HECToR to ARCHER

An Introduction from Cray

HECToR – High-End Computing Terascale Resource

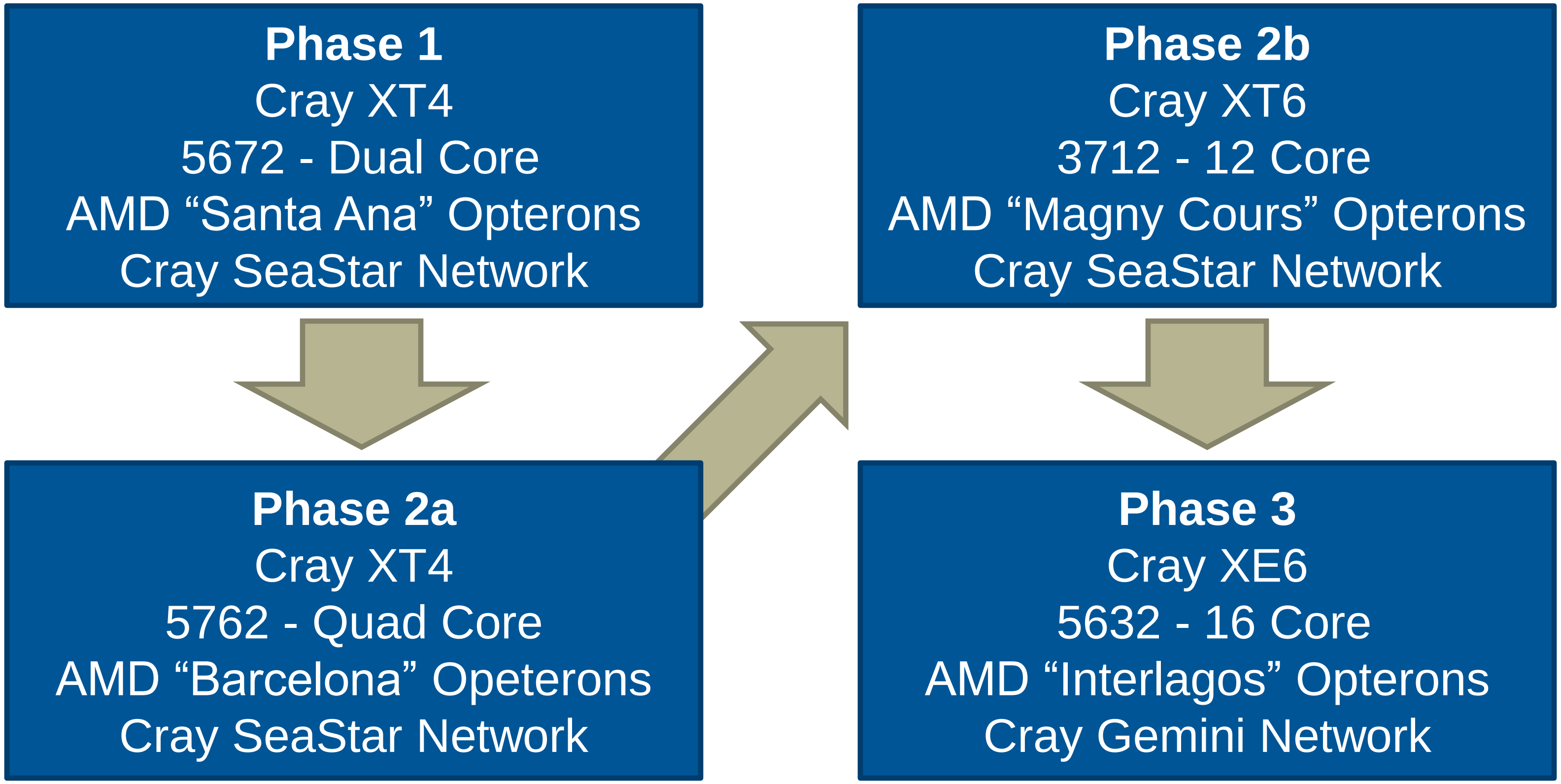
- **HECToR has been the UK’s front-line national supercomputing service since 2007.**
  - The service is due to end in March/April 2013.

- **Funded by the UK Research Councils and used by Academics across the world.**

- **Partners in HECToR:**
  - UK Research Councils
  - EPCC
  - NAG Ltd
  - Cray Inc
  - STFC Daresbury Laboratory
  - EPSRC

The Four Phases of HECToR

While the HECToR service has run continuously, the hardware has been upgraded throughout the lifetime of the service.



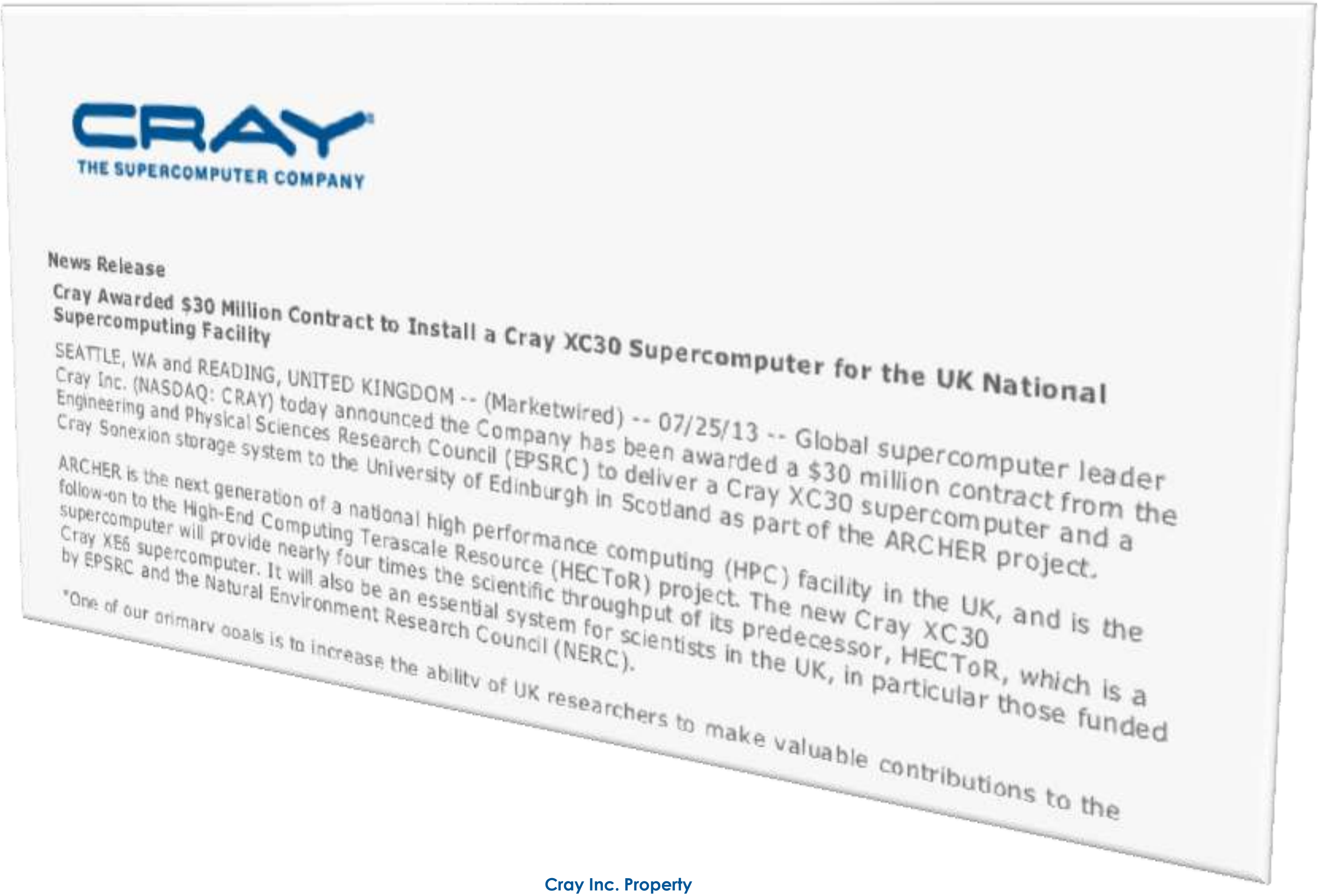
Phase 1
Cray XT4
5672 - Dual Core
AMD “Santa Ana” Opterons
Cray SeaStar Network

Phase 2a
Cray XT4
5762 - Quad Core
AMD “Barcelona” Opeterons
Cray SeaStar Network

Phase 2b
Cray XT6
3712 - 12 Core
AMD “Magny Cours” Opterons
Cray SeaStar Network

Phase 3
Cray XE6
5632 - 16 Core
AMD “Interlagos” Opterons
Cray Gemini Network

ARCHER – following on from HECToR



THE SUPERCOMPUTER COMPANY

News Release

Cray Awarded $30 Million Contract to Install a Cray XC30 Supercomputer for the UK National Supercomputing Facility

SEATTLE, WA and READING, UNITED KINGDOM -- (Marketwired) -- 07/25/13 -- Global supercomputer leader Cray Inc. (NASDAQ: CRAY) today announced the Company has been awarded a $30 million contract from the Engineering and Physical Sciences Research Council (EPSRC) to deliver a Cray XC30 supercomputer and a Cray Sonexion storage system to the University of Edinburgh in Scotland as part of the ARCHER project.

ARCHER is the next generation of a national high performance computing (HPC) facility in the UK, and is the follow-on to the High-End Computing Terascale Resource (HECToR) project. The new Cray XC30 supercomputer will provide nearly four times the scientific throughput of its predecessor, HECToR, which is a Cray XE6 supercomputer. It will also be an essential system for scientists in the UK, in particular those funded by EPSRC and the Natural Environment Research Council (NERC).

"One of our primary goals is to increase the ability of UK researchers to make valuable contributions to the

ARCHER

ARCHER is the follow on service to HECToR.

After official tendering Cray was awarded the contract for hardware and is supplying a Cray XC30 supercomputer.

Cray XC30 is a brand new architecture that offers users significant performance advantages over Cray XE6.

Cray XC30 users will be able to interact with the system using the same tools and interfaces as the Cray XE6.

But how is an Cray XC30 system put together?

The Cray XC30 System

Cray XC30 Systems

DARPA Demo system

NERSC7 - Edison

National Astronomical Observatory of Japan

Swiss National Supercomputer Centre (CSCS)

CSC Finland

XC30 Orders

- **CSIRO Pawsey Centre, Australia**
  - SKA pathfinder system
- **HLRN Germany**
  - Zuse Institute, Berlin
  - Leibniz University, Hannover
- **DWD Germany**
  - Operational weather centre
- **University of Tennessee, USA**
- **HLRS, Germany**
  - XC30 installed, larger system to be installed as part of phase 2
- **Kyoto University, Japan**
  - Research systems, Phase 2 will be an XC
- **JAIST, Japan**
  - Large scale simulations of nano-technology
- **ECMWF, UK**
  - Operational weather centre
- **ARCHER, UK**
  - Academic research system

Nodes: The building blocks

The Cray XC30 is a *Massively Parallel Processing (MPP)* supercomputer design. It is therefore built from many thousands of individual nodes.

There are two basic types of nodes in any Cray XC30:

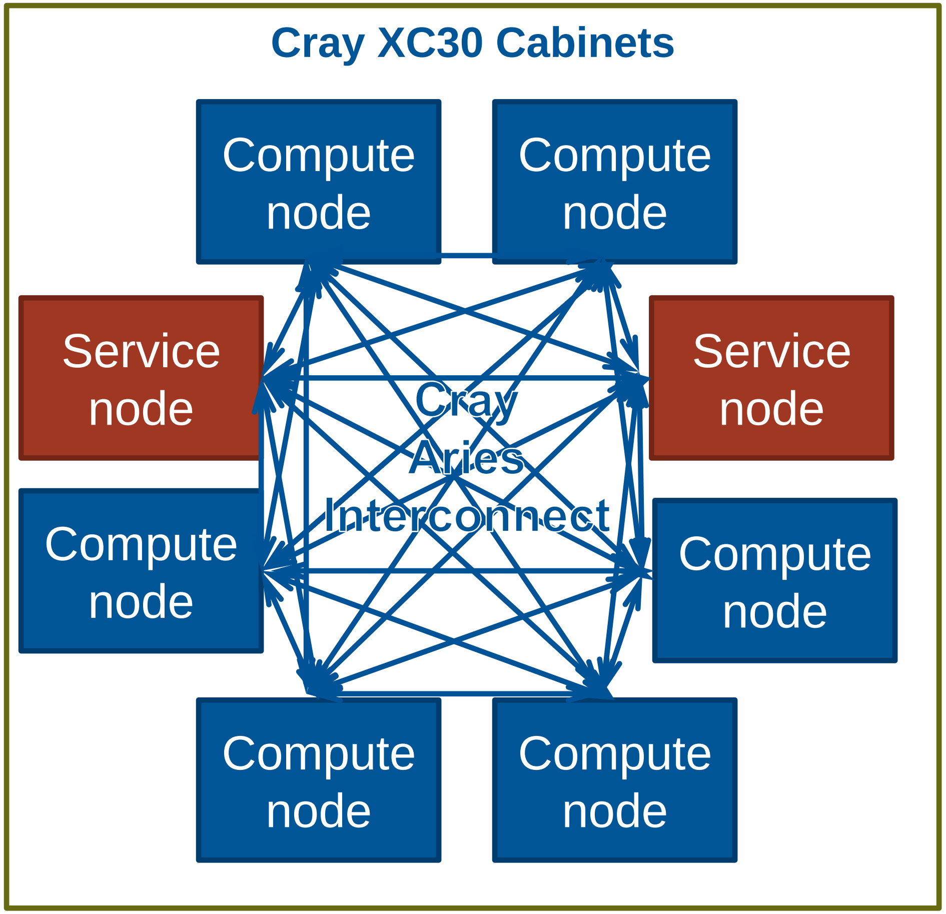
- **Compute nodes**
  - These only do user computation and are always referred to as “Compute nodes”
- **Service nodes**
  - These provide all the additional services required for the system to function, and are given additional names depending on their individual task, e.g. gateway, network, lnet router etc.

There are typically many more compute than service nodes

Connecting nodes together: Aries

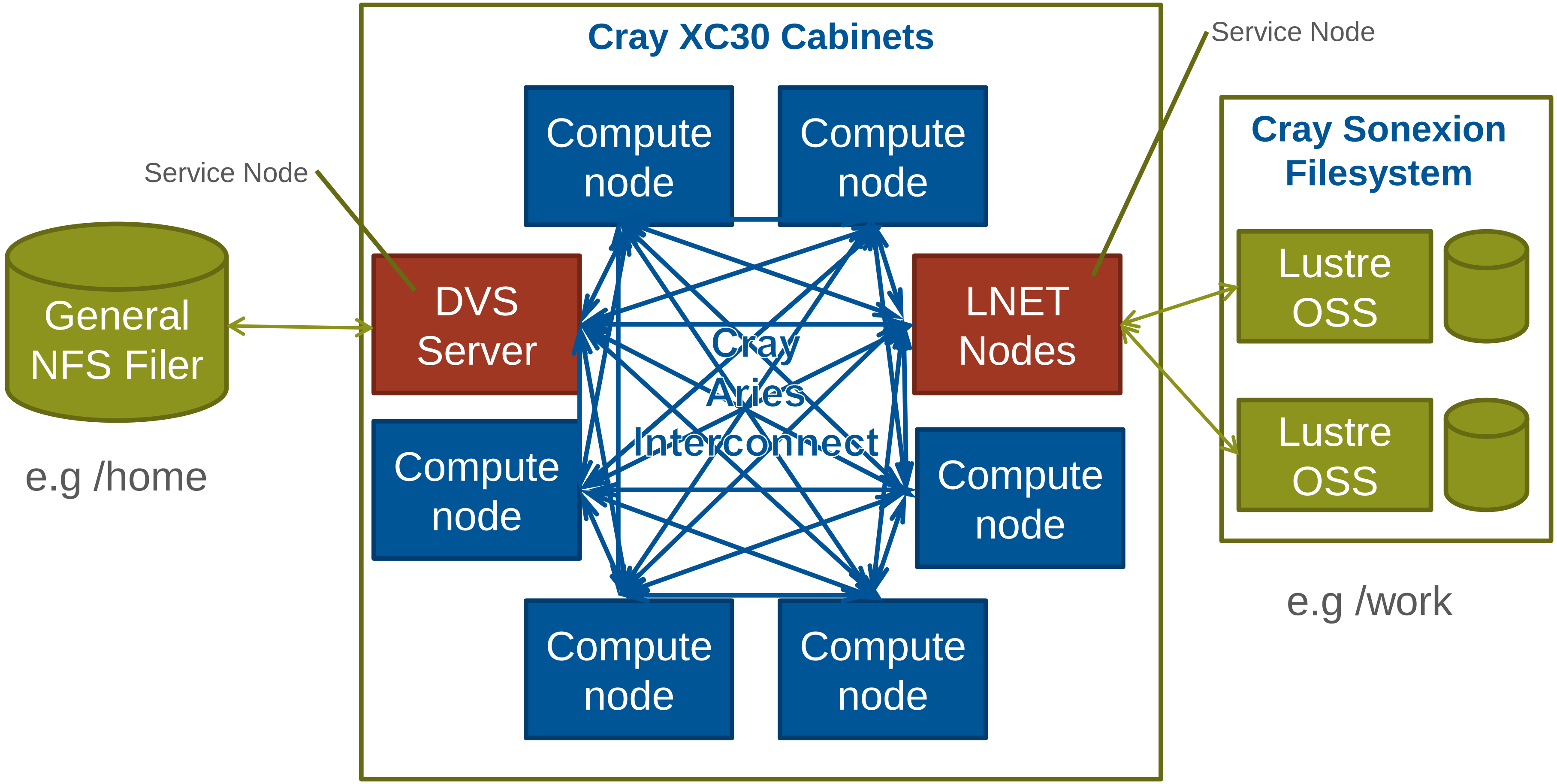
Obviously, to function as a single supercomputer, the individual nodes must have method to communicate with each other.

