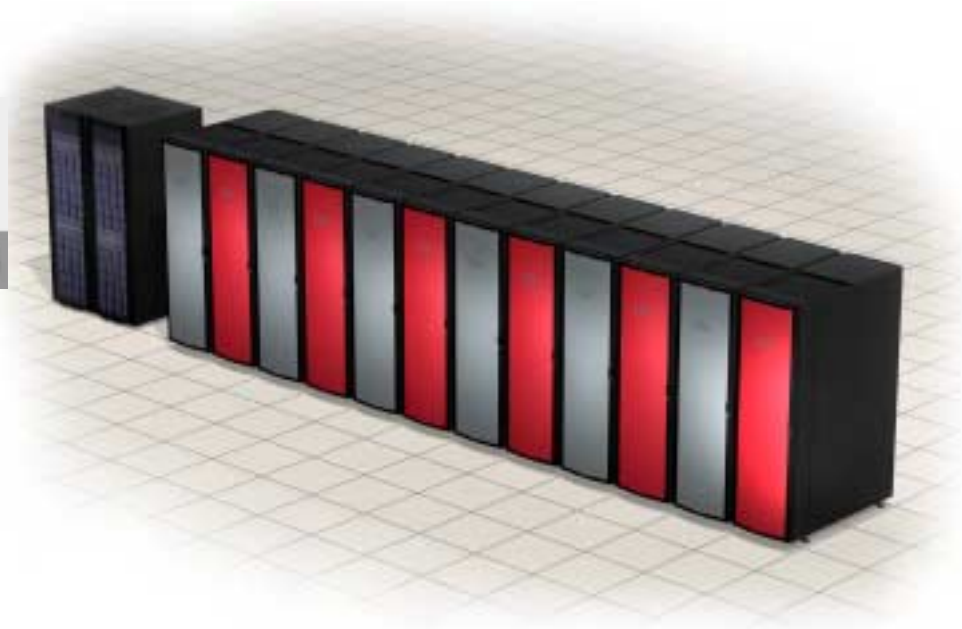
CRAY The Supercomputer Company

CSCS
Swiss National Supercomputing Centre



Cray XT3 architecture

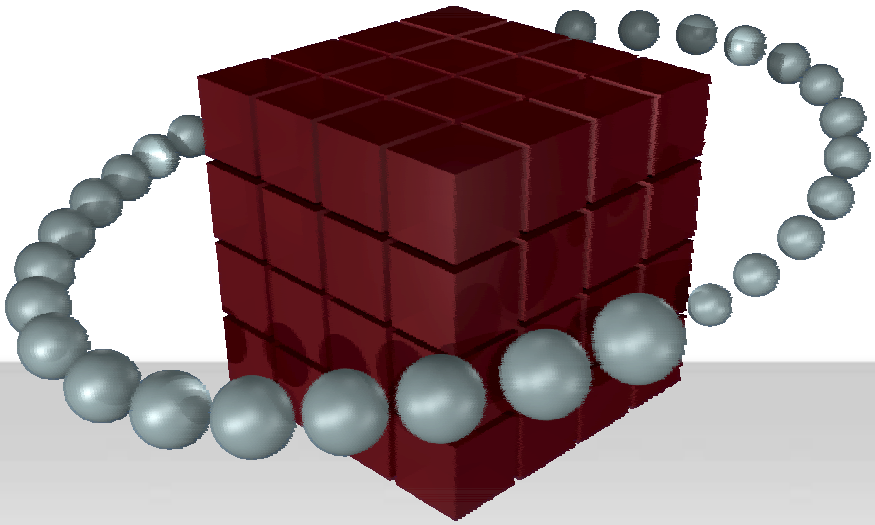
Roberto Ansaloni
roberto@cray.com

CSCS workshop May 3 / 4 2005

Agenda

- Cray MPP history
- Red Storm background and project status
- Cray XT3 architecture
  - Cray XT3 node
  - Cray XT3 blades, cages, cabinets
  - Cray XT3 configurations, topology
- CSCS stage0 and stage1 configurations
- Cray XT3 scalable software
  - Catamount
  - Application launching process
  - CRMS
  - lustre

Cray MPP history



MPP Computing at Cray

MPP Decision:

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

1991

1993

Cray T3D:

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

1996

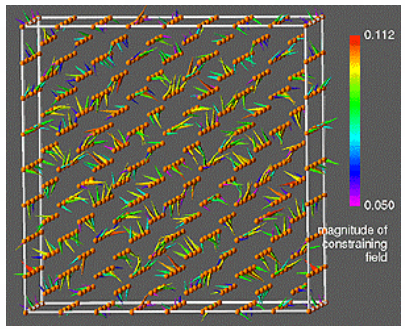
Cray T3E:

- MPI
- UNICOS mk
- Stream buffers
- Gigaring

MPP Computing at Cray

Cray T3E1200:

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



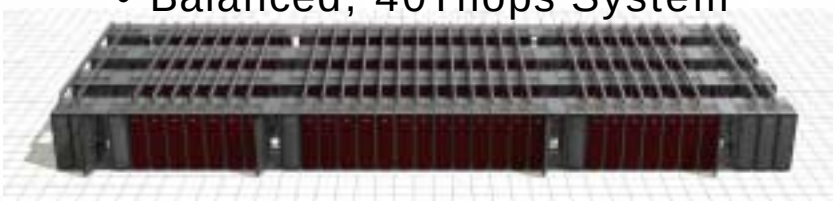
MPP: "Adams"

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

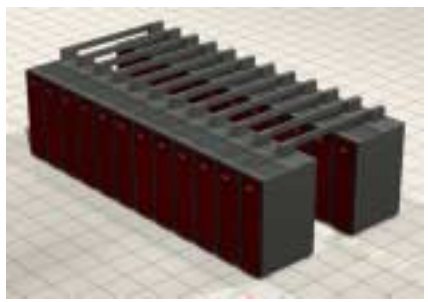


Cray XT3:

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

Cray XT3+:

- DDR2 Memory
- Faster Interconnect



Red Storm Background & Status

CSCS workshop May 3 / 4 2005

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
- 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."

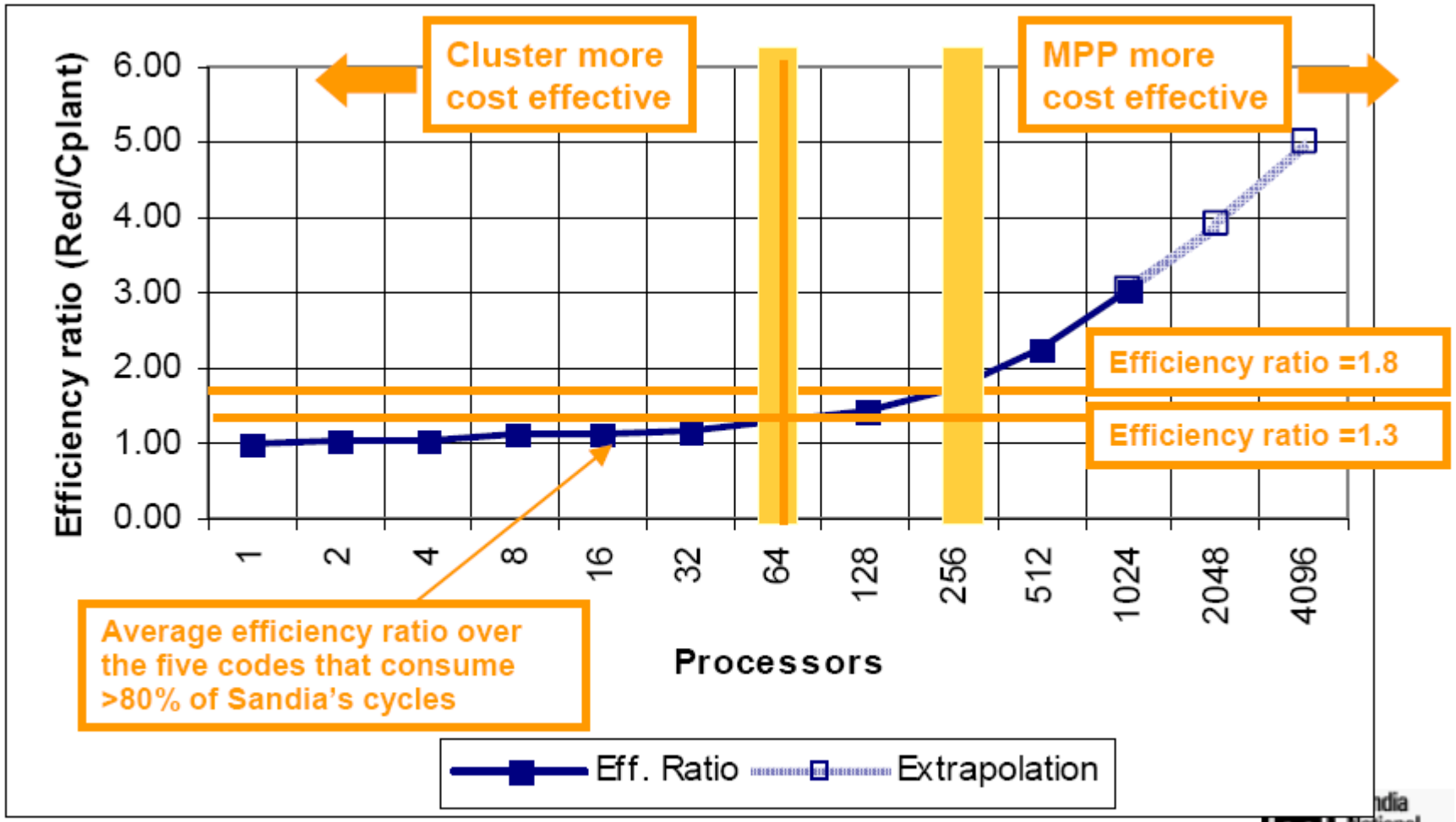
Bill Camp, Sandia Director of Computers, Computation, Information and Mathematics

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



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

Red Storm status

- SeaStar was checked out in September
- Assembling and testing of individual cabinets started in September
- First shipment to Sandia was October 8th
- First row of Red Storm was shipped at the end of October
- All Red Storm computer system hardware is at Sandia
- The system is integrated and has been booted repeatedly in three images: a 5760 processor large partition, a 2592 processor partition, and a 2016 processor partition.
- We have run all 7X applications with normal I/O on up to 4096 processors

Sandia 7X Applications Progress:

Code	Version	1/24/05	3/9/05	3/20/05
Alegra	4.5	8 PEs	256 PEs	256 PEs
CTH	6.0 Mar 02	44 PEs	1024 PEs	5120 PEs
ITS	6/23/04	16 PEs	670 PEs	3827 PEs
SAGE	3/10/03	16 PEs	1024 PEs	1900 PEs
Partisn	9/16/02	44 PEs	564 PEs	1900 PEs
UMT2000	1.2.2 (1/28/02)	4 PEs	1024 PEs	3000 PEs
sPPM	2.0 (1/1/04)	16 PEs	1331 PEs	3600 PEs
Salinas	1.2	ported	64 PEs	1000 PEs
Presto		Compile Issues	256 PEs	1536 PEs
Calore		Compile Issues	256 PEs	1024 PEs

Cray XT3 Orders

CSCS Swiss National Supercomputing Centre

ORNL Oak Ridge National Laboratory

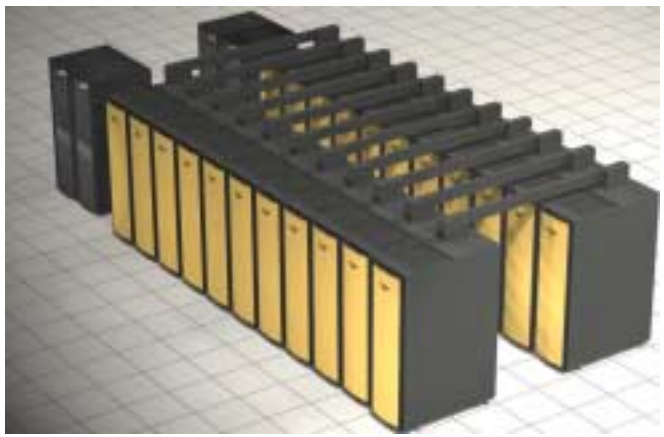
ERDC US Army Engineer Research and Development Center

JST Japan Science and Technology Agency

Some Early Cray XT3 Installs / Orders

We installed 4 Cray XT3 partial systems in 2004

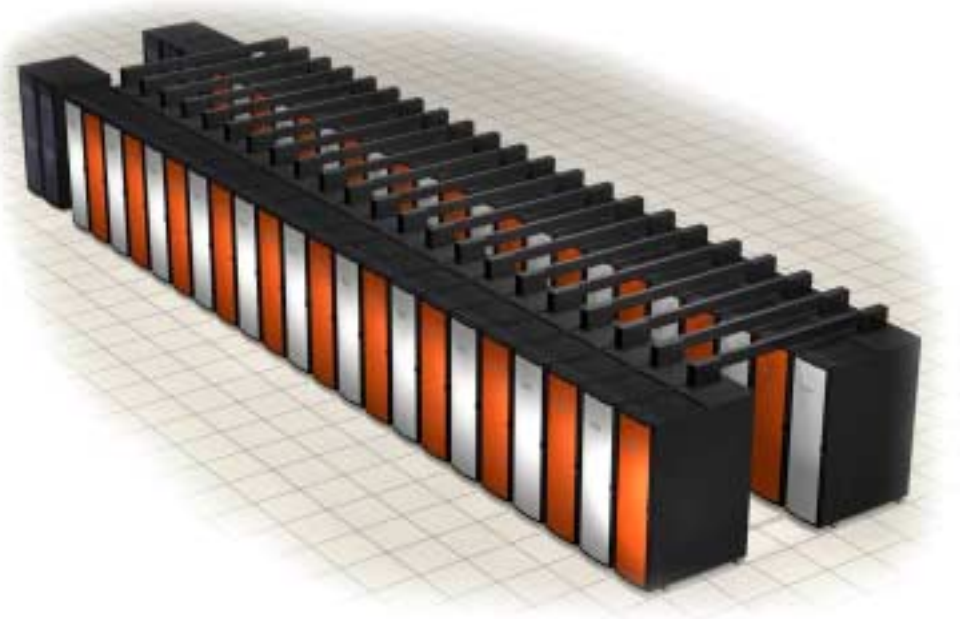
- Sandia
  - 124 Cabinet, 40 Tflop System
  - Almost all of system is delivered
  - Intention is to go to 100 Tflop in 2005
- PSC
  - 22 Cabinet System
  - 10 Tflops



PSC as of mid-February

Some Early Cray XT3 Orders

- Oak Ridge National Laboratory
  - 20 Tflop/s system in 2005
  - 20 cabinets already on site
- Classified Customers
  - 3 Systems
  - 2 US, one outside the US



Some Early Cray XT3 Orders

- U of Tokyo - Japan Science and Technology Agency
  - Final system will have two chassis
- Japan Advanced Institute of Science and Technology (JAIST)
  - First large memory machine (8GB per processor)
  - First Pink Cray

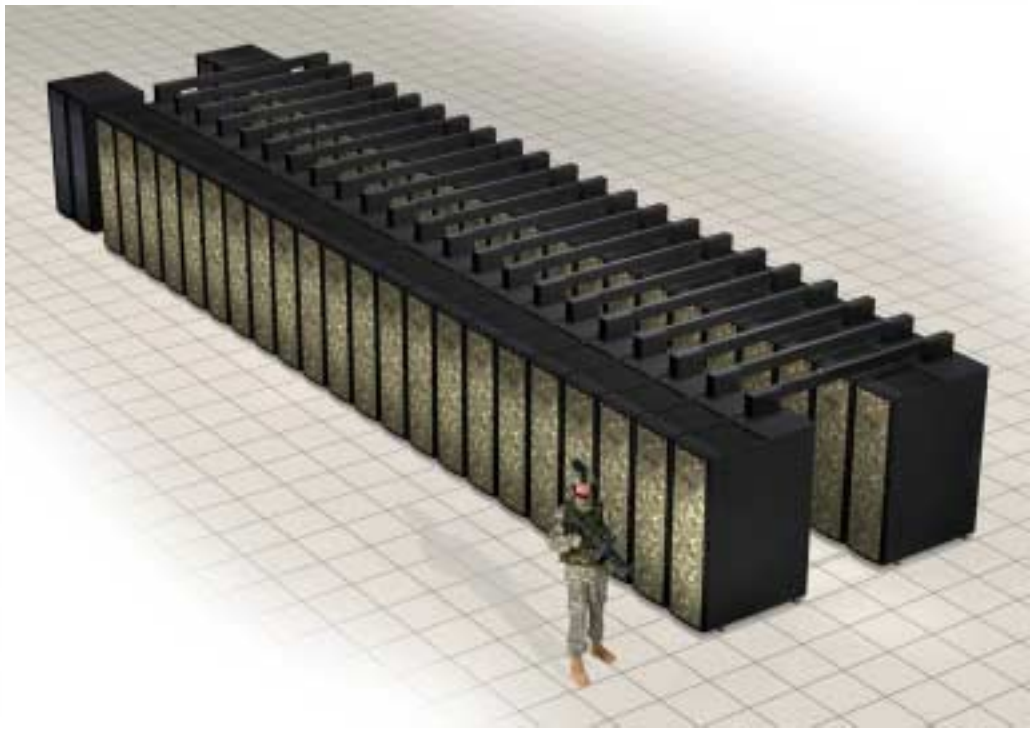
Japan Science and Technology Agency

JAIST
JAPAN ADVANCED INSTITUTE OF SCIENCE AND TECHNOLOGY
1990

Some Early Cray XT3 Orders

- US Army Engineer Research and Development Center (ERDC)
  - 44 Cabinet System
  - 2.6 Ghz parts
  - Q2 delivery

