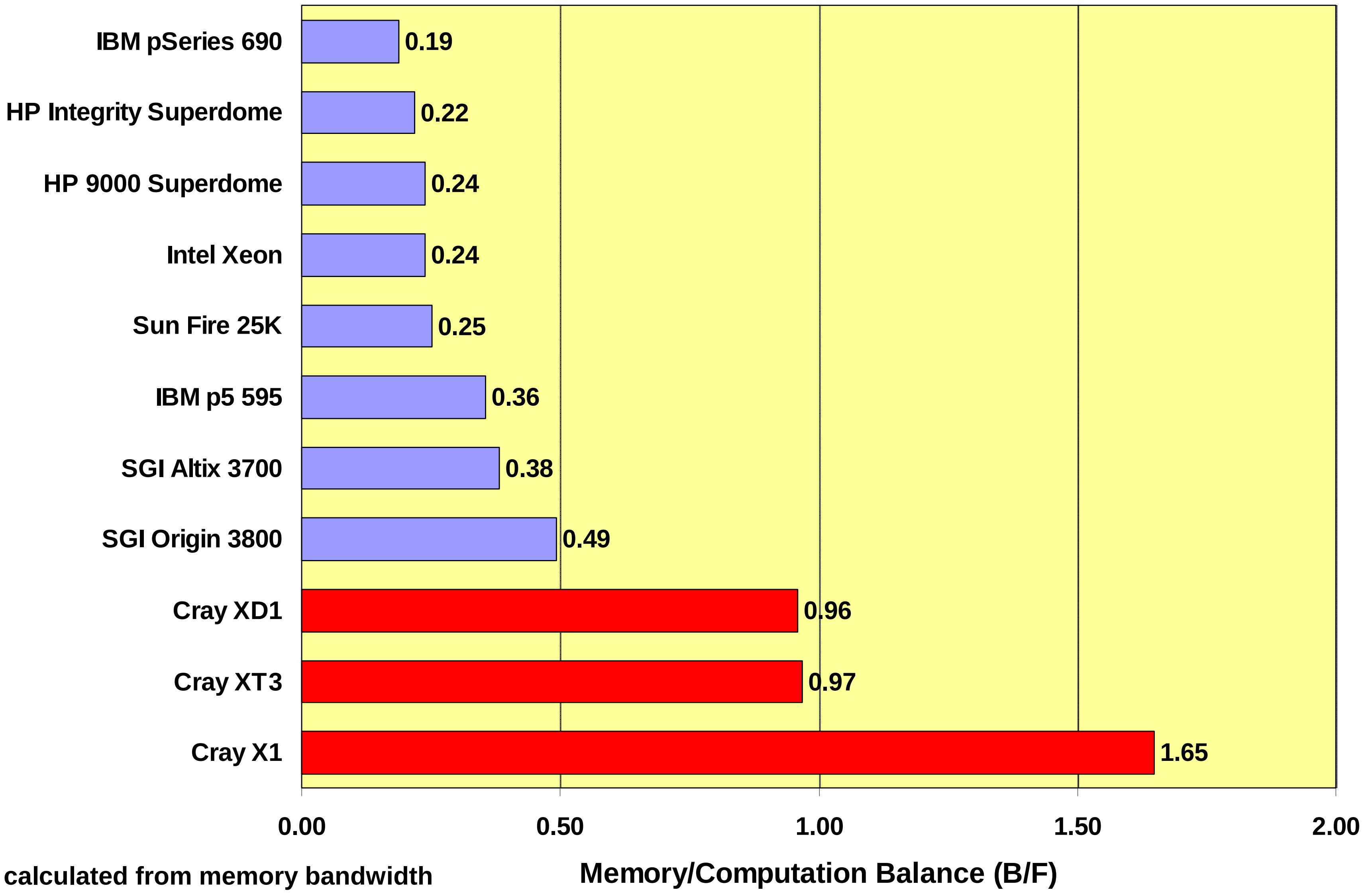


Single Processor architecture yields lowest memory latency

HPCC Random Access Benchmark

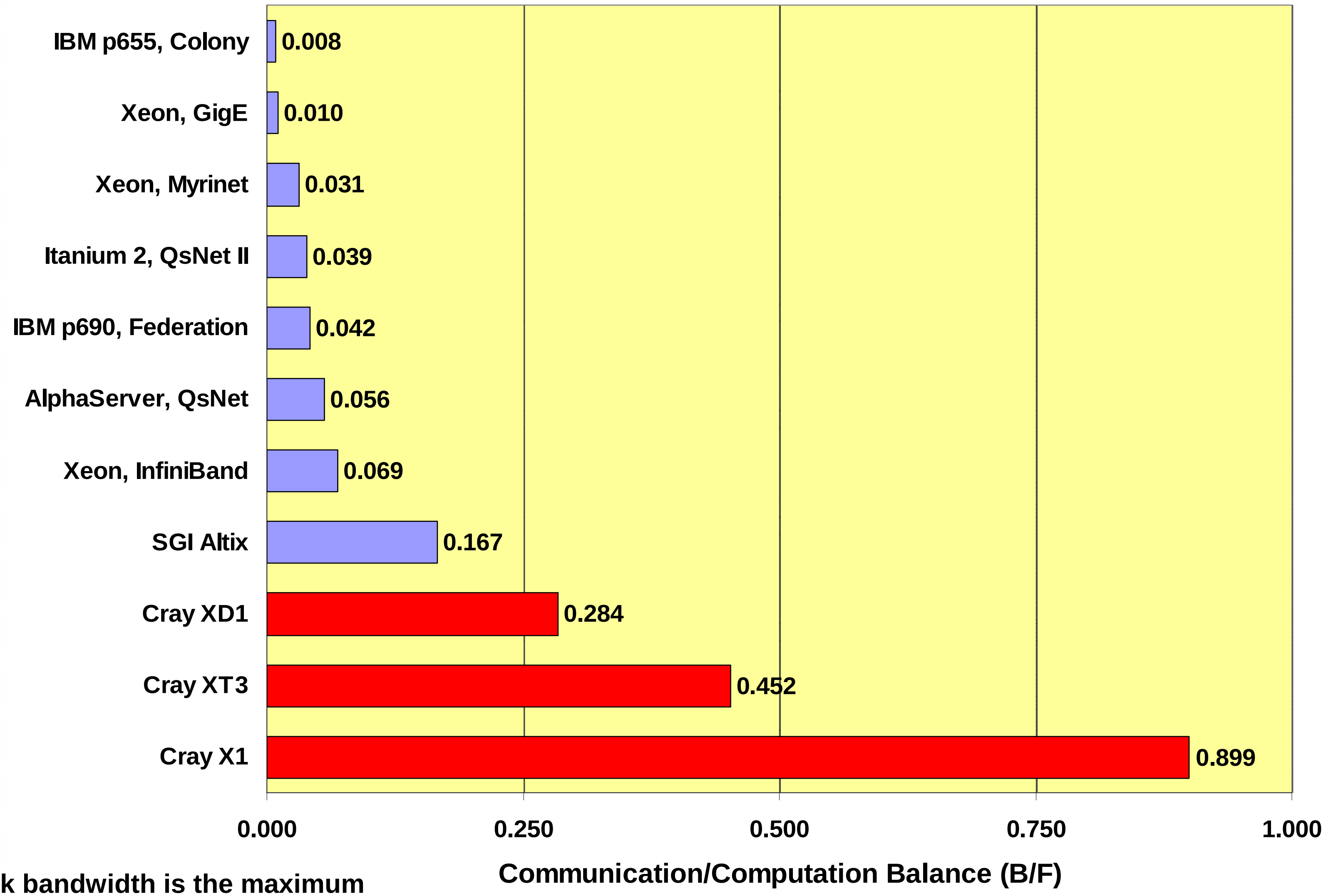


Measured Memory Balance



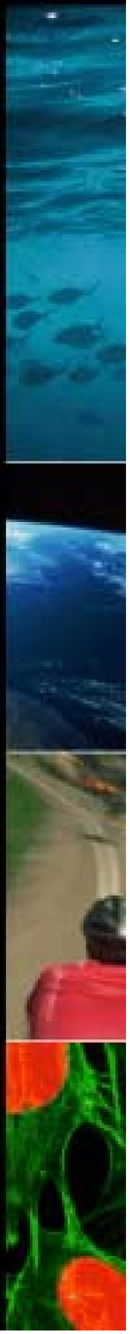
B/F calculated from memory bandwidth measured via STREAM Triad benchmark

Measured Network Balance

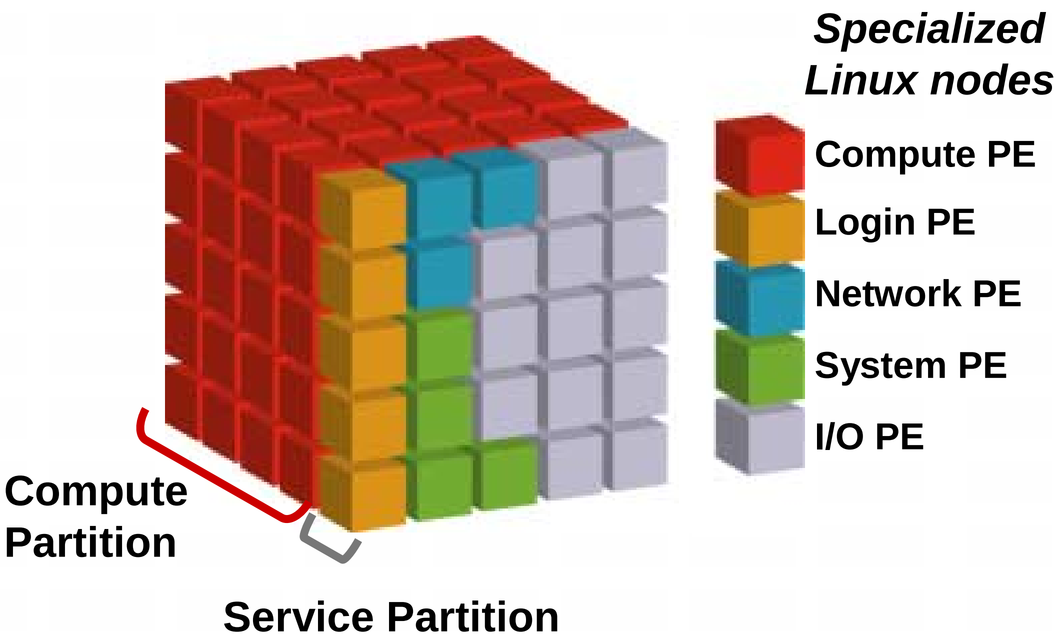


Network bandwidth is the maximum bidirectional data exchange rate between two nodes using MPI

Scalable Software



Scalable Software Architecture: UNICOS/lc



Specialized Linux nodes

🟥 **Compute PE**

🟧 **Login PE**

🟦 **Network PE**

🟩 **System PE**

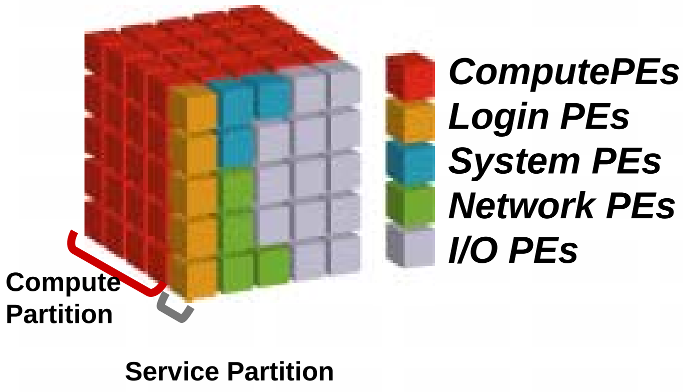
⬜ **I/O PE**

Compute Partition

Service Partition

- Microkernel on Compute PEs, full featured Linux on Service PEs.
- Contiguous memory layout used on compute processors to streamline communications
- Service PEs specialize by function
- Software Architecture eliminates OS “Jitter”