Every compute and service nodes in an Cray XC30 is interconnected via the high speed, low latency Cray Aries Network.



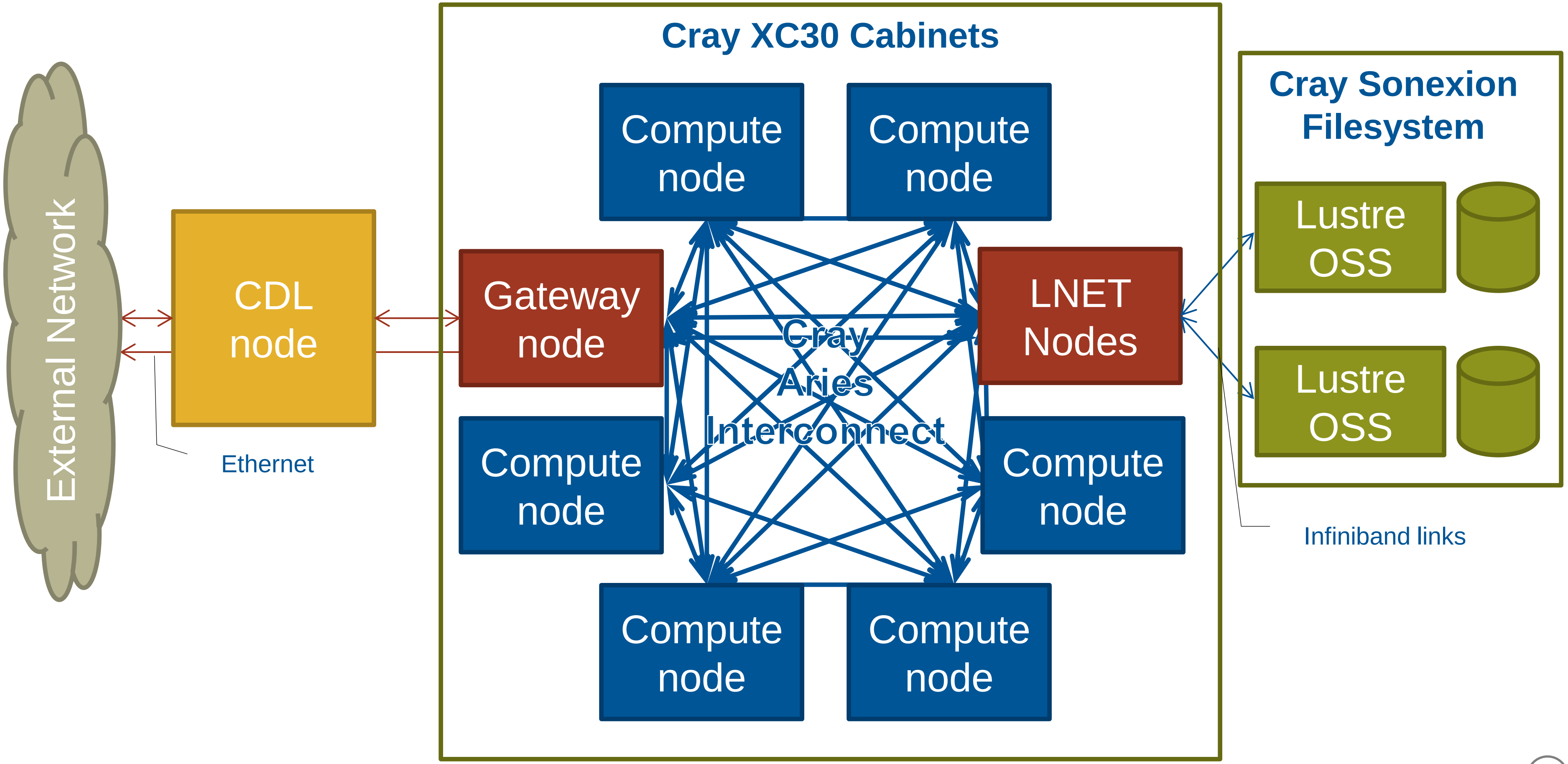
Adding Storage

Neither compute nor service nodes have storage of their own. It must be connected via the service node's native Lustre Client or projected using the Cray Data Virtualization Service (DVS)

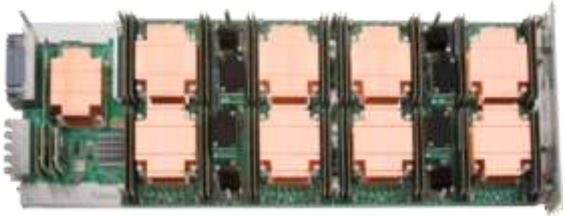


Interacting with the system

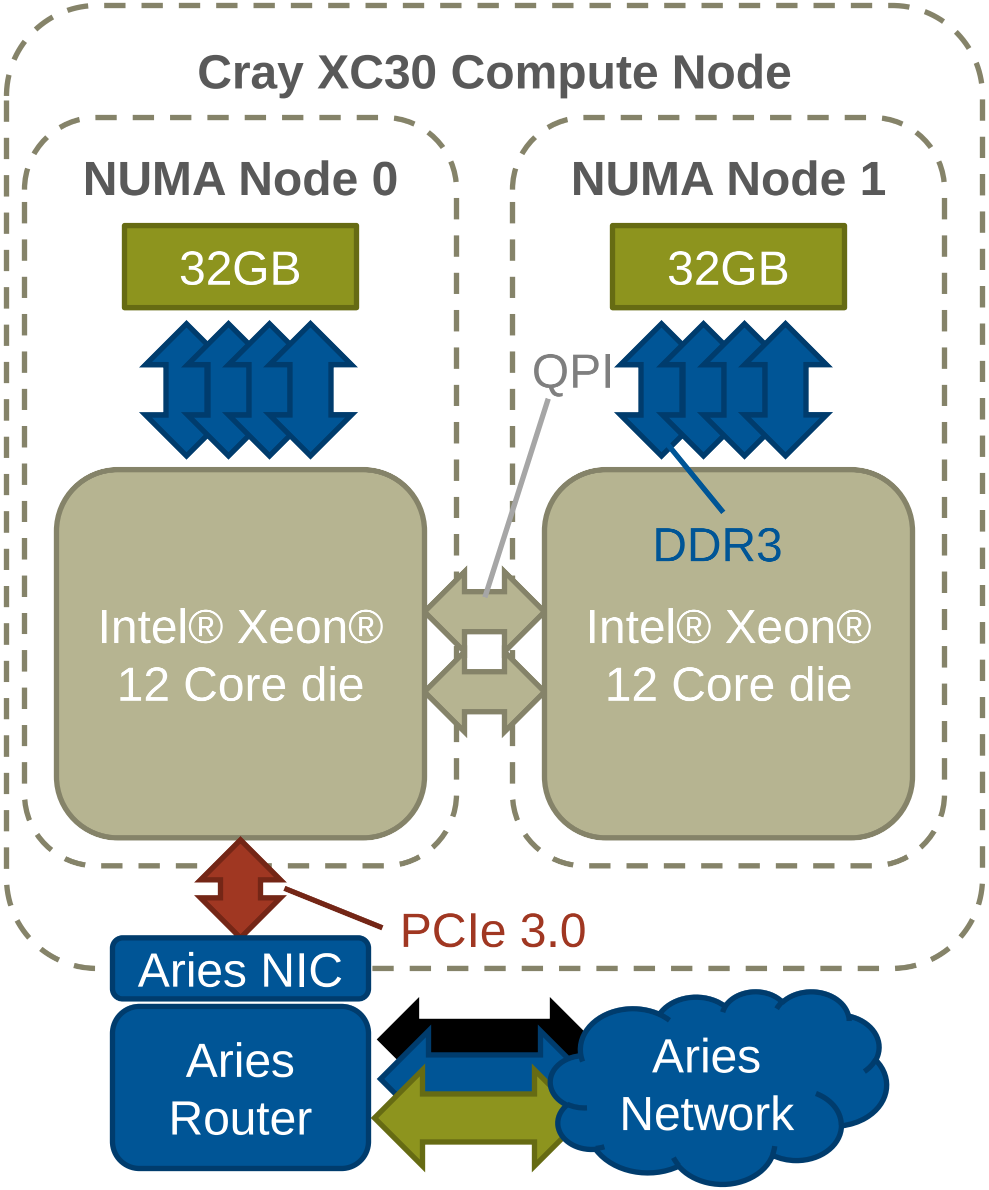
Users do not log directly into the system. Instead they run commands via a CDL server. This server will relay commands and information via a service node referred to as a "Gateway node"



Cray XC30 Compute Node Architecture



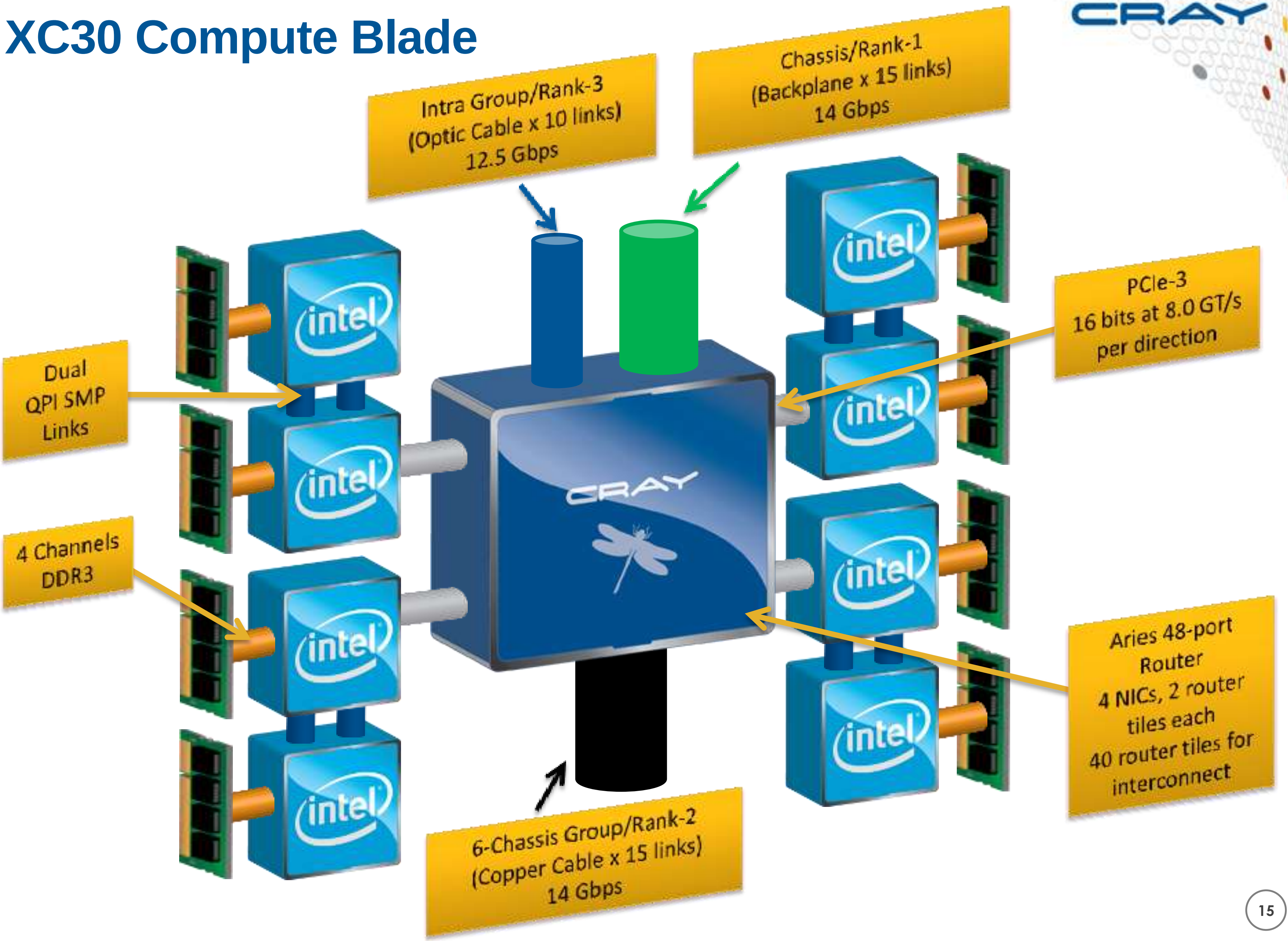
Cray XC30 Intel® Xeon® Compute Node



The XC30 Compute node features:

- **2 x Intel® Xeon® Sockets/die**
  - 12 core Ivybridge
  - QPI interconnect
  - Forms 2 NUMA nodes
- **8 x 1833MHz DDR3**
  - 8 GB per Channel
  - 64 GB total
- **1 x Aries NIC**
  - Connects to shared Aries router and wider network
  - PCI-e 3.0

XC30 Compute Blade



Intra Group/Rank-3
(Optic Cable x 10 links)
12.5 Gbps

Chassis/Rank-1
(Backplane x 15 links)
14 Gbps

PCIe-3
16 bits at 8.0 GT/s
per direction

Dual
QPI SMP
Links

4 Channels
DDR3

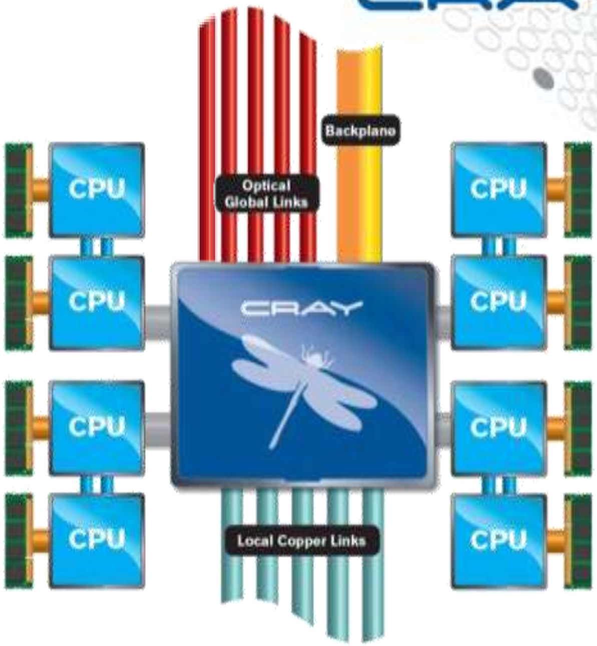
Aries 48-port
Router
4 NICs, 2 router
tiles each
40 router tiles for
interconnect

