ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

CSCS
Swiss National Supercomputing Centre

DAY 2: Introduction to Cray MPP Systems with Multi-core Processors

Multi-threaded Programming, Tuning and Optimization on Multi-core MPP Platforms

15-17 February 2011

CSCS, Manno

The Cray XE6 Architecture

Roberto Ansaloni
roberto.ansaloni@cray.com

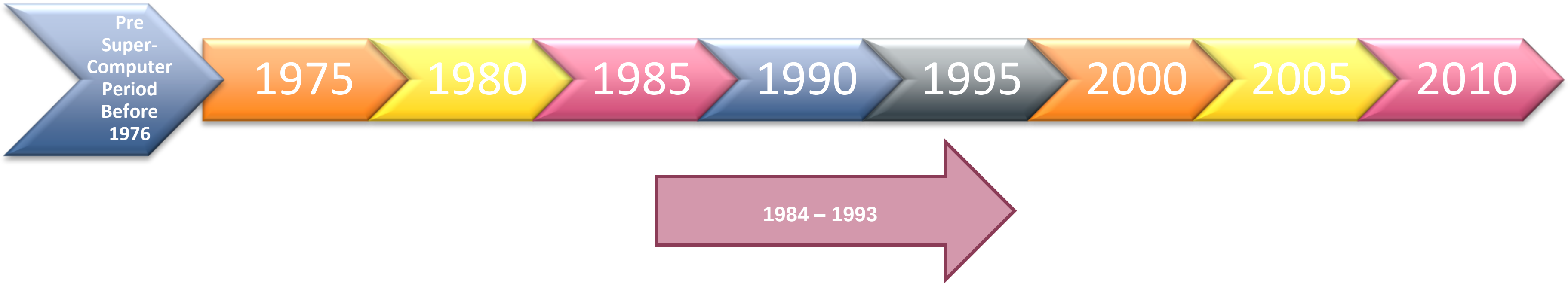
~course02/slides/day1/XE6Arch.pdf

Agenda

- Cray MPP product line
- Cray XE6 architecture
  - Cray XE6 node
  - Cray XE6 configurations, topology
- Cray XE6 scalable software
  - Service and compute nodes
  - CNL, Lustre, MPI
- Cray XE6 configuration
  - Cray XE6 blades, cages, cabinets
  - CSCS palu configuration
- Application launching process

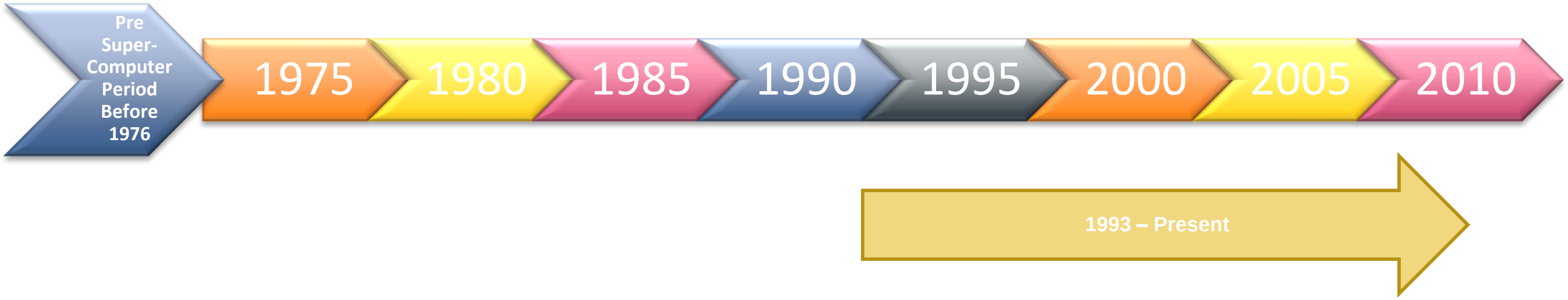
The Cray MPP History

Beginning of the MPP



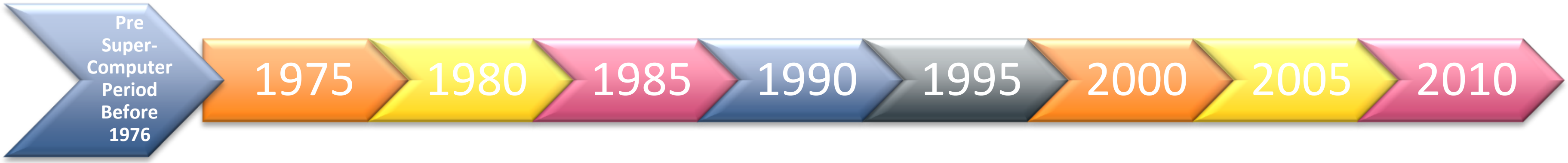
- Thinking Machines founded in 1984.
    - “The Connection Machine” contained 10,000s of very weak processors.
- Several other companies founded and failed.
- Not clear how to program these machines.
- Intel Paragon (1993)

The Cray MPP product line



- In 1993 Cray launched the Cray MPP product line
  - Powerful processor: DEC Alpha
  - Custom 3D Torus interconnect
  - Microkernel-based compute node OS

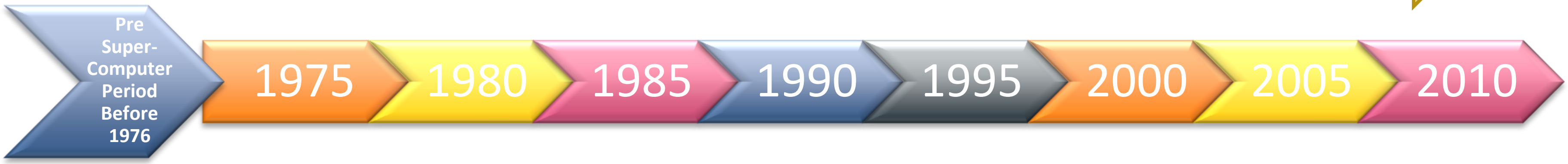
The Early Cray MPP systems



- Cray T3D (1993):
  - DEC Alpha EV4, UNICOS MAX (Mach based)
  - Relying on a Cray Y-MP front-end running UNICOS for all I/O and most system services.
- Cray T3E (1995):
  - DEC Alpha EV5/EV56, UNICOS/mk(Chorus based), self-hosted
  - A 1480-processor T3E-1200 was the first supercomputer to achieve a performance of more than 1 teraflops running a computational science application, in 1998.

MPPs Go Mainstream

1993 – Present


Pre Super-Computer Period Before 1976

1975 1980 1985 1990 1995 2000 2005 2010

- First MPI specification presented at Supercomputing 1994.
- IBM introduces the SP2 in 1995
- Community was really starting to figure out how to program MPPs.
- IBM Blue Gene Delivered ~ 2004.
- First Cray XT3 system delivered in 2004.
  - First Cray XT3 system in Europe delivered to CSCS in 2005.

The Cray recipe for a good MPP

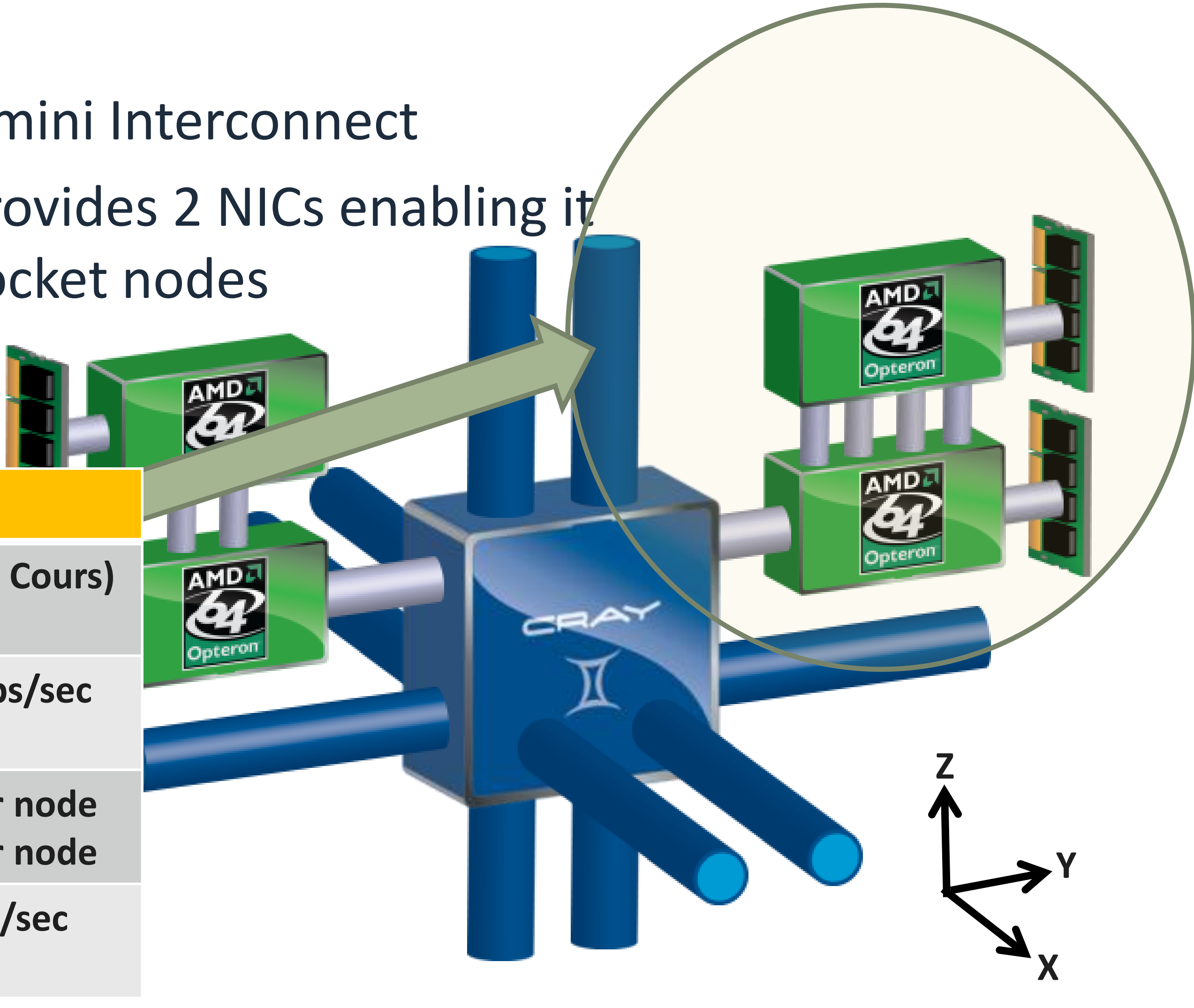
1. Select Best Microprocessor
2. Surround it with a balanced or “bandwidth rich” environment
3. “Scale” the System
   - Eliminate Operating System Interference (OS Jitter)
   - Design in Reliability and Resiliency
   - Provide Scalable System Management
   - Provide Scalable I/O
   - Provide Scalable Programming and Performance Tools
   - System Service Life (provide an upgrade path)