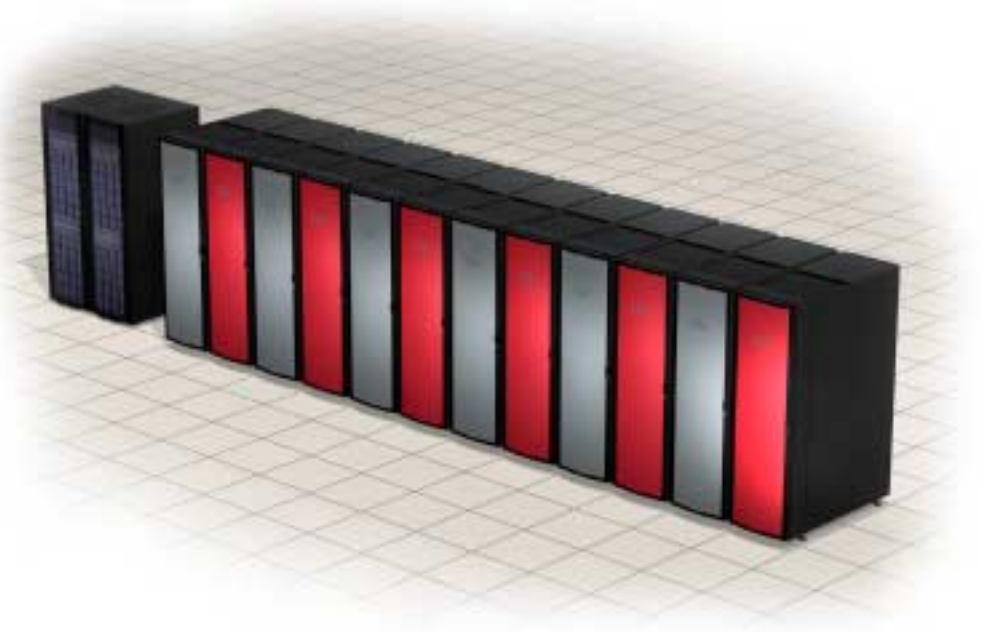
ERDC
US Army Engineer Research and Development Center



Some Early Cray XT3 Orders

- Swiss National Supercomputing Centre (CSCS)
  - 12 cabinet, ~6 Tflop/s system
  - First system in Europe
  - Q2 Delivery
  - 2.6 Ghz Opterons
  - Site will host Cray Workshop in Sept 2005 and CUG in 2006

CSCS
Swiss National Supercomputing Centre



CSCS Stage0 System – SN 2362

- Cabinets: 1 cpu, 1 I/O
- Topology: 3 x 4 x 8
- Compute blades: 21 (84 nodes)
  - AMD Opteron 146 - 2.0 GHz
  - Memory 2GB, 2 x 1GB PC3200
- Service blades: 3 (6 nodes)
  - AMD Opteron 146 - 2.0 GHz
  - Memory 4GB, 4 x 1GB PC3200
- I/O subsystem
  - System RAID 4 x (4+1) x 146 GB – total 2.4 TB

CSCS Stage1 System – SN 2366

- Cabinets: 12 cpu, 2 I/O
- Topology: 12 x 12 x 8
- Compute blades: 275 (1100 nodes)
  - AMD Opteron 152 - 2.6 GHz
  - Memory 2GB, 4 x 512MB PC3200
- Service blades: 13 (26 nodes)
  - AMD Opteron 152 - 2.6 GHz
  - Memory 4GB, 4 x 1GB PC3200
- I/O subsystem
  - System RAID 8 x (4+1) x 146 GB = total 4.7 TB
  - Parallel FS: 3 x 8 x (8+1) x 146 GB = total 28 TB

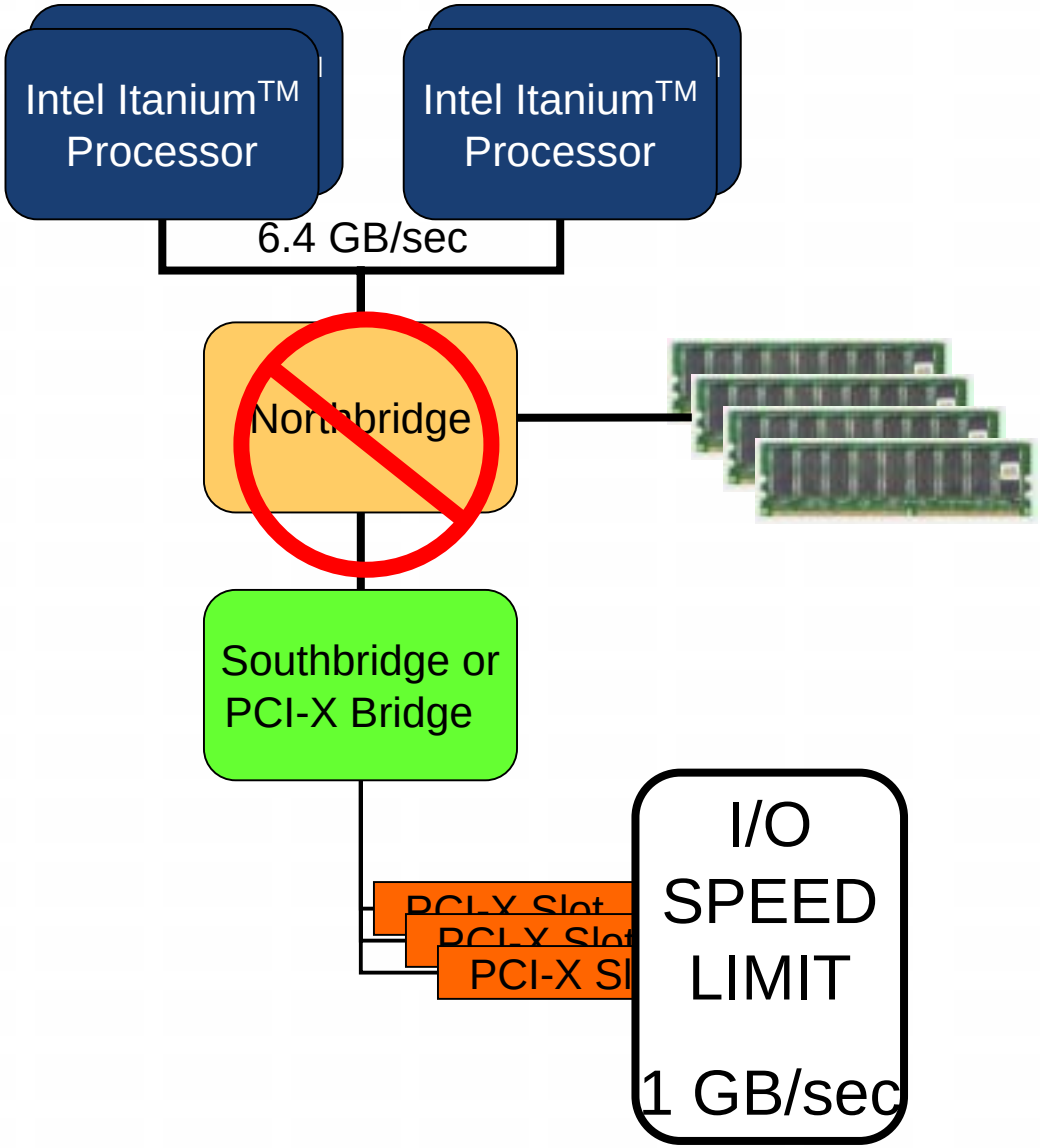
CRAY XT3
Balanced Architecture

CSCS workshop May 3 / 4 2005

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

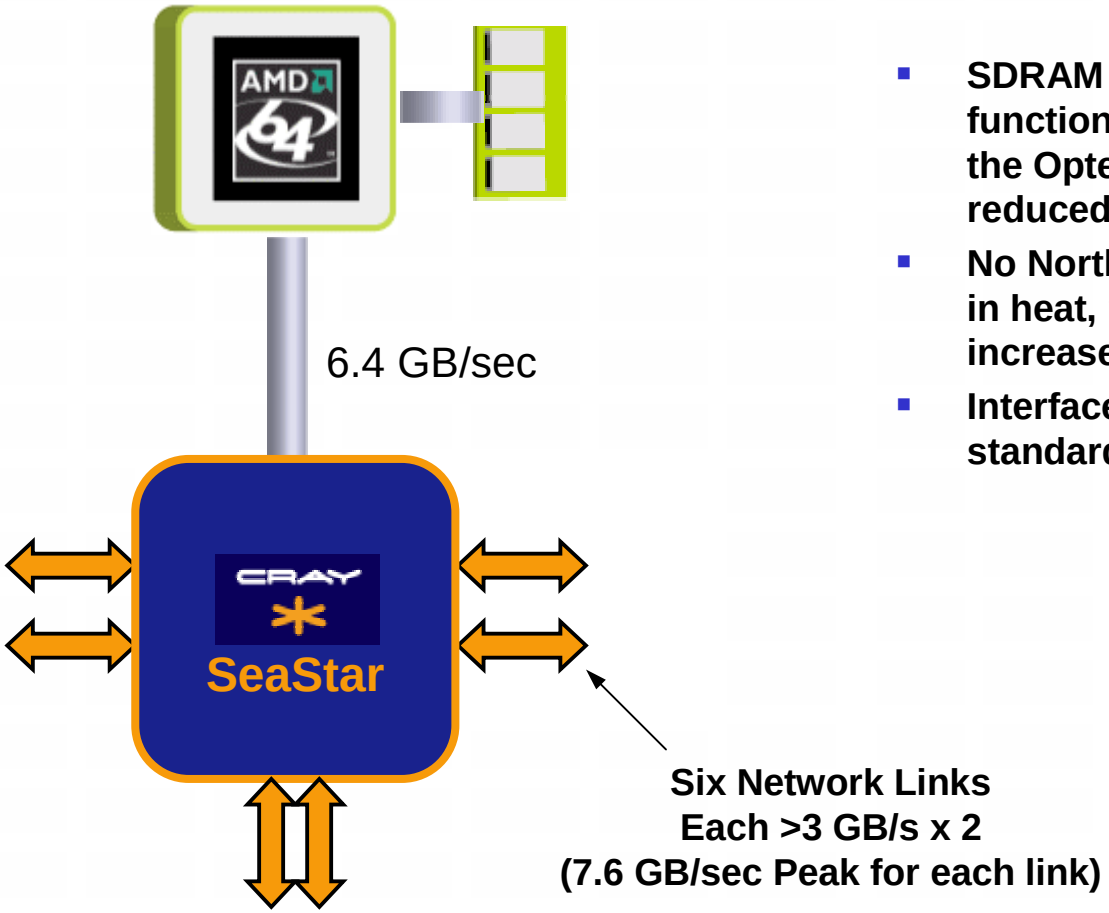
Picking the best Processor: Why not Intel?