- Software Architecture enables reproducible run times

Unicos/lc Status



🟥 ComputePEs
🟧 Login PEs
🟦 System PEs
🟩 Network PEs
⬜ I/O PEs

Compute Partition

Service Partition

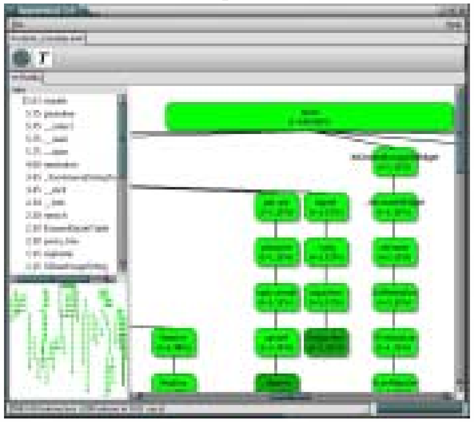
- Cray and Sandia have successfully demonstrated the Cray XT3 OS and MPI stack on 3342 compute PEs
- The Sandia ASCI Red system was used as a testbed system (called “Redshift”)
- Several Applications have been successfully run and demonstrated scalability including:
  - CTH on 3200 processors
  - MPI Barrier testing up to 3342 compute PEs
  - Bisection bandwidth benchmarks up to 3342 compute PEs
  - HPL on 121 PEs and 3339 PEs

Programming Environment

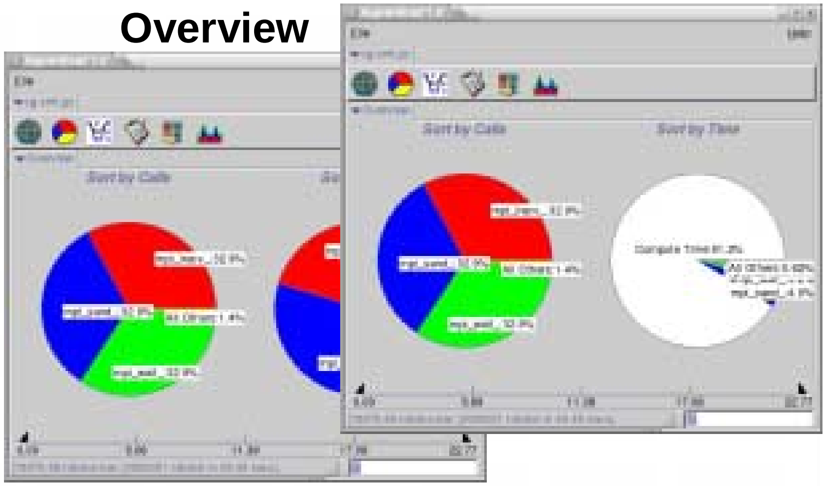
- The Portland Group compilers (unmodified from Linux version)
- High Performance MPI library (tuned collectives)
- Shmem Library
- AMD Math Libraries
- CrayPat & Apprentice^2 performance tools
- Etnus TotalView debugger available
- X86-64 *only*
- Static Binaries *only*