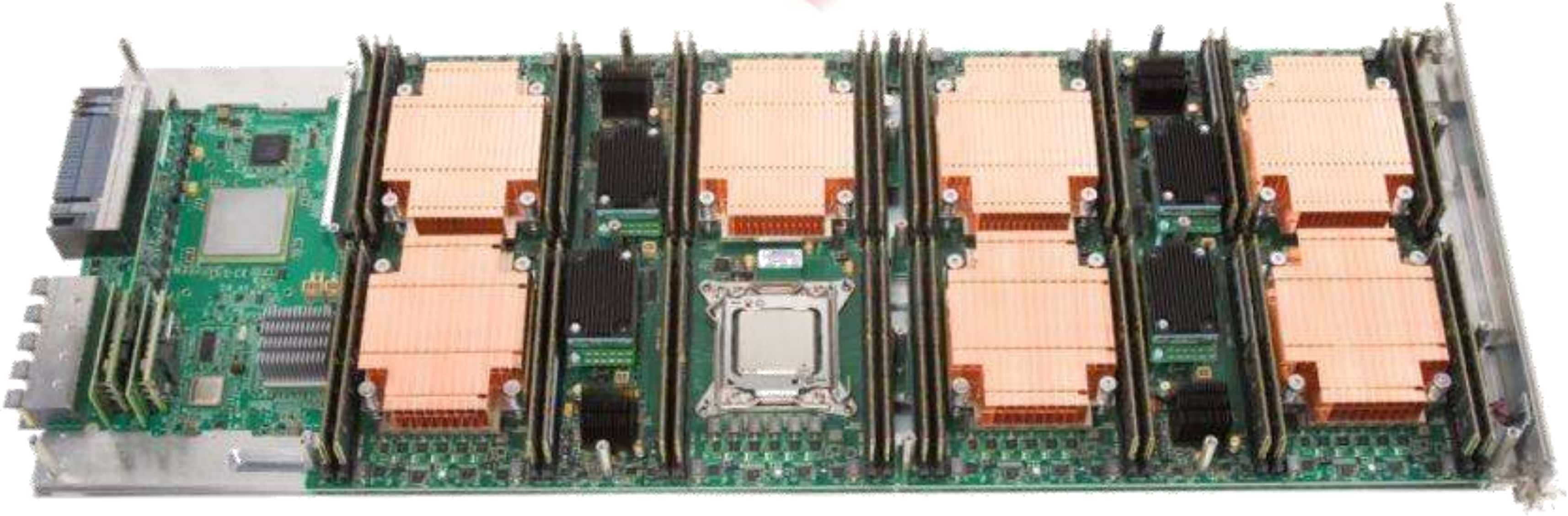
6-Chassis Group/Rank-2
(Copper Cable x 15 links)
14 Gbps

Cray XC30 Fully Populated Compute Blade

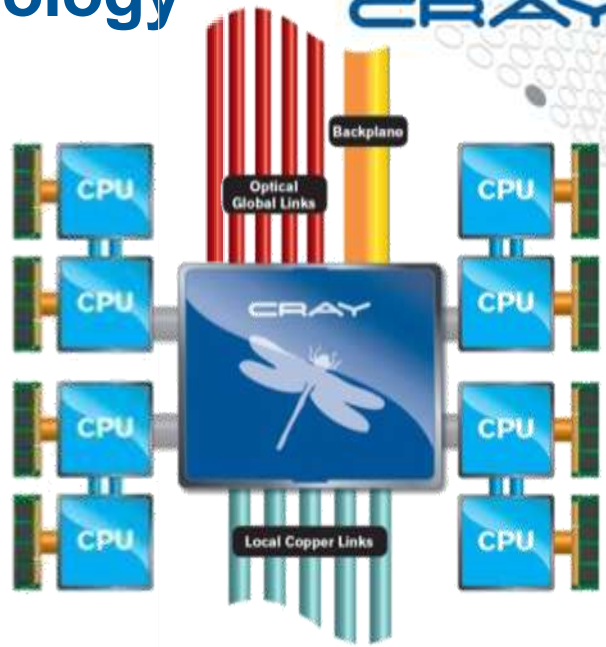
SPECIFICATIONS

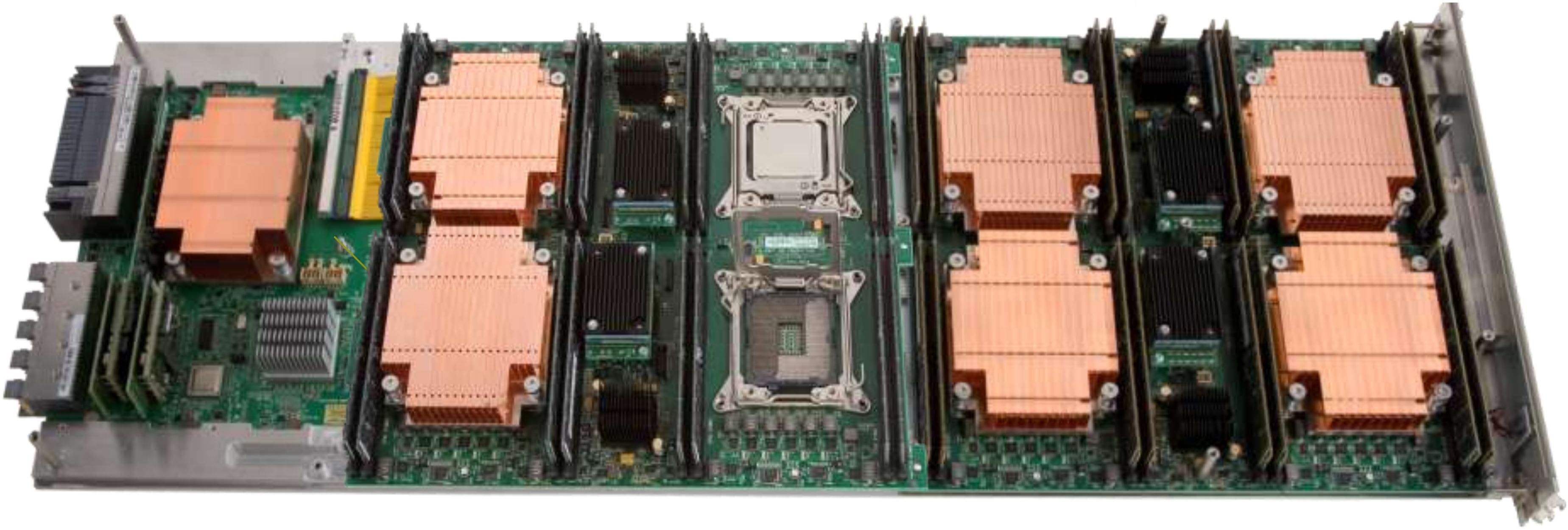
Module power:	2014 Watts
PDC max. power:	900 Watt
Air flow req.:	275 cfm
Size:	2.125 in x 12.95 in x 33.5 in
Weight:	<40 lbm



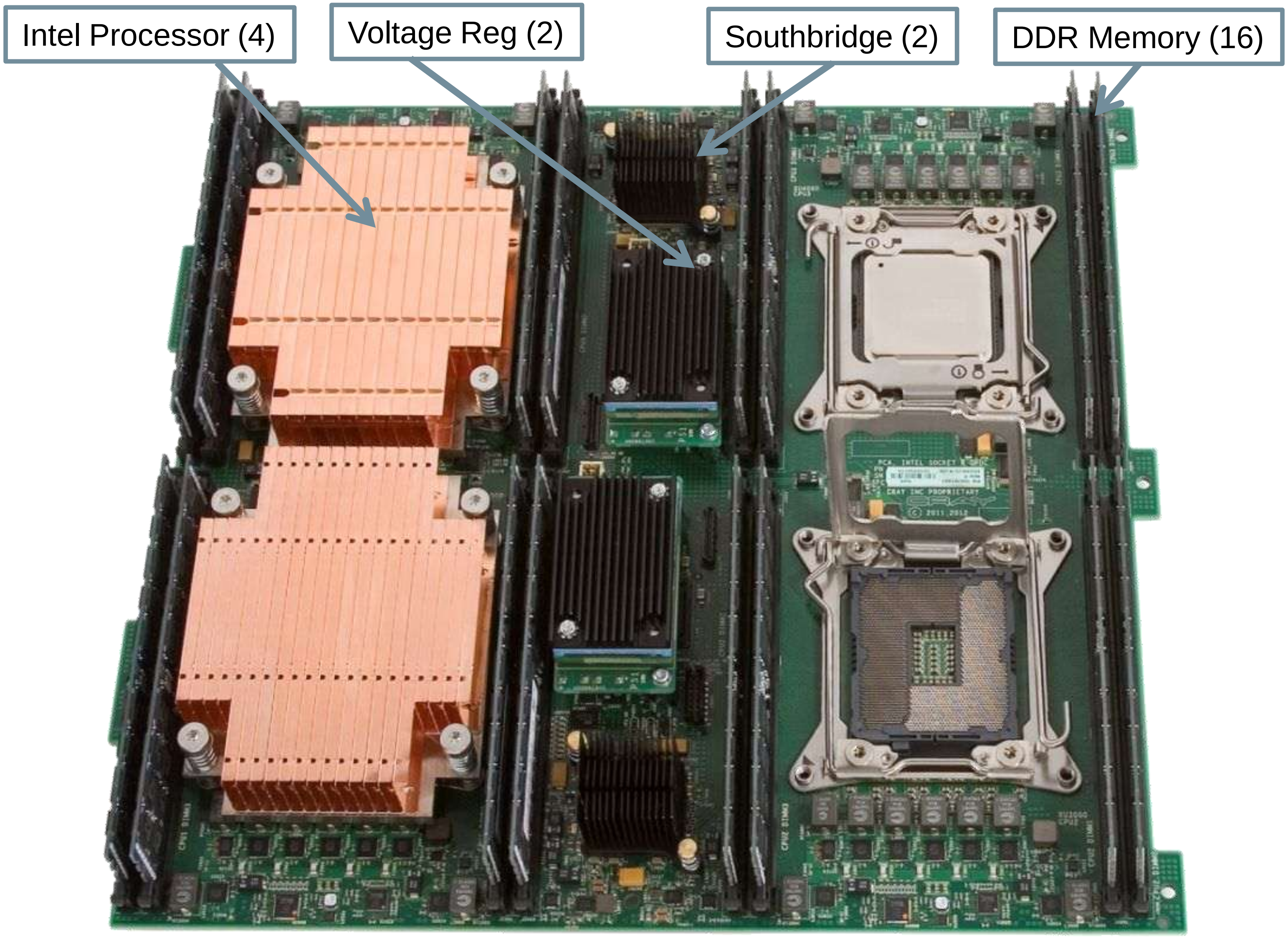


PDC’s are Upgradeable to New Technology





Cray XC30 Quad Processor Daughter Card



Intel Processor (4)

Voltage Reg (2)

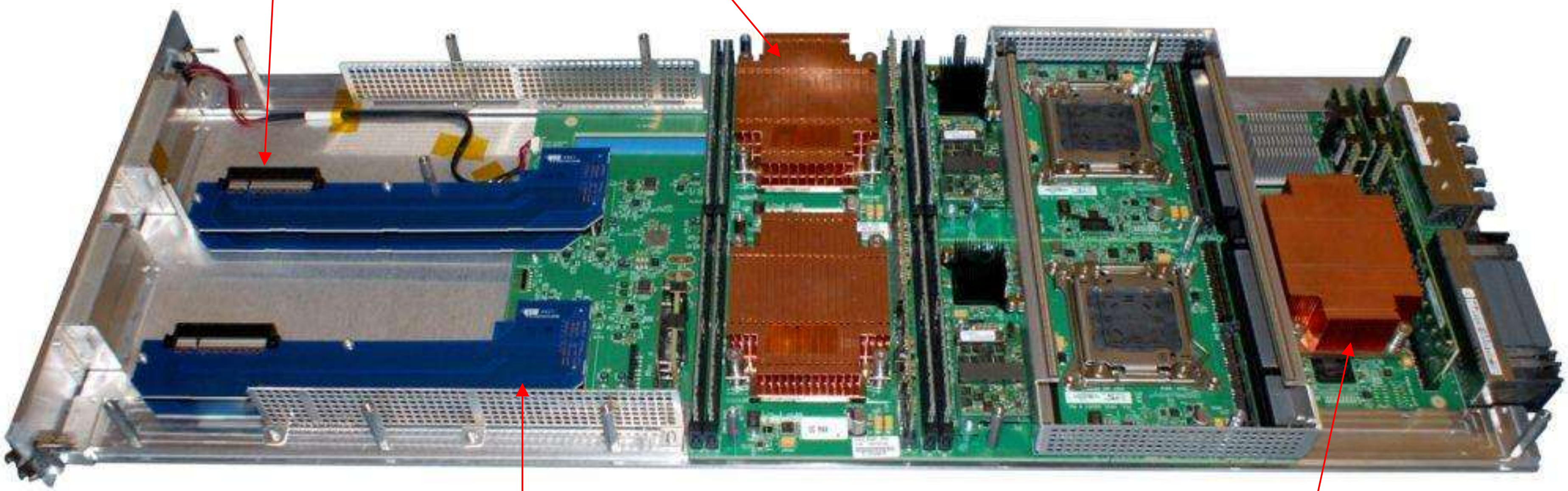
Southbridge (2)

DDR Memory (16)

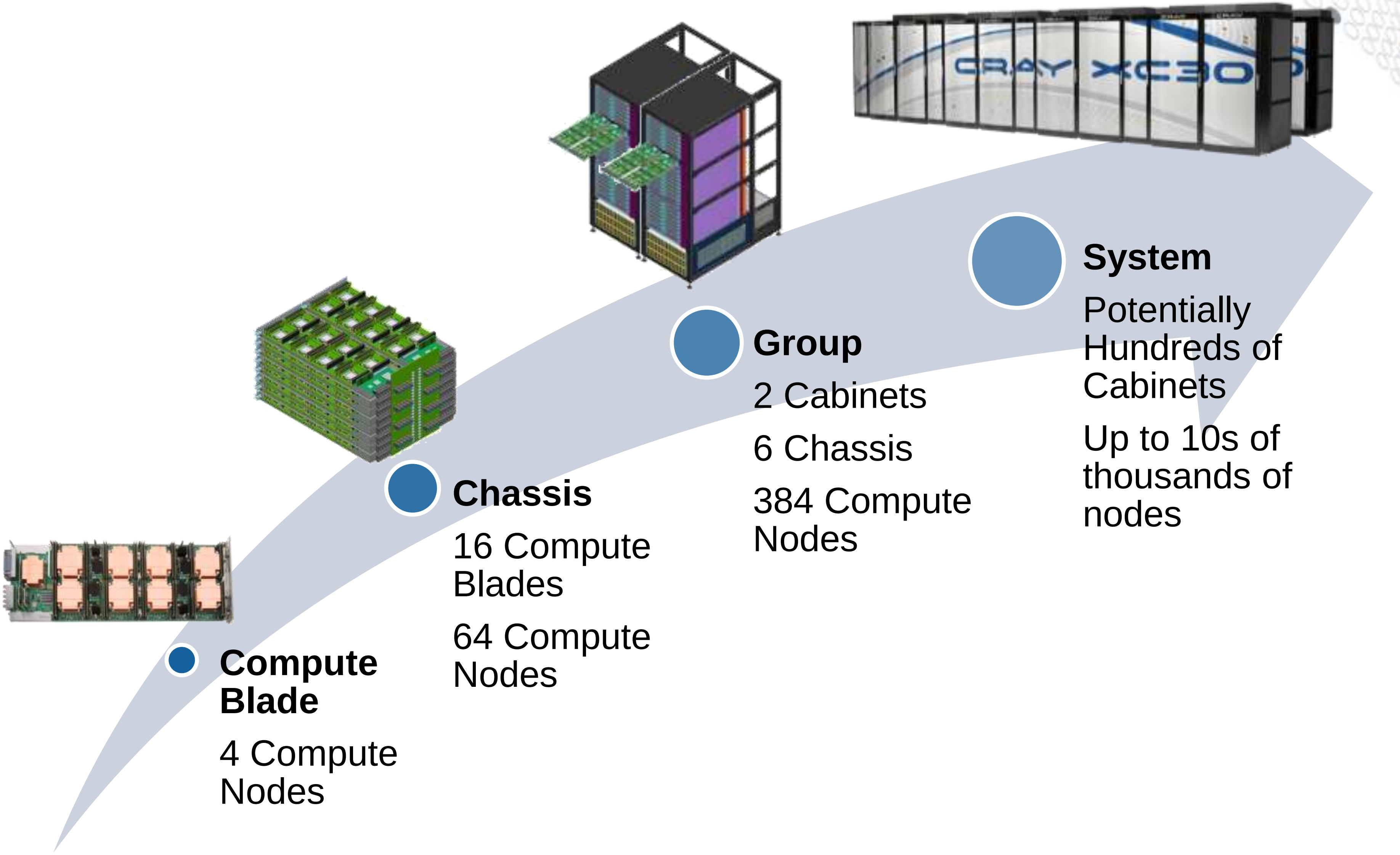
Cray XC Service Node

SPECIFICATIONS

Module power:	1650 Watts
PDC max. power:	225 Watt
Air flow req.:	275 cfm
Size:	2.125 in x 12.95 in x 33.5 in
Weight:	35 lbs