- **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**

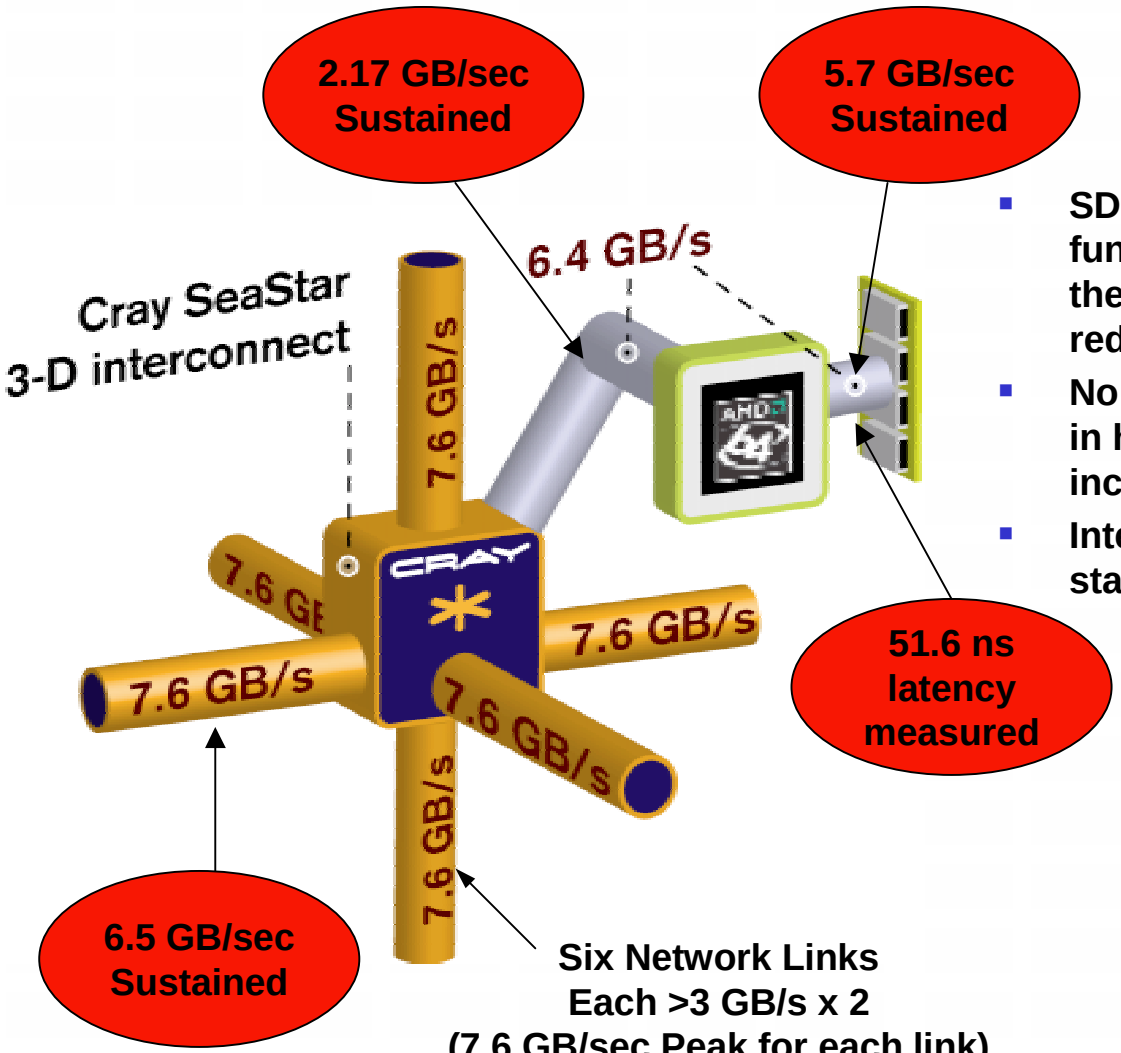
AMD Opteron Generic System

CRAY XT3 PE



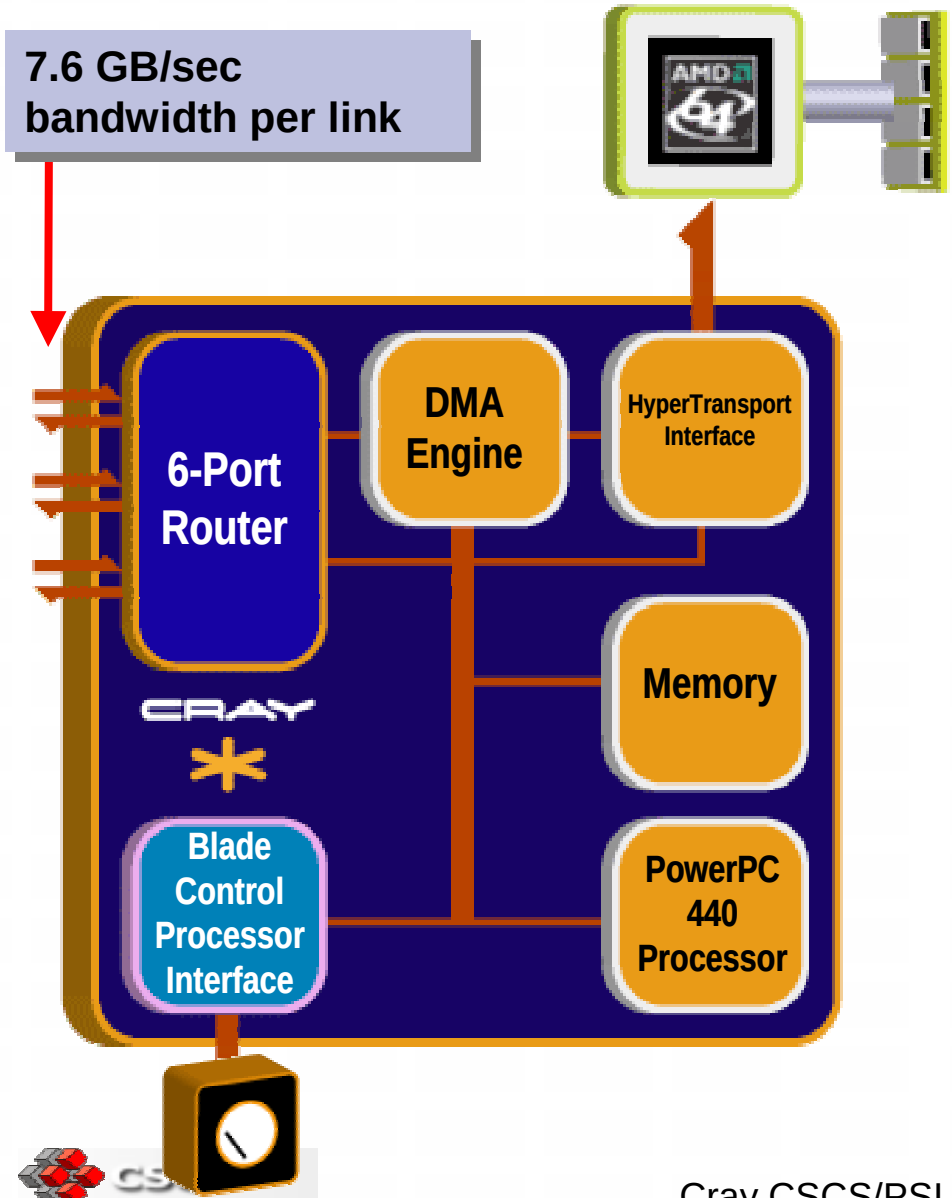
- SDRAM memory controller and function of Northbridge is pulled onto the Opteron die. Memory latency reduced to 60-90 ns
- No Northbridge chip results in savings in heat, power, complexity and an increase in performance
- Interface off the chip is an open standard (HyperTransport)

Cray XT3 Processing Element: Measured Performance



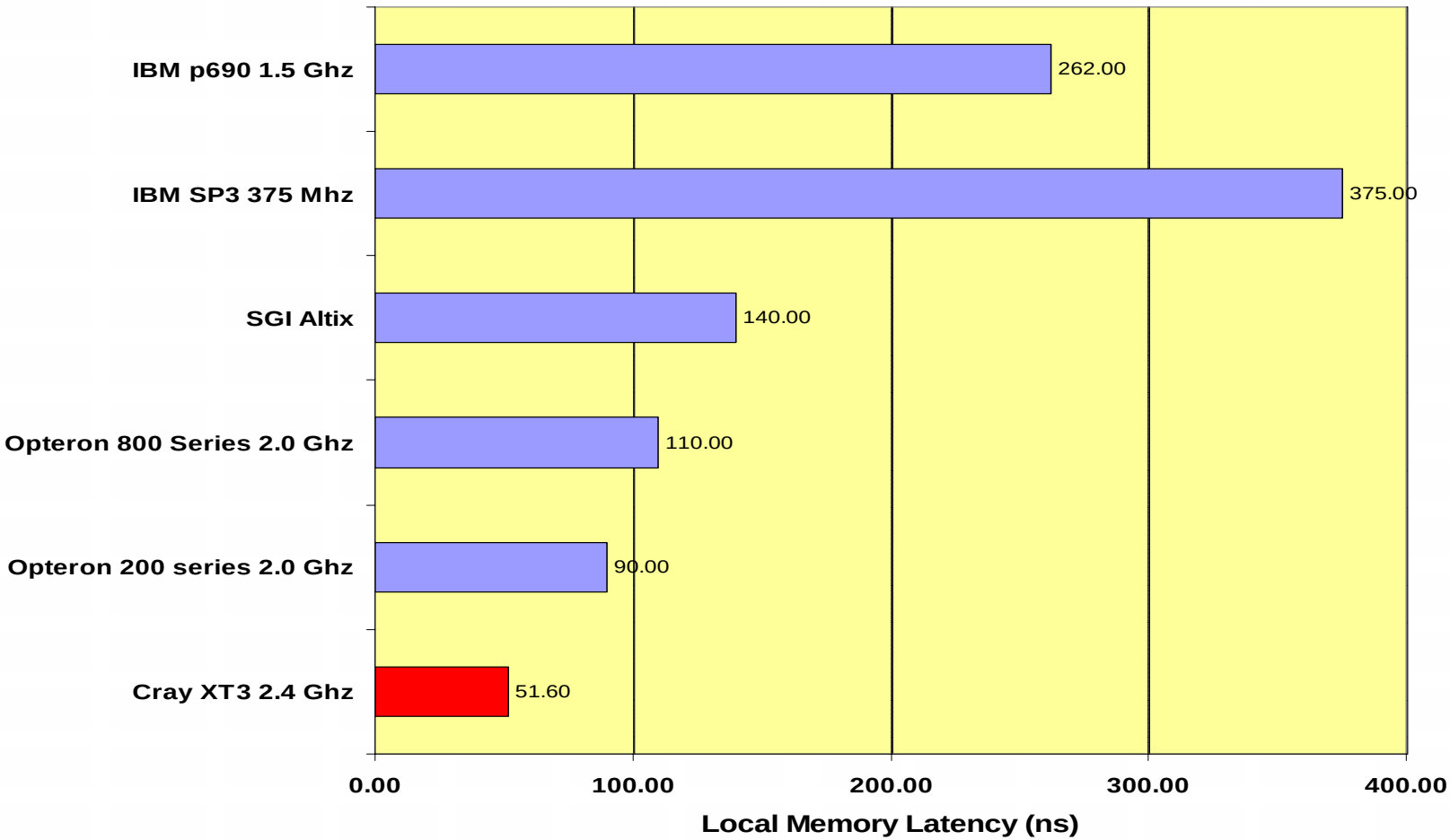
- SDRAM memory controller and function of Northbridge is pulled onto the Opteron die. Memory latency reduced to <60 ns
- No Northbridge chip results in savings in heat, power, complexity and an increase in performance
- Interface off the chip is an open standard (HyperTransport)

Cray SeaStar Internals



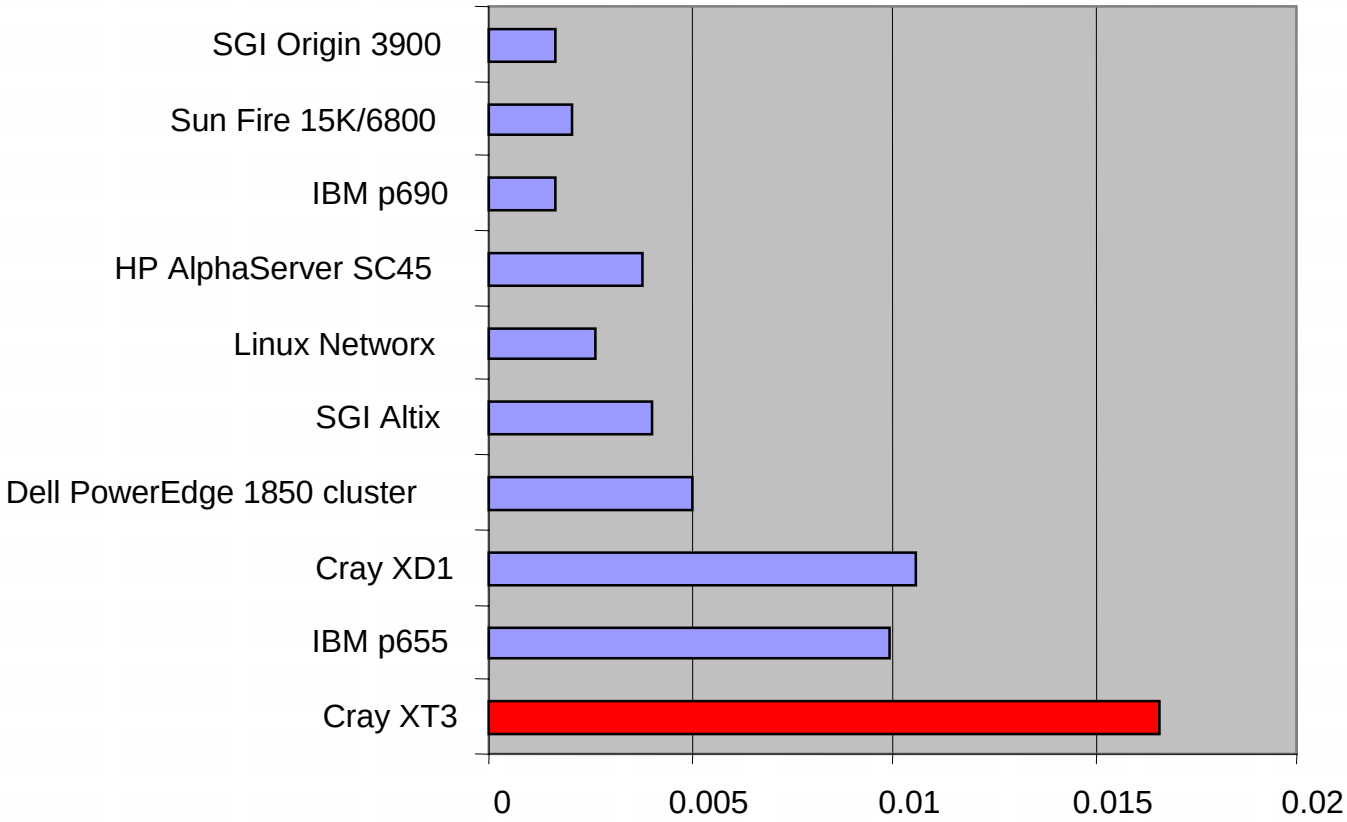
- Each Processor is directly connected to a dedicated SeaStar
- Each SeaStar contains a 6-Port router *and* communications engine
- Provides serial connection to the Cray RAS and Management System