Cray Apprentice2

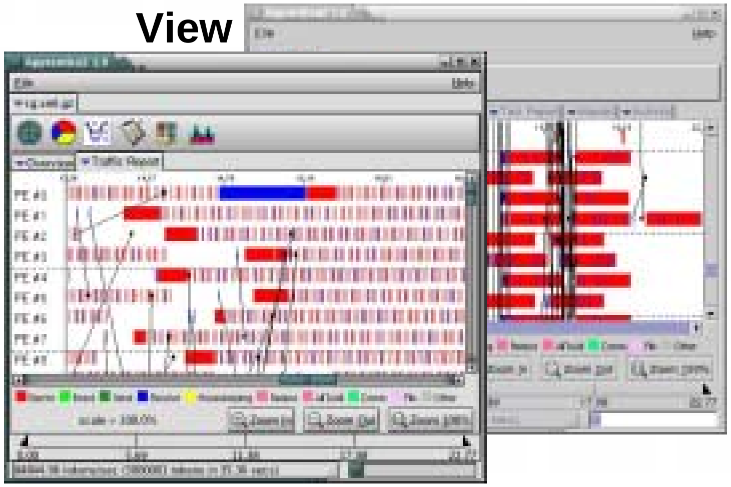
Call Graph Profile

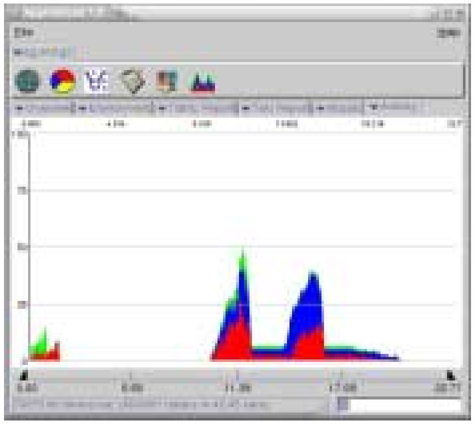


Communication Overview

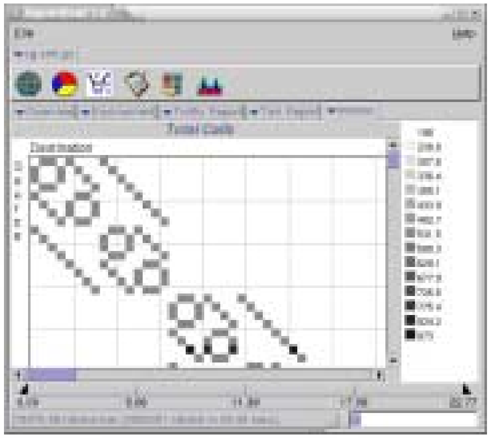


Time Line View



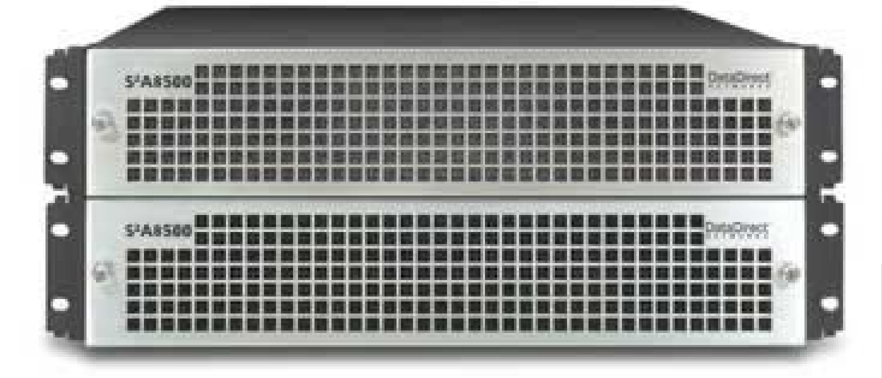


Communication Activity View

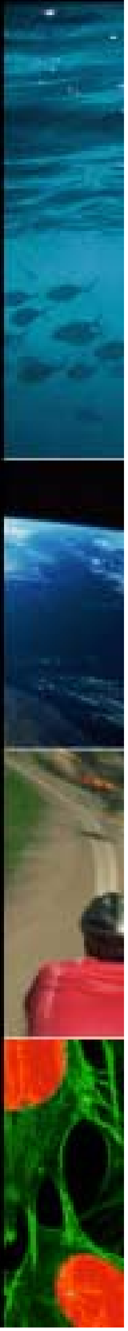


