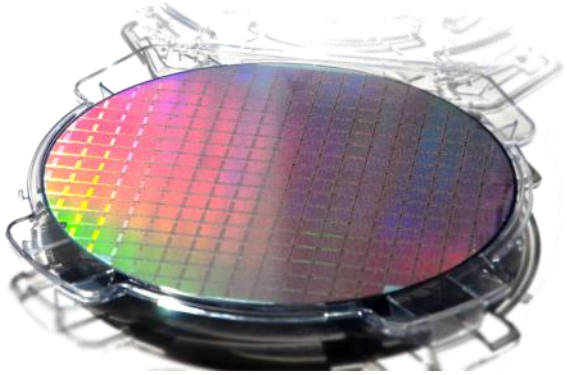
Cray XC30 System

Cray Aries Custom Interconnect

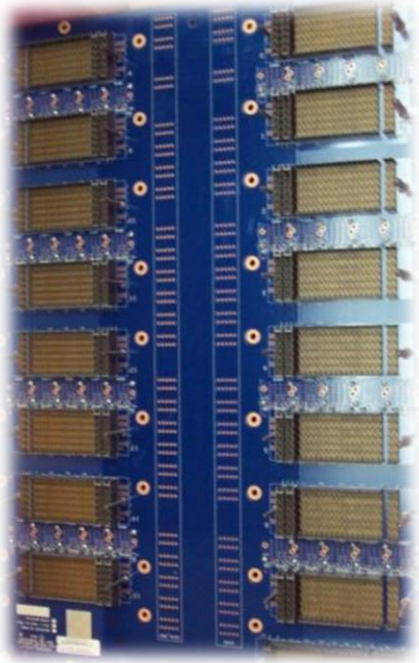


Aries vs. Gemini

Feature	Benefit
3X increase in sustained injection bandwidth to 10 GB/sec	Improves the efficiency of a wide range of communication intensive applications
Large increase in global bandwidth, from 3X at the low end to 20X or more at the high end	Benefits applications with complex communication patterns (unstructured meshes, adaptive mesh refinement, search and sort, data mining etc.)
Hardware support for collectives	Especially helpful on large jobs (1000s of cores) which utilize collective operations. HW support is about a 2X improvement over our best SW algorithms
3X increase in the rate of small puts and gets to 120M/sec	Key for new programming environments (PGAS), important to our DARPA mission partners and some of our more advances customers
PCI-Express Gen3 interface	Benefits Cray, reduces our dependency on a particular CPU vendor