The Cray XE6 Node Architecture

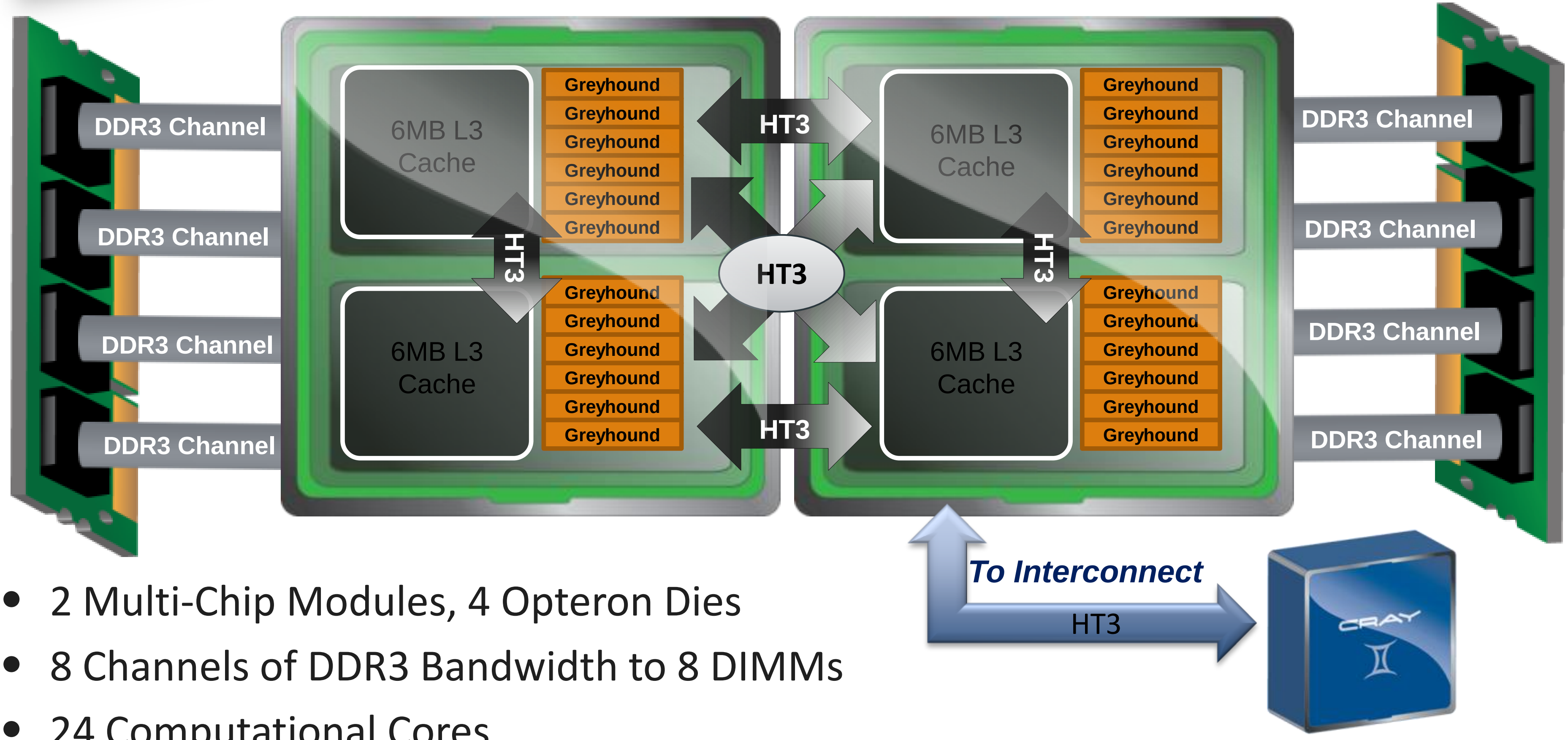
Cray XE6 Compute Node

- Built around the Gemini Interconnect
- Each Gemini ASIC provides 2 NICs enabling it to connect 2 dual-socket nodes

Node Characteristics	
Number of Cores	24 (Magny Cours)
Peak Performance MC-12 (2.2)	211 Gflops/sec
Memory Size	32 GB per node 64 GB per node
Memory Bandwidth (Peak)	83.5 GB/sec



Cray XE6 Node Details: 24-core Magny-Cours



- 2 Multi-Chip Modules, 4 Opteron Dies
- 8 Channels of DDR3 Bandwidth to 8 DIMMs
- 24 Computational Cores
- Dies are fully connected with HT3
- Snoop Filter Feature Allows 4 Die SMP to scale well

AMD Magny-Cours Data Caches (G34 MCM)

Conway et al."Cache Hierarchy and Memory Subsystem of the AMD Opteron Processor,“ IEEE Micro, pp. 16-29, March/April, 2010

- L1 Data Cache
  - 64KB, 64B cacheline, 2-way associative
  - Load-to-use latency: 3 clock cycles
- L2 Cache
  - 512KB, 64B cacheline , 16-way associative
  - Load-to-use latency: 12 clock cycles
  - Victim / Copy-Back from L1
  - Hits are invalidated from L2 and placed into L1
- L3 Cache, shared
  - 6 MB per die, 64B cacheline, 16-way associative
  - Victim / Copy-Back from L2
  - Hits can be removed or stay on L3 if needed by other threads

The Cray Gemini Interconnect

Cray Network Evolution

SeaStar (Cray XT)

- Built for scalability to 250K+ cores
- Very effective routing and low contention switch

Gemini (Cray XE)

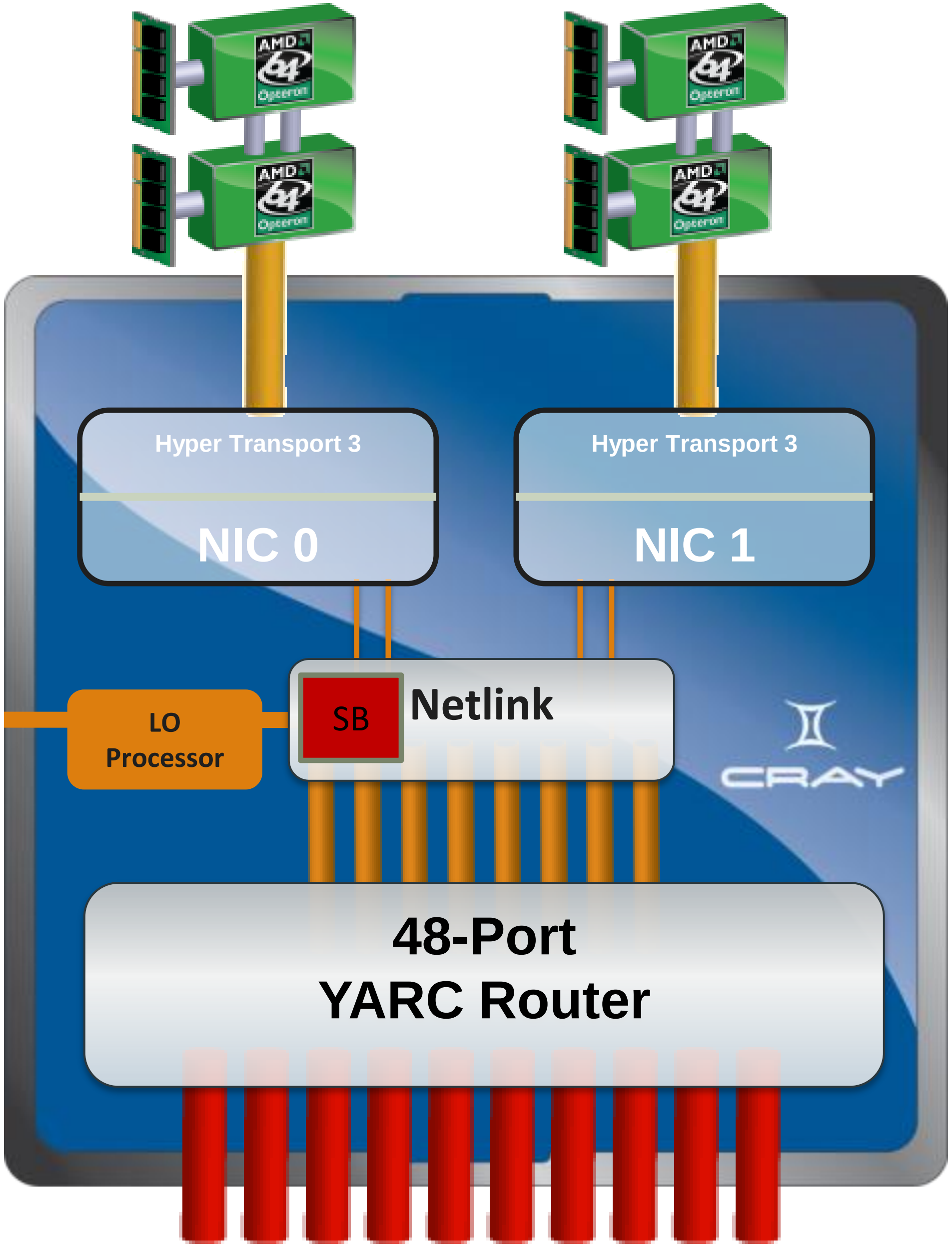
- 100x improvement in message throughput
- 3x improvement in latency
- PGAS Support, Global Address Space
- Scalability to 1M+ cores