Cray XC30 – Grouping the nodes



Compute Blade

4 Compute Nodes

Chassis

16 Compute Blades

64 Compute Nodes

Group

2 Cabinets

6 Chassis

384 Compute Nodes

System

Potentially Hundreds of Cabinets

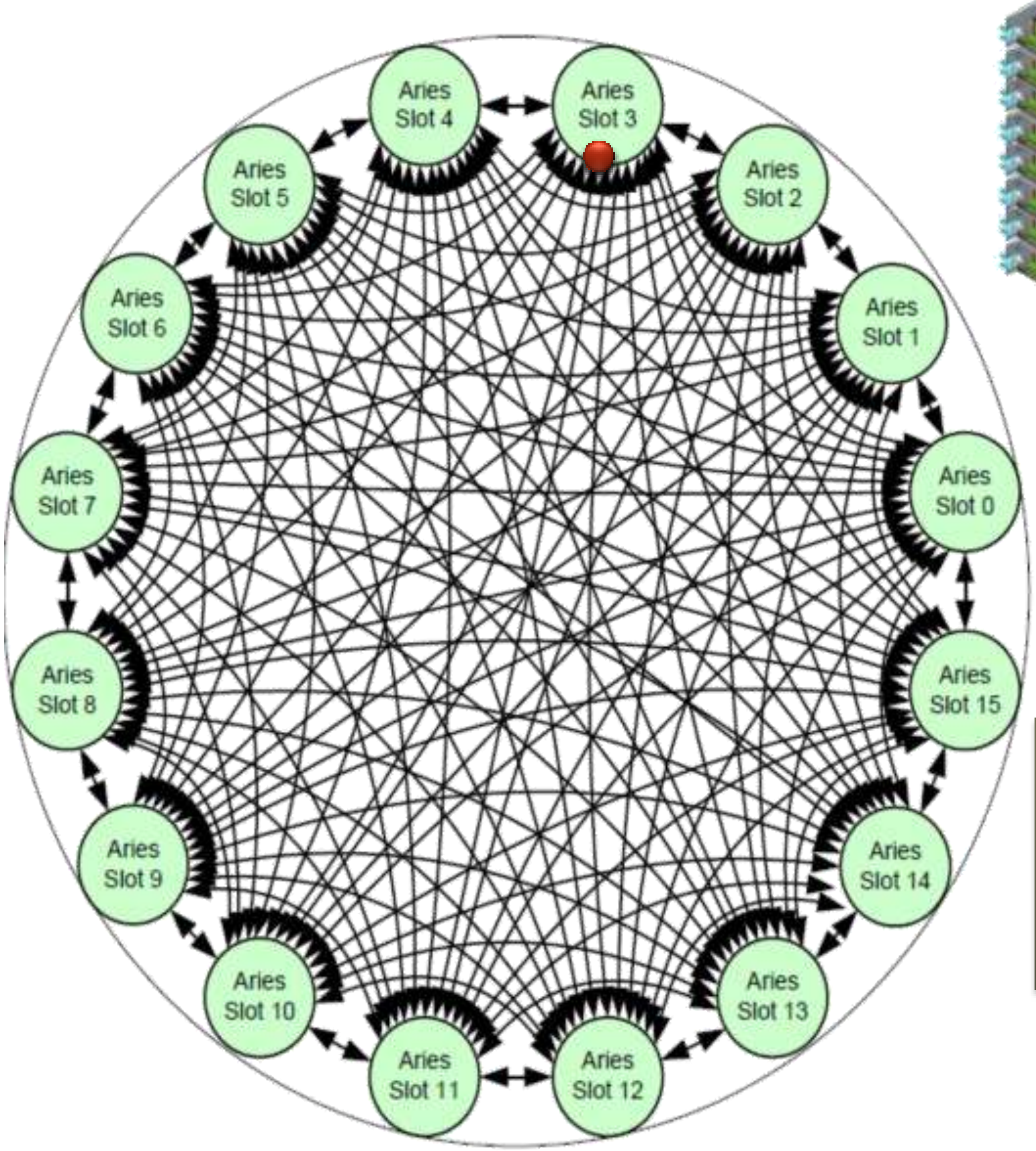
Up to 10s of thousands of nodes

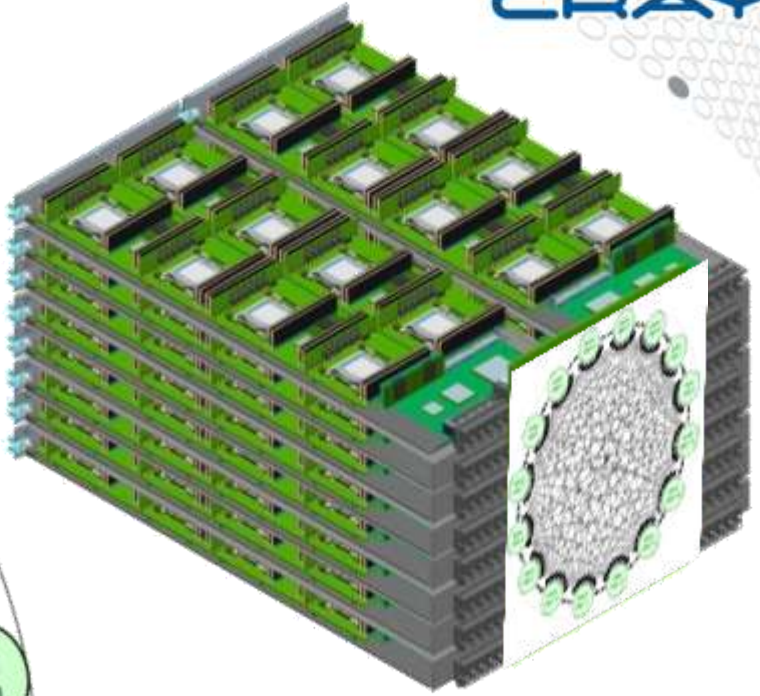
Cray XC30 Dragonfly Topology + Aries

Cray Aries Features

- **Scalability to > 500,000 X86 Cores**
  - Cray users run large jobs – 20-50% of system size is common
  - Many examples of 50K-250K MPI tasks per job
  - Optimized collectives MPI_Allreduce in particular
- **Optimized short transfer mechanism (FMA)**
  - Provides global access to memory, used by MPI and PGAS
  - High issue rate for small transfers: 8-64 byte put/get and amo in particular
- **HPC optimized network**
  - Small packet size 64-bytes
  - Router bandwidth >> injection bandwidth
  - Adaptive Routing & Dragonfly topology
- **Connectionless design**
  - Doesn’t depend on a connection cache for performance
  - Limits the memory required per node
- **Fault tolerant design**
  - Link level retry on error
  - Adaptive routing around failed links
  - Network reconfigures automatically (and quickly) if a component fails
  - End to end CRC check with automatic software retry in MPI

Cray XC30 Rank1 Network

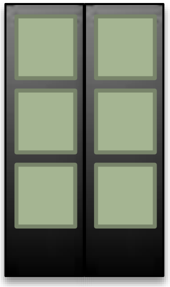


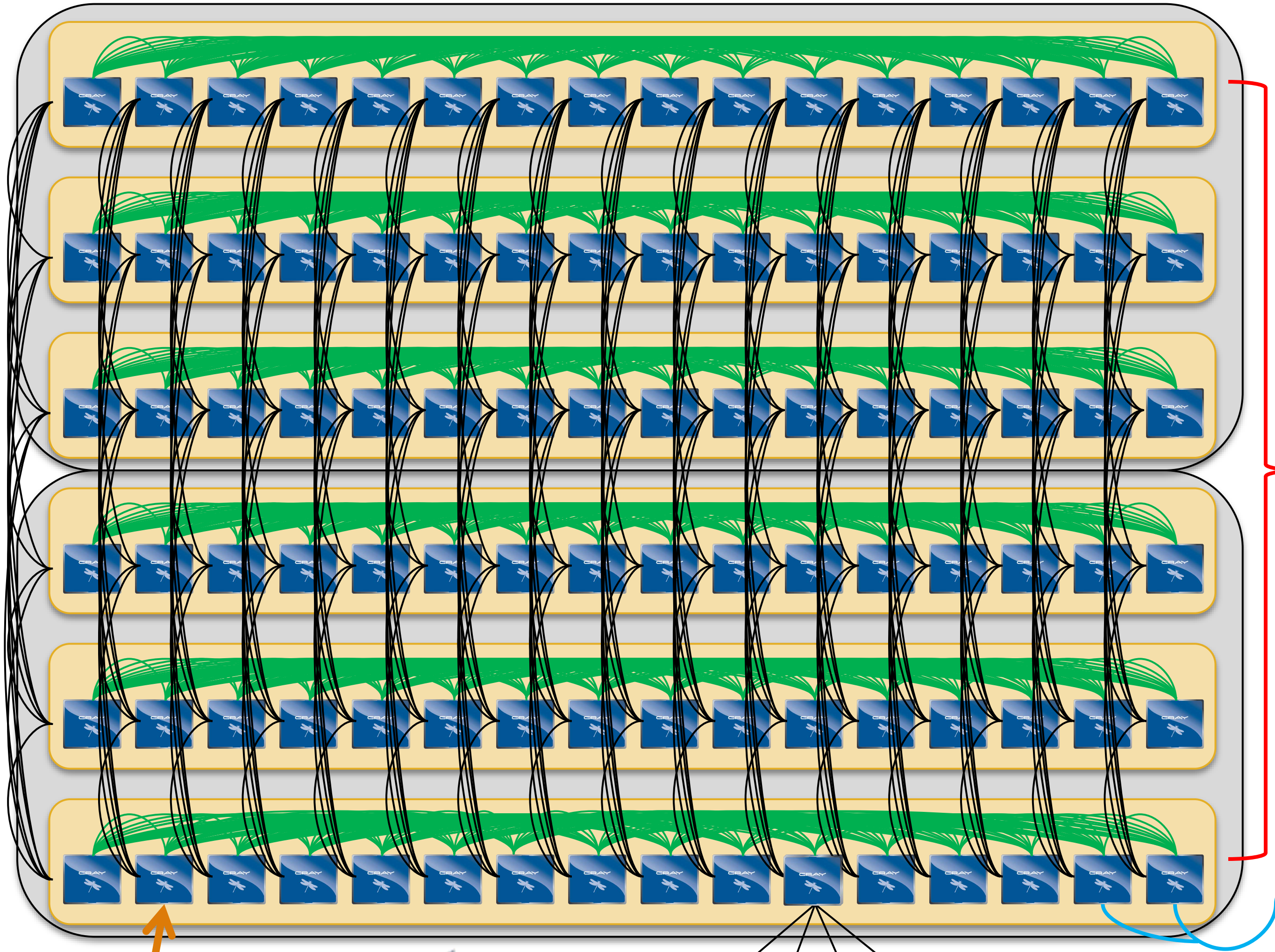


- Chassis with 16 compute blades
- 128 Sockets
- Inter-Aries communication over backplane
- Per-Packet adaptive Routing

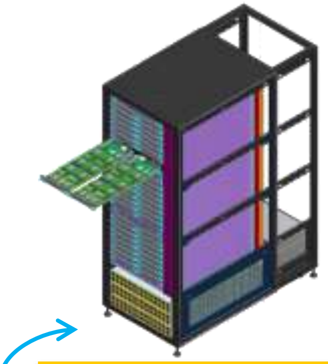
Cray XC30 Rank-2 Copper Network

2 Cabinet Group
768 Sockets



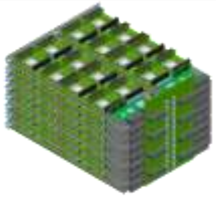


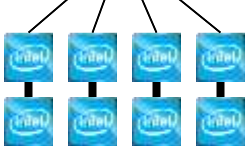
6 backplanes connected with copper cables in a 2-cabinet group: “Black Network”



Active optical cables interconnect groups
“Blue Network”

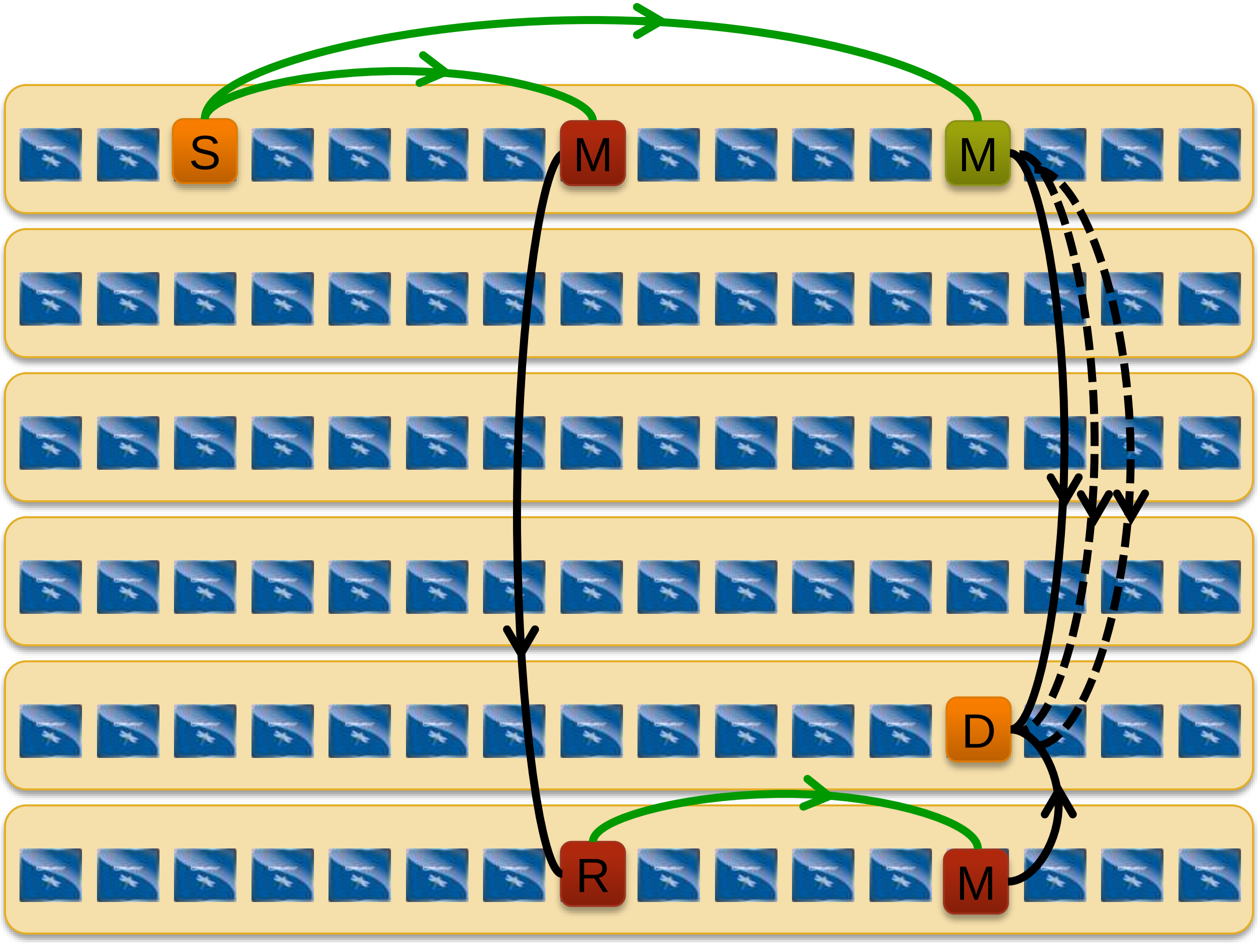
16 Aries connected by backplane
“Green Network”





4 nodes connect to a single Aries

Cray XC30 Routing



Minimal routes between any two nodes in a group are just two hops

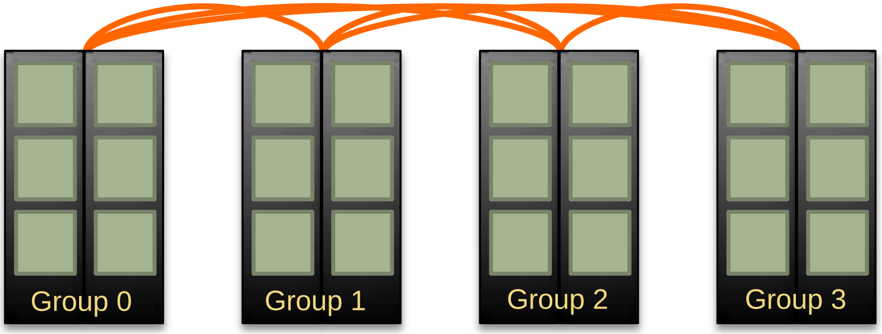
Non-minimal route requires up to four hops.