Cray XC30 Network

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

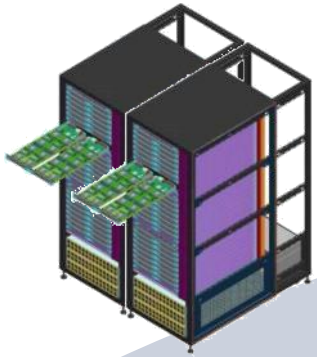


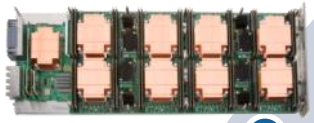
Rank-1
PC Board: ¢¢¢

Rank-2
Passive CU: $

Rank-3
Active Optics: $$$$

Cray XC30 System Building Blocks





Compute Blade

4 Compute Nodes

Chassis

Rank 1 Network

16 Compute Blades

No Cables

64 Compute Nodes

Group

Rank 2 Network

Passive Electrical Network

2 Cabinets

6 Chassis

384 Compute Nodes

System

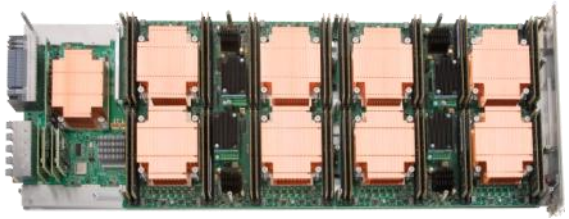
Rank 3 Network