Pair-wise Communication View

Scalable I/O

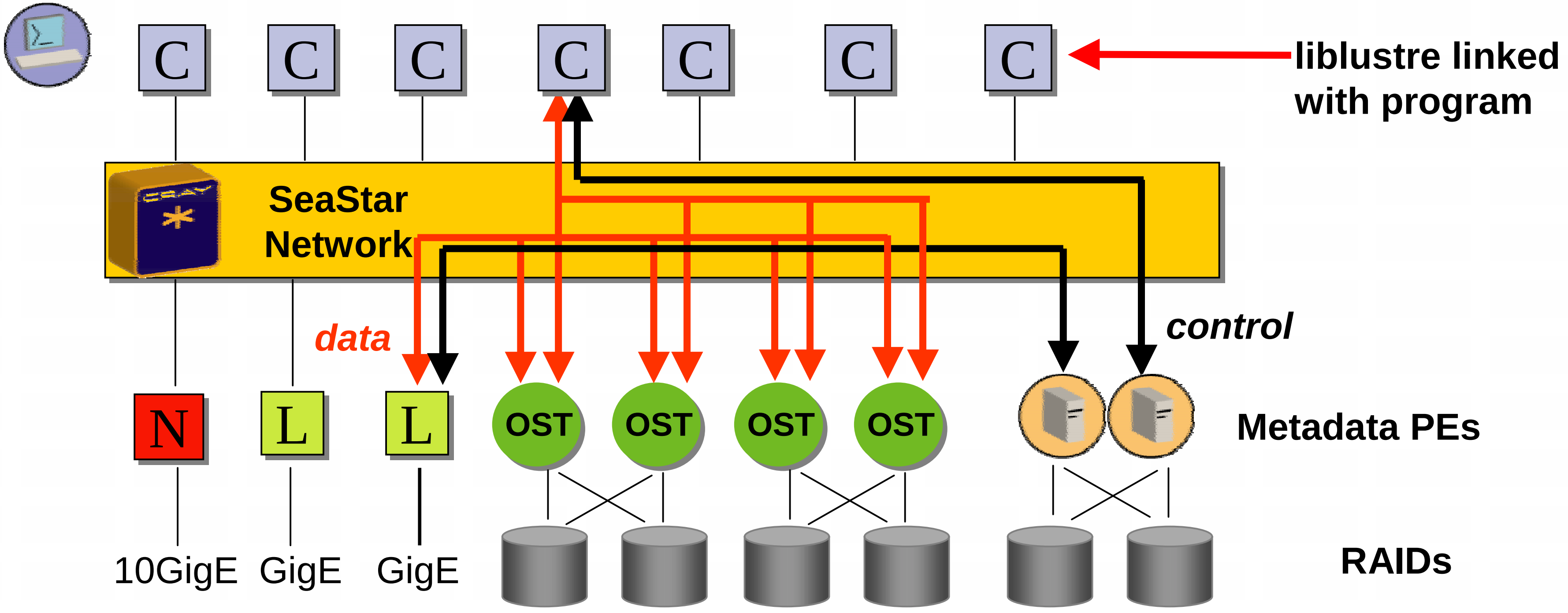


lustre



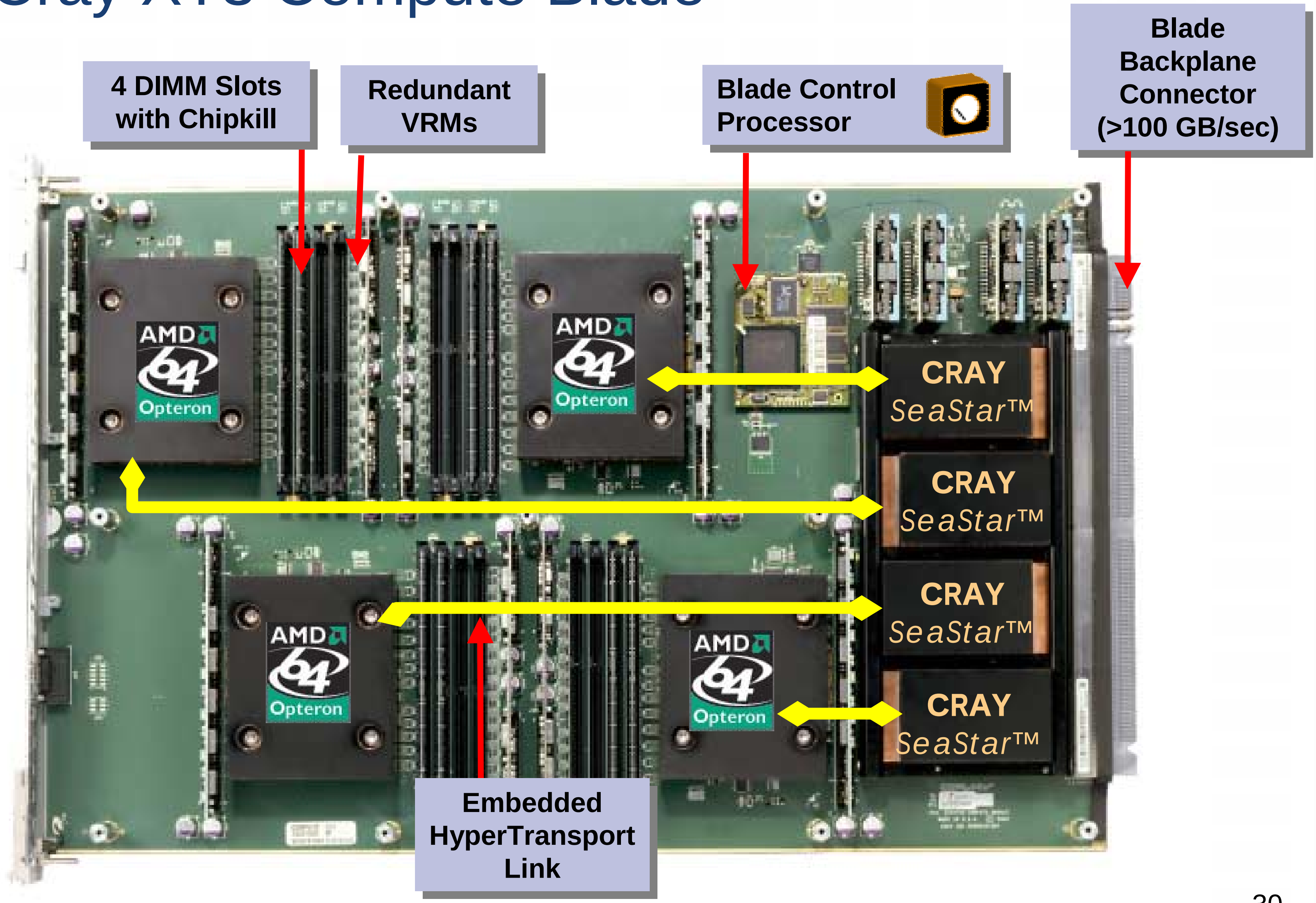
Scalable I/O

- Global Parallel File System: Lustre
  - Open Source, Vendor Neutral
  - Highly Scalable, block allocation NOT serialized
  - Liblustre for MPPs
  - OST Software Failover, Dual Path controllers