Memory Latency

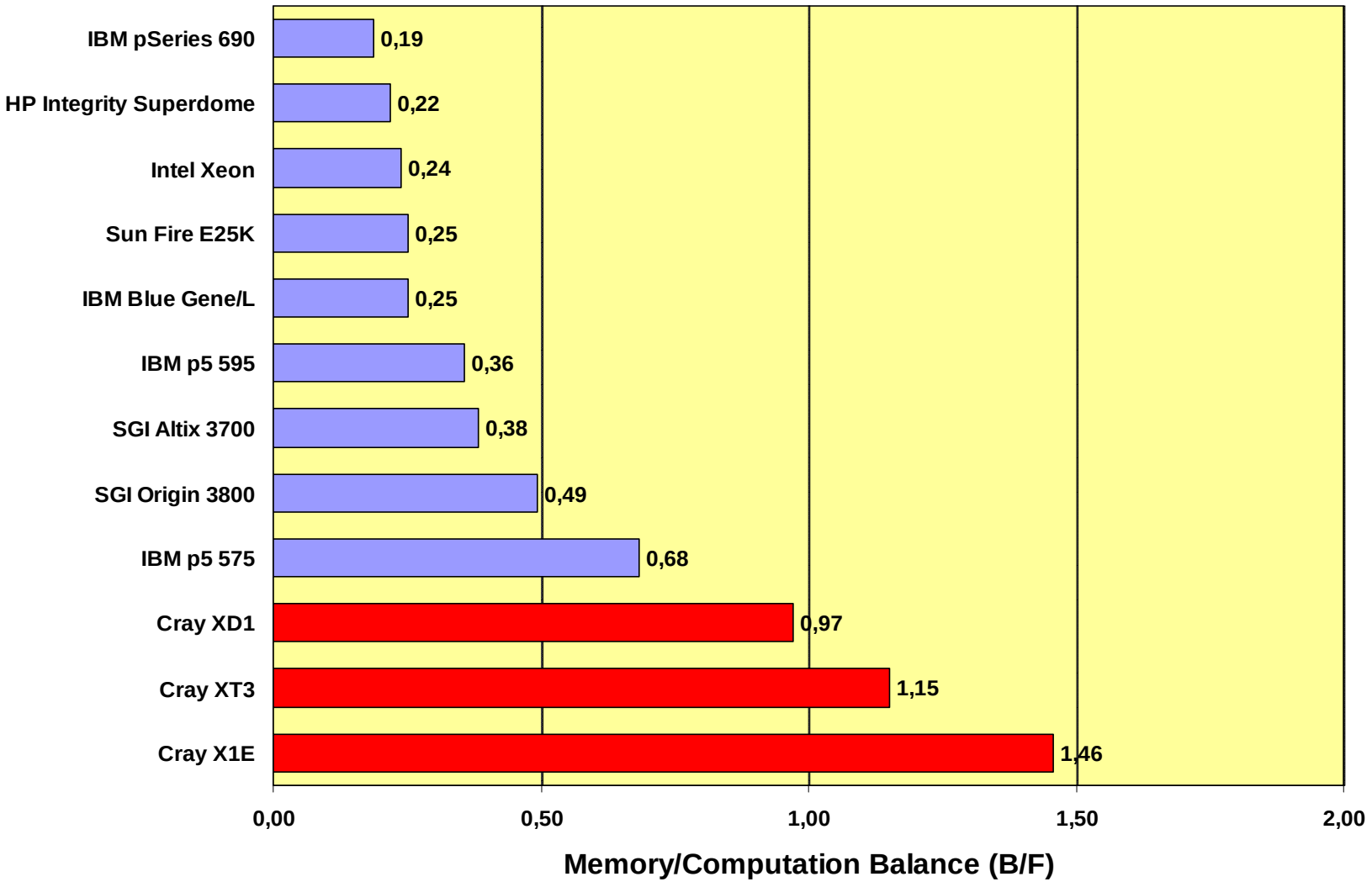


Single Processor architecture yields lowest memory latency

HPCC Random Access Benchmark

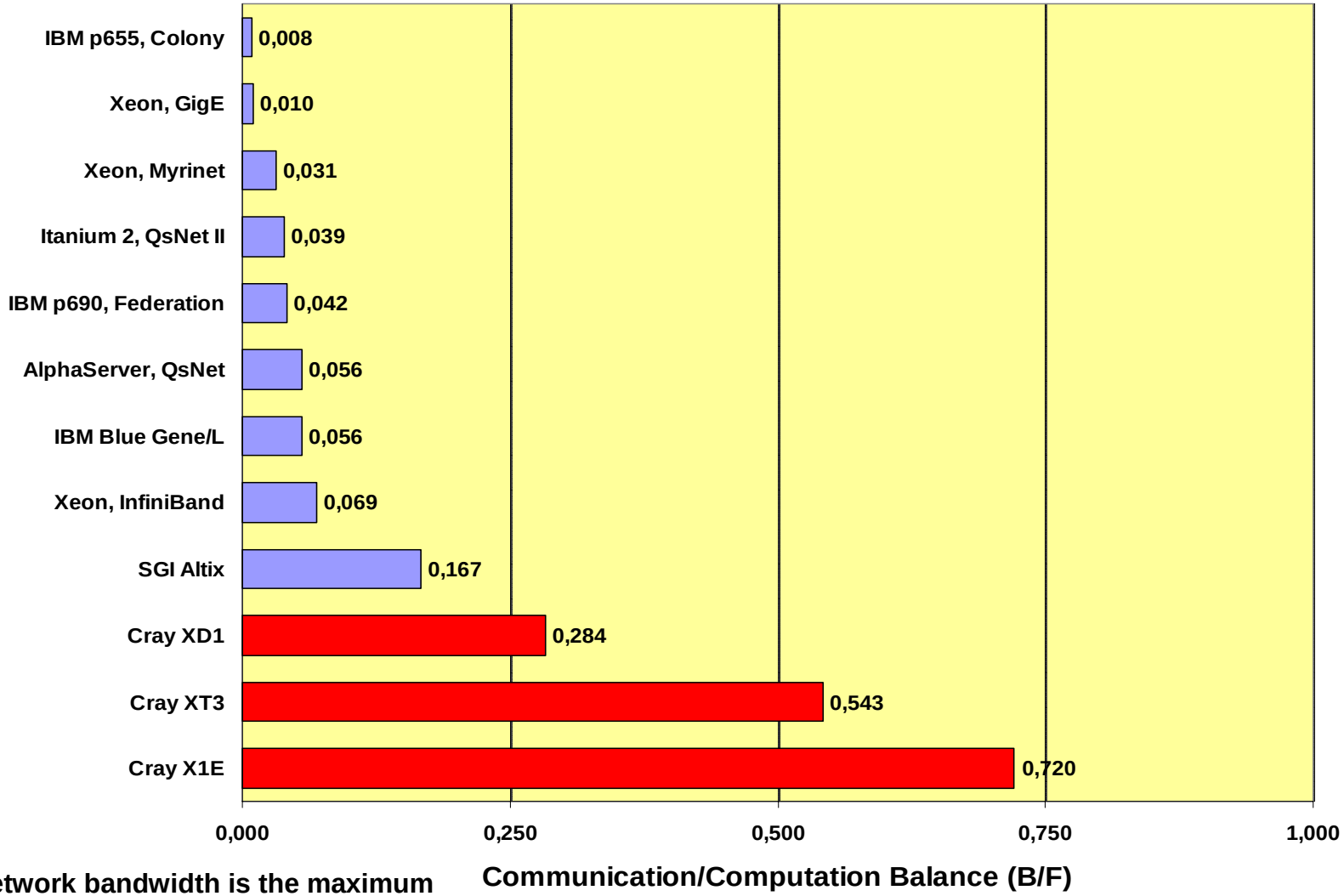


Measured Memory Balance



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

Measured Network Balance

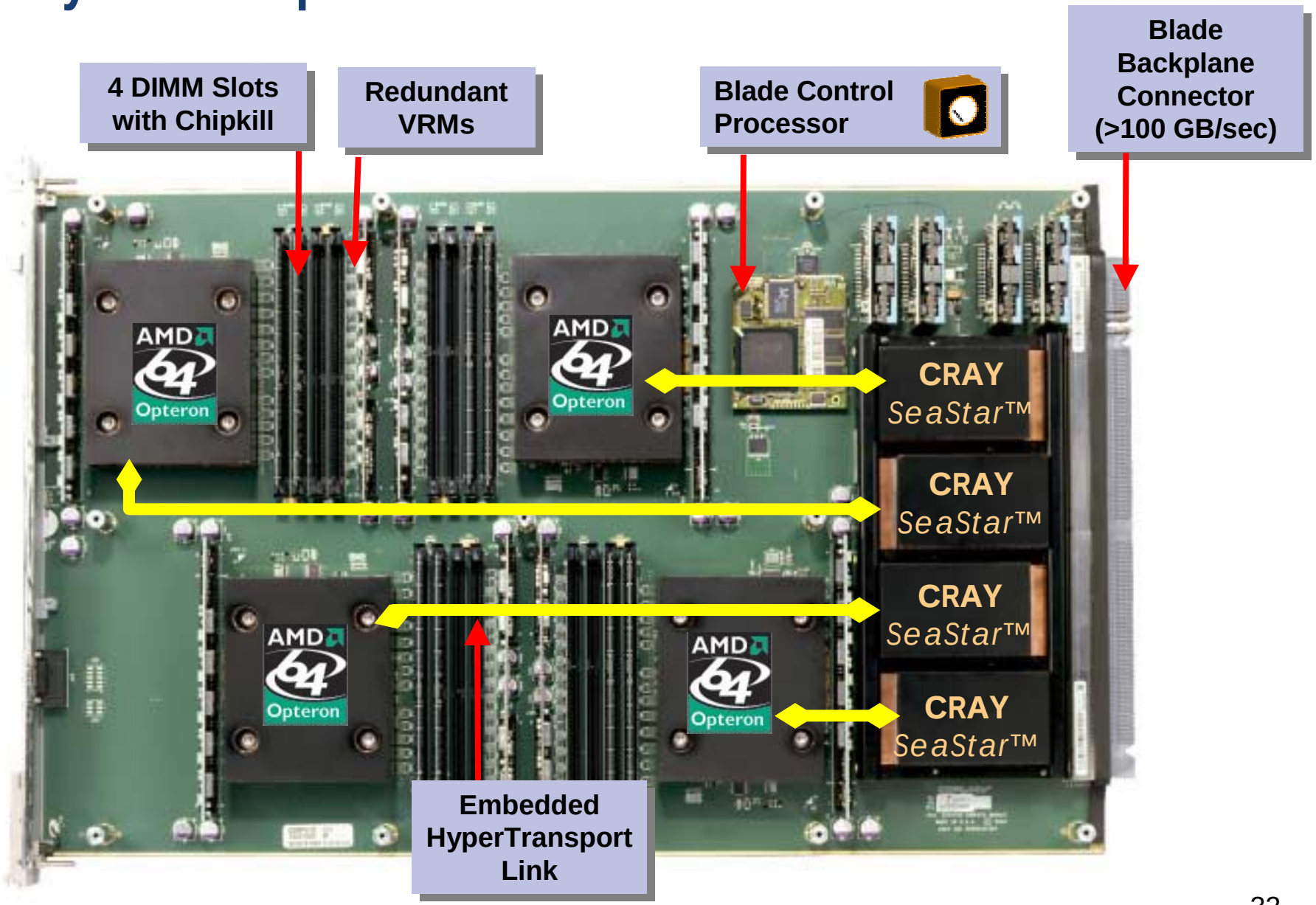


Network bandwidth is the maximum bidirectional data exchange rate between two nodes using MPI

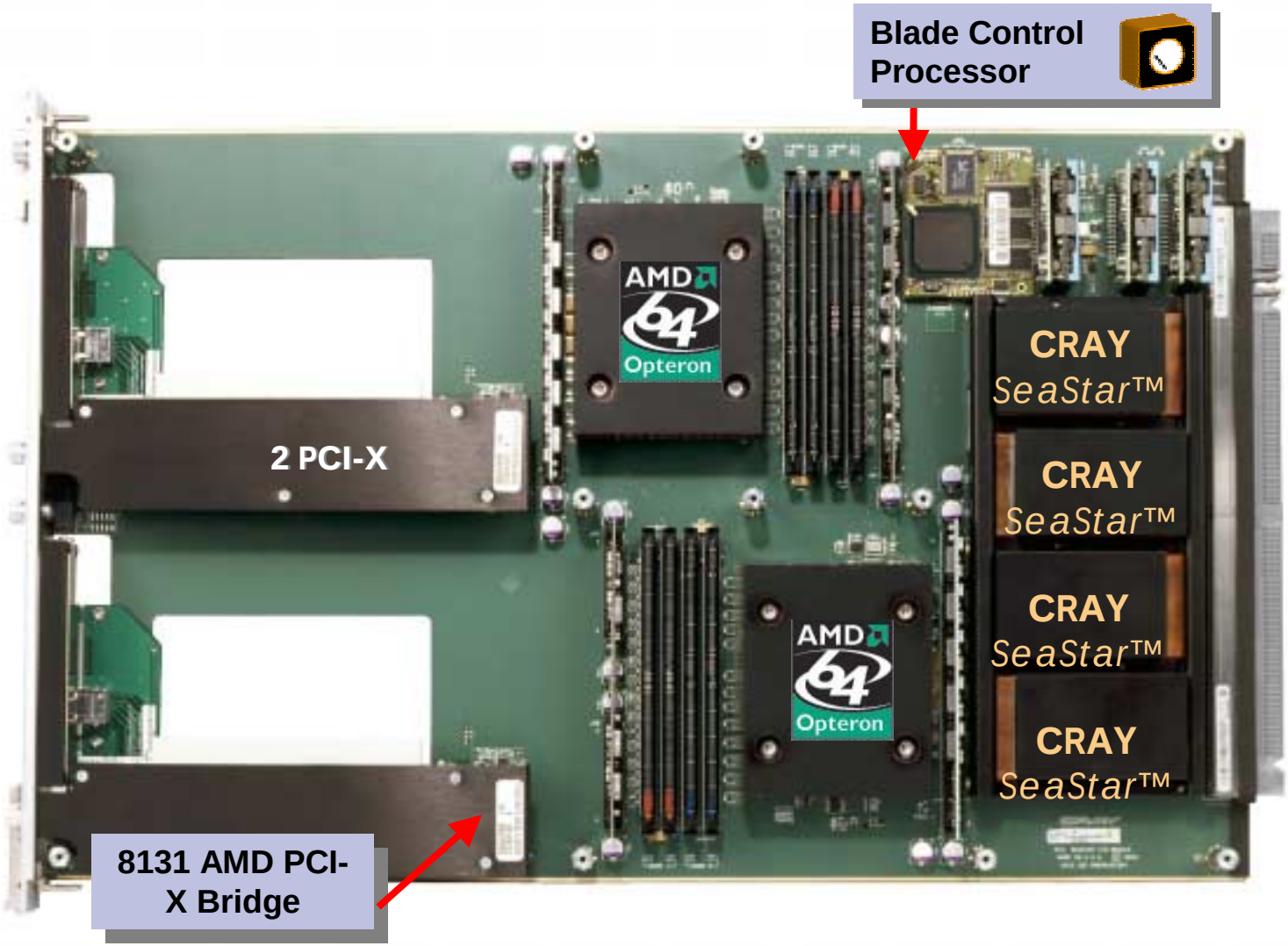
Cray XT3 Architecture Engineering

CSCS workshop May 3 / 4 2005

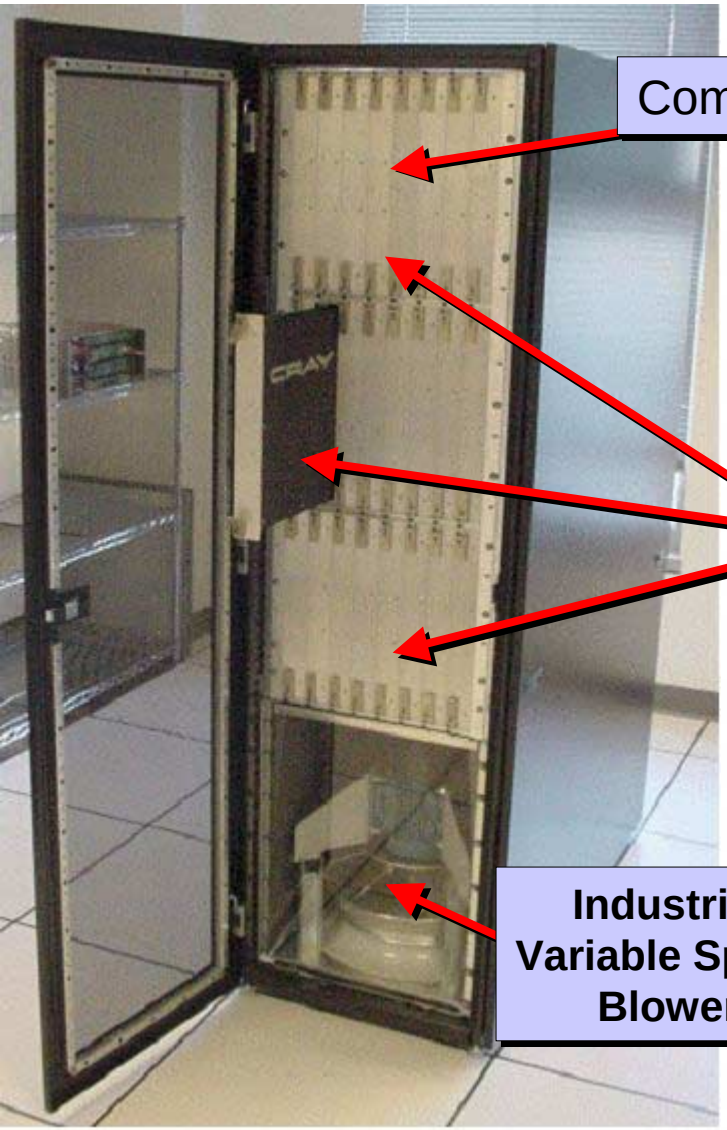
Cray XT3 Compute Blade



Cray XT3 Service and I/O Blade

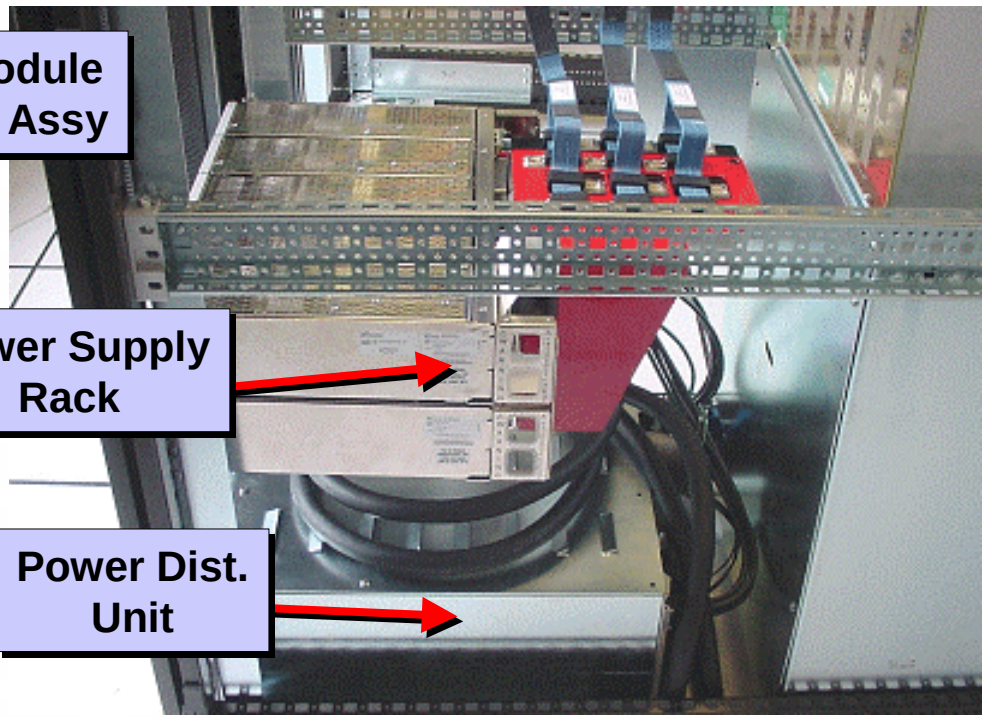


CRAY XT3 Compute Cabinet



Compute Modules - enclosed

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



24 Module Cage Assy

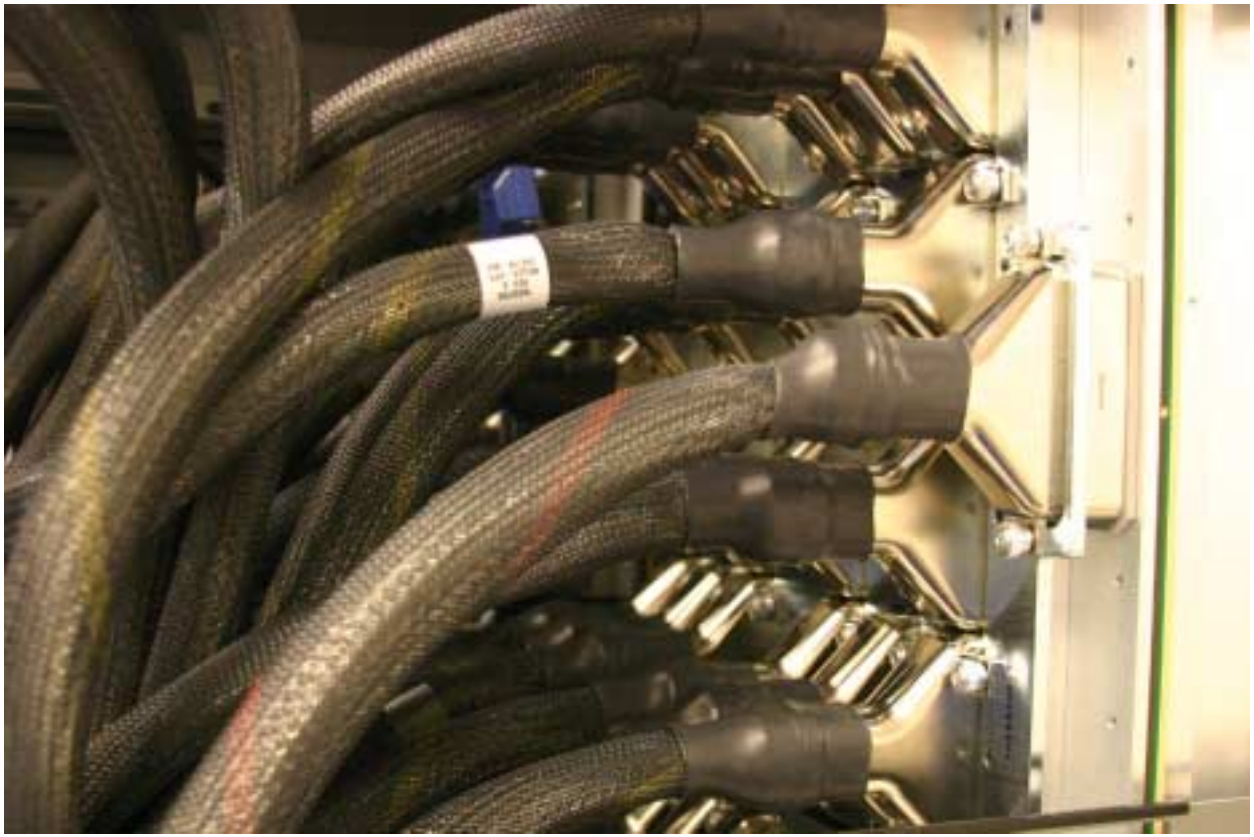
Power Supply Rack

Industrial Variable Speed Blower

Power Dist. Unit

Seastar Cables

- Each Seastar Cable carries 4 torus connections or about 30 GB/sec



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

System Interconnect & Topologies

CSCS workshop May 3 / 4 2005

Topology Tutorial

Class 0 Topology (CSCS stage0)

For a system of up to 3 cabinets, with 1 to 9 chassis, the topology is a full 3D Torus of size N x 4 x 8, where N is the number of chassis.