Aries

- DON’T ask me about it

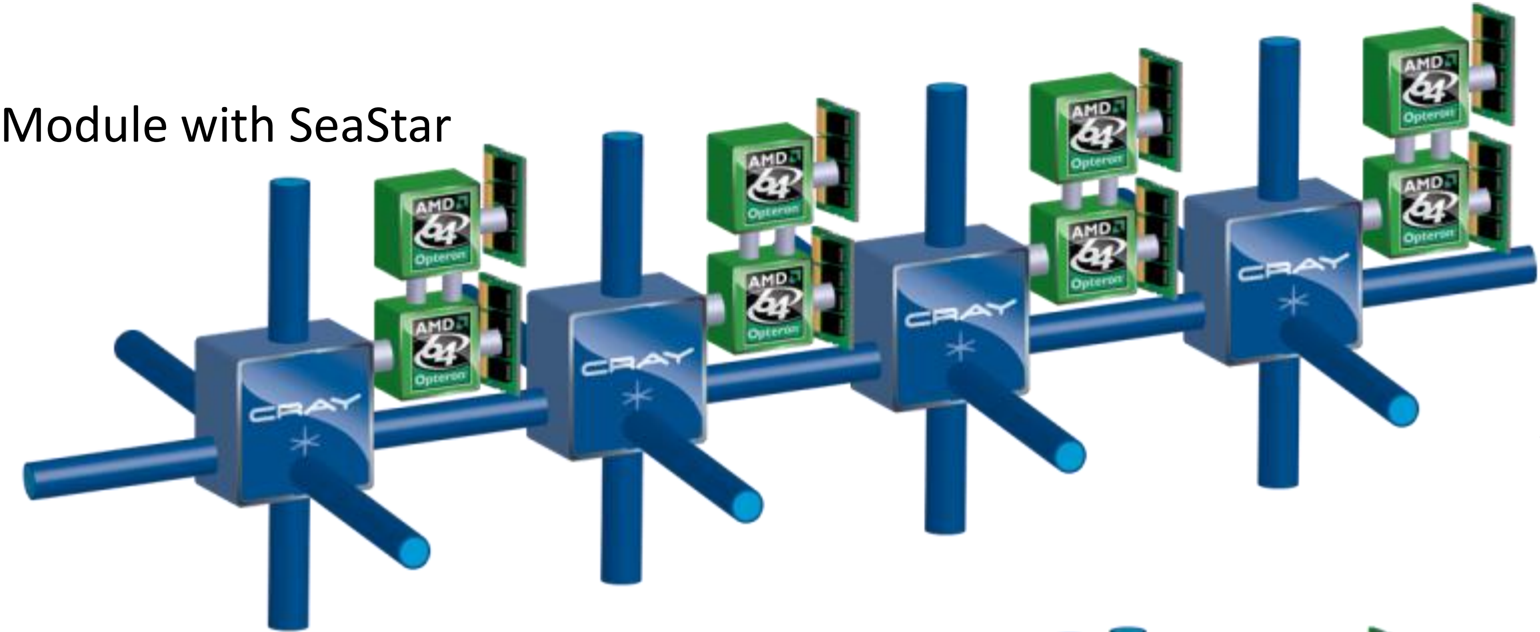
Cray Gemini

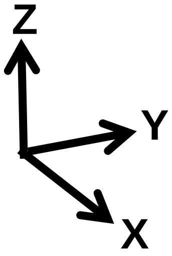
- 3D Torus network
- Supports 2 Nodes per ASIC
- 168 GB/sec routing capacity
- Scales to over 100,000 network endpoints
  - Link Level Reliability and Adaptive Routing
  - Advanced Resiliency Features
- Provides global address space
- Advanced NIC designed to efficiently support
  - MPI
  - One-sided MPI, SHMEM
  - FORTRAN 2008 with coarrays,UPC
  - Global Atomics



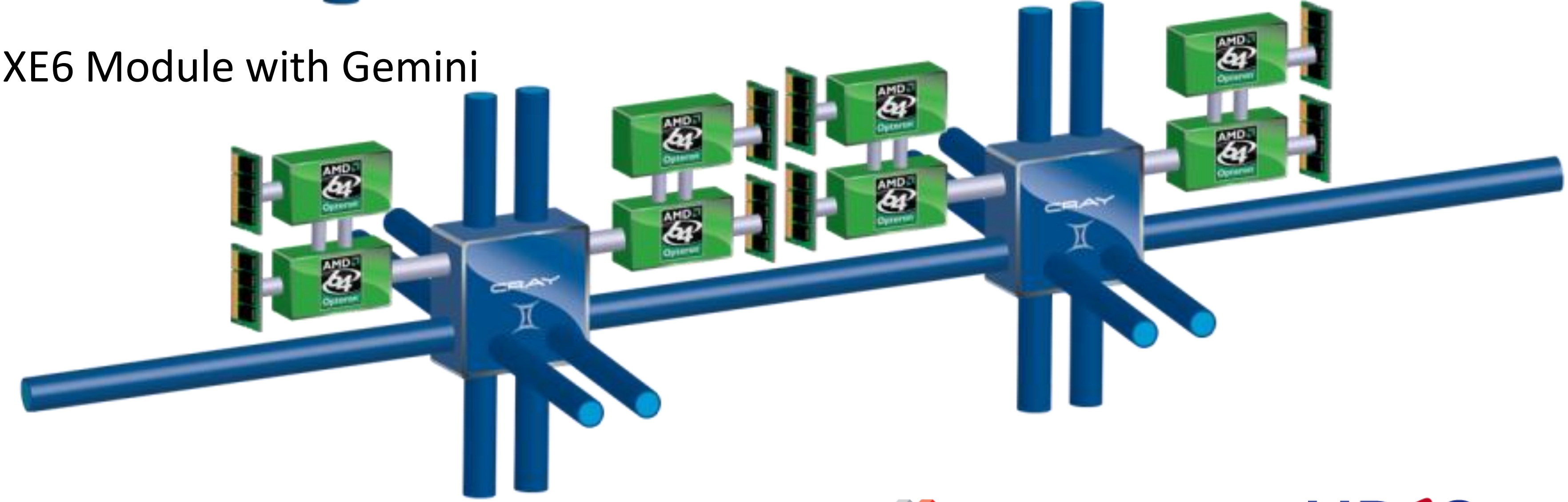
Gemini vs SeaStar – Topology

XT6 Module with SeaStar





XE6 Module with Gemini



Gemini MPI Features

- FMA (Fast Memory Access)
  - Mechanism for most MPI transfers
  - Supports tens of millions of MPI requests per second
- BTE (Block Transfer Engine)
  - DMA offload engine , supports *asynchronous* block transfers between local and remote memory, in either direction
- Gemini provides low-overhead OS-bypass features for short transfers
  - MPI latency around 1.4us in current release
  - NIC provides for many millions of MPI messages per second 20 times better than Seastar
- Much improved injection bandwidth – 6 GB/s user data injection with HT3 link

Gemini Advanced Features

- Globally addressable memory provides efficient support for UPC, FORTRAN 2008 with Coarrays, Shmem and Global Arrays
  - Much improved one-sided communication mechanism with hardware support
  - Cray compiler targets this capability directly
- Pipelined global loads and stores
  - Allows for fast irregular communication patterns
- Atomic memory operations
  - AMOs provide a faster synchronization method for barriers
  - Provides fast synchronization needed for one-sided communication models

Scalable Software Architecture

CLE3, An Adaptive Linux OS designed specifically for HPC

CRAY
LINUX ENVIRONMENT CLE3

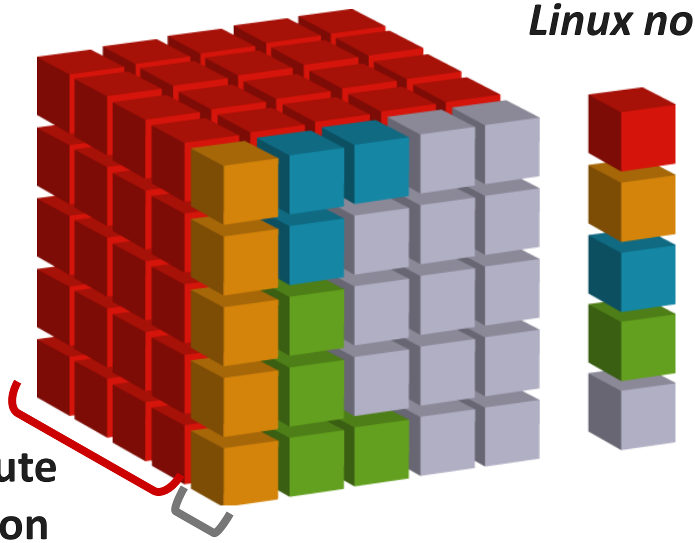
ESM – *Extreme Scalability Mode*

- No compromise *scalability*
- Low-Noise Kernel for scalability
- Native Comm. & Optimized MPI
- Application-specific performance tuning and scaling

CCM –*Cluster Compatibility Mode*

- No compromise *compatibility*
- Fully standard x86/Linux
- Standardized Communication Layer
- Out-of-the-box ISV Installation
- ISV applications simply install and run

Scalable Software Architecture: CLE



Microkernel on Compute nodes, full featured Linux on Service nodes.

Service PEs specialize by function

Software Architecture eliminates OS “Jitter”

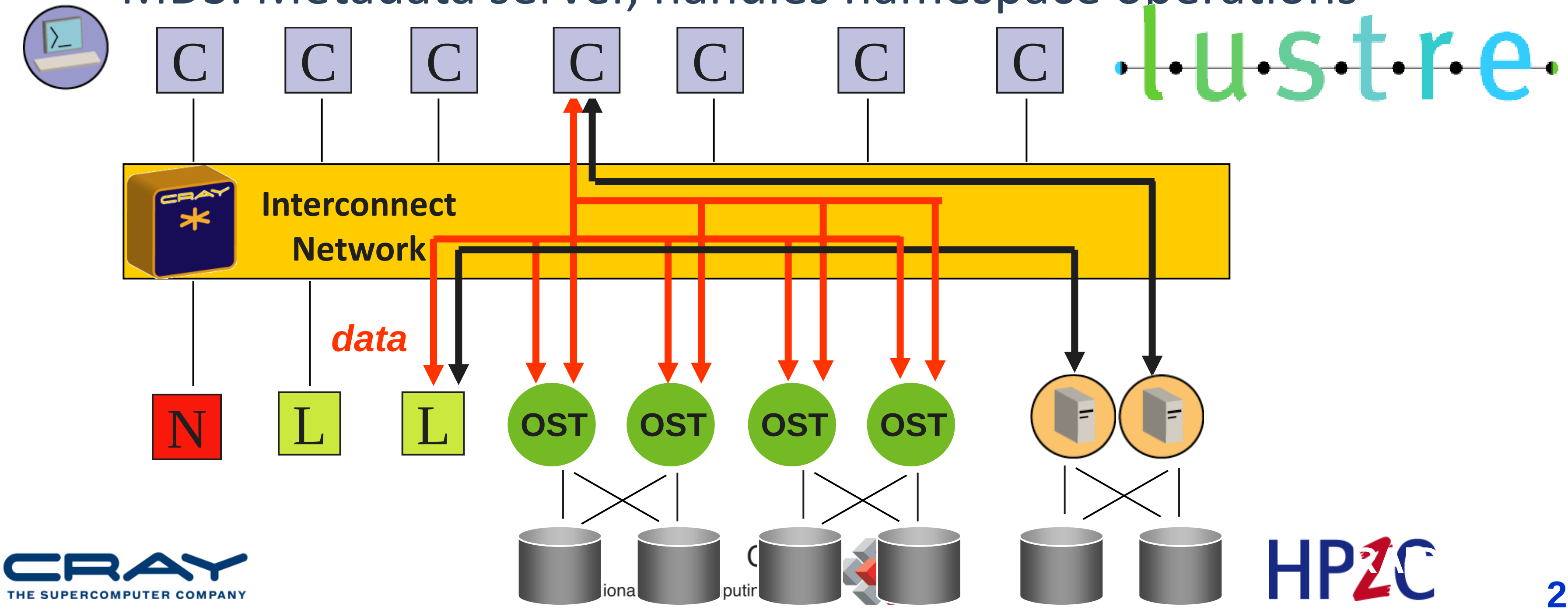
Software Architecture enables reproducible run times