Engineered Reliability

Cray XT3 Compute Blade



4 DIMM Slots with Chipkill

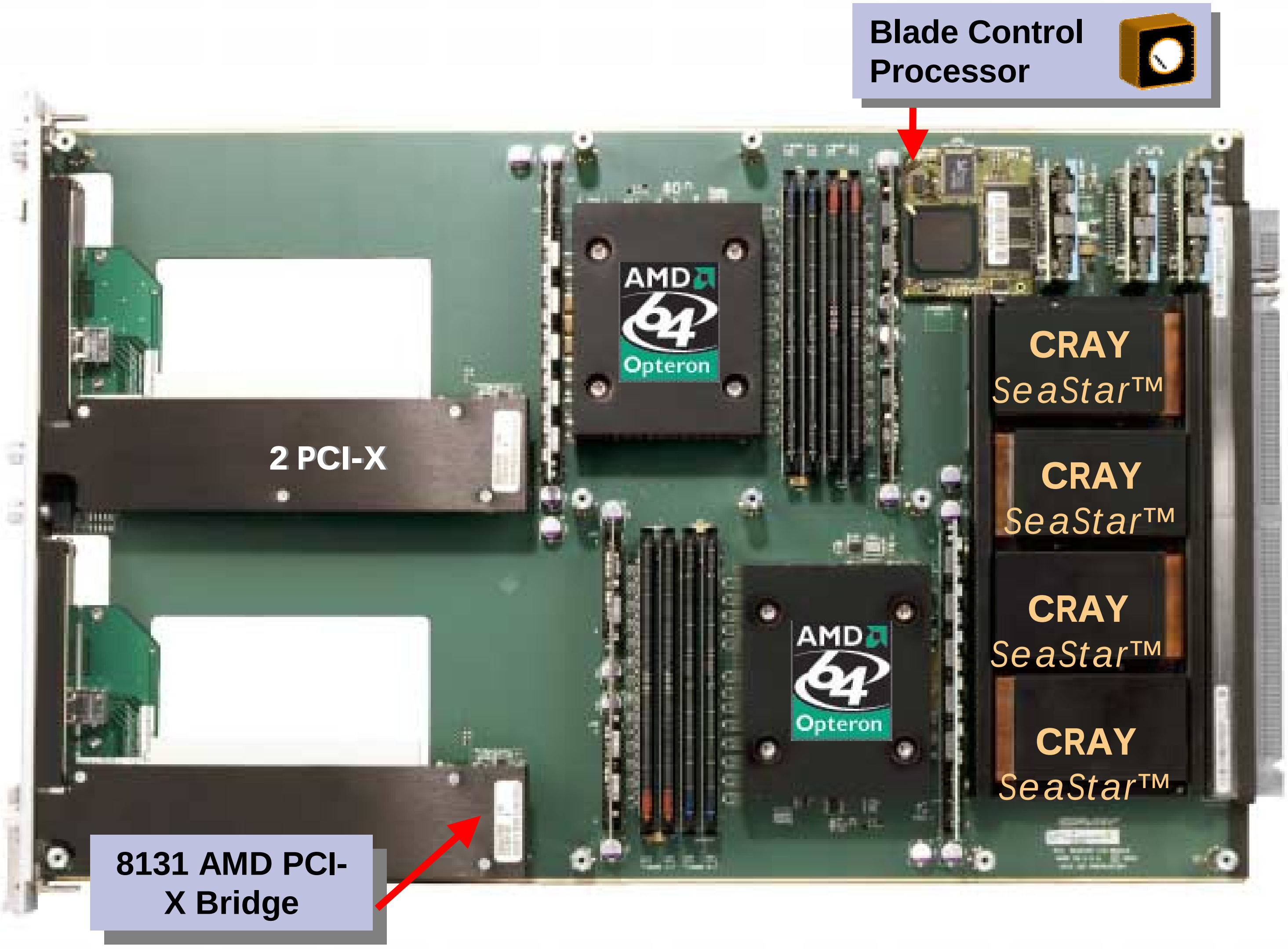
Redundant VRMs

Blade Control Processor

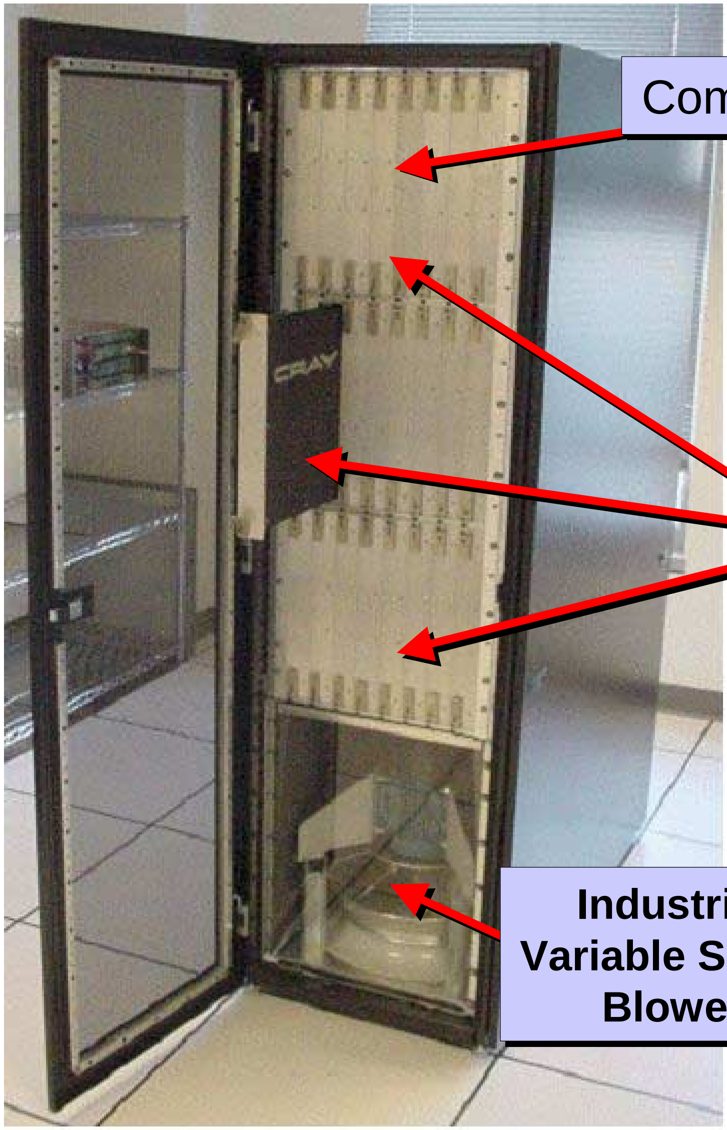
Blade Backplane Connector (>100 GB/sec)

Embedded HyperTransport Link

Cray XT3 Service and I/O Blade



CRAY XT3 Compute Cabinet

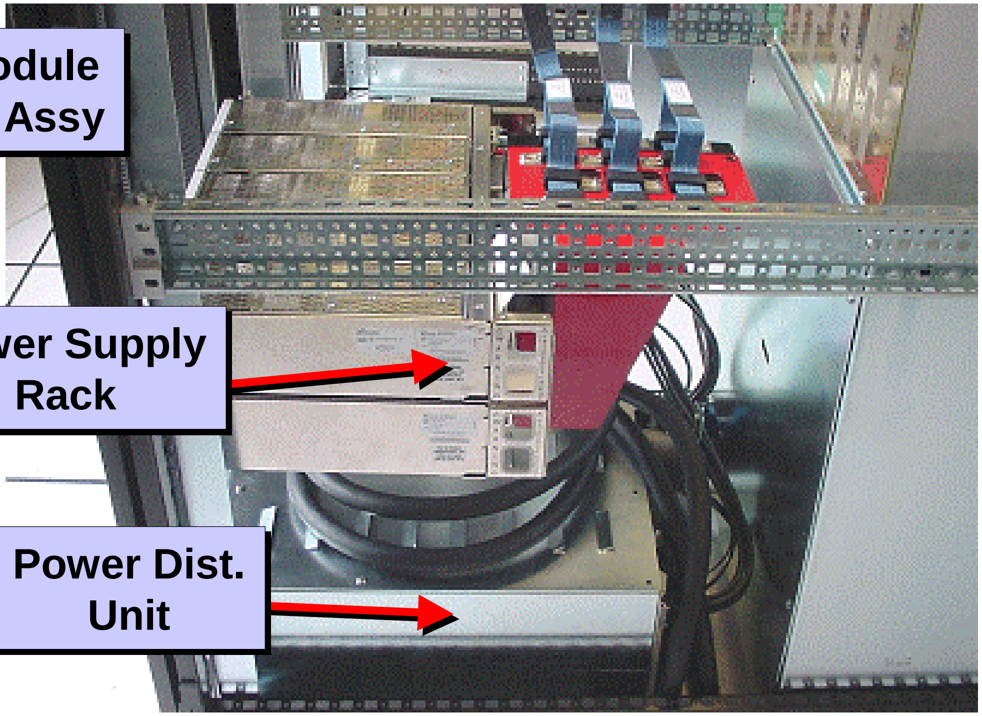


Compute Modules - enclosed

24 Module Cage Assy

Industrial Variable Speed Blower

- Cabinets are 1 floor tile wide
- Cold air is pulled from the floor space
- Room can be kept at a comfortable temp