Class 1 Topology (CSCS stage1)

For a system that is a single row of 4 or more and up to 16 cabinets, the topology is a full 3D Torus of size N x 12 x 8, where N is the number of cabinets.

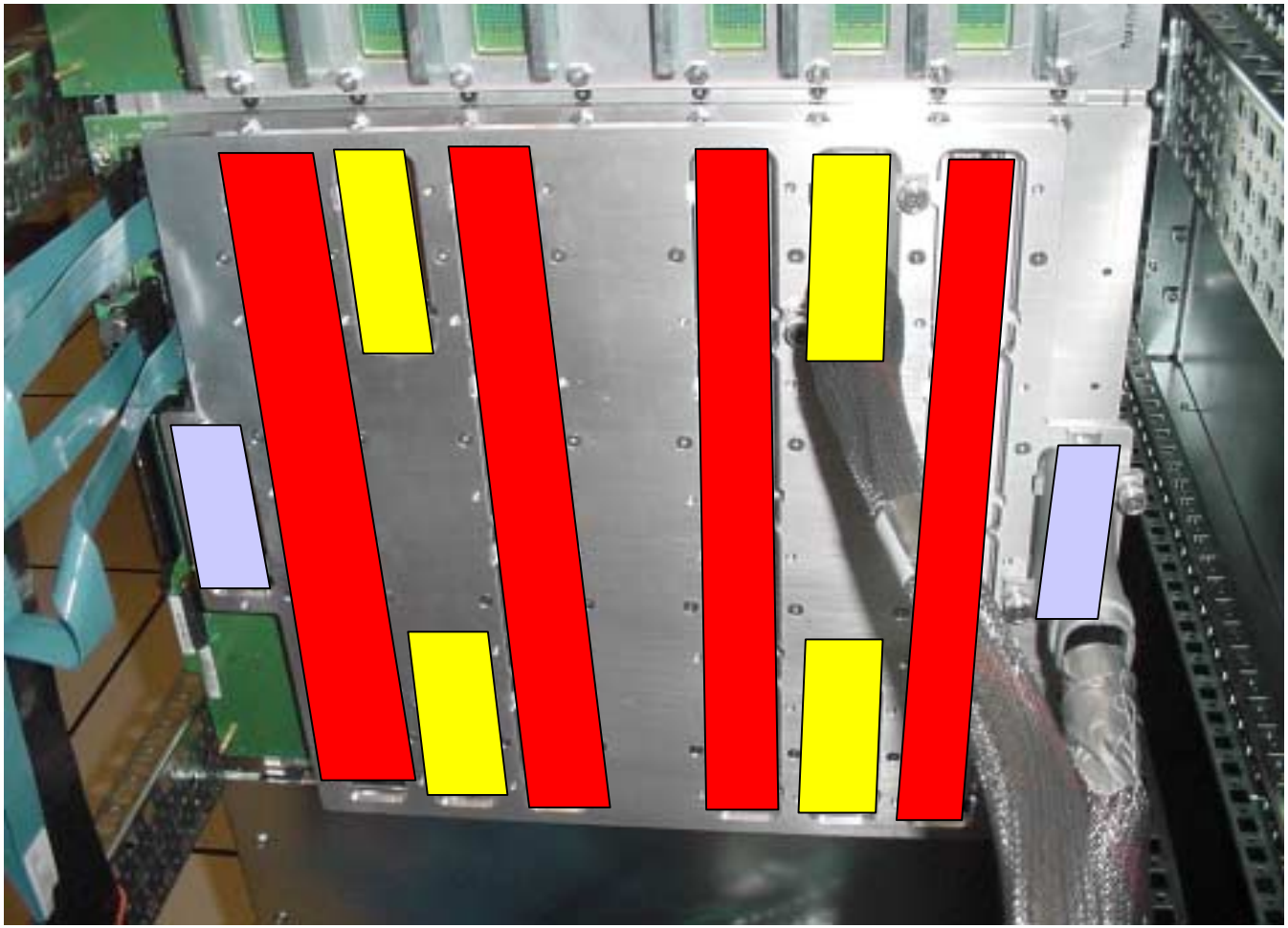
Class 2 Topology

For systems comprised of two rows, the topology is a full 3D torus of size N x 12 x 16, where N is the number of cabinets in a row (total 2 x N cabinets). This class covers configurations from 16 (N=8) to 48 (N=24) cabinets.

Class 3 Topology (Subject to final engineering testing and verification)

For larger, multi-row systems, with an even number of rows, the topology is a full 3D torus of size N x (4 x number of rows) x 24. This class covers configurations from 48 (4 rows, 12 cabinets per row) to 576 (12 rows, 48 cabinets per row) cabinets.

Backplane Docking Plate

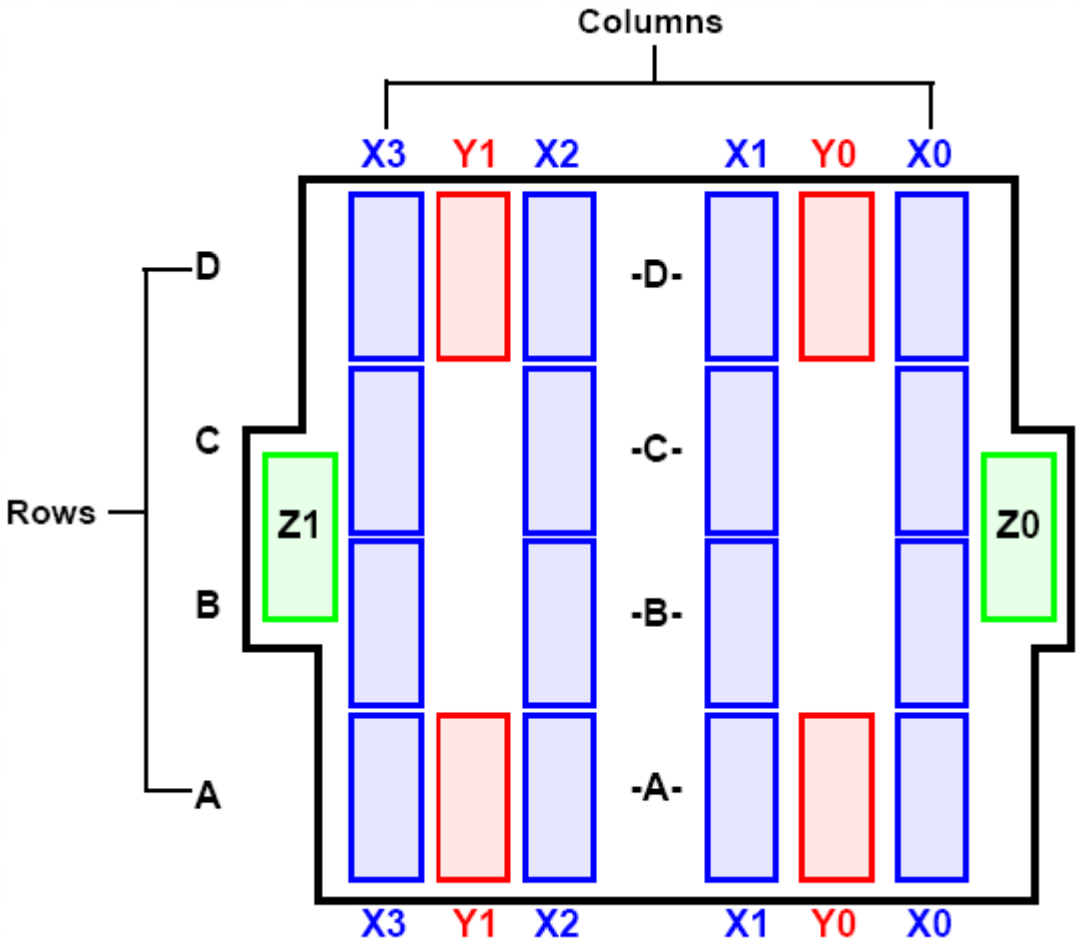


X: 64 Seastar connections

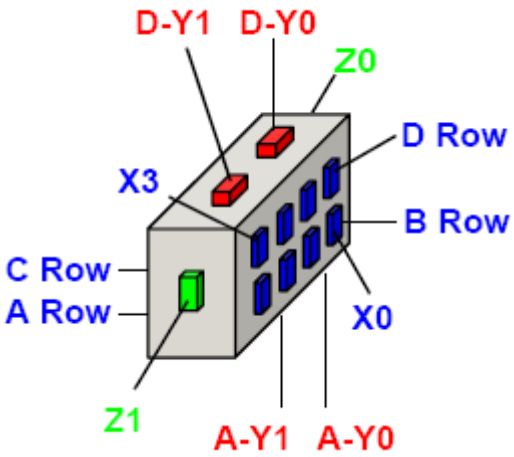
Y: 16 Seastar connections

Z: 8 Seastar connections

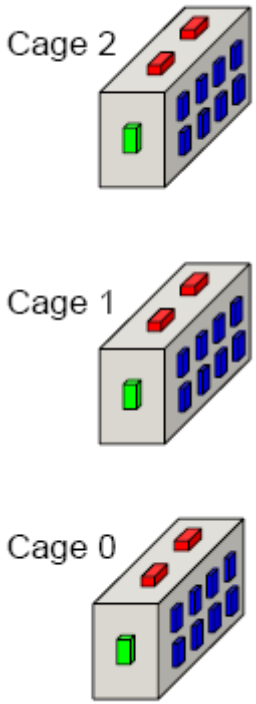
Backplane Connectors



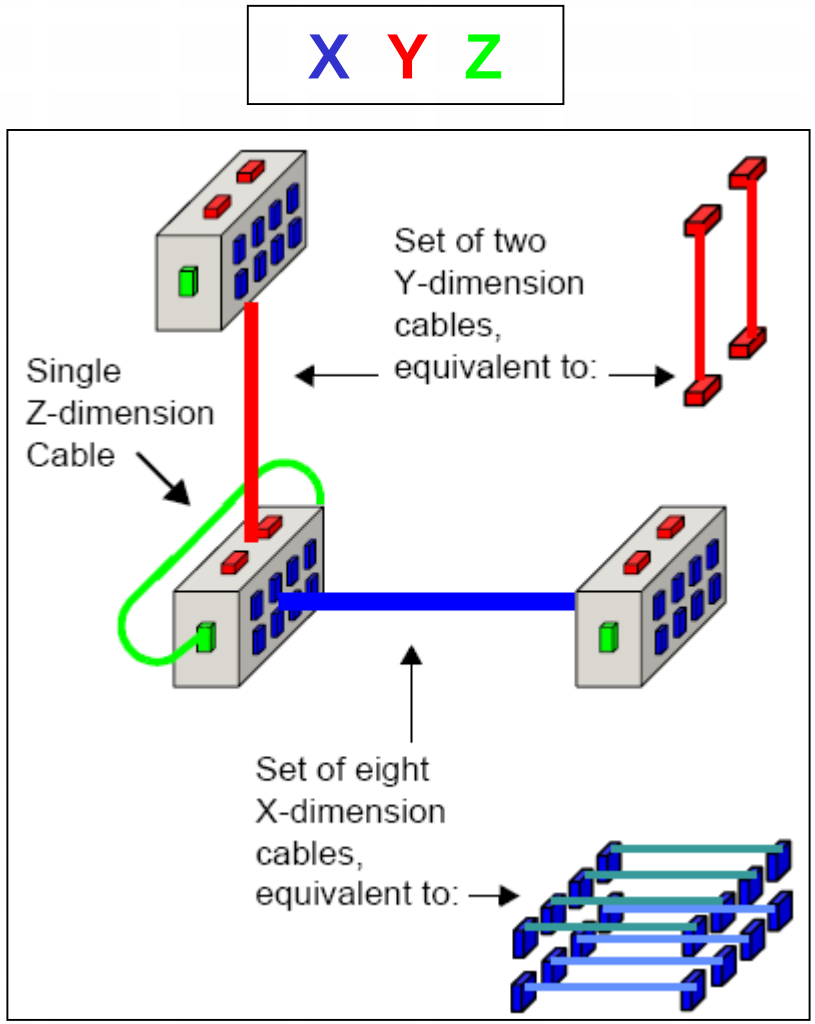
Topology Conventions



Chassis or Cage



Cabinet

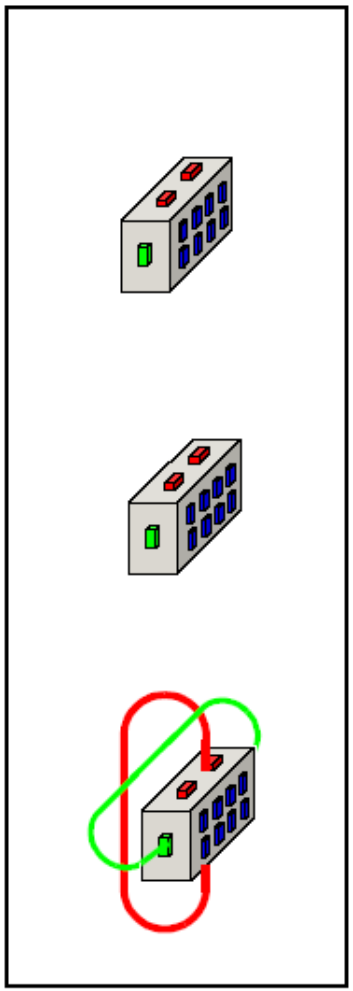


Cables

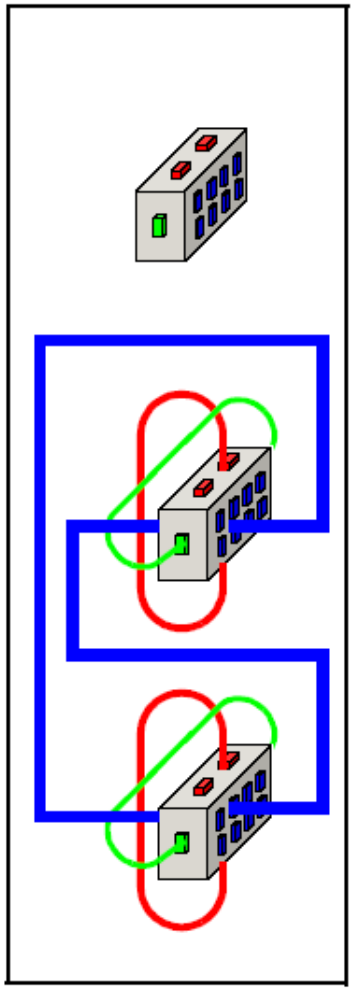
Single Cabinet Configurations

- 1 chassis collaboration
  - 8 blades
  - 6 service and IO PEs
  - 20 compute PEs
  - 1 x 4 x 8
  - Class 0 topology
- 3 chassis single cabinet
  - 24 blades
  - 6 service and IO PEs
  - 84 compute PEs
  - 3 x 4 x 8
  - Class 0 topology