Active Optical Network

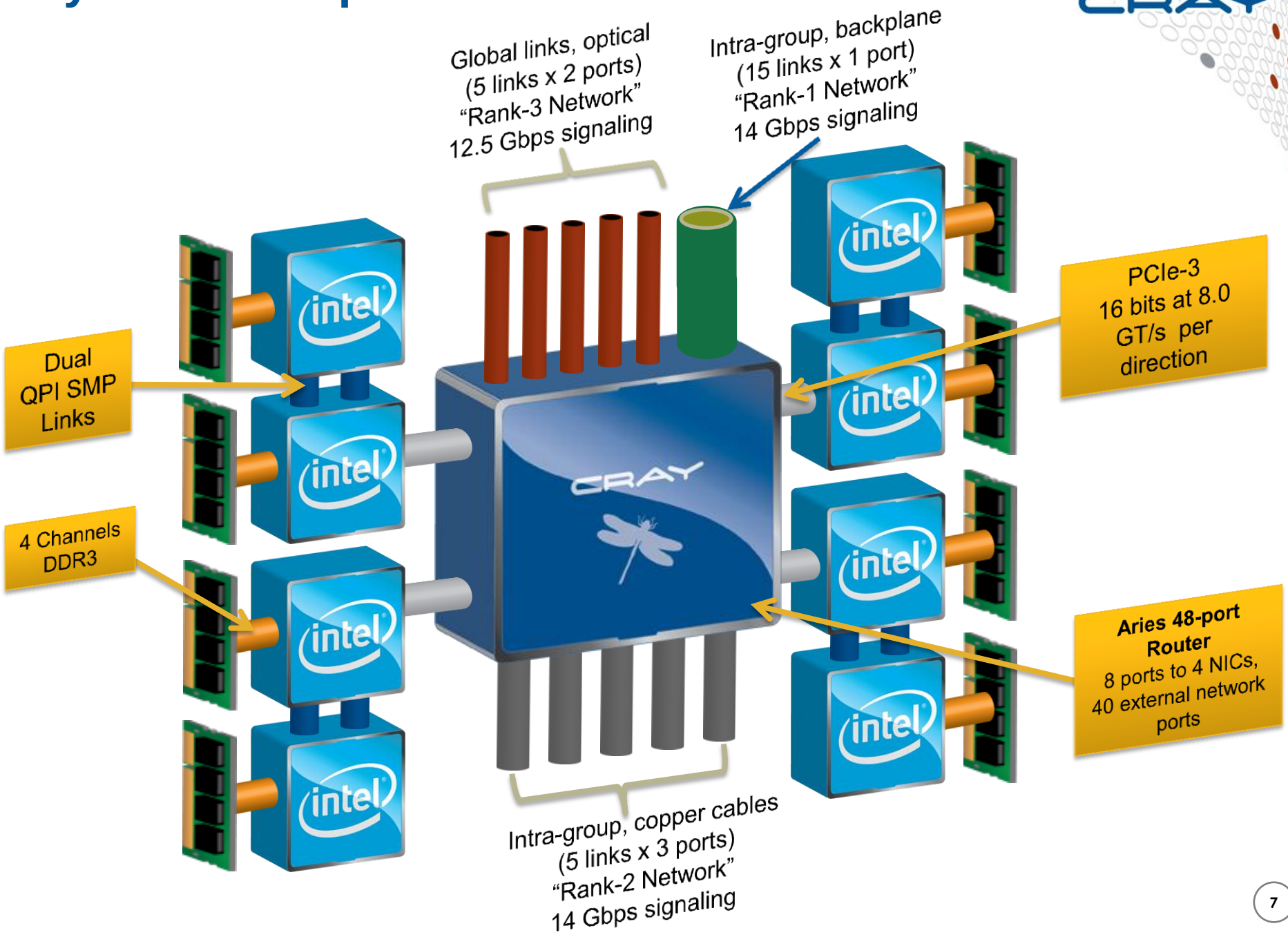
Hundreds of Cabinets

Up to 10s of thousands of nodes

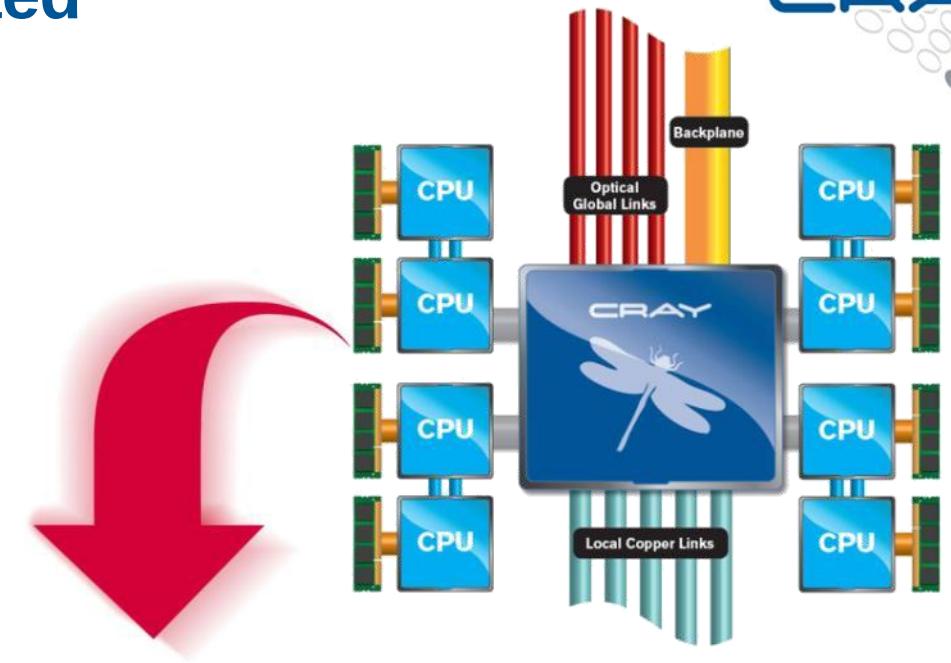
Cray XC30 Modular Blades

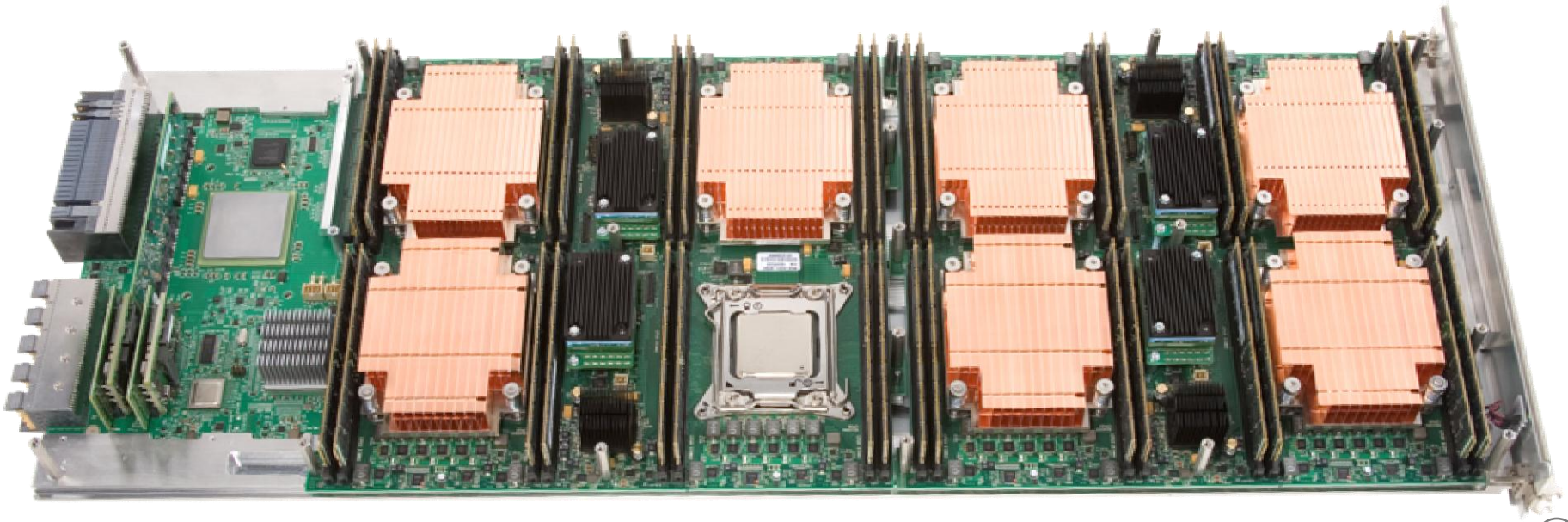


Cray XC30 Compute Blade Architecture