Service nodes

- Overview
  - Run full Linux (SuSe SLES), 2 nodes per service blade
- Boot node
  - first XE6 node to be booted: boots all other components
- SDB node
  - hosts MySQL database
  - processors, allocation, accounting, PBS information
- Login nodes
  - User login and code preparation activities: compile, launch
  - Partition allocation: ALPS (Application Level Placement Scheduler)

I/O nodes

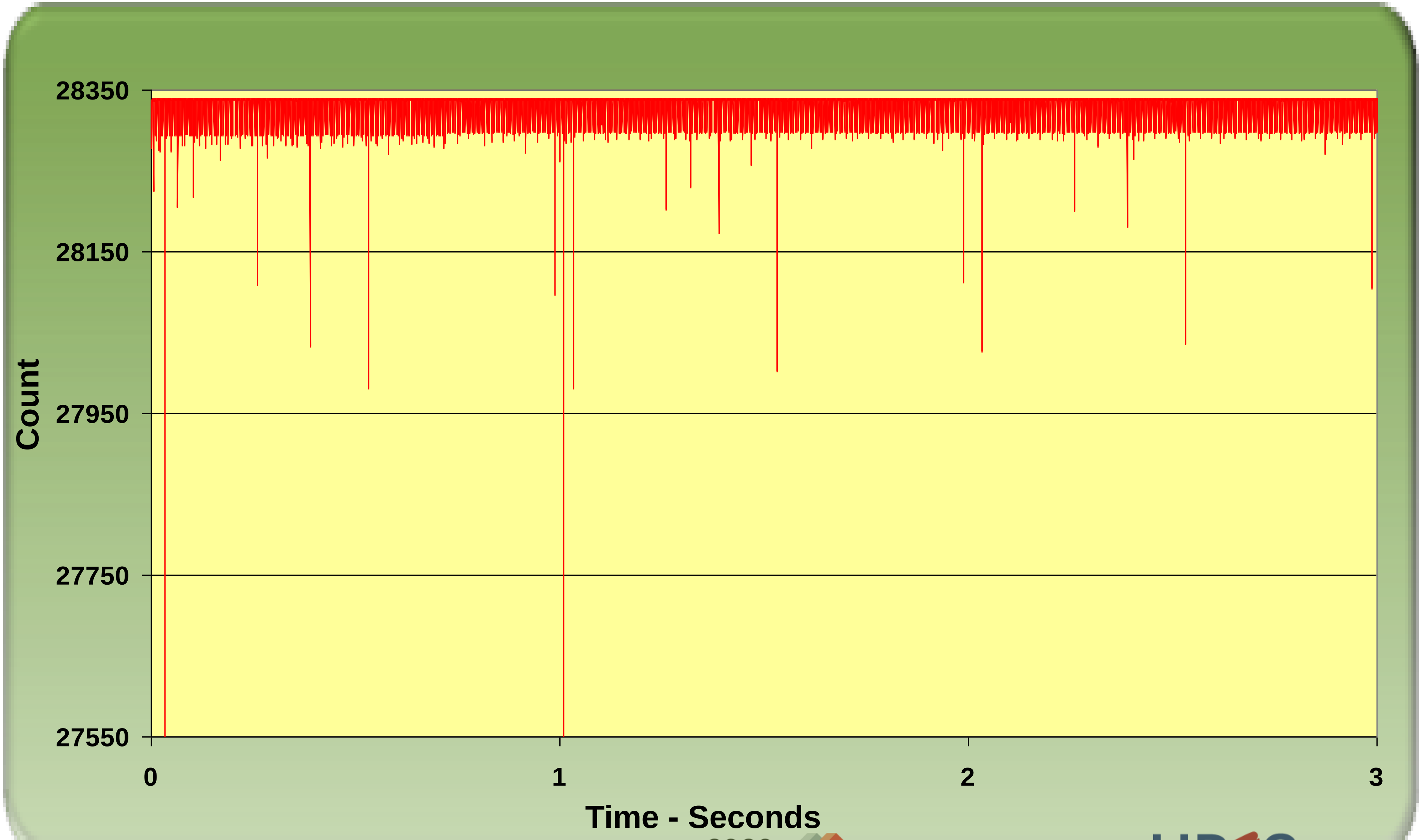
- I/O nodes
  - Run Lustre processes (OST, metadata server)
- Lustre terminology:
  - OST: Object Storage Target, software interface to back-end storage volumes
  - OSS: Object Storage Server, I/O node that host OSTs
  - MDS: Metadata server, handles namespace operations



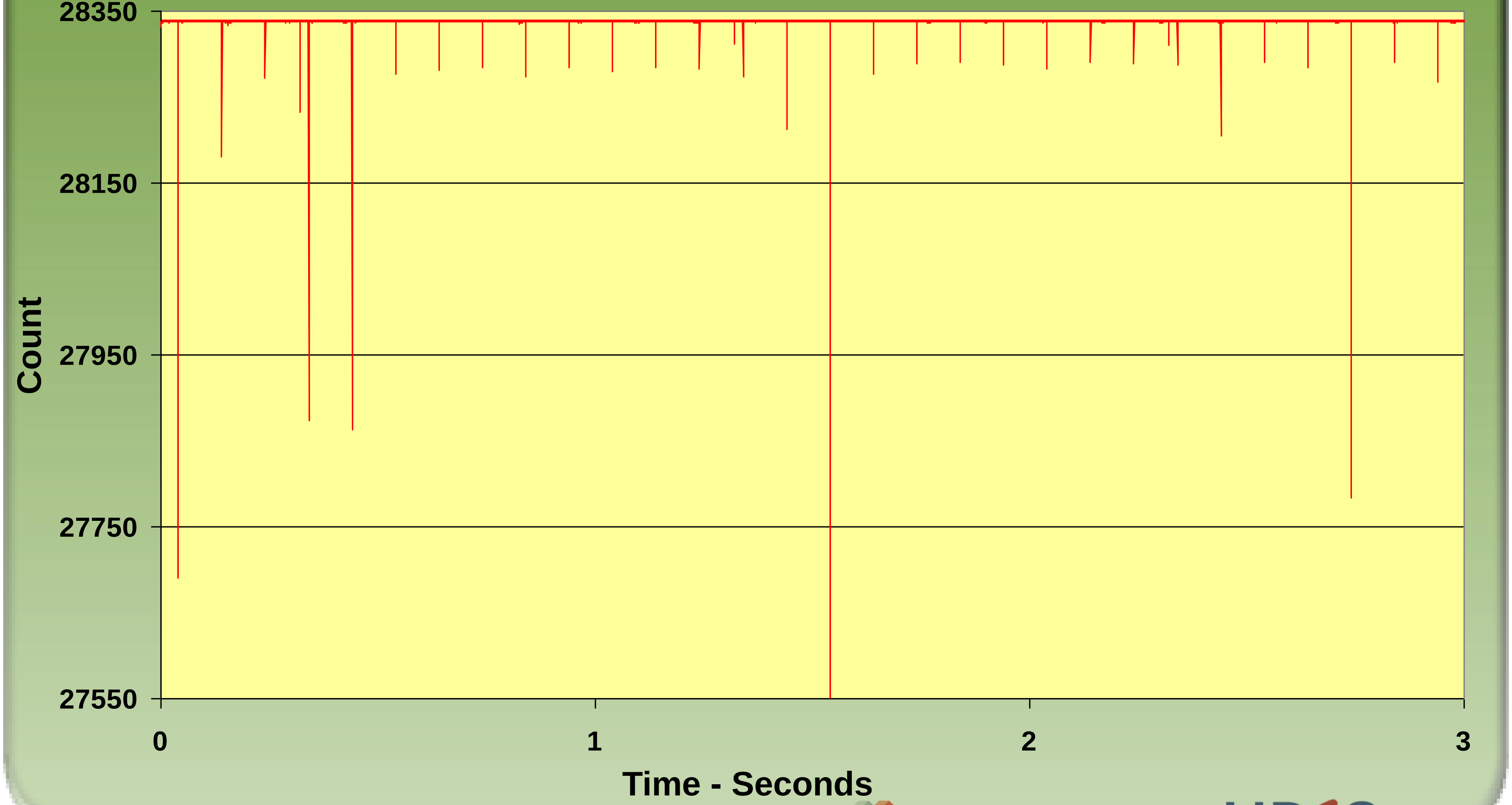
Compute nodes

- Configuration
  - 4 compute nodes per compute blade
  - Each compute node has 2 Opteron sockets
  - Each socket hosts a 12-core Magny-Cours MCM
- Runs Cray Linux Environment (CLE)
  - Linux-based operating system
  - designed to run large, complex applications and scale efficiently to hundreds of thousands of processor cores