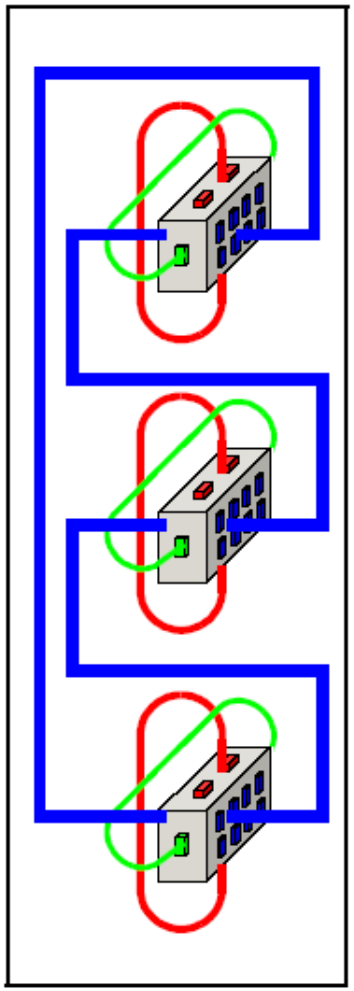
Single Cabinet, Class 0



One Cage
(1 x 4 x 8)



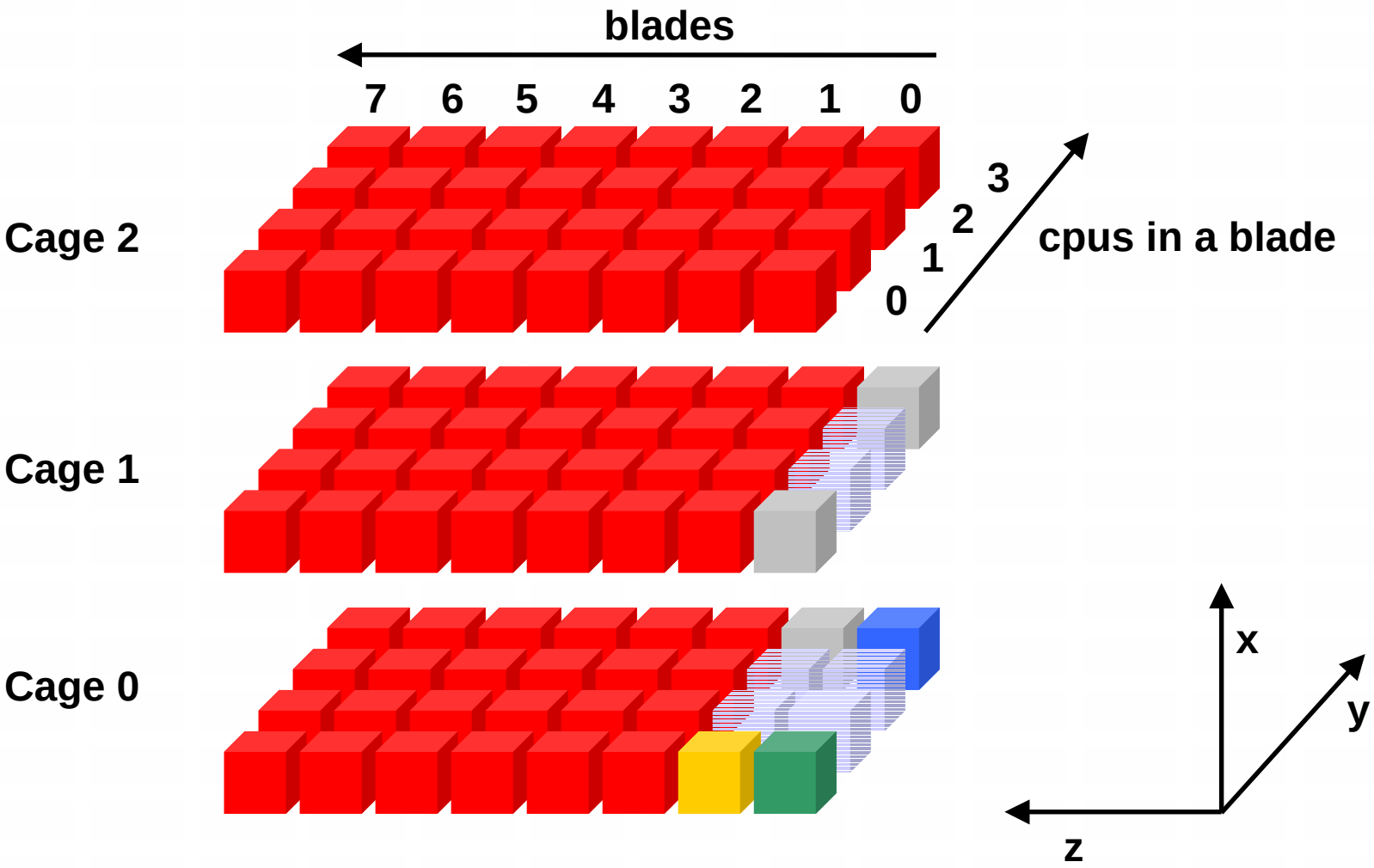
Two Cages
(2 x 4 x 8)



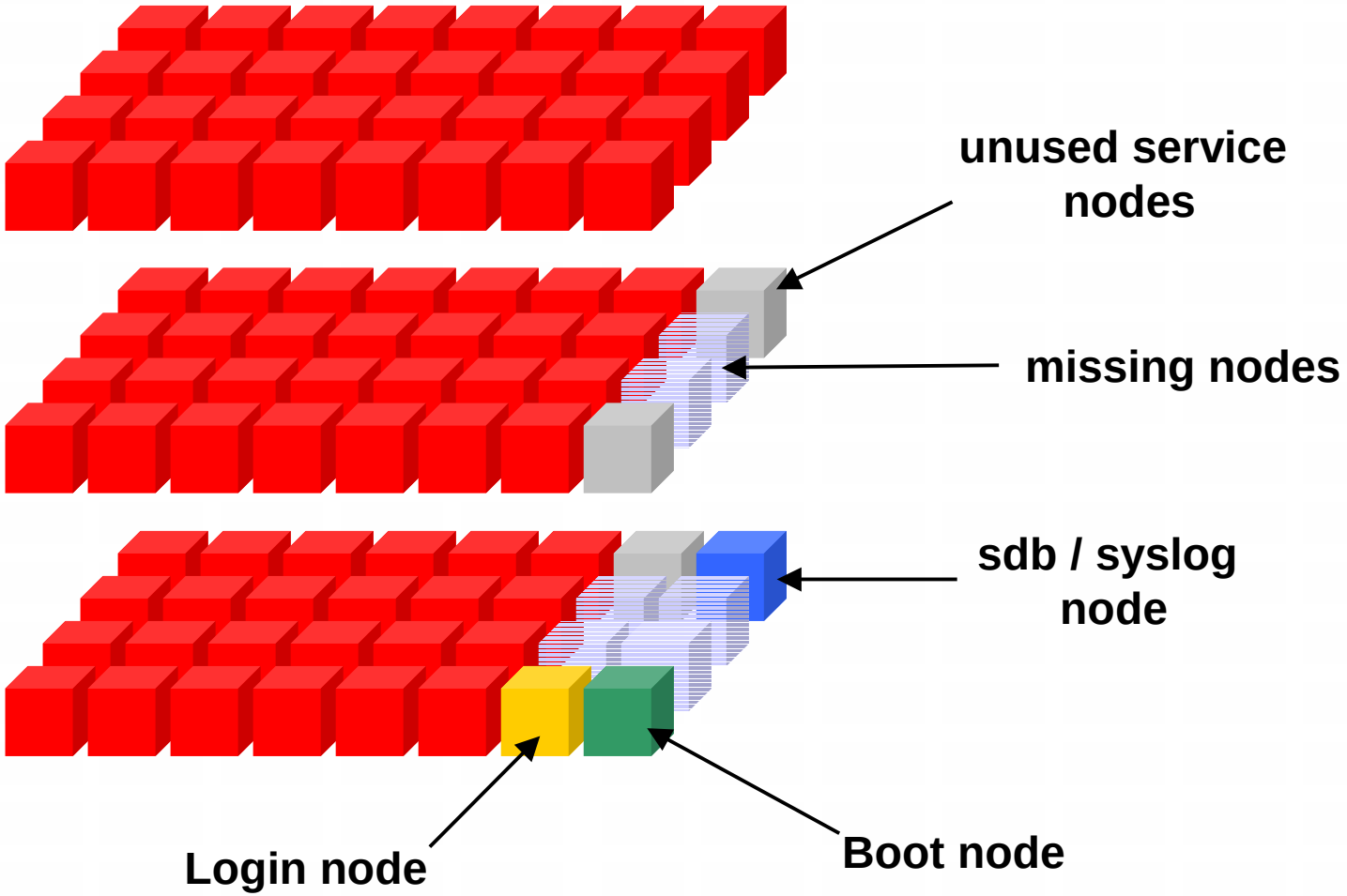
Three Cages
(3 x 4 x 8)

X
Y
Z

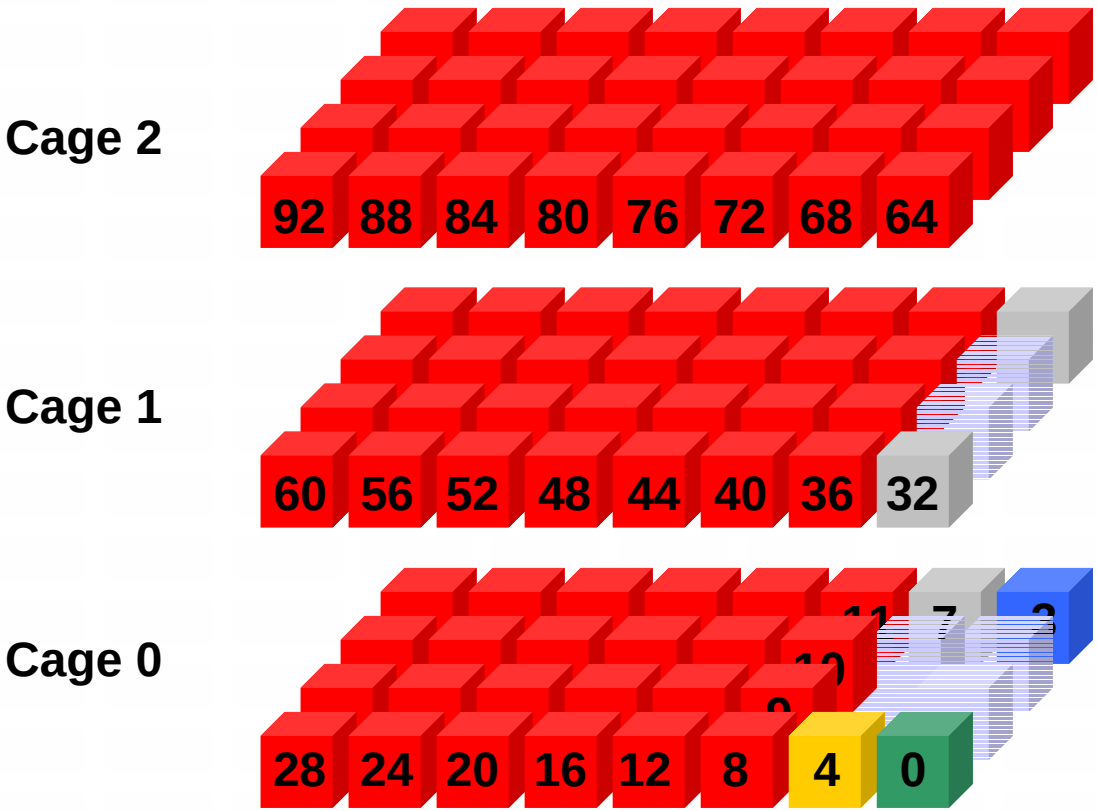
CSCS stage0 topology: cages, slots, blades



CSCS stage0 topology: 3 x 4 x 8 = 84 + 6



CSCS stage0 topology: node id numbering



CSCS stage0 nodes: xtprocadmin

NID	CABX	CABY	CAGE	SLOT	CPU	TYPE	STATUS	MODE	PSLOTS	FREE
0	0	0	0	0	0	service	up	interactive	4	4
3	0	0	0	0	3	service	up	interactive	4	4
4	0	0	0	1	0	service	up	interactive	4	4
7	0	0	0	1	3	service	up	interactive	4	4
8	0	0	0	2	0	compute	up	interactive	4	0
9	0	0	0	2	1	compute	up	interactive	4	0
10	0	0	0	2	2	compute	up	interactive	4	0
11	0	0	0	2	3	compute	up	interactive	4	0
12	0	0	0	3	0	compute	up	interactive	4	4
13	0	0	0	3	1	compute	up	interactive	4	4
14	0	0	0	3	2	compute	up	interactive	4	4
15	0	0	0	3	3	compute	up	interactive	4	4
16	0	0	0	4	0	compute	up	interactive	4	4
88	0	0	2	6	0	compute	up	interactive	4	4
89	0	0	2	6	1	compute	up	interactive	4	4
90	0	0	2	6	2	compute	up	interactive	4	4
91	0	0	2	6	3	compute	up	interactive	4	4
92	0	0	2	7	0	compute	up	interactive	4	4
93	0	0	2	7	1	compute	up	interactive	4	4
94	0	0	2	7	2	compute	up	interactive	4	4
95	0	0	2	7	3	compute	up	interactive	4	4

CSCS stage0 topology: node id mapping

- From xtprocadmin generate

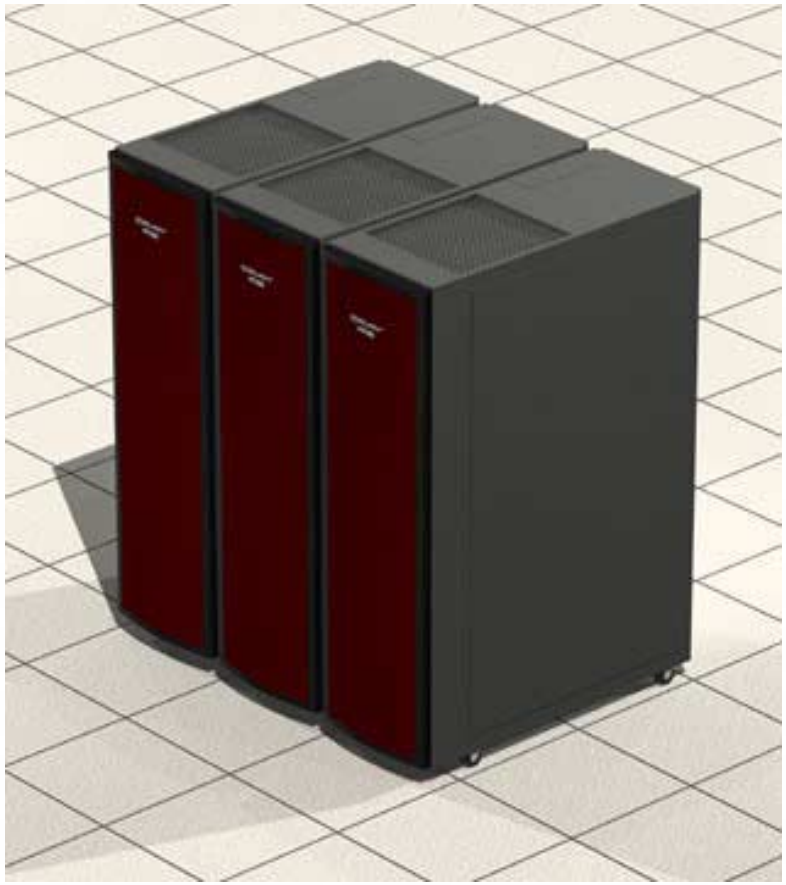
```
struct processor_entry {
  int nid;
  int cabx;
  int caby;
  int cage;
  int slot;
  int cpu;
};
struct processor_entry processor_table[];
```