With adaptive routing we select between minimal and non-minimal paths based on load

The Cray XC30 Class-2 Group has sufficient bandwidth to support full injection rate for all 384 nodes with non-minimal routing

Cray XC30 Network Overview – Rank-3 Network

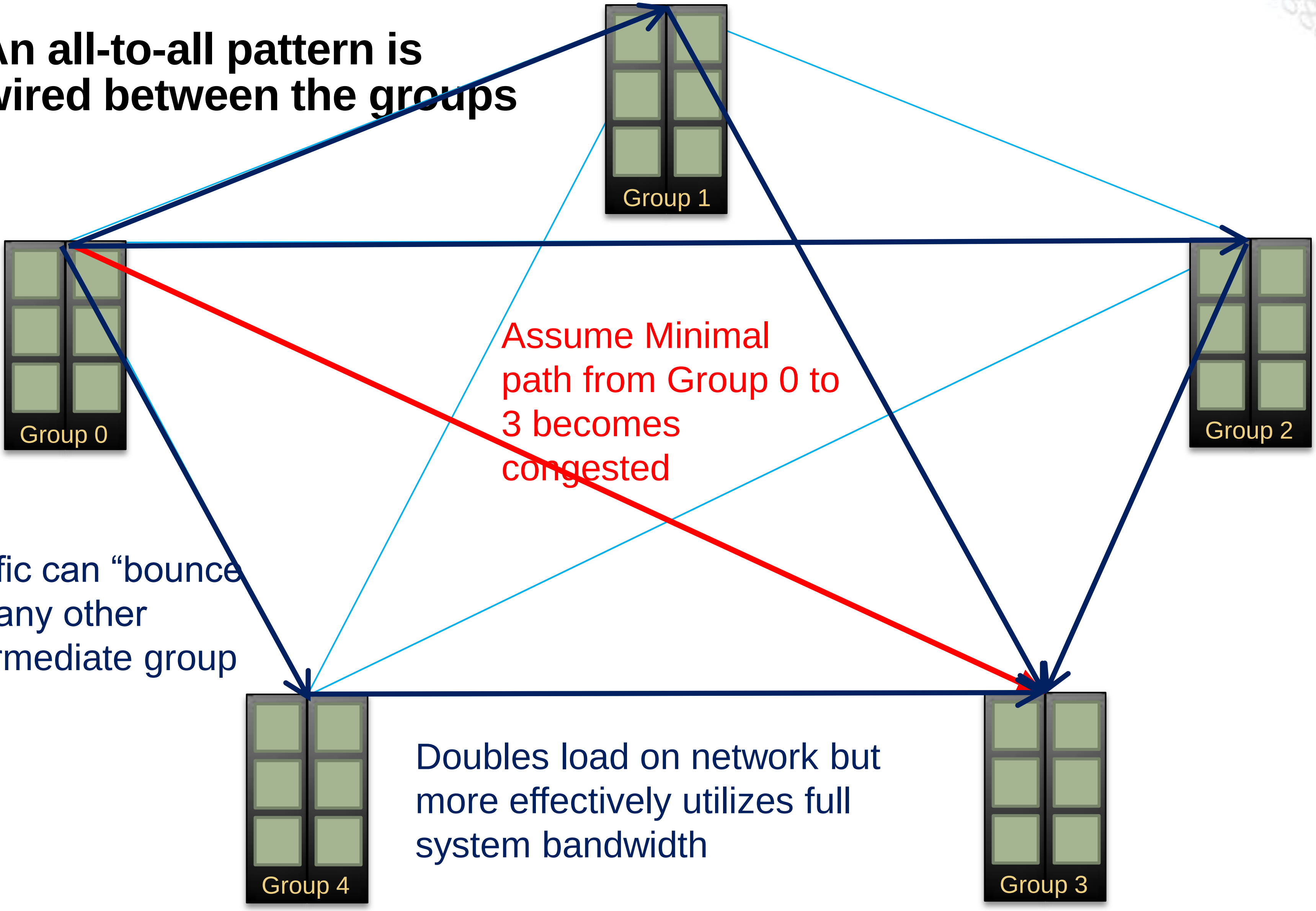
- An all-to-all pattern is wired between the groups using optical cables (blue network)
- Up to 240 ports are available per 2-cabinet group
- The global bandwidth can be tuned by varying the number of optical cables in the group-to-group connections



Example: An 4-group system is interconnected with 6 optical “bundles”. The “bundles” can be configured between 20 and 80 cables wide

Adaptive Routing over the Blue Network

- **An all-to-all pattern is wired between the groups**



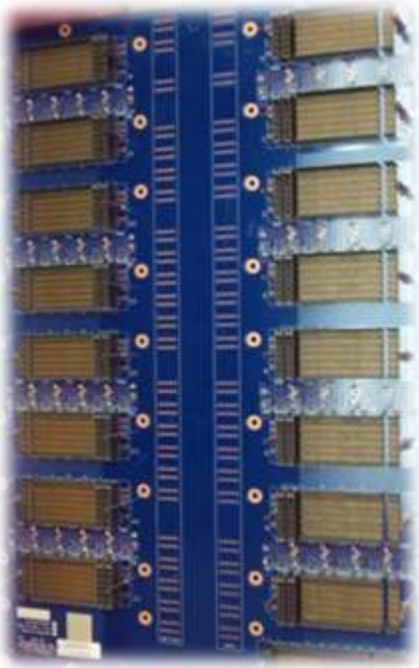
Assume Minimal path from Group 0 to 3 becomes congested

Traffic can “bounce off” any other intermediate group

Doubles load on network but more effectively utilizes full system bandwidth

Cray XC30 Network

- The Cray XC30 system is built around the idea of optimizing interconnect bandwidth and associated cost at every level



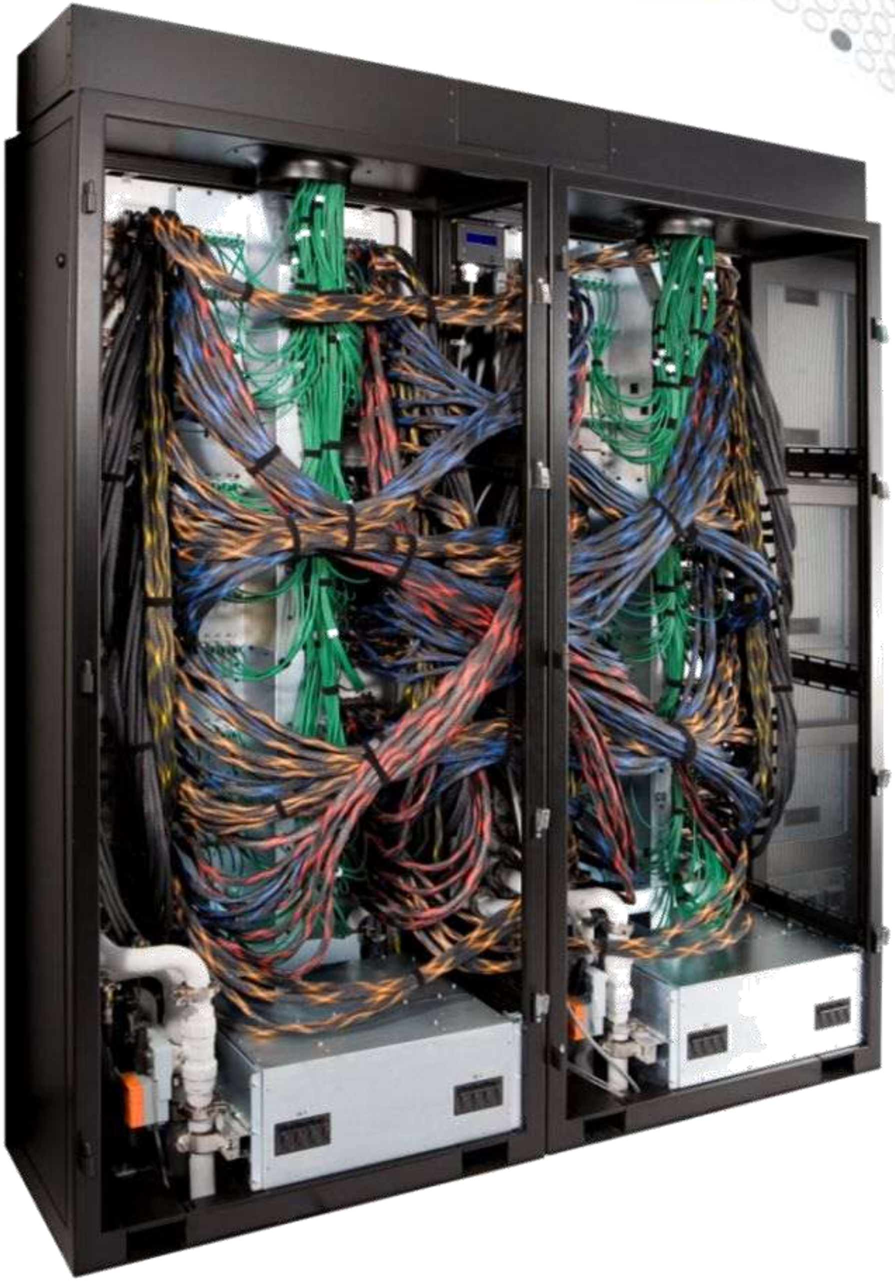
Rank-1
PC Board

Rank-2
Passive CU

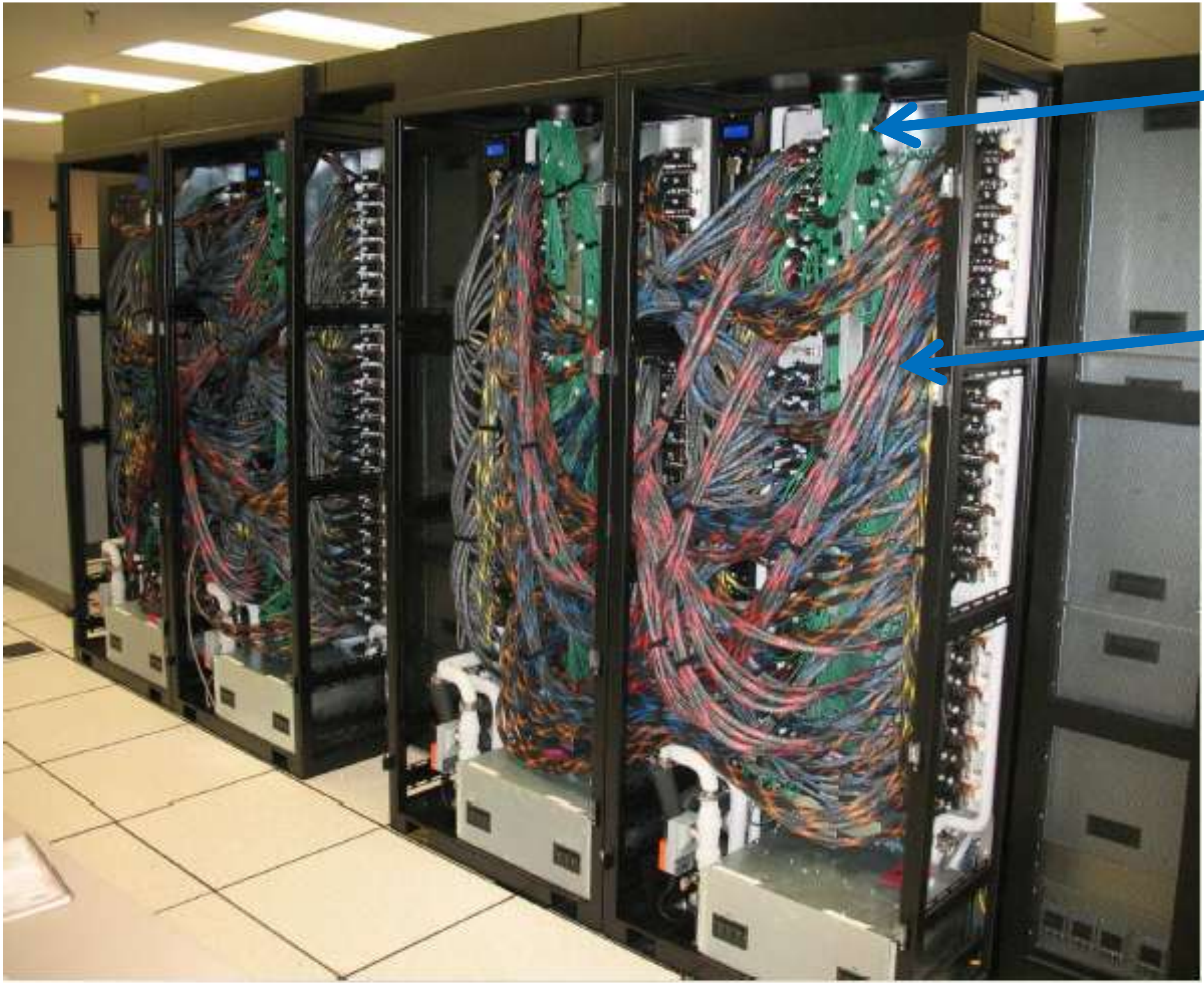
Rank-3
Active Optics

Cray XC30 Rank-2 Cabling

- Cray XC30 two-cabinet group
  - 768 Sockets
  - 96 Aries Chips
- All copper and backplane signals running at 14 Gbps



Copper & Optical Cabling



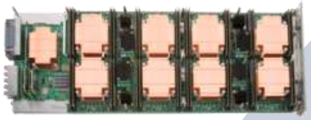
Optical Connections

Copper Connections

Why is the Dragonfly topology a good idea?

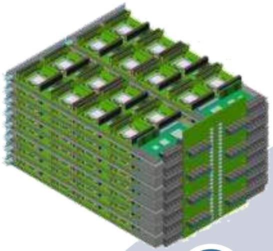
- **Scalability**
  - Topology scales to very large systems
- **Performance**
  - More than just a case of clever wiring, this topology leverages state-of-the-art adaptive routing that Cray developed with Stanford University
  - Smoothly mixes small and large messages eliminating need for a 2nd network for I/O traffic
- **Simplicity**
  - Implemented *without* external switches
  - No HBAs or separate NICs and Routers
- **Price/Performance**
  - Dragonfly maximizes the use of backplanes and passive copper components
  - Dragonfly minimizes the use of active optical components