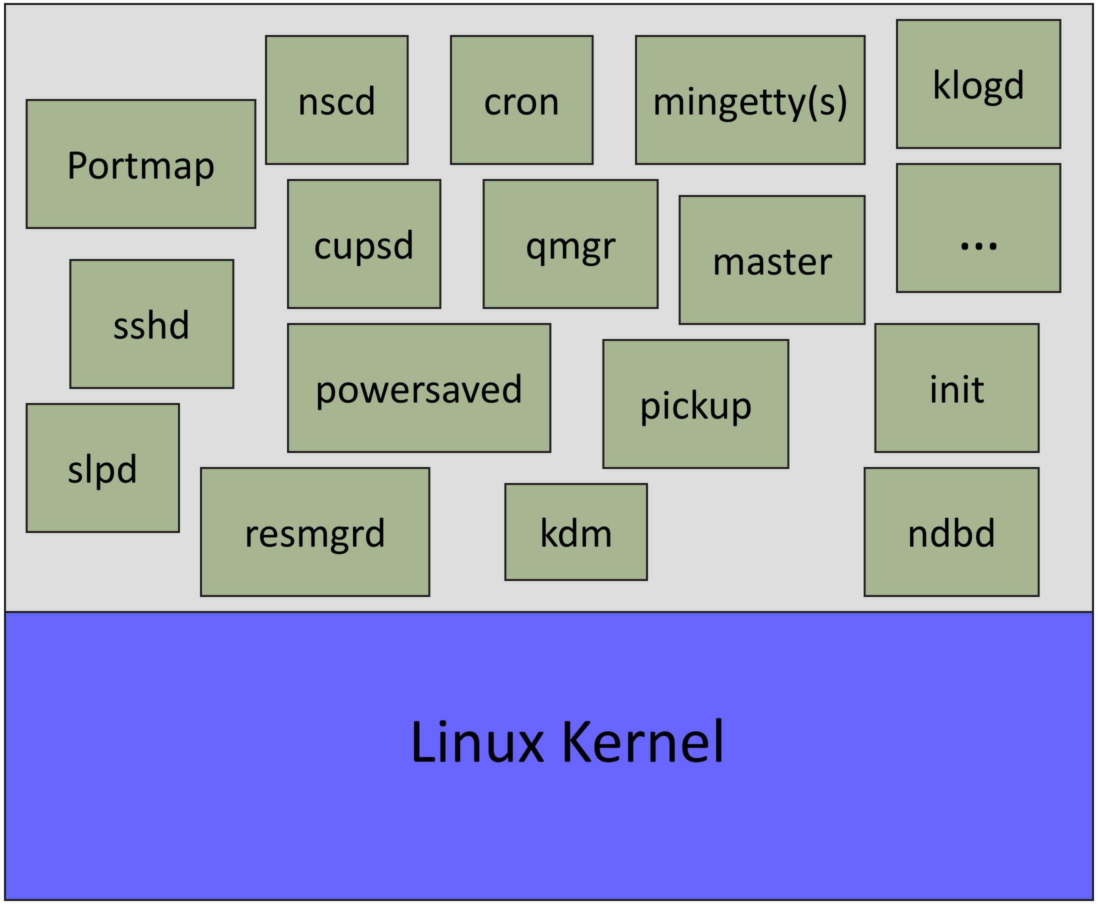
FTQ Plot of Stock SuSE (most daemons removed)



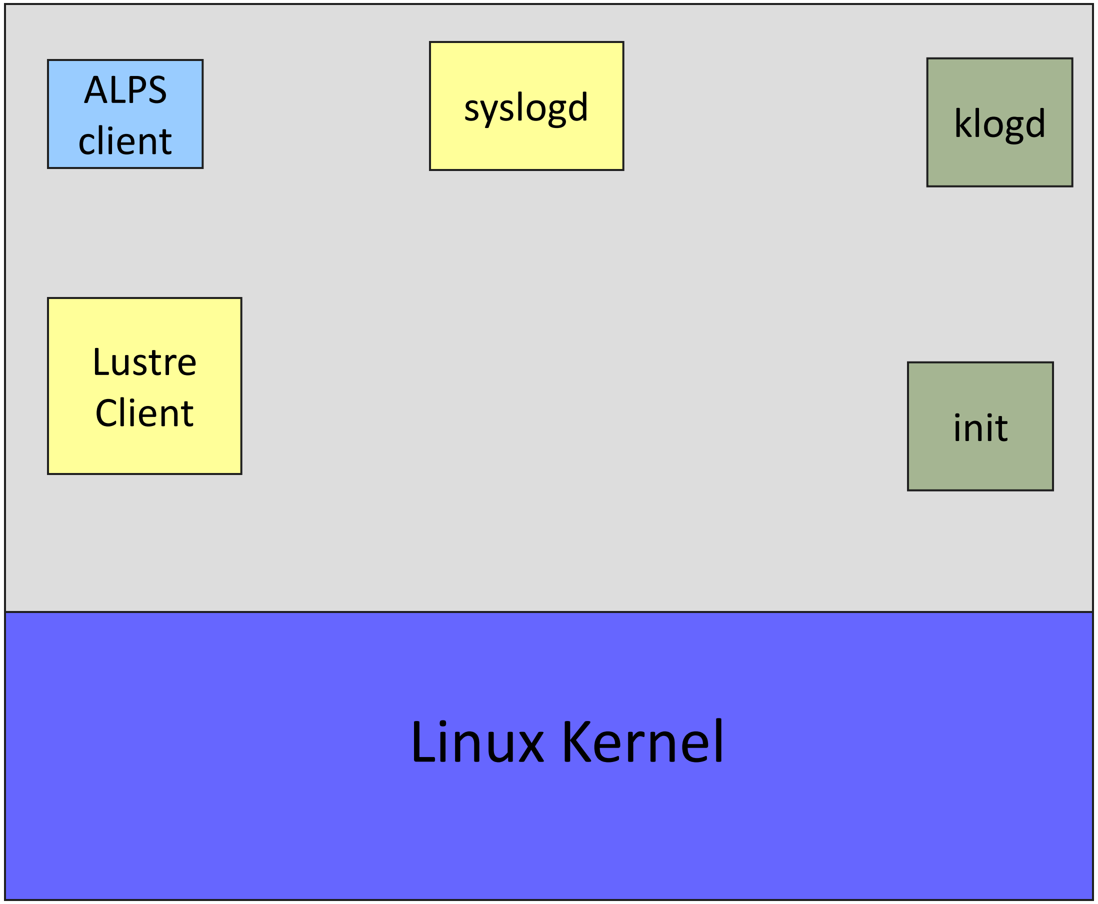
FTQ plot of CNL



Trimming OS – *Standard Linux Server*



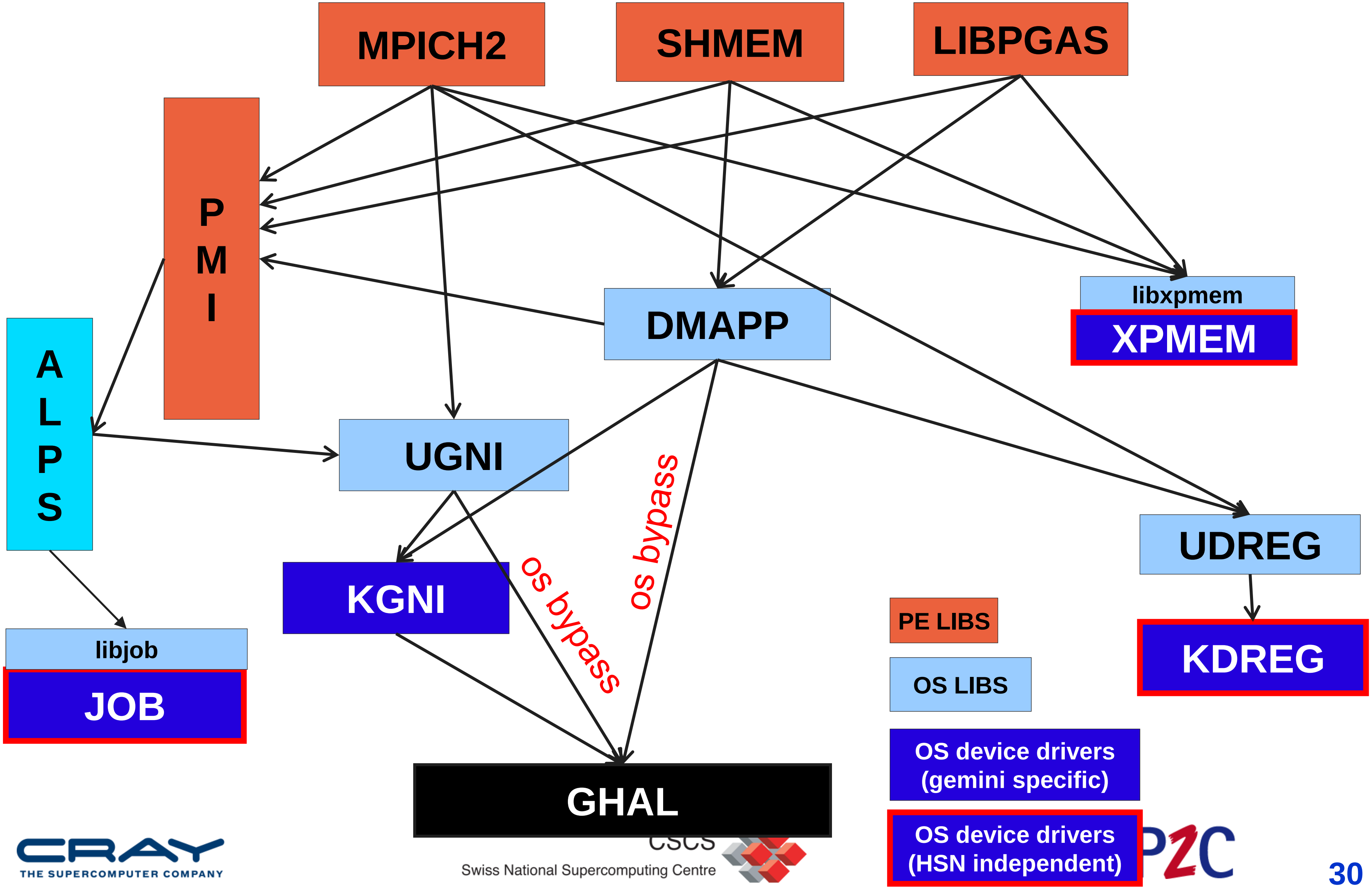
Linux on a Diet – *CNL*



Gemini Software

- Cray MPI uses MPICH2 distribution from Argonne
  - CH3 device Nemesis: multi-method device with a highly optimized shared memory sub-method
- MPI device for Gemini based on
  - User level Gemini Network Interface (uGNI)
  - Distributed Memory Applications (DMAPP) library
- FMA
  - In general used for small transfers
  - FMA transfers are lower latency
- BTE
  - BTE transfers take longer to start but can transfer large amount of data without CPU involvement

MPICH2 in Gemini Software Stack



MPICH2 GNI Netmod Message Protocols

- Eager Protocol
  - For a message that can fit in a GNI SMSG mailbox (E0)
  - For a message that can't fit into a mailbox but is less than MPICH_GNI_MAX_EAGER_MSG_SIZE in length (E1)
- Rendezvous protocol (LMT: Large Message Transfer)
  - RDMA Get protocol – up to 512 KB size messages by default
  - RDMA Put protocol – above 512 KB
- Several MPICH environment variables available
  - Described on mpi man page
  - More on this later…