- Get nid and coordinates

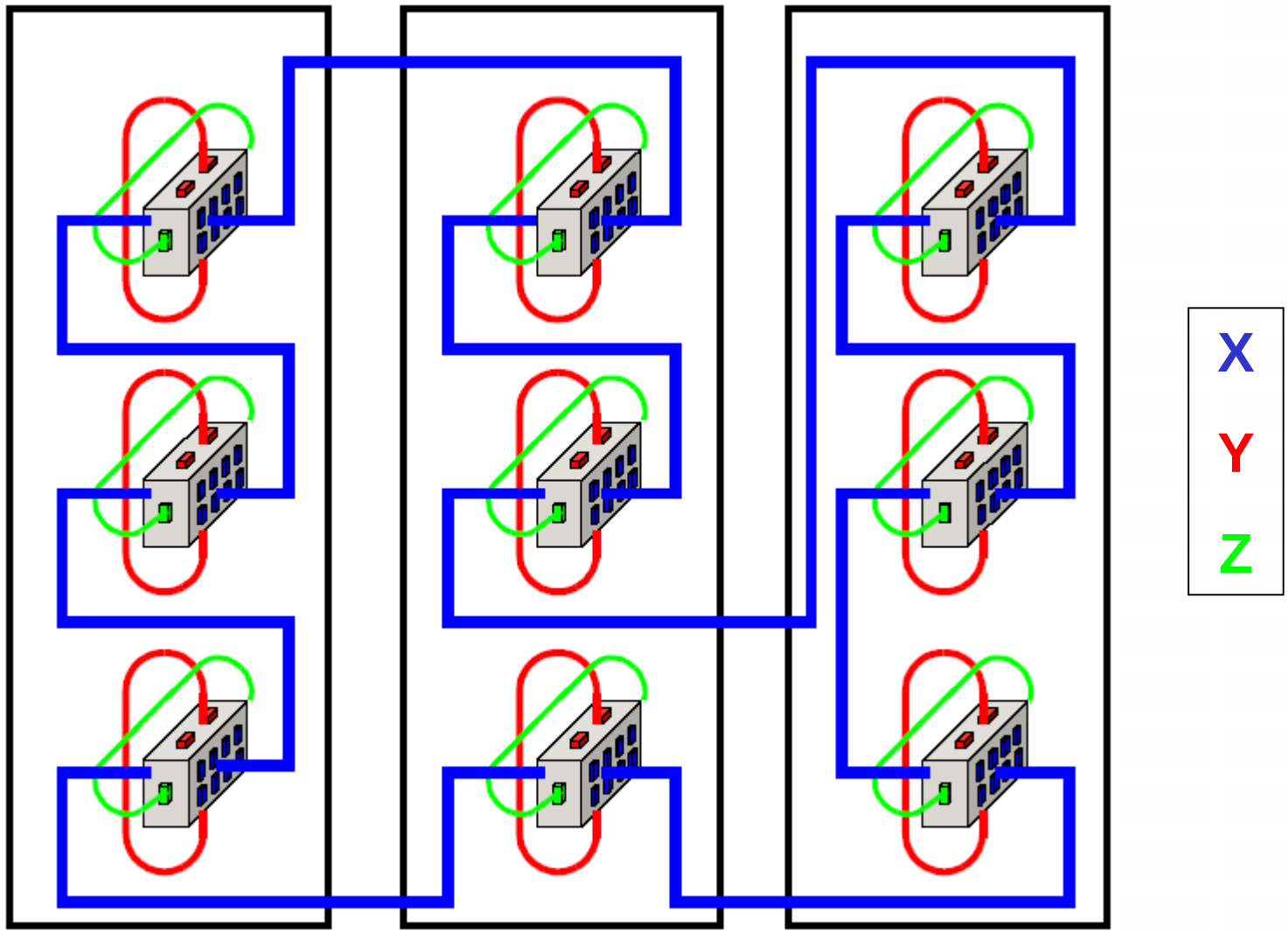
```
#include <catamount/data.h>
nid = _my_pnid;

x = processor_table[i].cage;
y = processor_table[i].cpu;
z = processor_table[i].slot;
```

3 Cabinet Configuration

- 3 cabinet system
  - 72 blades
  - 8 service and IO PEs
  - 272 compute PEs
  - 9 x 4 x 8
  - Class 0 topology



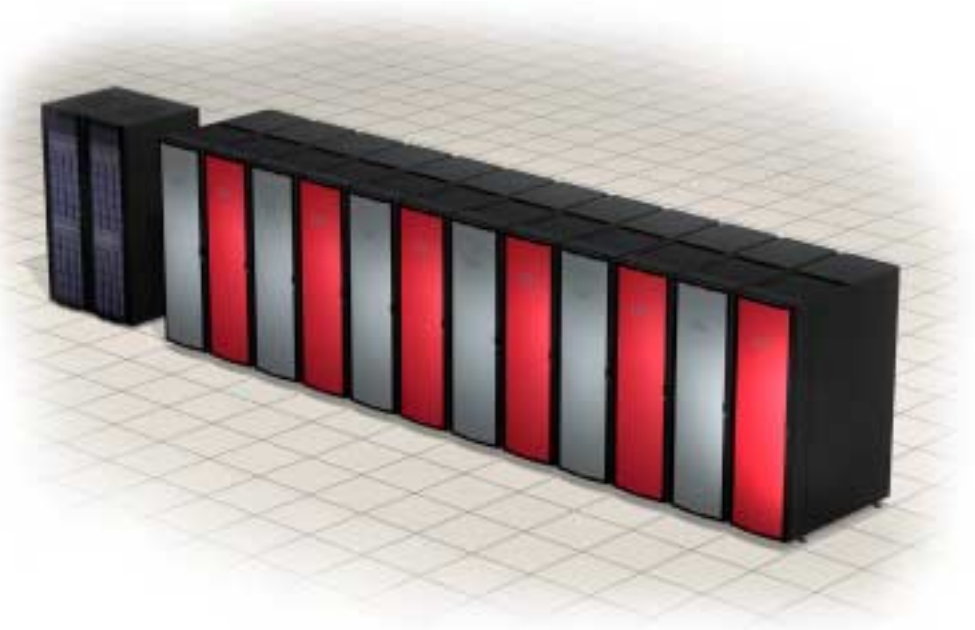
3 Cabinets, Class 0



Three Cabinets
(9 x 4 x 8)

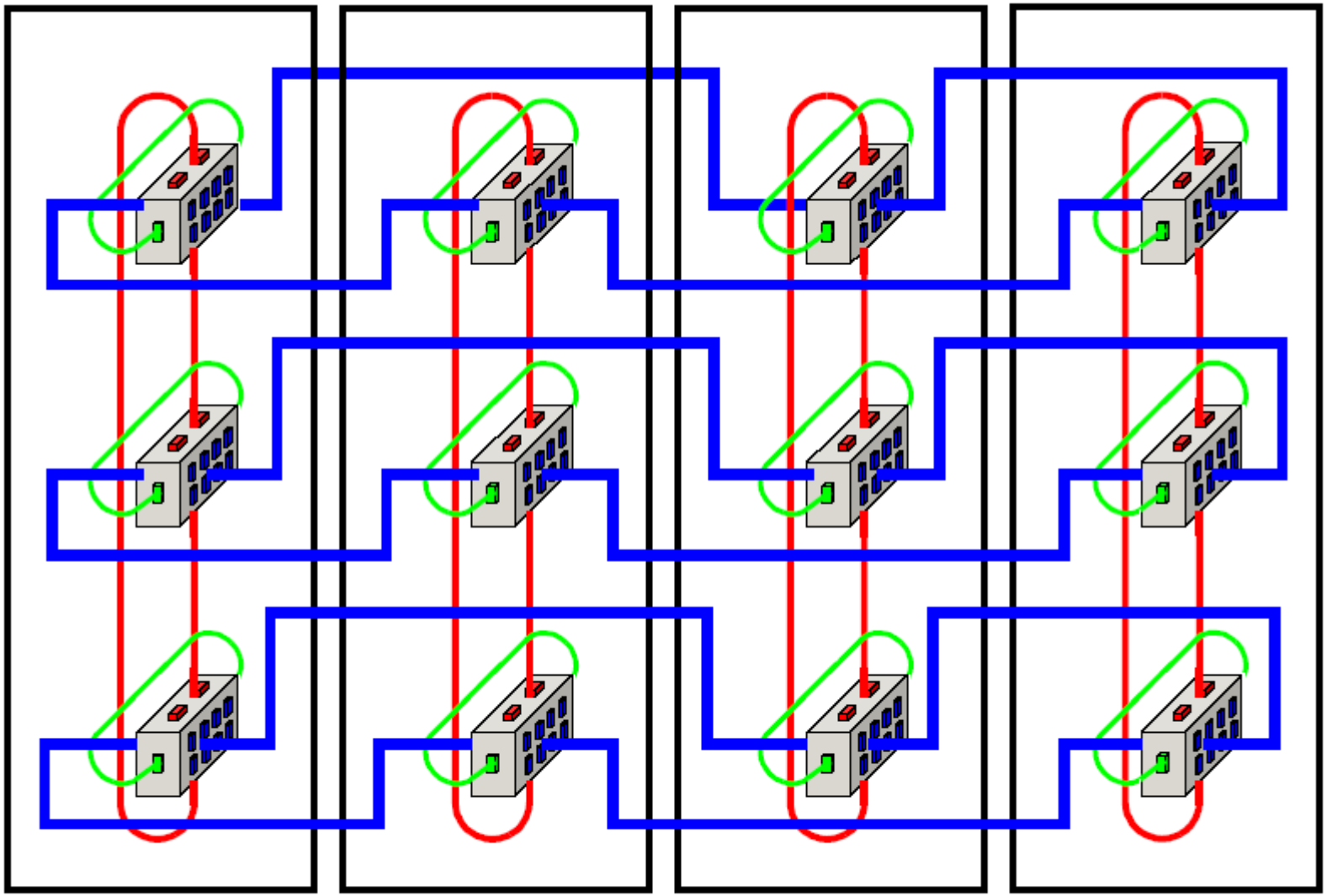
Class 1 Topologies: 4-16 Cabinets

- Example:12 Cabinets
  - 288 Blades
  - 1100 Compute PEs
  - 26 Service PEs
  - X Wired Across Row
  - 12 x 12 x 8
  - Class 1 Topology
- Can add to the system in increments of full cabinets



4 Cabinet, Class 1

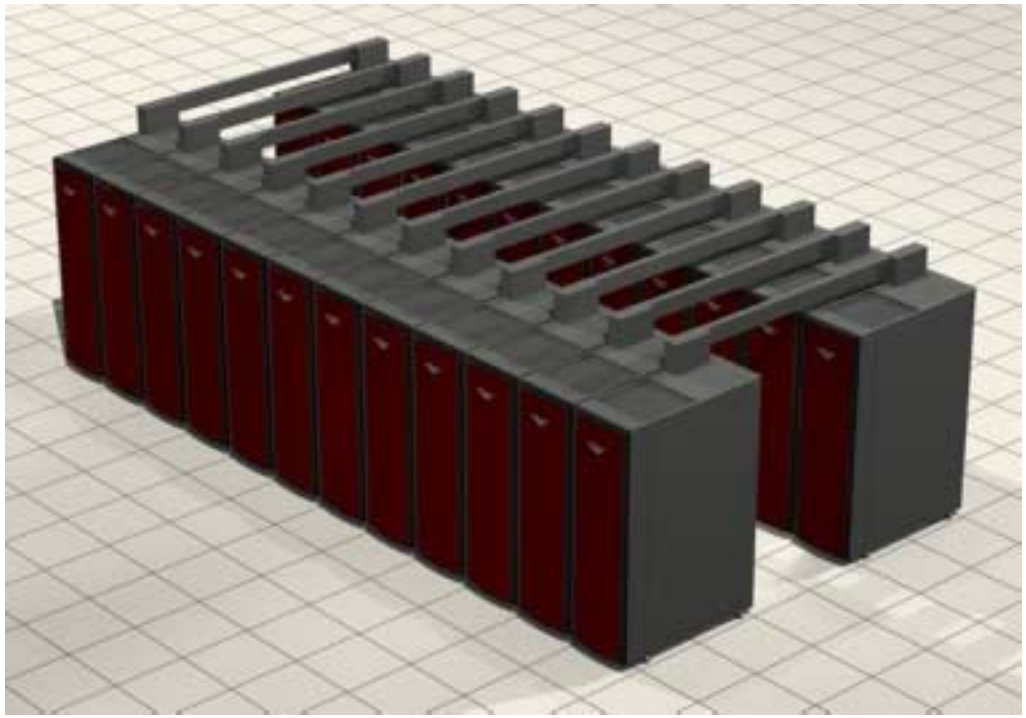
X Y Z



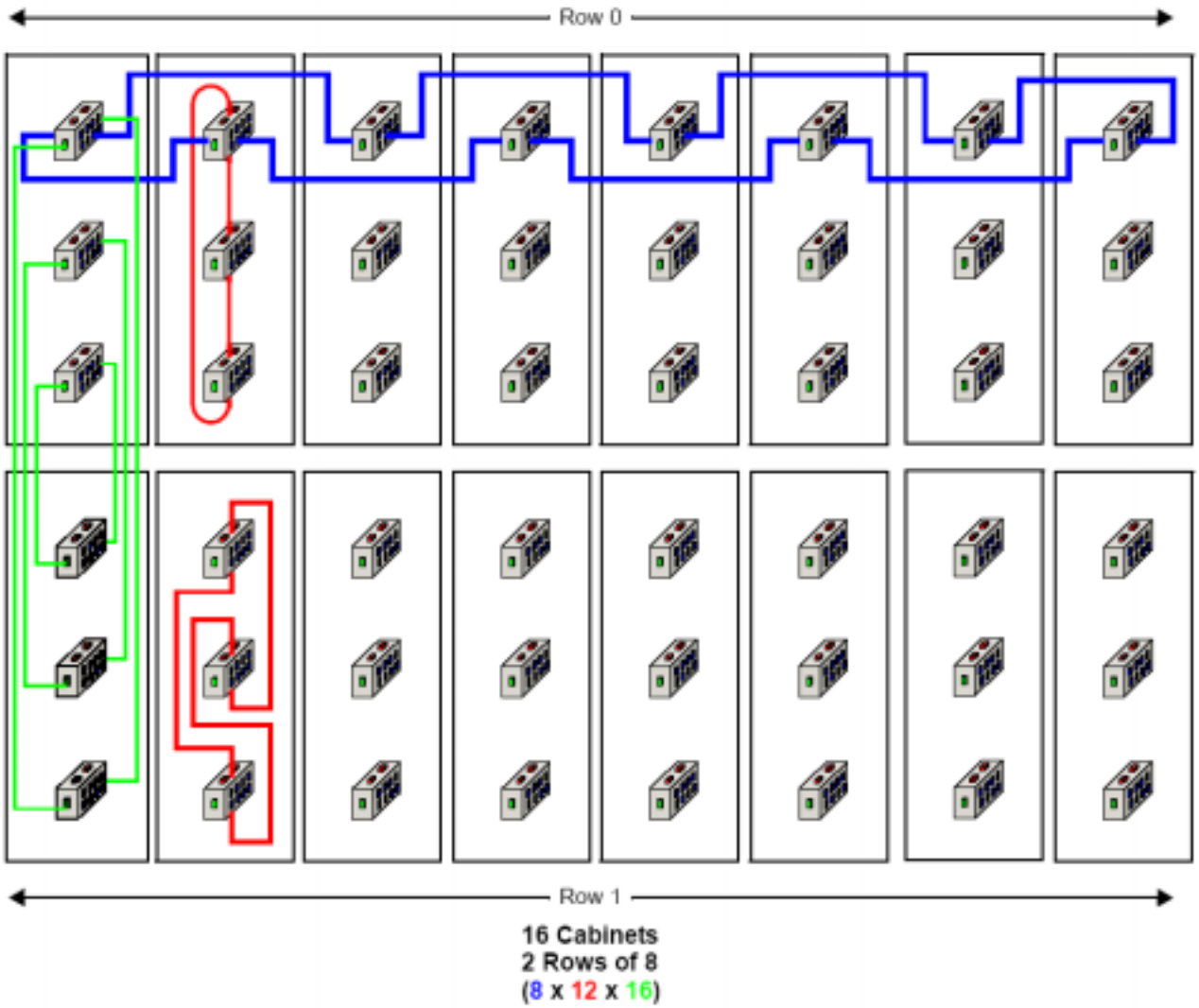
Four Cabinets
(4 x 12 x 8)

16-32 Cabinet Configurations

- 2 Rows of Cabinets
- Example: 24 cabinets
  - 576 blades
  - 40 service and IO PEs
  - 2224 compute PEs
  - 12 x 12 x 16
  - Class 2 Topology
  - Can add to the system in increments of 2 cabinets