Cray XC30 – Grouping the nodes



Compute Blade

4 Compute Nodes



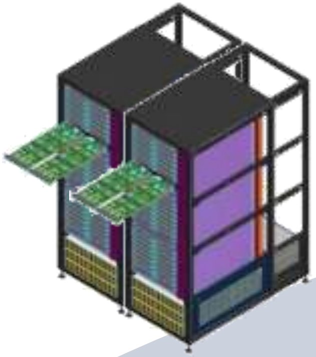
Chassis

16 Compute Blades

64 Compute Nodes

Rank 1 Network

No cables



Group

2 Cabinets

6 Chassis

384 Compute Nodes

Rank 2 Network

Passive Electrical Connections

System

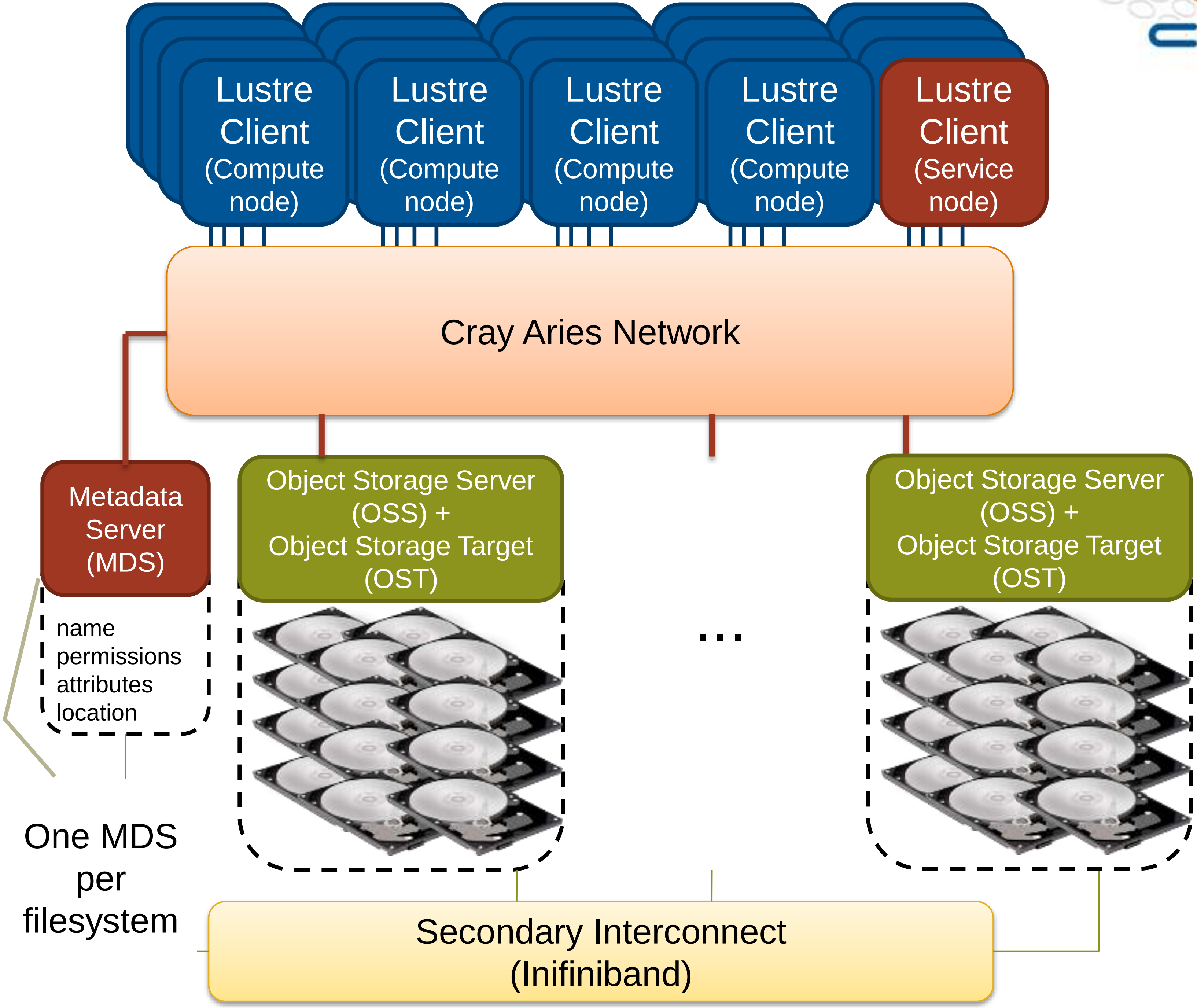
Potentially Hundreds of Cabinets

Up to 10s of thousands of nodes

Rank 3 Network

Active Optical Connections

Storage



Lustre Client (Compute node)

Lustre Client (Compute node)

Lustre Client (Compute node)

Lustre Client (Compute node)

Lustre Client (Service node)

Cray Aries Network

Metadata Server (MDS)

name
permissions
attributes
location

Object Storage Server (OSS) +
Object Storage Target (OST)

…

Object Storage Server (OSS) +
Object Storage Target (OST)

One MDS per filesystem

Secondary Interconnect
(Inifiniband)

Sonexion: Only Three Components

1

MMU: *Metadata Management Unit*

Fully integrated metadata module

- Lustre Metadata software
- Metadata disk storage
- Dual redundant management servers
- Metadata storage target RAID

2

SSU: *Scalable Storage Unit*

Fully integrated storage module

- Storage controller, Lustre server
- Disk controller, RAID engine
- High speed storage
- Provides both capacity and performance

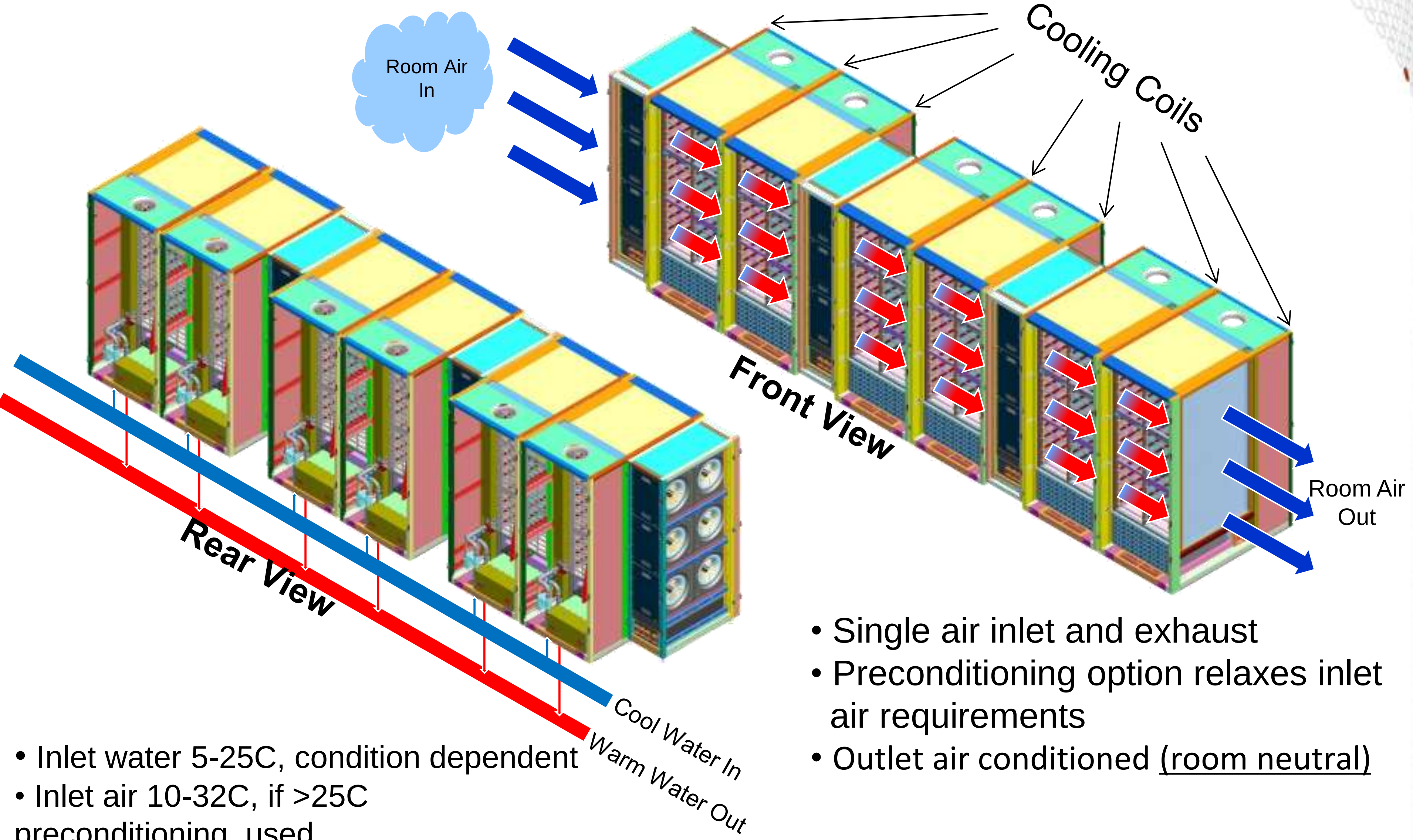
3

Fully prepared rack

- Prewired for InfiniBand, Ethernet and power
- Ready for instant expansion

Power & Cooling

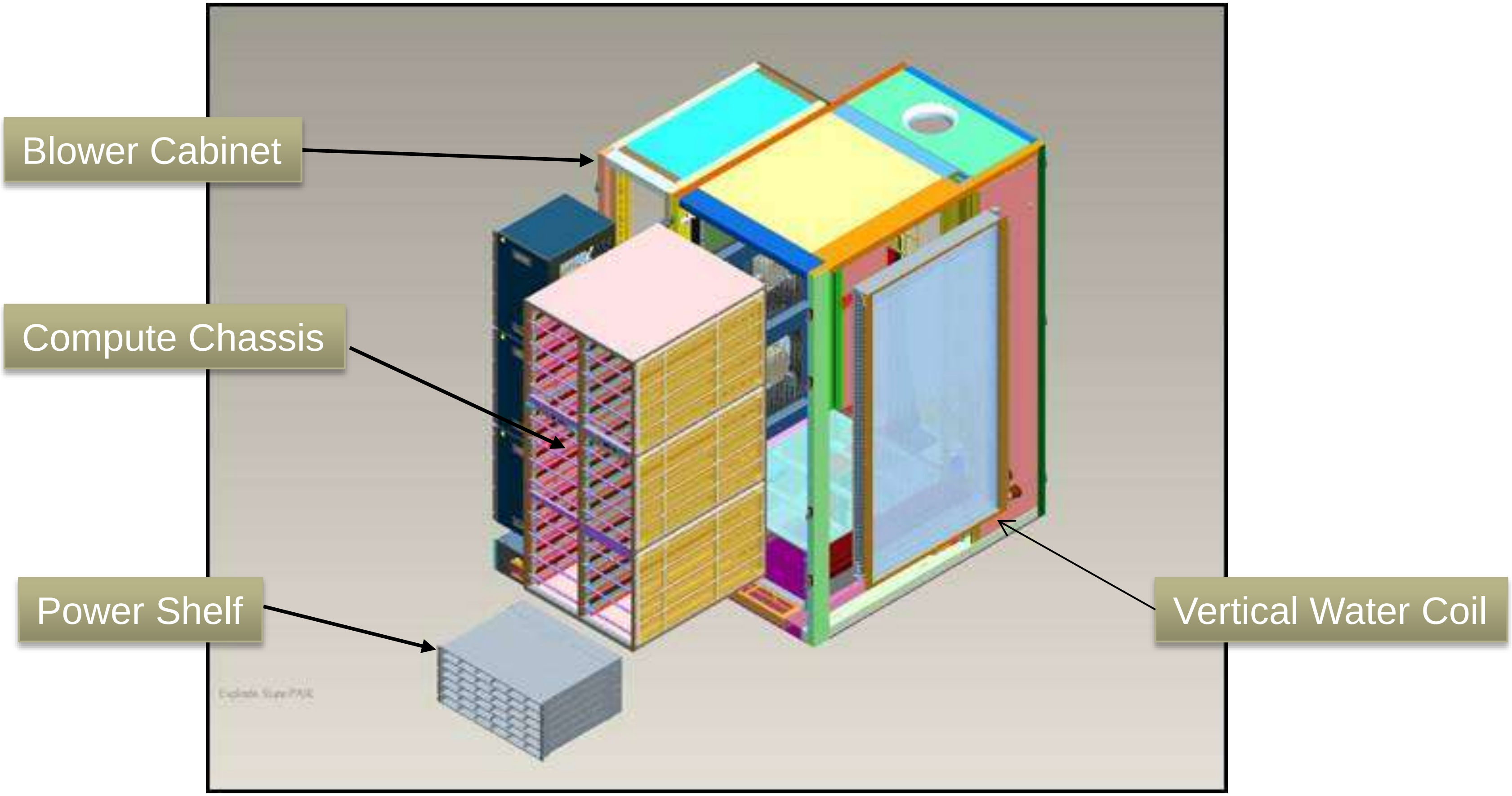
XC30 Cooling Overview



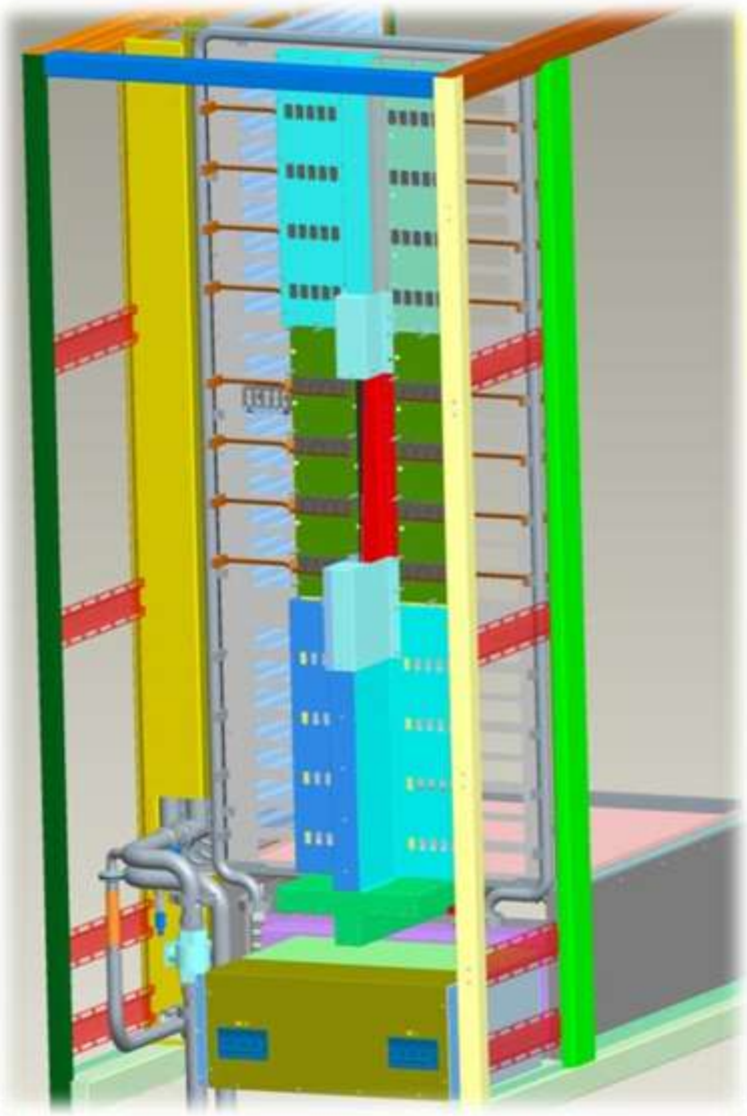
- Single air inlet and exhaust
- Preconditioning option relaxes inlet air requirements
- Outlet air conditioned (room neutral)

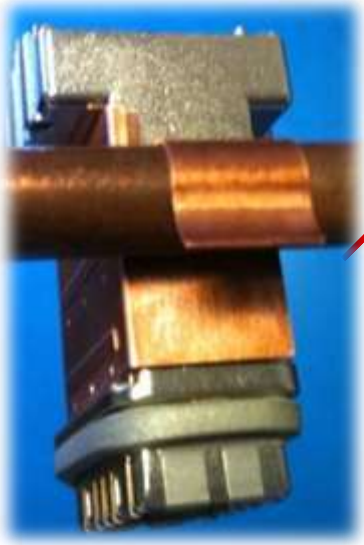
- Inlet water 5-25C, condition dependent
- Inlet air 10-32C, if >25C preconditioning used

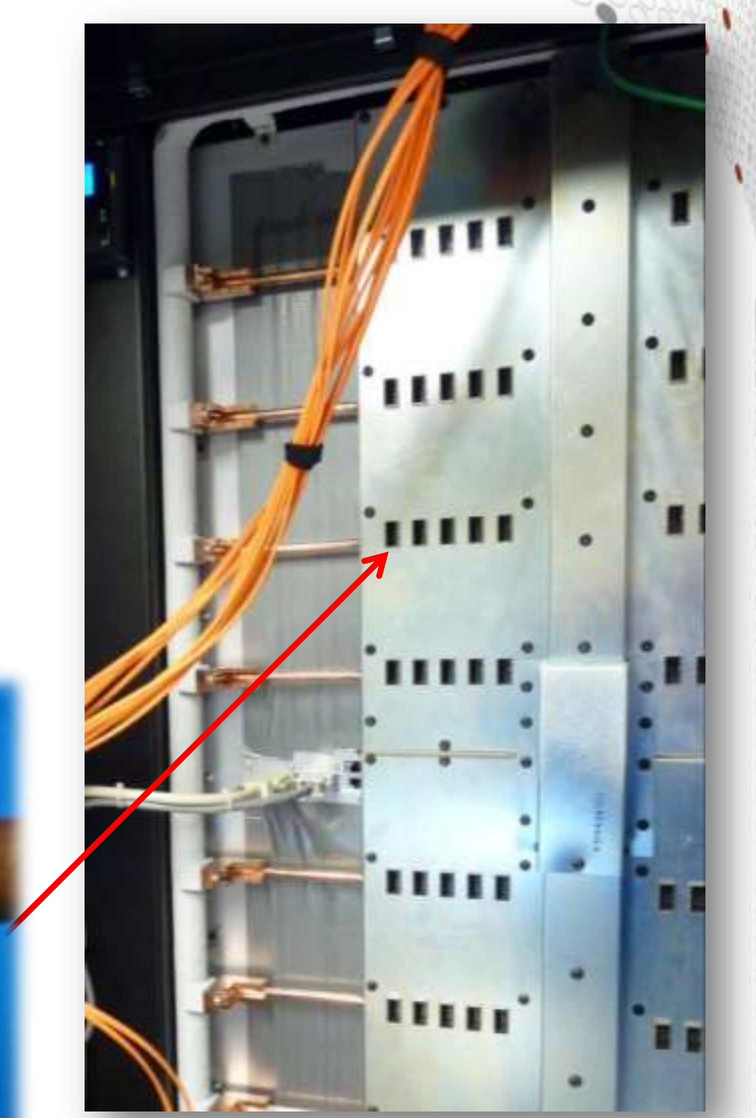
XC30 Cabinet Assembly



Liquid Cooling for Active Optics







Cray XC30 Transverse Cooling Advantages

- **Performance**
  - Transverse cooling and graduated heat sink pitch ensure that all processors operate in the same thermal envelope
  - “Turbo mode” works like it should in a parallel job
- **Simplicity**
  - No airflow issues to manage or adjust
  - System is 100% water-cooled
  - No pumps, refrigerant, treated water, or plumbing on the blades
- **Cost of Ownership**
  - Excellent PUE characteristics
  - 25% better density than other ‘direct’ water cooled solutions
  - All cooling infrastructure is retained across multiple generations of computing technology
- **Maintainability**
  - Blades can be warm-swapped without disturbing any plumbing
  - Blowers can be hot-swapped if required and can provide N+1 redundancy