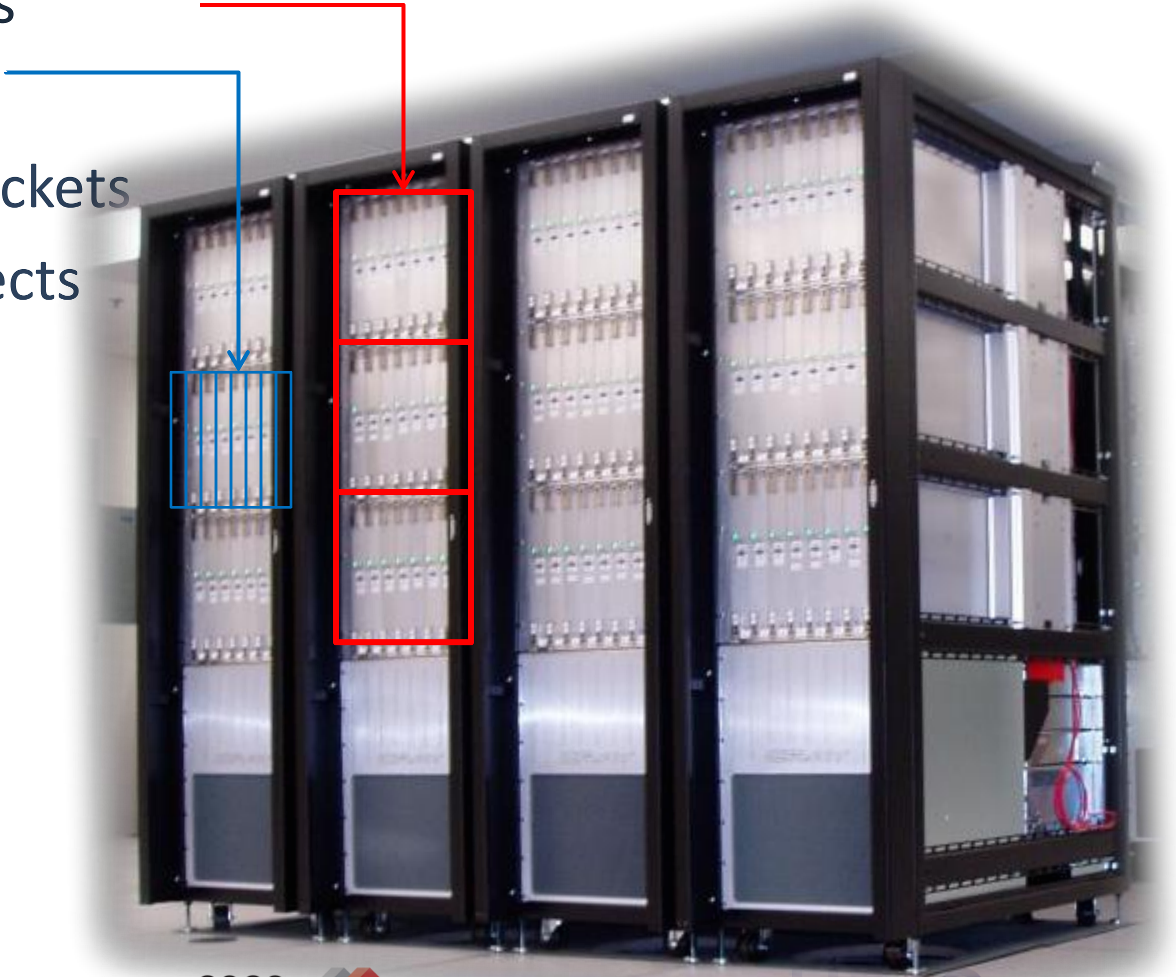
ALPS

- ALPS (Application Level Placement Scheduler)
  - Handles the execution of applications on compute nodes
  - aprun is ALPS application launcher
  - The algorithm used by ALPS to allocate compute nodes for the applications is configurable at ALPS startup
- Compute Node allocation
  - Nodes are allocated in configurable topology aware sequence, according to 3D torus dimensions
  - Max dimension is the “outer” dimension
  - Smallest dimension will change most quickly
  - On rosa (10x12x16) this means: X first, then Y, then Z

The Cray XE6 Configuration

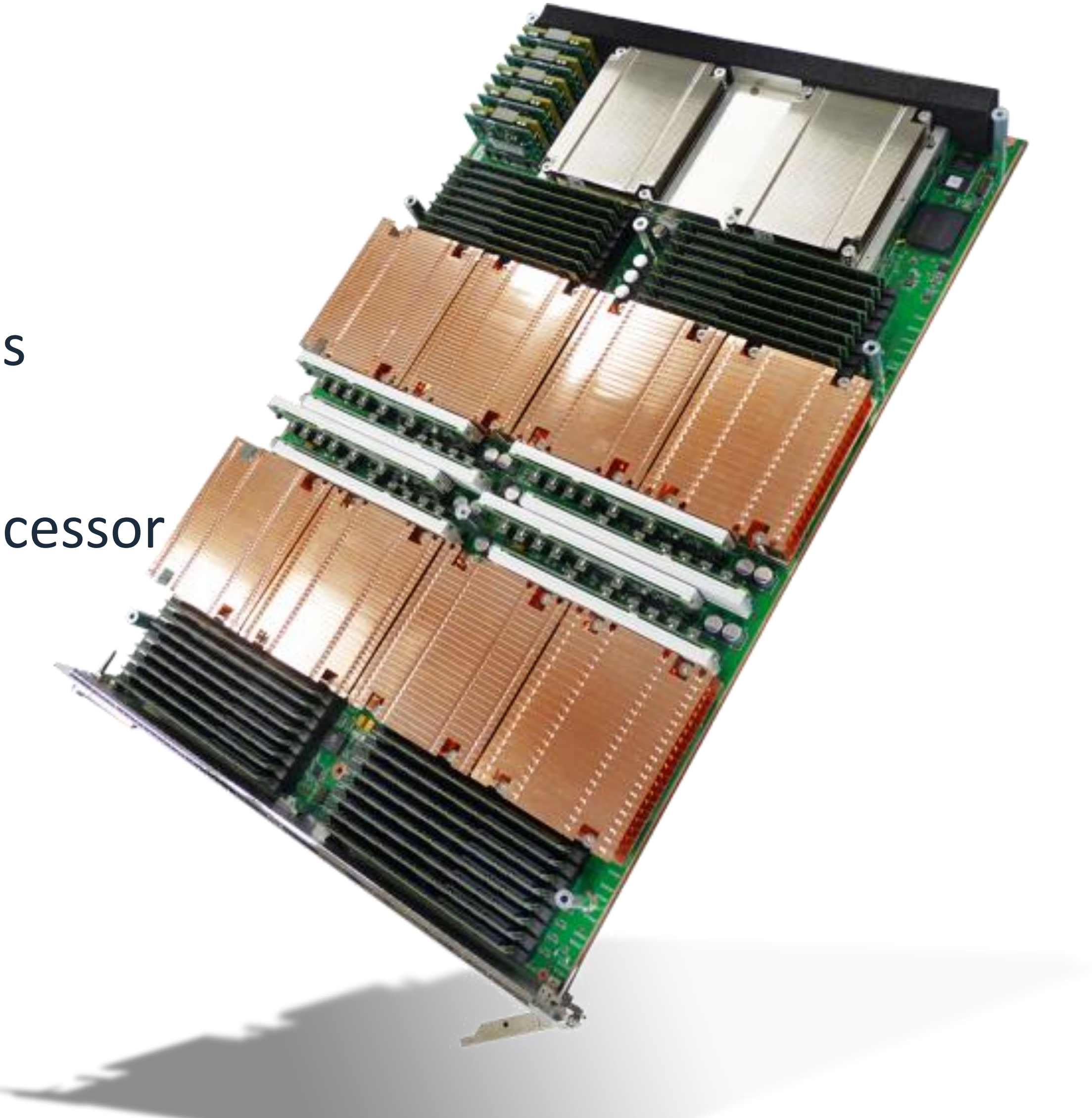
XE6 configuration details

- A XE6 cabinet contains 3 cages
- A cage contains 8 blades
- A blade contains
  - 4 dual-processor sockets
  - 1 Gemini interconnects



Cray XE6 Compute Blade

- 8 Magny Cours Sockets
- 96 Compute Cores
- 32 DDR3 Memory DIMMS
- 32 DDR3 Memory channels
- 2 Gemini ASICs
- L0 Blade management processor



XE6 Topology Tutorial

Class 0 Topology (CSCS palu, XE6m is a bit different)

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

Class 1 Topology

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 (CSCS rosa)

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.

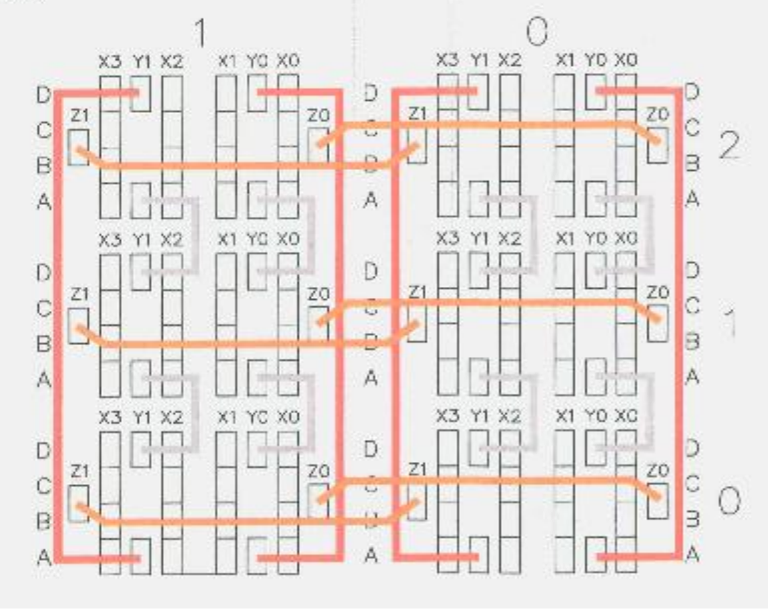
Class 3 Topology (ORNL JaguarPF)

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.