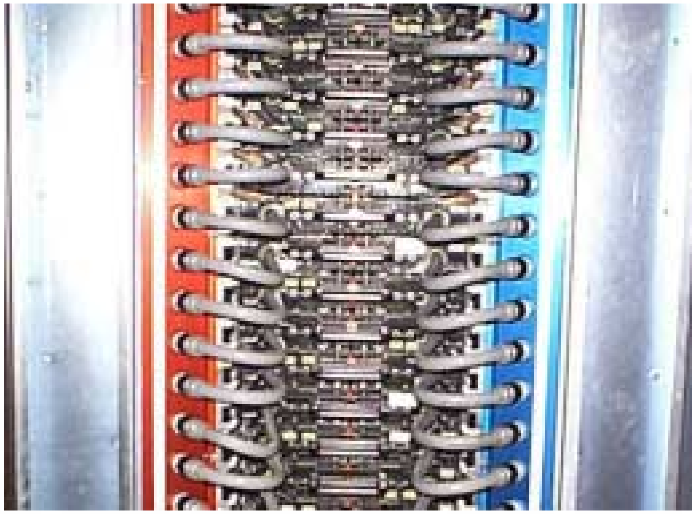


Power Supply Rack

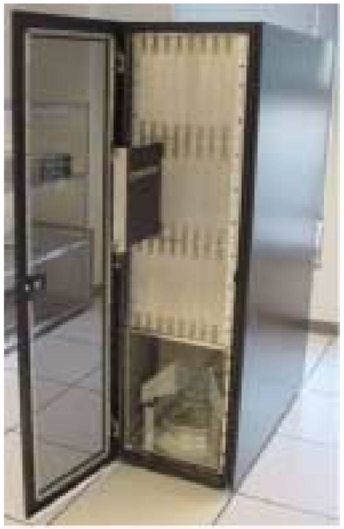
Power Dist. Unit

Pre-Prototype Cabinet

System Packaging: Compared with T3E



- Cold Plate cooling with Flourinert: $300 / Gallon
- 8 PEs per double-sided module
- 272 Processors per cabinet



- Air Cooled with variable speed blower: (air is free)
- 4 PEs per module
- 96 Processors per cabinet

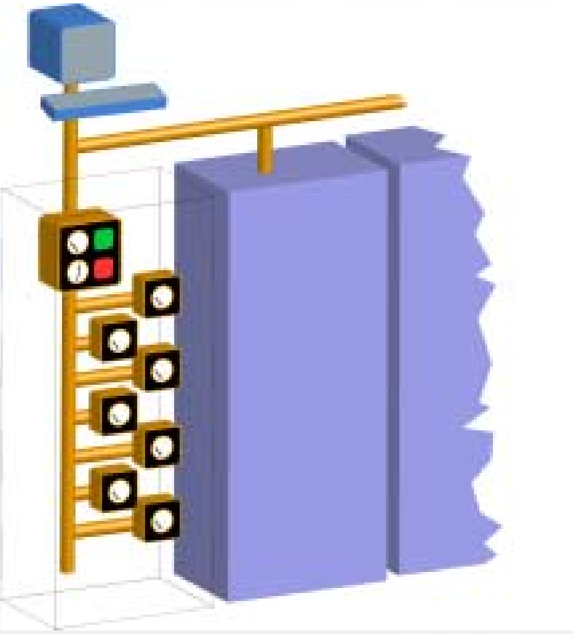
Which machine has higher density at 1000 processors?

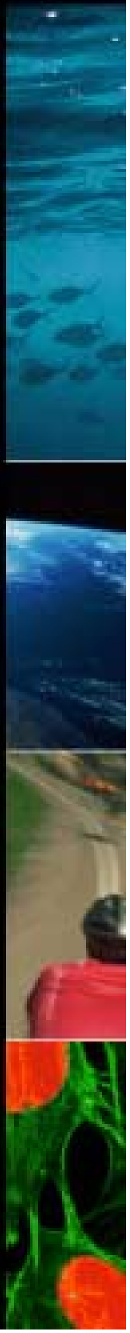
- **Cray T3E: 40 floor tiles, 14.4 Sq Meters**
- **Cray RS: 24 floor tiles, 8.6 Sq Meters**

Cray XT3 Reliability Features

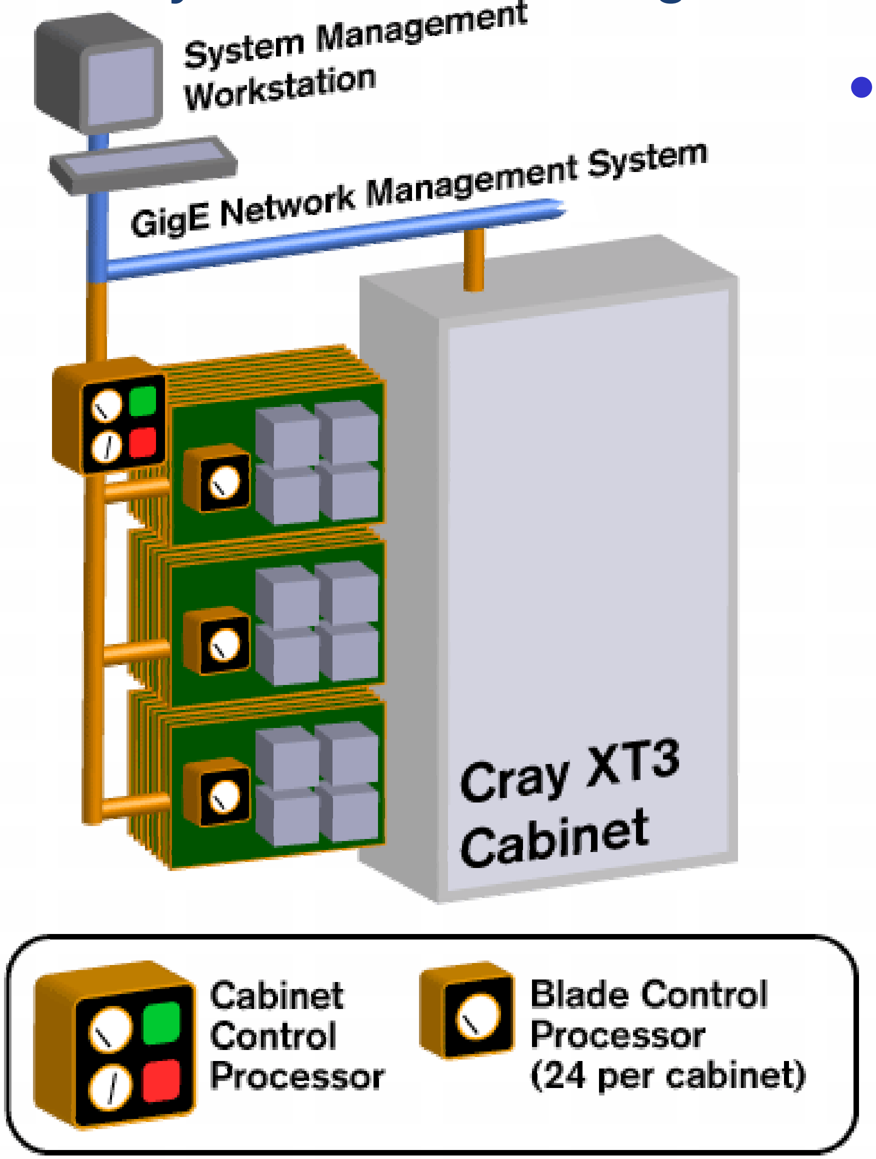
- Simple, microkernel-based software design
- Redundant Power Supplies and Voltage Regulator Modules (VRMs)
- Chipkill Memory protection
- Small number of moving parts
- Limited surface-mount components
- All RAID devices connected with dual paths to survive controller failure
- Seastar Engineered to Provide Reliable Interconnect
- No-Single-Point-of-Failure software design