Power Monitoring and Management

- **Monitoring power is the key first step**
- **Control power at the system level, at the job level, and at the application level**
- **Cray XC30 GA (Mar 2013) will contain a base set of features for power monitoring and system power capping**
  - GUI and CLI to manage system-wide power use
  - Run more nodes within fixed power budget
  - Adapt to site power/cooling events
- **Additional features in later Cray XC30 releases**
  - Ability to charge based on power (or reserved power), not just node-hours
  - Lower power consumption when idle
  - Manage system power consumption at the job level

ARCHER's Nodes

ARCHER hardware on site today has the following:

- **16 Cabinets = 8 Groups**
- **3008 Compute Nodes**
  - Dual socket 12 core Intel® Xeon Ivybridge @2.7GHz
    - 2632 x 64 GB 1866MHz Memory
    - 376 x128GB 1866MHz Memory (1 group)
- **32 Service Nodes**
- **8 Cray Development Logins**
  - 256 GB Memory available
- **2 Pre/Post Processing Servers**
  - 1TB Memory per server
- **20 Sonexion SSUs**
  - 160 Lustre Object Storage Targets (distributed over multiple filesystems)
  - 4.34 PB of storage (distributed over multiple filesystems)

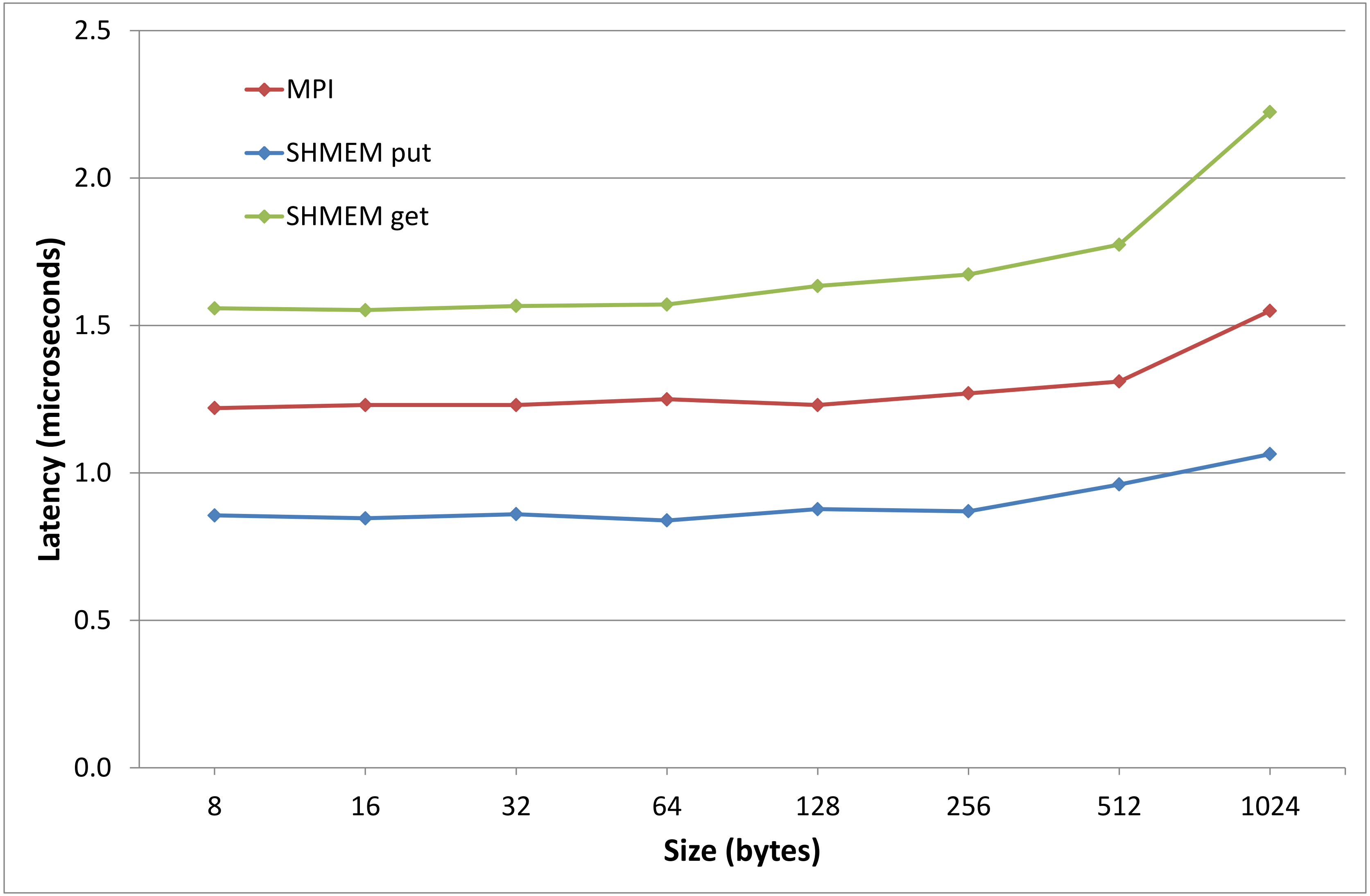
XC30 Performance Data

Performance comparison: XE-IL vs XC30

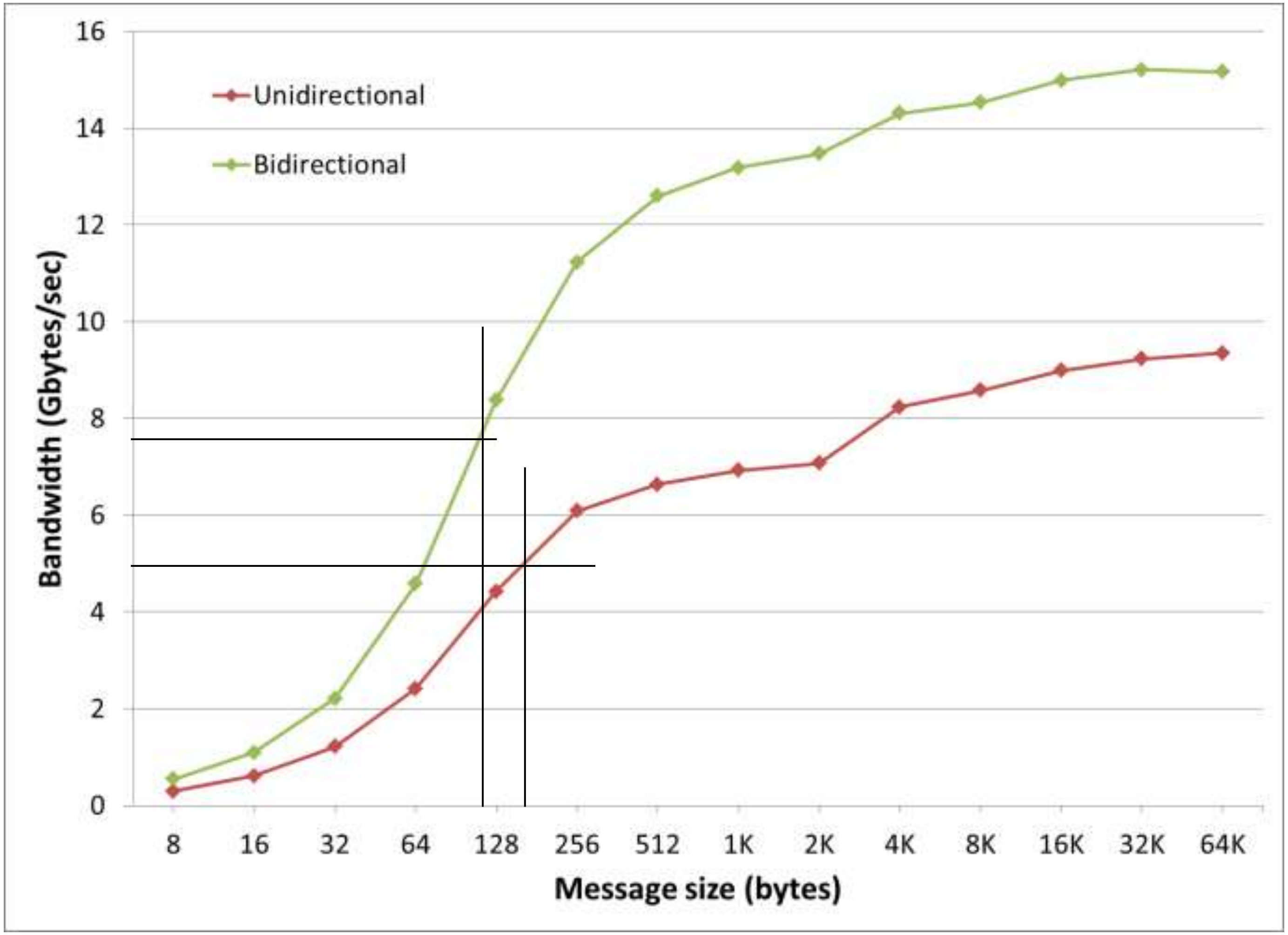
- Typical XE-IL vs. XC-Sandybridge for 256-512 compute nodes
- Actually results will depend on system size and configurations and problem size and definition

Tests (Units)	XE-Interlagos	XC	XC/XE
HPL (Tflops)	~81%	~86%	106%
Star DGEMM (Gflops)	~87%	~102%	117%
STREAMs (Gbytes/s/node)	72	78	108%
RandomRing (Gbytes/s/rank)	~0.055	~0.141	256%
Point-to-Point BW (Gbytes/s)	2.8-5.6	>8.5	157% - 314%
Nearest Node Point-to-Point Latency (usec)	1.6-2.0	<1.4	116% - 145%
GUPs	2.66	15.6	525%
GFFT (Gflops)	628	2221	354%
HAMR Sort (GiElements/sec)	9.4	36.6	390%