CSCS Palu system

- Cabinets: 2 cabinets
- Topology: 1 x 6 x 16
- Compute
  - 175 nodes
  - 2x12-core (Magny-Cours) @ 2.1 GHz
  - Memory 32 GB/node, 1.3GB/core
  - Peak/node: 202 Gflop/s
  - Peak total: 35 Tflop/s
- Service
  - 4+ blades (17 nodes)
  - 6-core 2.2 GHz
  - Memory 16GB/node

Palu: XE6m 2 Cabinets, Class 1 Topology, 1x6x16



Y

2

2

2

Y: along the blades within the cabinet
2 per blade
6 in total for the 3 chassis

<--- 8 --- Z --- 8 --->

Z: across the blades
8 per chassis
16 in total for the 2 cabs

CRAY THE SUPERCOMPUTER COMPANY

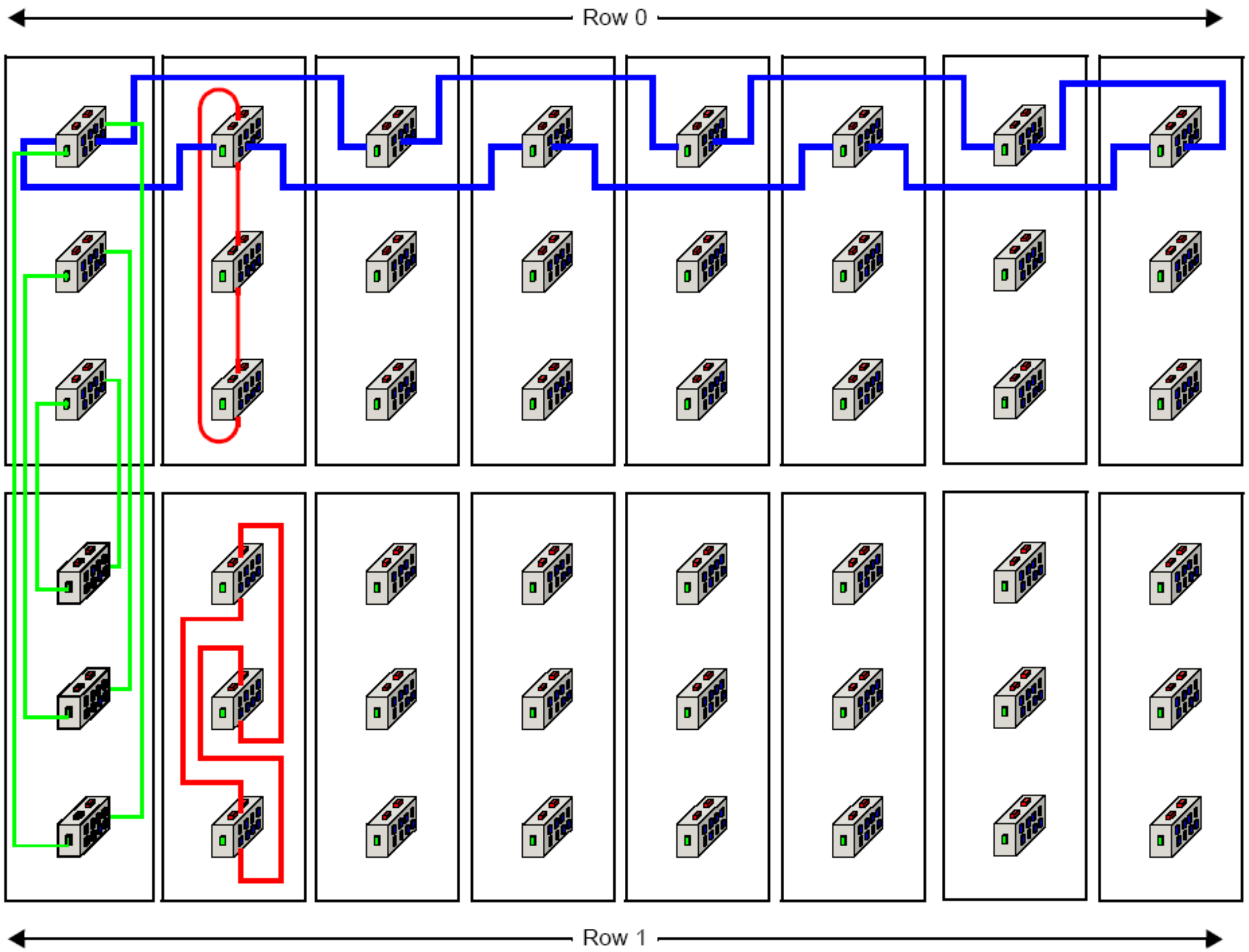
CSCS Swiss National Supercomputing Centre

HP2C

CSCS Rosa system

- Cabinets: 20 cabinets, 2 rows
- Topology: 10 x 12 x 16
- Compute
  - 461 blades (1844 nodes, 22128cores)
  - 2x6-core (Istanbul) @ 2.4 GHz
  - Memory 16 GB/node, 1.3GB/core
  - Peak/node: 115 Gflop/s
  - Peak total: 212 Tflop/s
- Service
  - 19 blades (38 nodes)
  - Dual-core 2.6 GHz
  - Memory 8GB/node