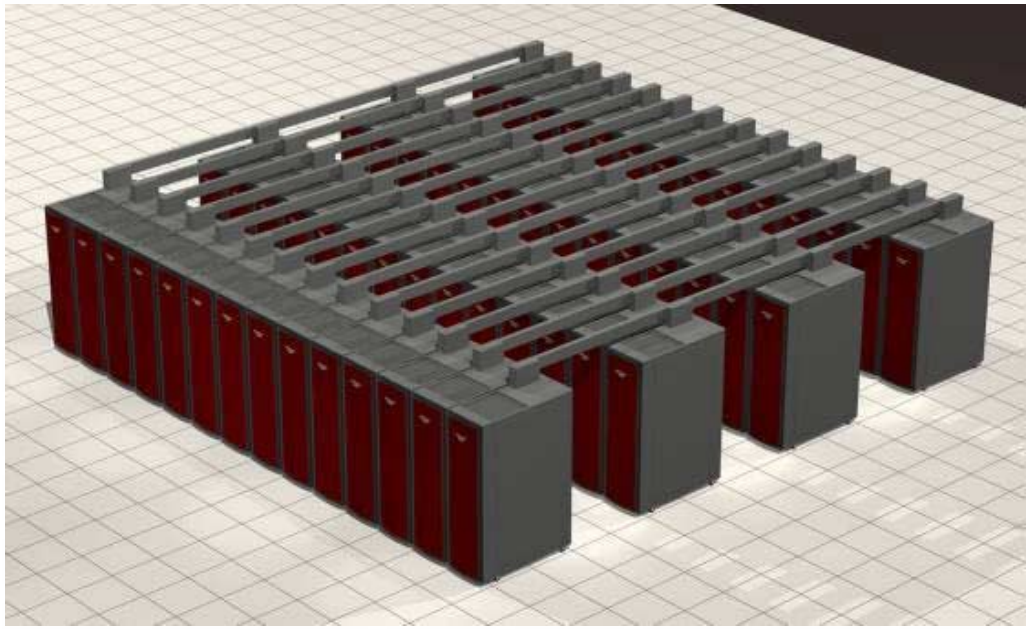


16 Cabinets, Class 2 Topology



36-144 Cabinet Configurations

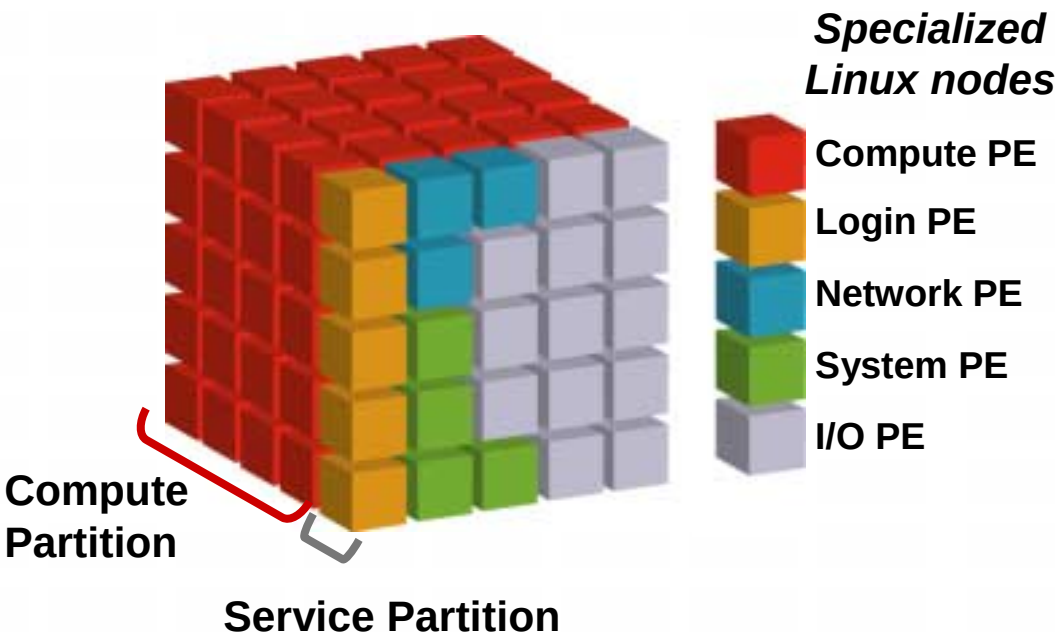
- Example: 56 Cabinets in 4 rows
  - 1344 blades
  - 176 service and IO PEs
  - 5024 compute PEs
  - 14 x 16 x 24
  - Class 3 Topology
  - Even number of Rows Only due to cable restrictions



Scalable Software

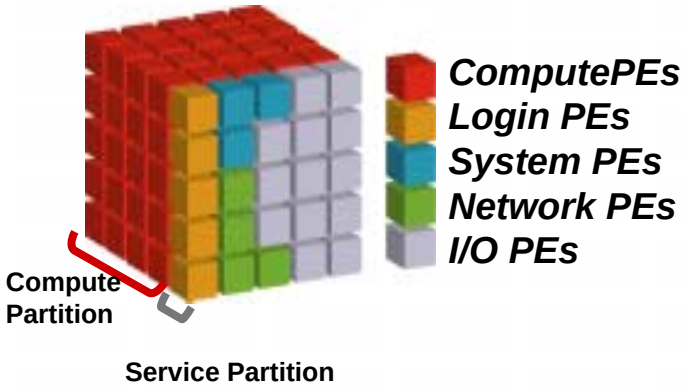
CSCS workshop May 3 / 4 2005

Scalable Software Architecture: UNICOS/lc



- 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”
  - 100 ms interrupt times
  - Will be synchronized if required
  - OS heartbeat checked once per second.
- Software Architecture enables reproducible run times

Unicos/lc Status

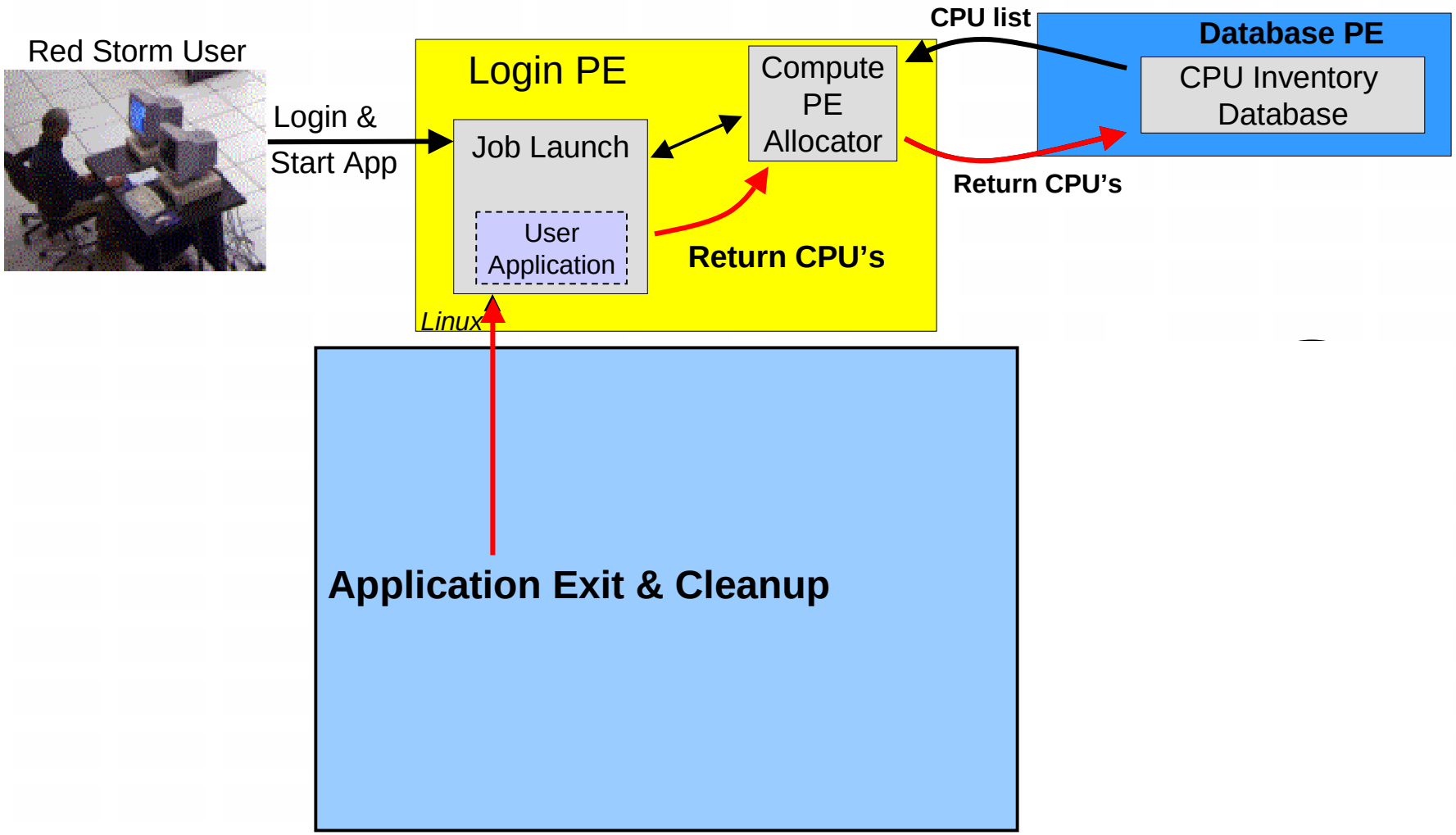


- Cray and Sandia demonstrated the Cray XT3 OS and MPI stack on 3342 compute PEs in the fall of 2004
- The Sandia ASCI Red system was used as a testbed system (called “Redshift”)
- Currently booting over 5000 processors on XT3 Hardware
- Size of compute microkernel
  - Less than 40MB, including all text, data, and I/O mappings.
  - Does not grow, all the rest is for the user.

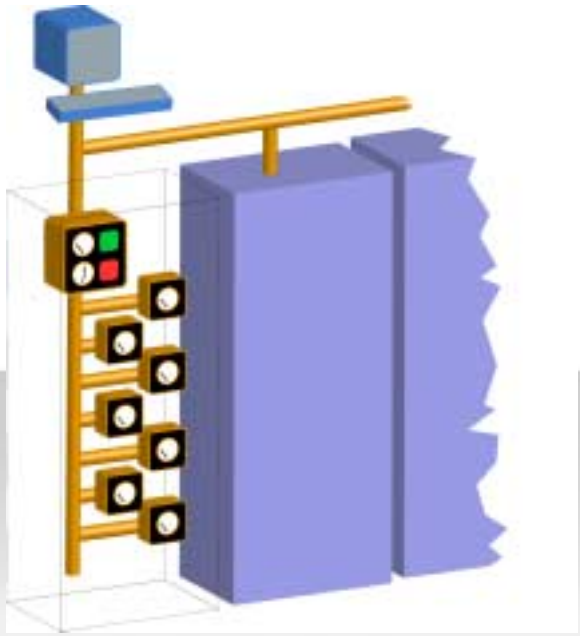
Why Catamount?

- Minimal “jitter”
- Simple memory system
  - Contiguous physical allocations
  - No demand paging
- Catamount limitations
  - No threads
  - No sockets
  - No /proc
  - No mmap
  - No native I/O
  - No fork/exec
- Catamount components
  - Quintessential Kernel (QK)
  - Process Control Thread (PCT)
  - YOD (X+1, N+1, C+1 eXecute Network Computer)
  - Libraries

Job Launch

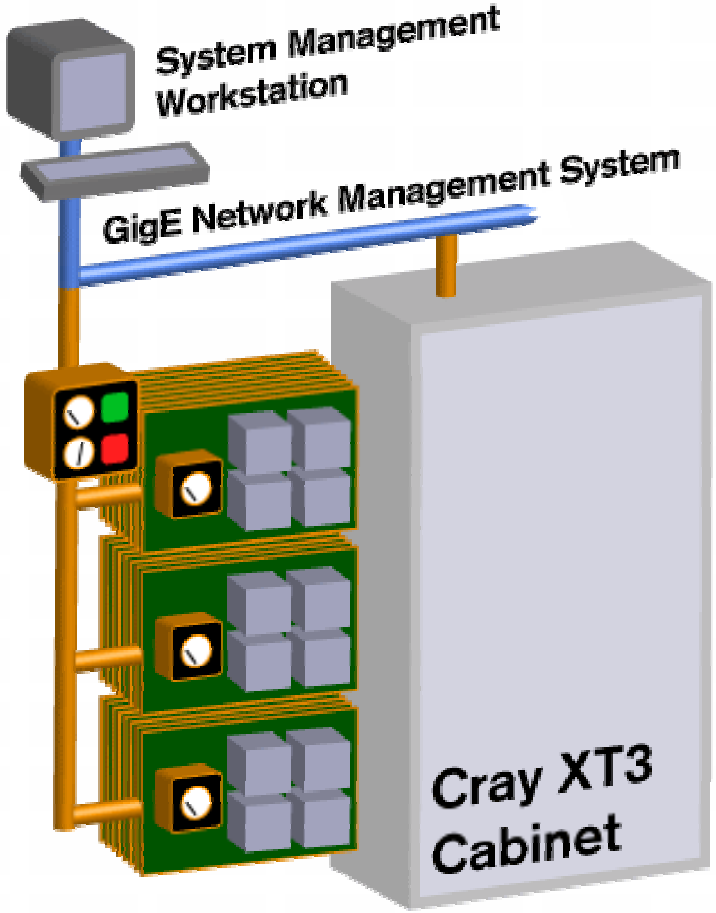

Red Storm User
Login &
Start App
Login PE
Job Launch
User Application
Compute PE Allocator
CPU list
Database PE
CPU Inventory Database
Return CPU's
Return CPU's
Linux
Application Exit & Cleanup

Cray RAS and Management System (CRMS)



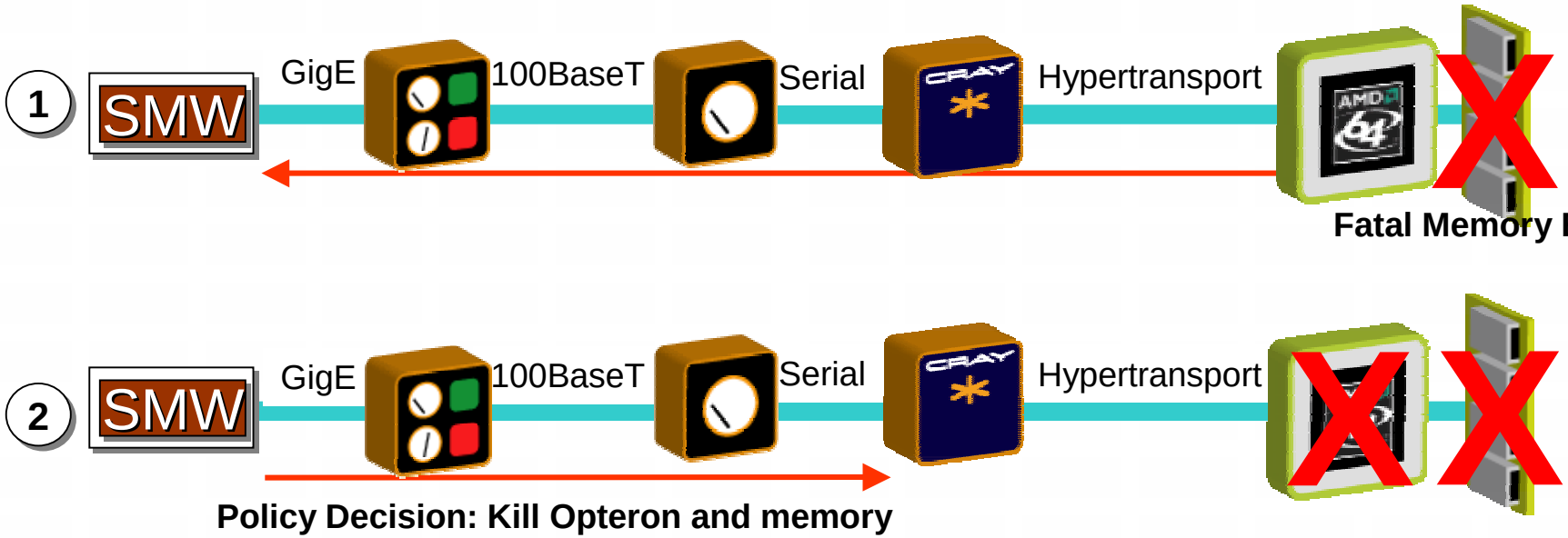
CSCS workshop May 3 / 4 2005

Cray RAS and Management System

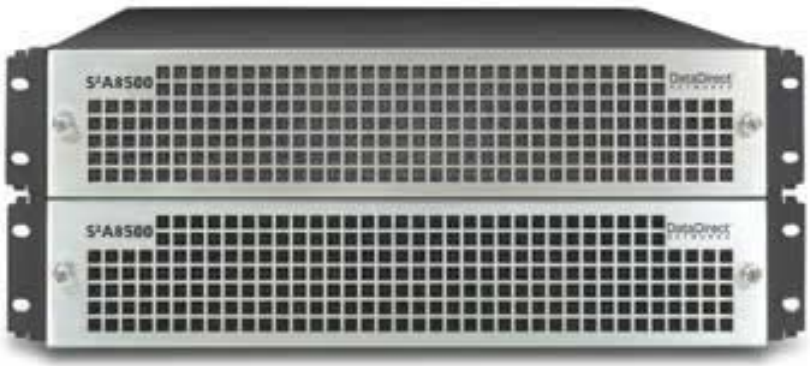


- CRMS provides Scalable System Management
  - An independent system with a separate control processors and management network
  - Single System View
  - Software failover management for critical functions
  - Real Time failure monitoring
  - Hot Swap module support

Error Handling Example



Scalable I/O



lustre

CSCS workshop May 3 / 4 2005

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