XC Network Performance – Latency

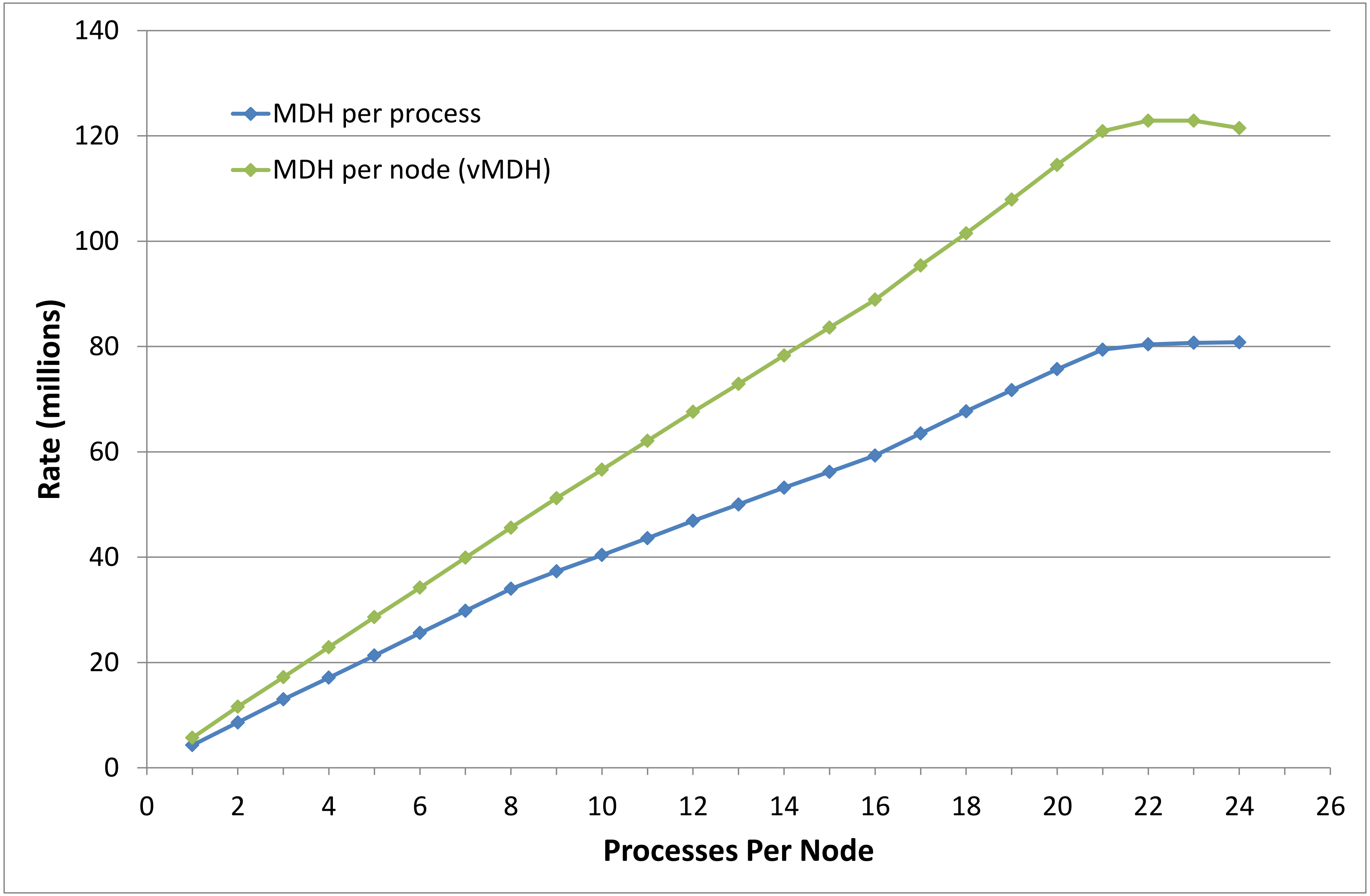


XC Network Performance – Put Bandwidth

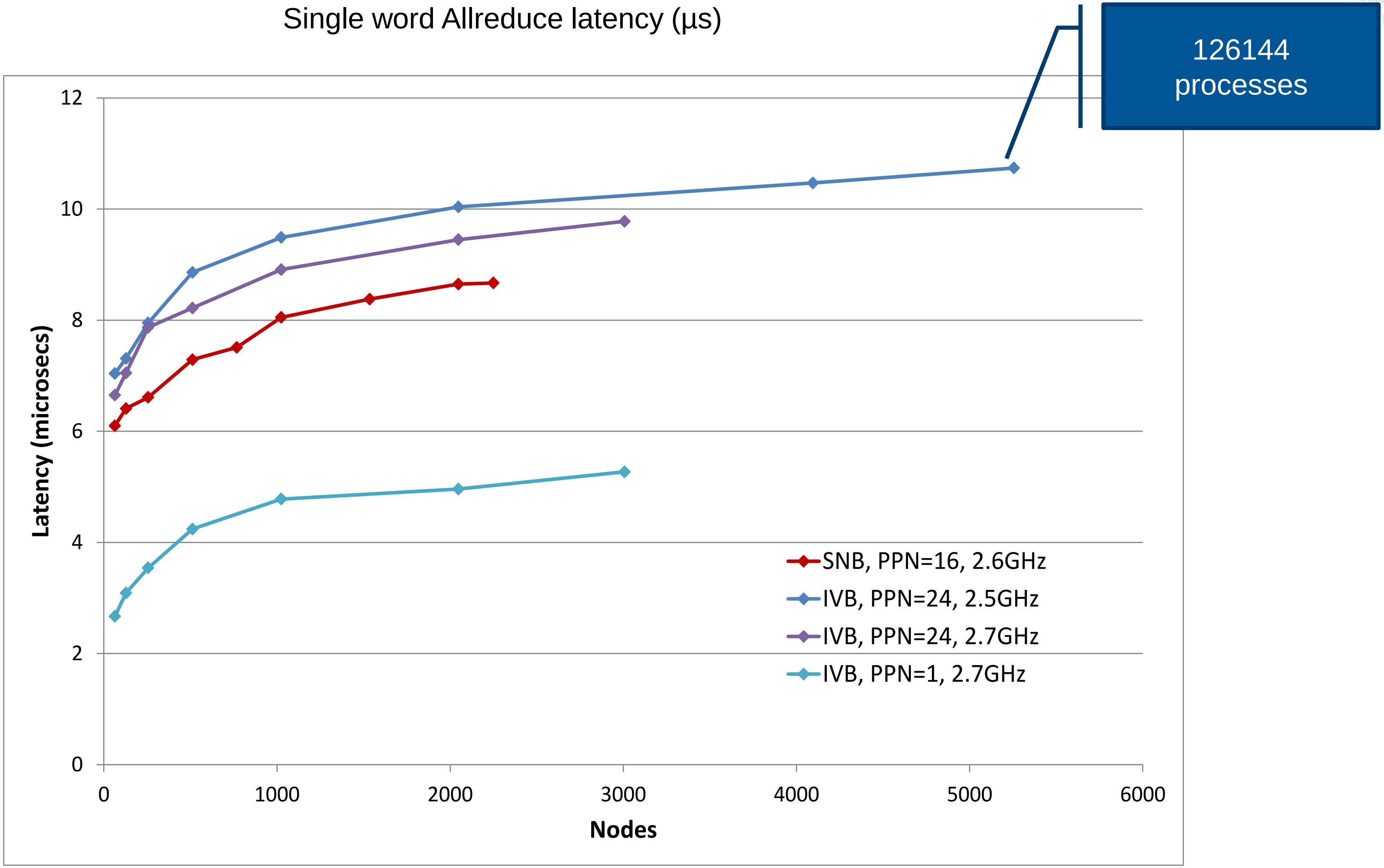


Aries point to point Put bandwidth as a function of transfer size measured on Cascade prototype hardware and software

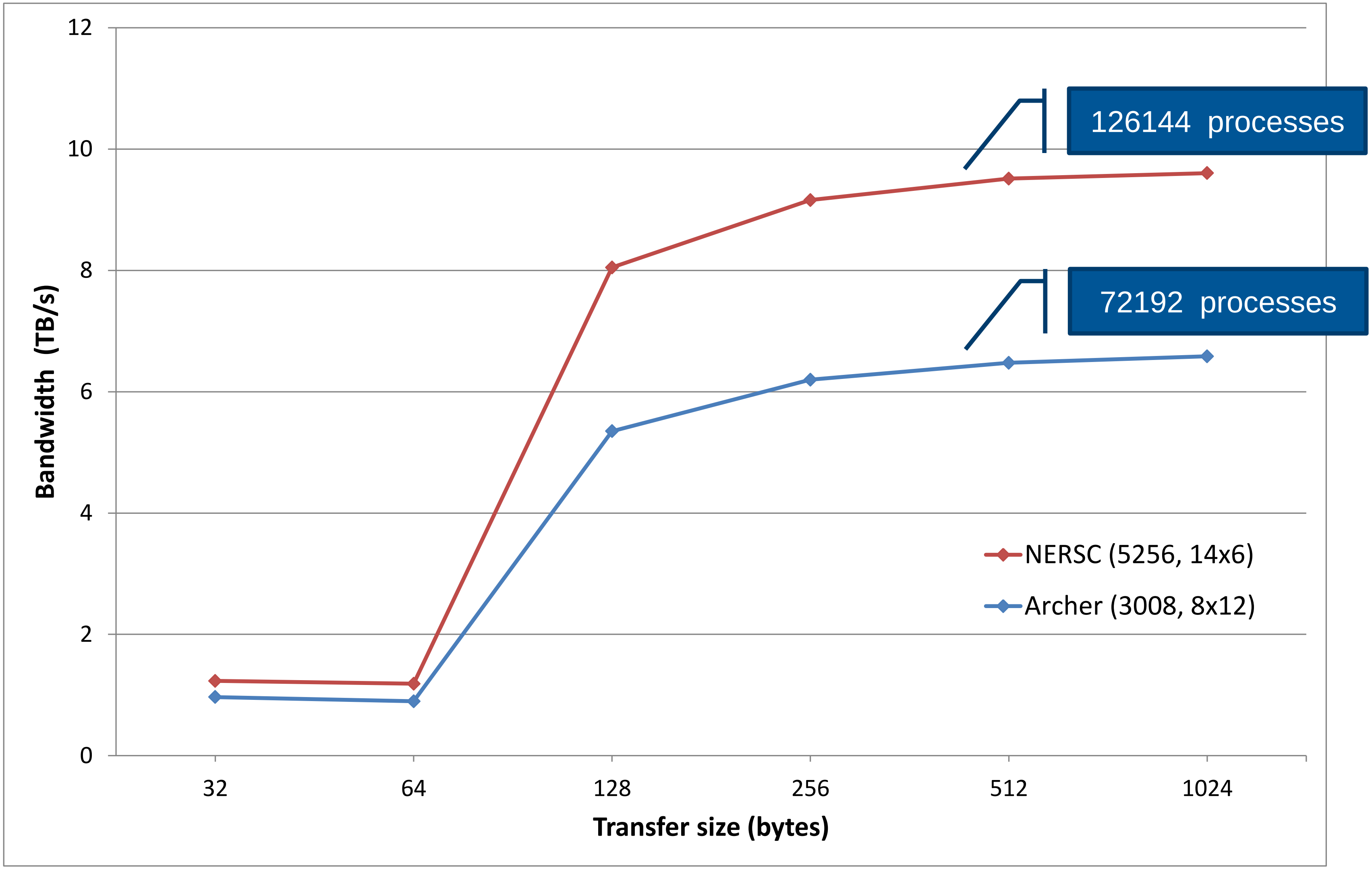
XC Network Performance – Put Rate



Collective offload

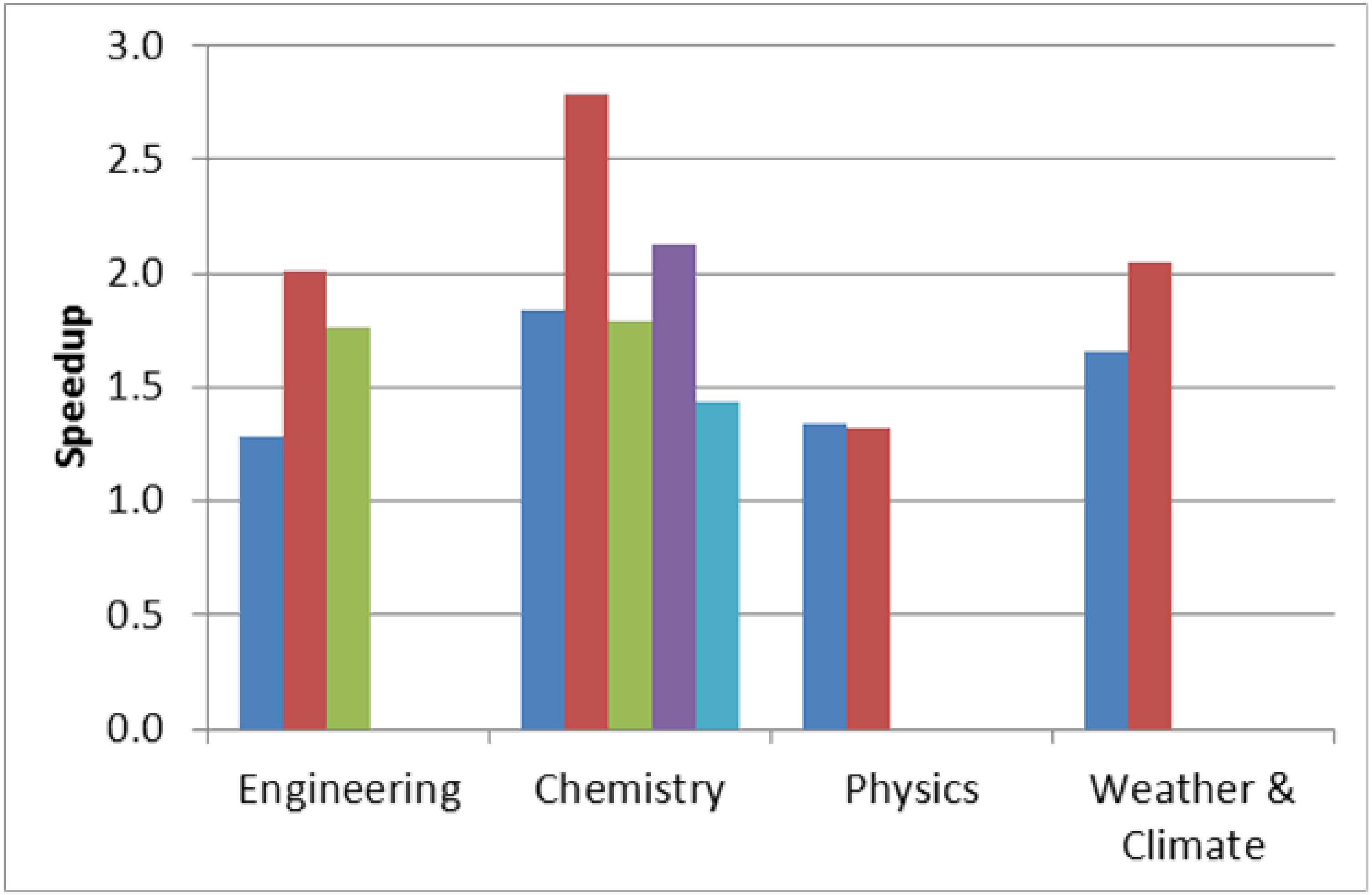


All-to-all bandwidth (TB/s)

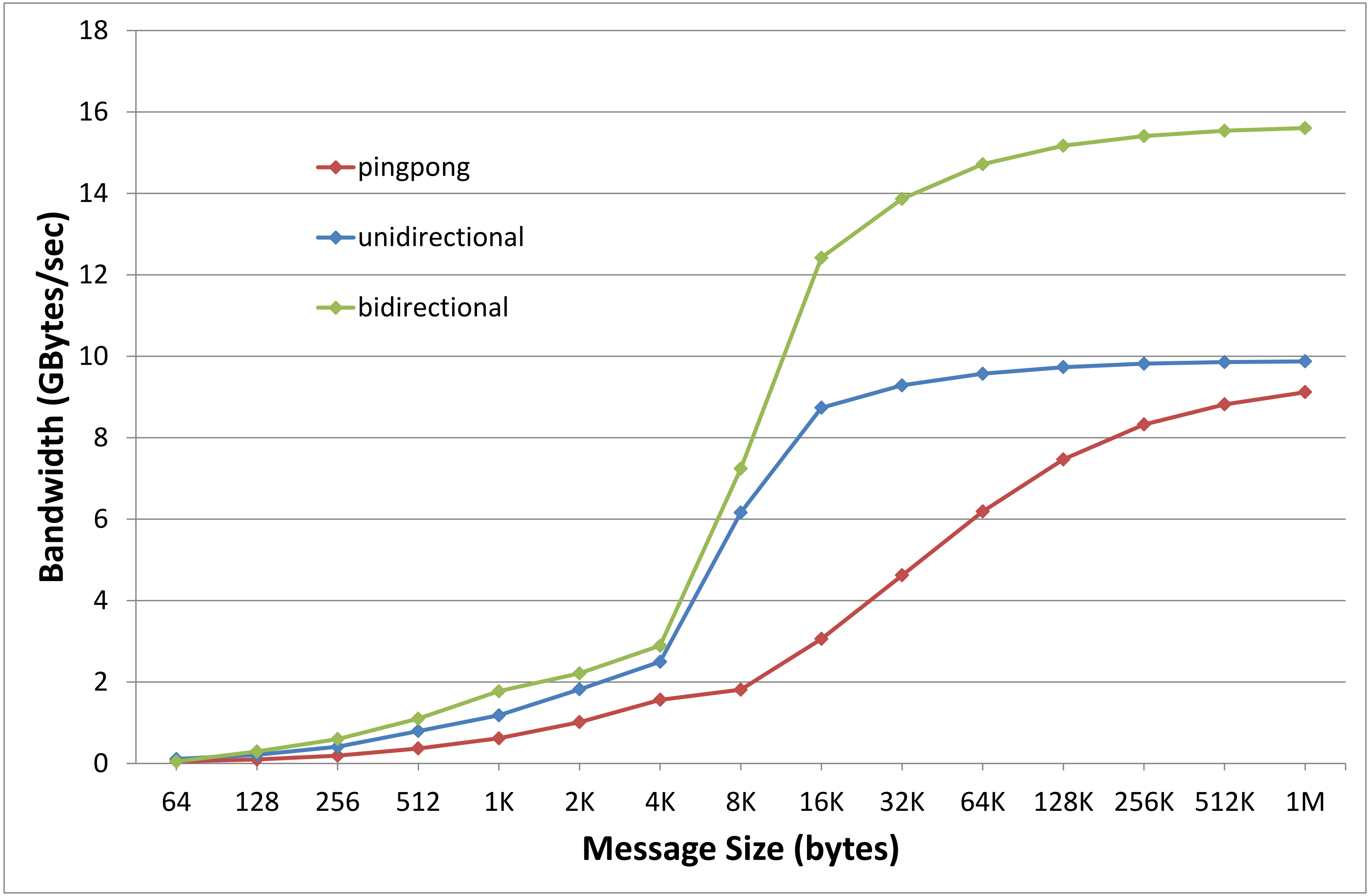


Application Codes

- Cascade Sandy Bridge per-node performance relative to XE6 with Interlagos



MPI Bandwidths



Further Information

- **SC12 paper**

**Cray Cascade - A Scalable HPC System
Based on a Dragonfly Network**

**AUTHOR(S):Gregory Faanes, Abdulla Bataineh,
Duncan Roweth, Tom Court,
Edwin Froese, Bob Alverson,
Tim Johnson, Joe Kopnick,
Michael Higgins, James Reinhard**

SC12

Salt Lake City, Utah

Conference Dates:
November 10-16, 2012

Exhibition Dates:
November 12-15, 2012

The International Conference for
High Performance Computing,
Networking, Storage and Analysis

SC12 Home > SC12 Schedule > SC12 Presentation - Cray Cascade - A Scalable HPC System Based on a Dragonfly Network

SCHEDULE: NOV 10-16, 2012

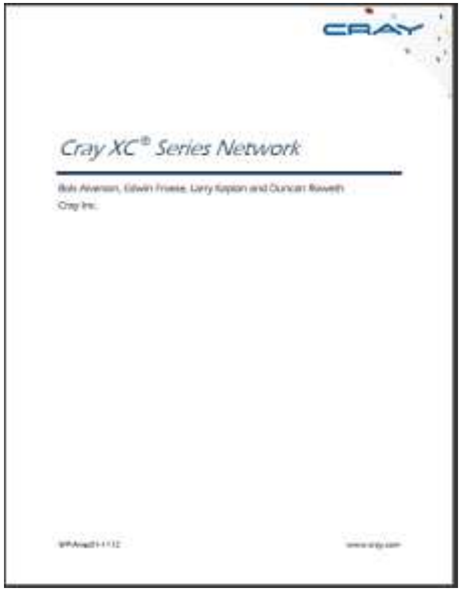
When viewing the Technical Program schedule, on the far righthand side is a column labeled "PLANNER." Use this planner to build your own schedule. Once you select an event and want to add it to your personal schedule, just click on the calendar icon of your choice (outlook calendar, ical calendar or google calendar) and that event will be stored there. As you select events in this manner, you will have your own schedule to guide you through the week.

Cray Cascade - A Scalable HPC System Based on a Dragonfly Network

SESSION: New Computer Systems

- **Cascade networking whitepaper**

http://www.cray.com/Products/XC



Cray XC® Series Network

Bob Alverson, Edwin Froese, Larry Kaplon and Duncan Roweth
Cray Inc.

WP-Aries01-1112

www.cray.com