Cray RAS and Management System (CRMS)



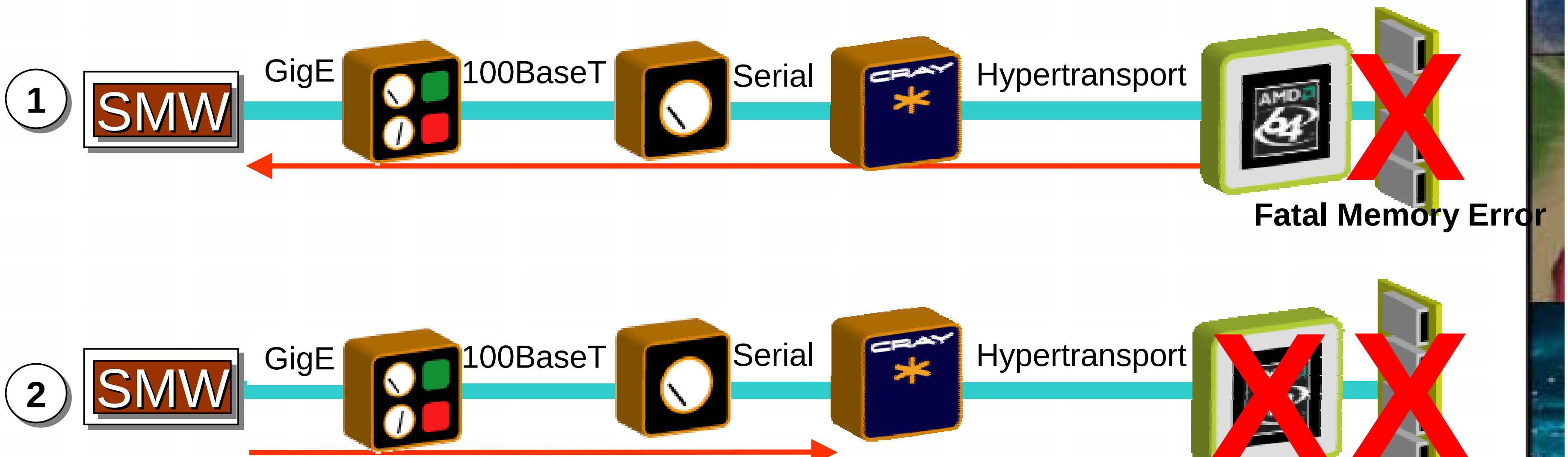


Cray RAS and Management System

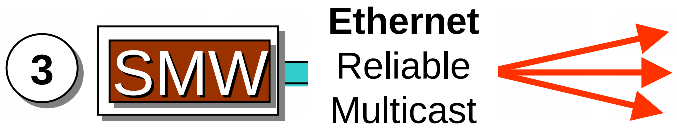

System Management Workstation
GigE Network Management System
Cray XT3 Cabinet
Cabinet Control Processor
Blade Control Processor (24 per cabinet)

- CRMS provides Scalable System Management
  - An independent system with a separate control processors and management network
  - Single System View
  - Software failover management for critical functions
  - Real Time failure monitoring
  - Hot Swap module support

Error Handling Example

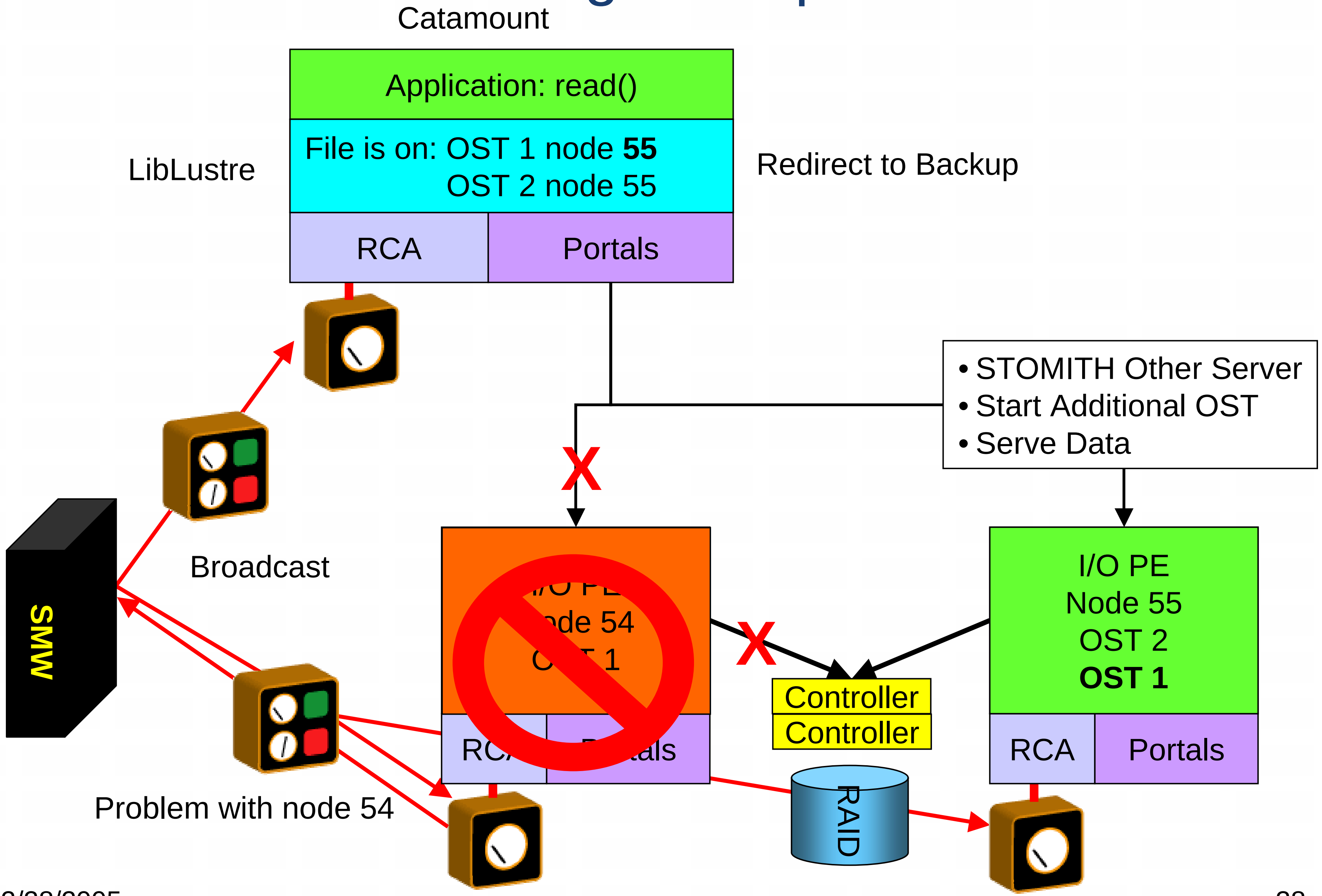


Policy Decision: Kill Opteron and memory



Broadcast failure and handling to all RCAs
Multi-PE job halted, PE removed from service