Rosa: 20 Cabinets, Class 2 Topology, 10 x 12 x 16



16 Cabinets
2 Rows of 8
(8 x 12 x 16)

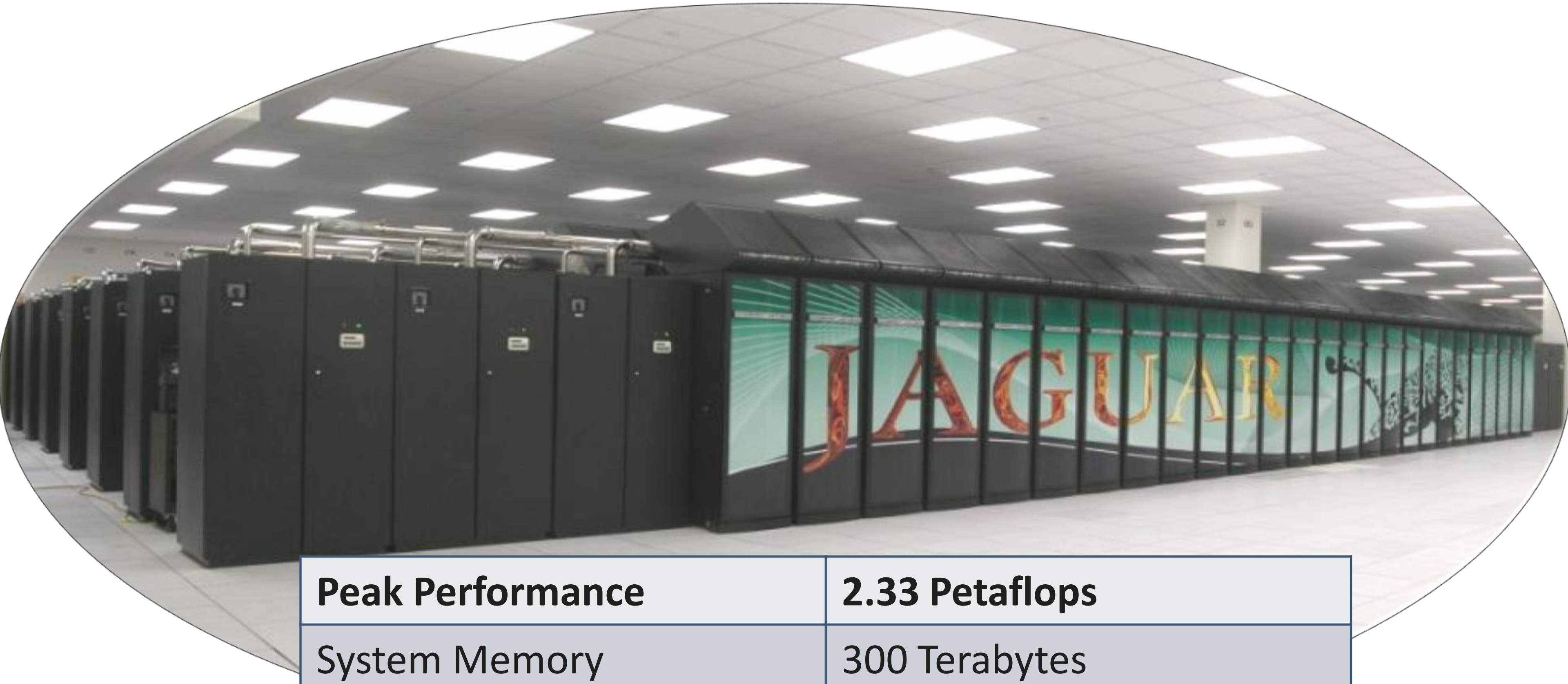
X
Y
Z

THE SUPERCOMPUTER COMPANY

CSCS
Swiss National Supercomputing Centre

HPZC

Class 3: ORNL JaguarPF

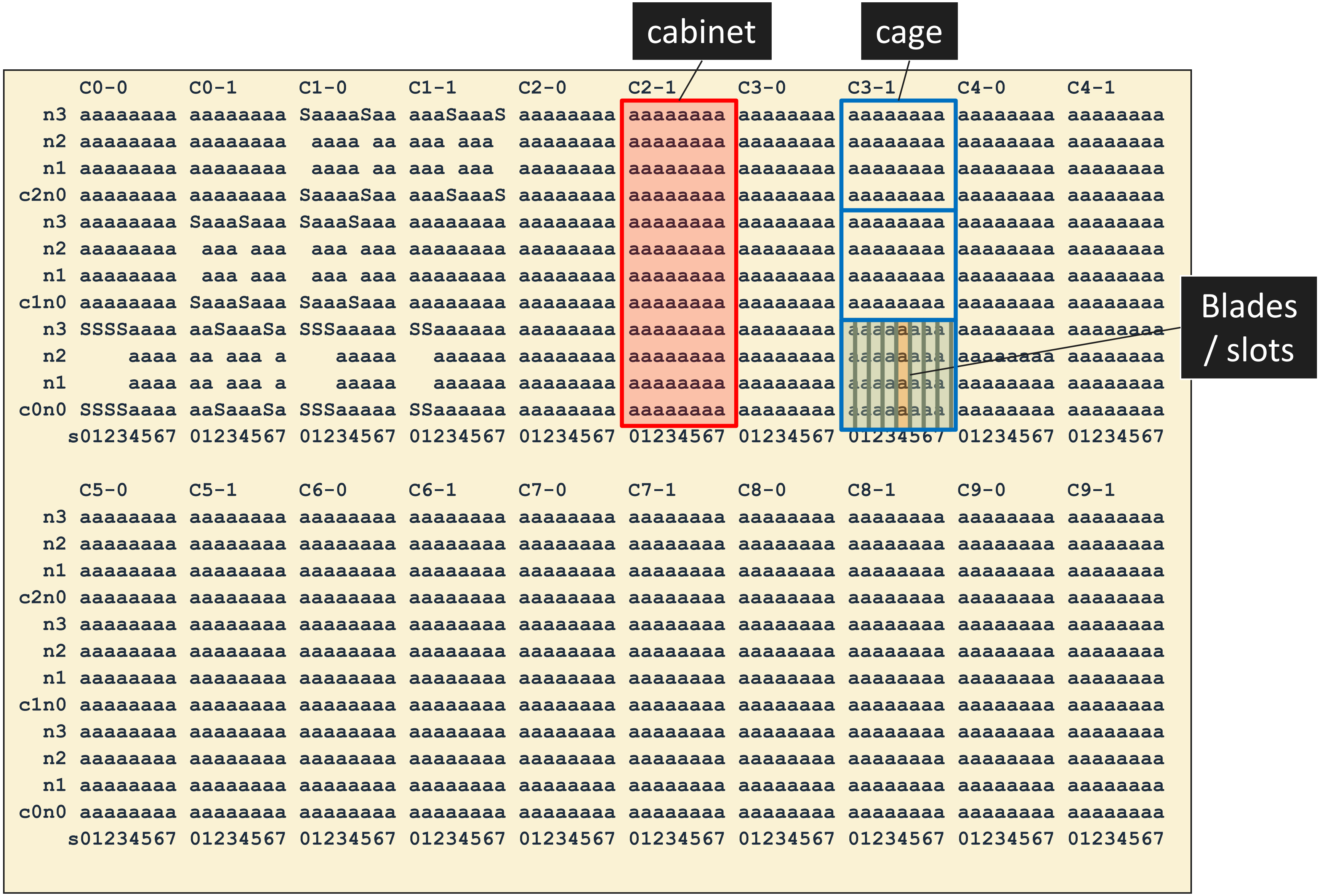


Peak Performance	2.33 Petaflops
System Memory	300 Terabytes
Disk Space	10.7 Petabytes
Interconnect	3D Torus 25x32x24
Processor Cores	224,256

Nodes description: xtprocadmin -A

```
   NID     (HEX)    NODENAME     TYPE ARCH        OS CORES AVAILMEM   PAGESZ CLOCKMHZ
     0       0x0  c0-0c0s0n0  service   xt (service)     2     8000     4096     2600
     3       0x3  c0-0c0s0n3  service   xt (service)     2     8000     4096     2600
     4       0x4  c0-0c0s1n0  service   xt (service)     2     8000     4096     2600
     7       0x7  c0-0c0s1n3  service   xt (service)     2     8000     4096     2600
     8       0x8  c0-0c0s2n0  service   xt (service)     2     8000     4096     2600
    11       0xb  c0-0c0s2n3  service   xt (service)     2     8000     4096     2600
    12       0xc  c0-0c0s3n0  service   xt (service)     2     8000     4096     2600
    15       0xf  c0-0c0s3n3  service   xt (service)     2     8000     4096     2600…
    16      0x10  c0-0c0s4n0  compute   xt       CNL     8    16000     4096     2400
    17      0x11  c0-0c0s4n1  compute   xt       CNL     8    16000     4096     2400
    18      0x12  c0-0c0s4n2  compute   xt       CNL     8    16000     4096     2400
    19      0x13  c0-0c0s4n3  compute   xt       CNL     8    16000     4096     2400
……

  2520     0x9d8  c9-1c2s6n0  compute   xt       CNL     8    16000     4096     2400
  2521     0x9d9  c9-1c2s6n1  compute   xt       CNL     8    16000     4096     2400
  2522     0x9da  c9-1c2s6n2  compute   xt       CNL     8    16000     4096     2400
  2523     0x9db  c9-1c2s6n3  compute   xt       CNL     8    16000     4096     2400
  2524     0x9dc  c9-1c2s7n0  compute   xt       CNL     8    16000     4096     2400
  2525     0x9dd  c9-1c2s7n1  compute   xt       CNL     8    16000     4096     2400
  2526     0x9de  c9-1c2s7n2  compute   xt       CNL     8    16000     4096     2400
  2527     0x9df  c9-1c2s7n3  compute   xt       CNL     8    16000     4096     2400
```

xtnodestat



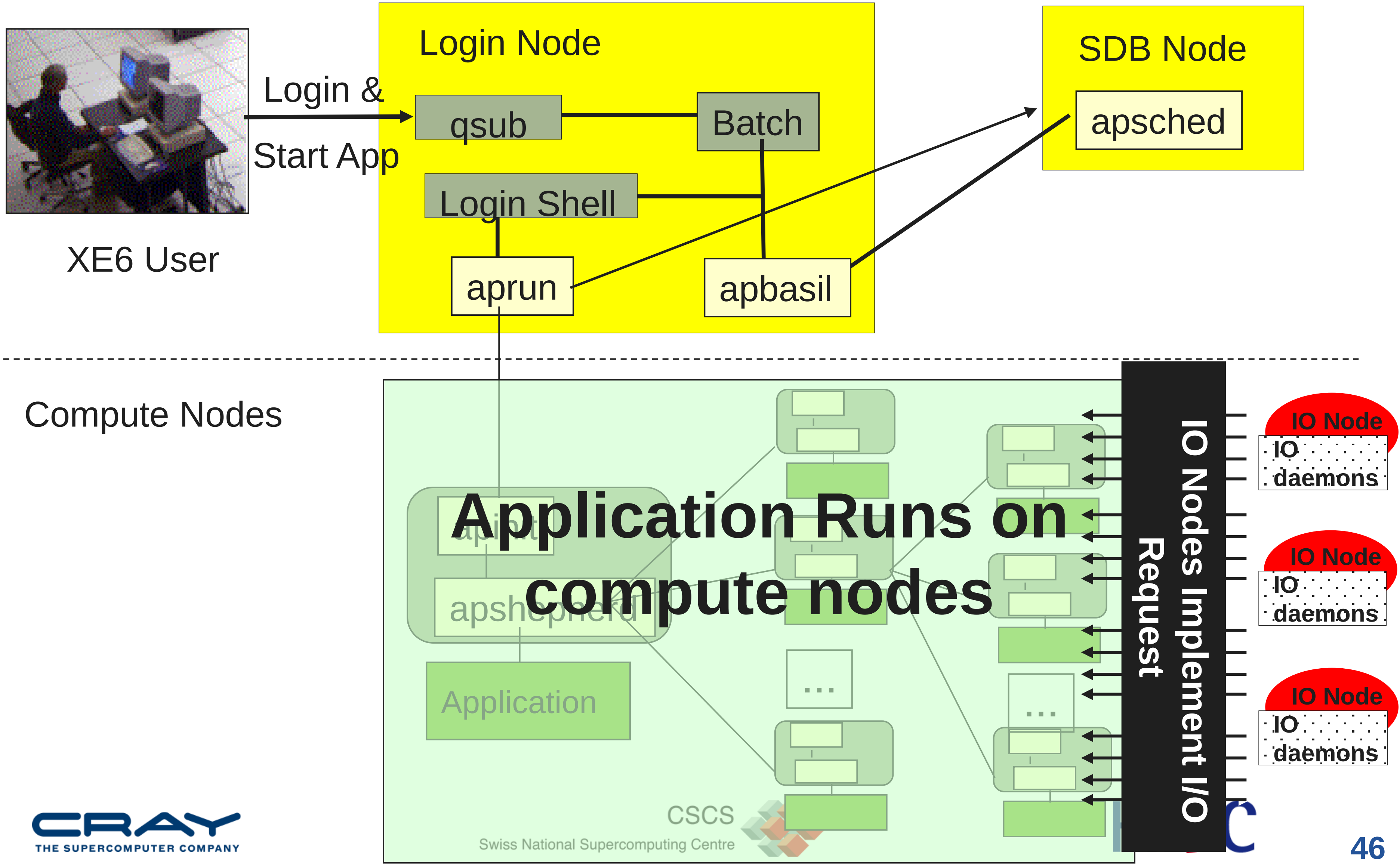
```
     C0-0     C0-1     C1-0     C1-1     C2-0     C2-1     C3-0     C3-1     C4-0     C4-1
  n3 aaaaaaaa aaaaaaaa SaaaaSaa aaaSaaaS aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa
  n2 aaaaaaaa aaaaaaaa  aaaa aa aaa aaa  aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa
  n1 aaaaaaaa aaaaaaaa  aaaa aa aaa aaa  aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa
c2n0 aaaaaaaa aaaaaaaa SaaaaSaa aaaSaaaS aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa
  n3 aaaaaaaa SaaaSaaa SaaaSaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa
  n2 aaaaaaaa  aaa aaa  aaa aaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa
  n1 aaaaaaaa  aaa aaa  aaa aaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa
c1n0 aaaaaaaa SaaaSaaa SaaaSaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa
  n3 SSSSaaaa aaSaaaSa SSSaaaaa SSaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa
  n2     aaaa aa aaa a    aaaaa   aaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa
  n1     aaaa aa aaa a    aaaaa   aaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa
c0n0 SSSSaaaa aaSaaaSa SSSaaaaa SSaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa
    s01234567 01234567 01234567 01234567 01234567 01234567 01234567 01234567 01234567 01234567

     C5-0     C5-1     C6-0     C6-1     C7-0     C7-1     C8-0     C8-1     C9-0     C9-1
  n3 aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa
  n2 aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa
  n1 aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa
c2n0 aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa
  n3 aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa
  n2 aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa
  n1 aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa
c1n0 aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa
  n3 aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa
  n2 aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa
  n1 aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa
c0n0 aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa aaaaaaaa
    s01234567 01234567 01234567 01234567 01234567 01234567 01234567 01234567 01234567 01234567
```

Short interactive session

- xtprocadmin
  - List the node configuration
- xtnodestat
  - List the applications and the node partitions
- apstat
  - Status of applications
- cnselect
  - Provides a list of nodes querying the sdb database

The application launching process

Job Launch



XE6 User

Login & Start App

Login Node

qsub

Batch

Login Shell

aprun

apbasil

SDB Node

apsched

Compute Nodes

apinit

apshepherd

Application

. . .

IO Node

IO daemons

IO Node

IO daemons

IO Node

IO daemons

CSCS

Swiss National Supercomputing Centre

Application Runs on compute nodes

IO Nodes Implement I/O Request

THE SUPERCOMPUTER COMPANY