Lustre Error Handling Example



Catamount

Application: read()

LibLustre

File is on: OST 1 node **55**
OST 2 node 55

Redirect to Backup

RCA

Portals

- STOMITH Other Server
- Start Additional OST
- Serve Data

SMW

Broadcast

I/O PE
Node 54
OST 1

RCA

Portals

Controller

Controller

RAID

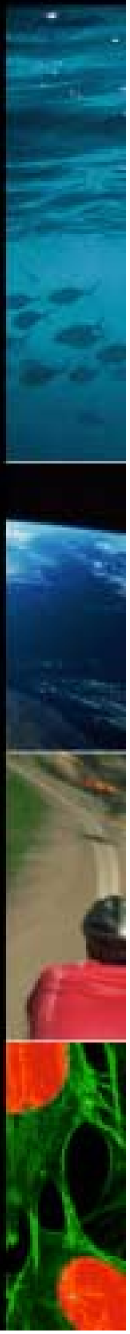
I/O PE
Node 55
OST 2
OST 1

RCA

Portals

Problem with node 54

Cray XT3
Early Results



We Won some Awards…

- HPCwire 2004 Reader's Choice Awards
  - *Most Important Emerging Technology*
  - *Most Innovative HPC Technology*
- HPCwire 2004 Editor's Choice Awards
  - *Most Important Emerging Technology*
  - *Most Innovative HPC Technology*

- "Cray put an industry-standard microprocessor into a bandwidth-rich environment to create an extraordinary high-performance computing system"
  D.H. Brown Associates, Inc.

D.H. BROWN
ASSOCIATES, INC.

Stream Benchmark

Function	T3E1200E	CRAY XT3	Ratio
	(MB/sec)	(MB/sec)	
Copy:	520	5755	11.1
Scale:	517	4464	8.6
Add:	611	4142	6.8
Triad:	622	5549	8.9

Measured on a 2.4 Ghz Opteron with PC3200 DDR DIMMS. Tuned assembler code

Stream Benchmark (parallel)

Function	CRAY XT3 (MB/sec)
Copy :	1.927 TB/s
Scale:	2.085 TB/s
Add :	2.212 TB/s
Triad:	2.212 TB/s

Measured on 559 PEs at Pittsburgh Supercomputing Center.
PGI Generated code, -fastsse -Mnotemporal

NAS Kernels

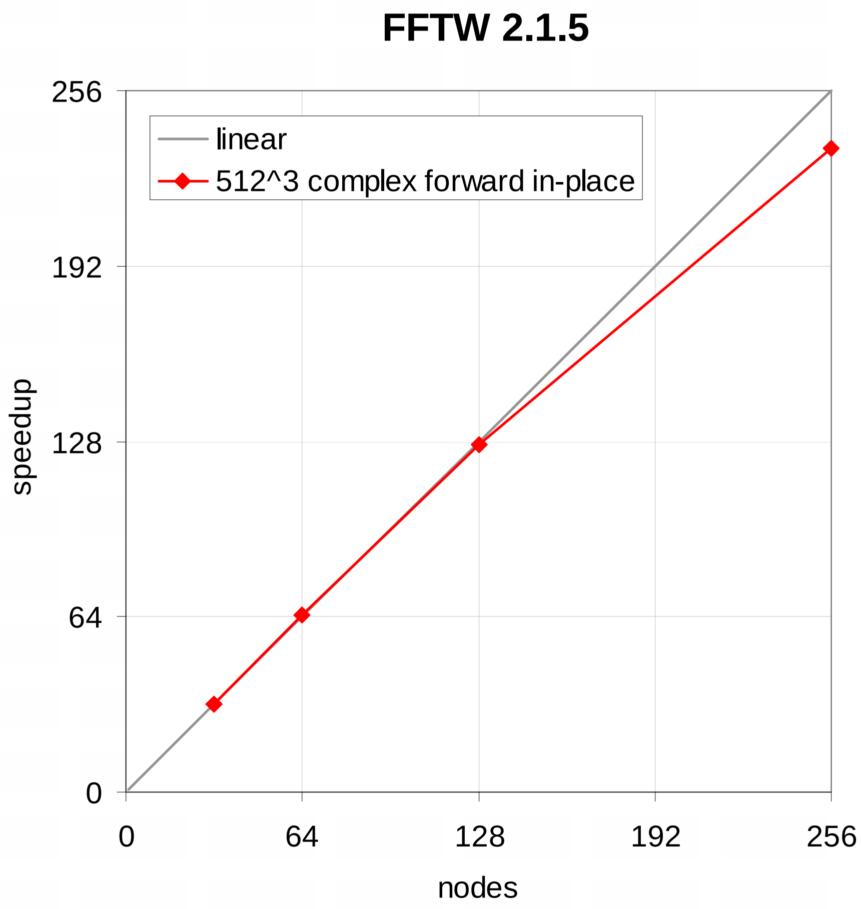
- **All results in Mflops/second (64-bit)**
- **No source code changes**

Kernel	Cray T3E900	CRAY XT3 2.4Ghz	Speedup
MXM	174	1847	10.6
CFFT2D	23	775	33.7
CHOLSKY	26	578	22.2
BTRIX	48	1017	21.2
GMTRY	73	472	6.5
EMIT	246	825	3.4
VPENTA	26	146	5.6
Average			14.7

Interconnect Performance

- Full N x N network run on two cabinets
- Network Topology was 2 x 4 x 24
- Bi-Section Bandwidth across a 2 x 4 “plane” measured at 52.5 GB/sec
- This nets out to 6.5 GB/sec payload bandwidth per link

FFTW Performance



- **Favorable scaling on FFTs and other transpose-intensive operations is essential to numerous applications**

nodes	efficiency
32	1
64	1.01
128	0.990
256	0.918

Standard Benchmarks

- *Performance numbers are extremely preliminary and will improve as the system matures*
- HPCC (552 nodes)
  - HPL : 1,463 GFlop/s (55% of theoretical peak)
  - PTRANS : 49.6 GB/s
  - EP DGEMM : 4.26 GFlop/s per processor
  - EP GUPS : .016 billion updates / sec per processor
- Pallas MPI Benchmarks
  - ping-pong bandwidth: 1094 MB/s
  - Send-receive benchmark: 2170 MB/sec

Codes Ported and running by Feb 2005:

- **<u>Sandia 7x Apps</u>**
  - Alegra
  - CTH
  - ITS
  - SAGE
  - Partisn
  - UMT2000
  - sPPM
  - Salinas
  - Presto
  - Calore
- **TI-05 Apps**
  - Aero
  - AVUS (Cobalt-60)
  - GAMESS
  - Hycom
  - RF CTH
  - WRF
  - Overflow
- **Research and Academic Chemistry**
  - Gromacs
  - NAMD
  - Amber 8
  - CPMD
- **Material Science**
  - LSMS
- **Weather/Climate**
  - ARPS
  - CAM
- **Other**
  - Quake
  - Gasoline
- **Benchmarks**
  - LINPACK
  - HPCC
  - NPB
  - STREAM
  - OSU Bi-section Bandwidth

CRAY XT3
SCALABLE BY DESIGN