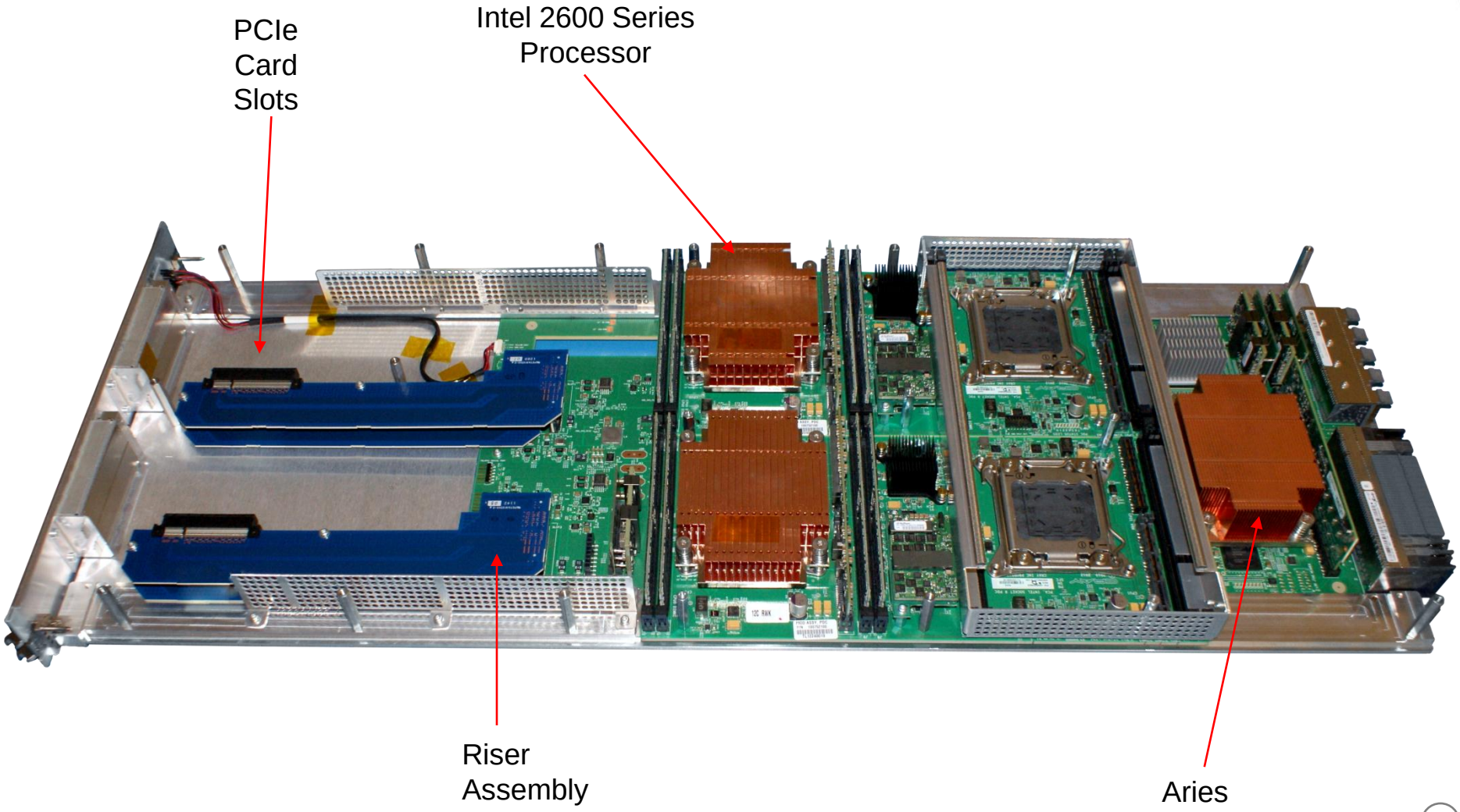


Cray XC30 Fully Populated Compute Blade



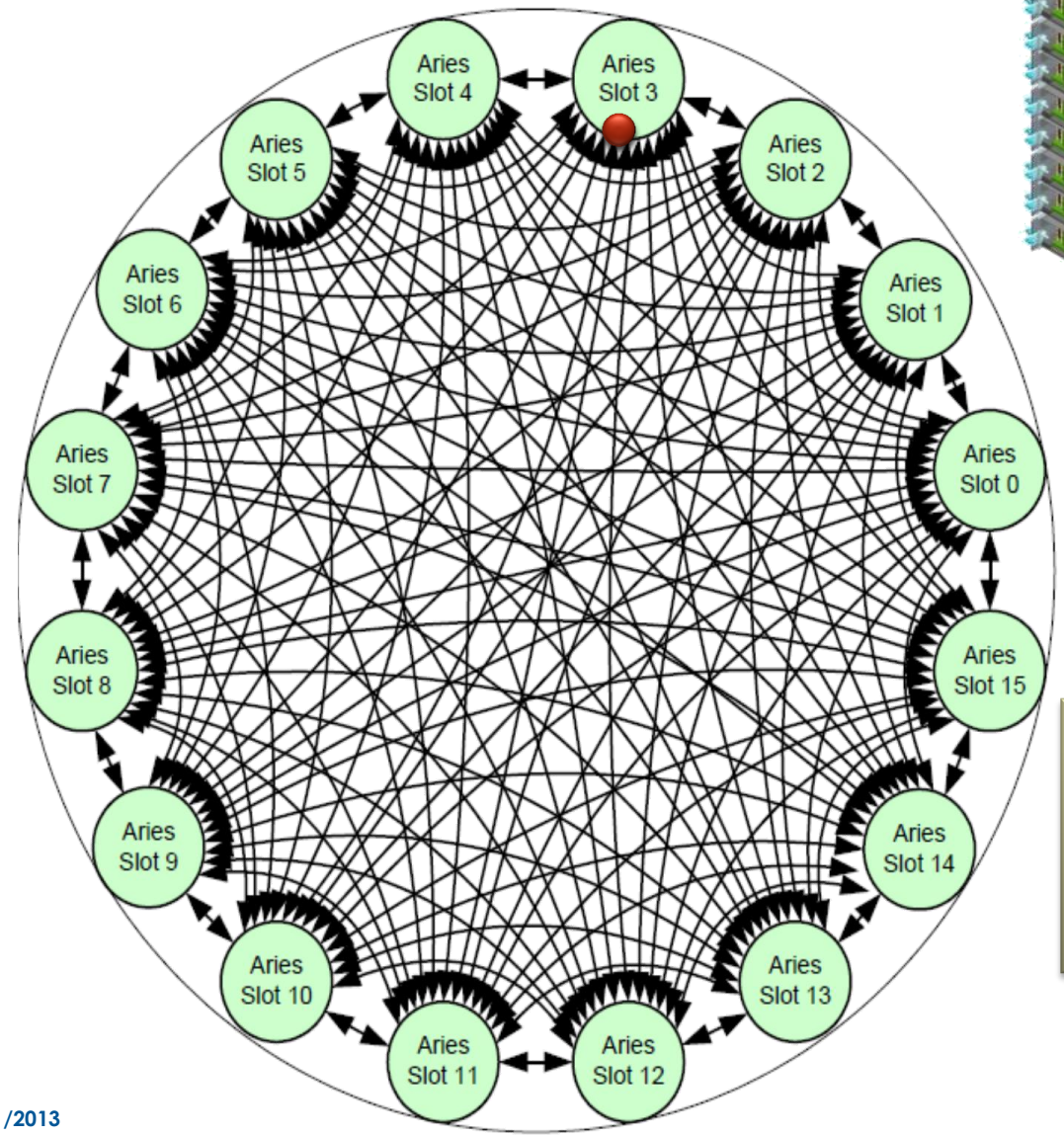


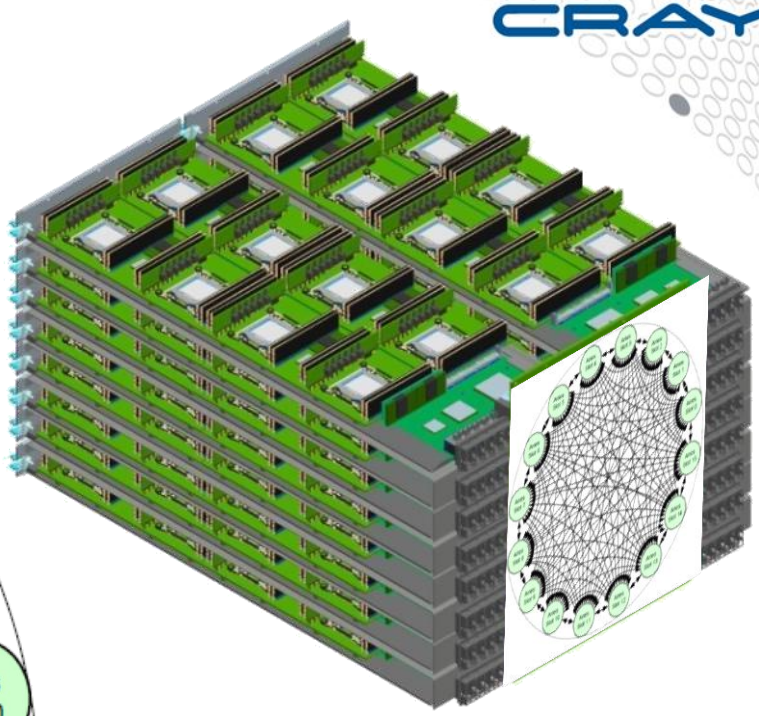
Cray XC I/O Module



Cray XC30 Dragonfly Topology

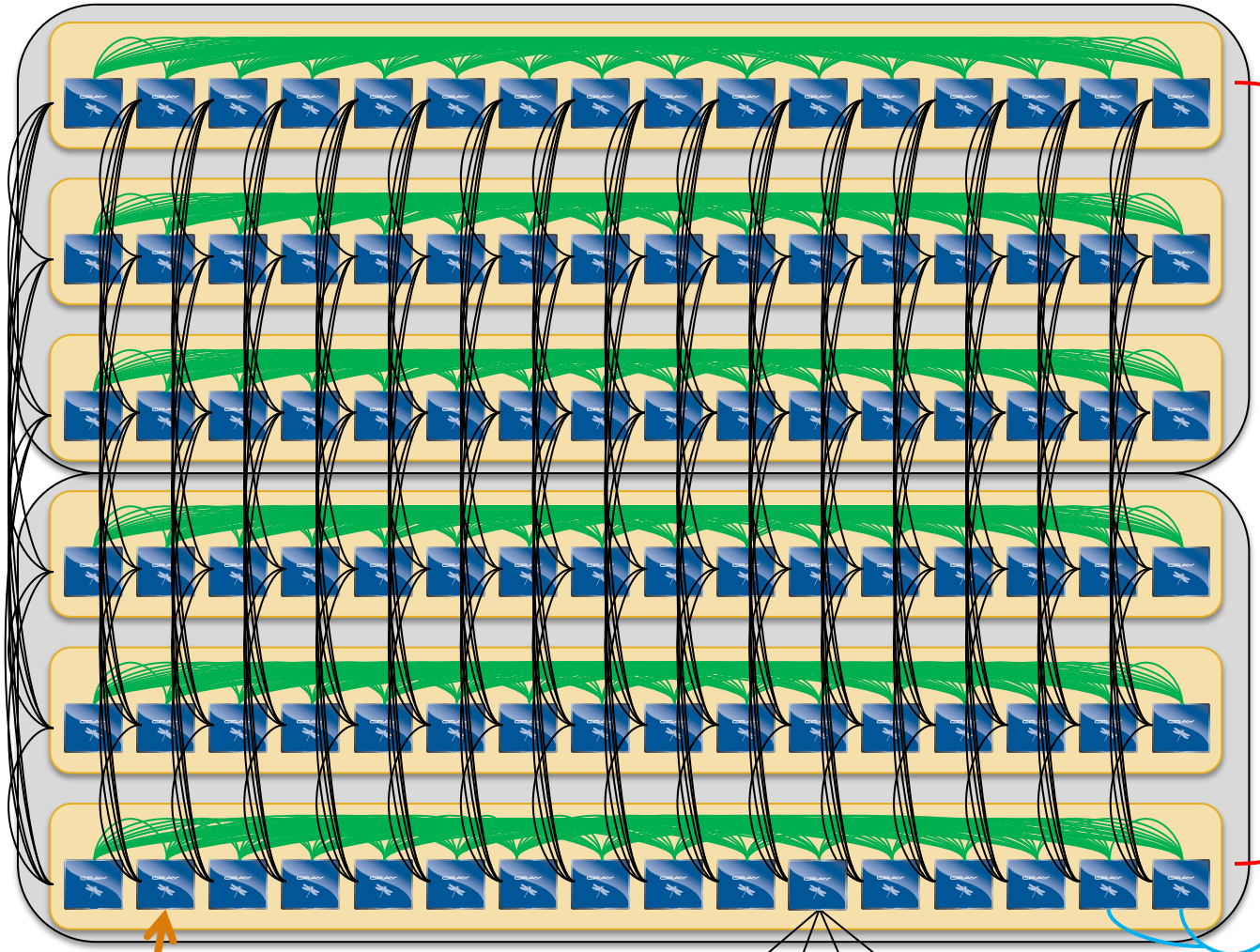
Cray XC30 Rank1 Network





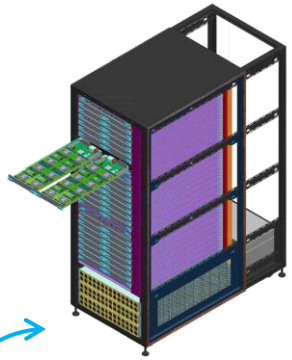
- Chassis with 16 compute blades
- 128 Sockets
- Inter-Aries communication over backplane
- Per-Packet adaptive Routing

Cascade – Local Electrical Network



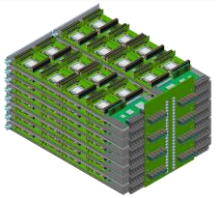
2 Cabinet Group
768 Sockets

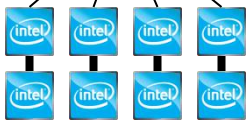
6 backplanes connected with copper cables in a 2-cabinet group: "Rank-2 Network"



Active optical cables interconnect groups "Rank-3 Network"

16 Aries connected by backplane "Green Network"



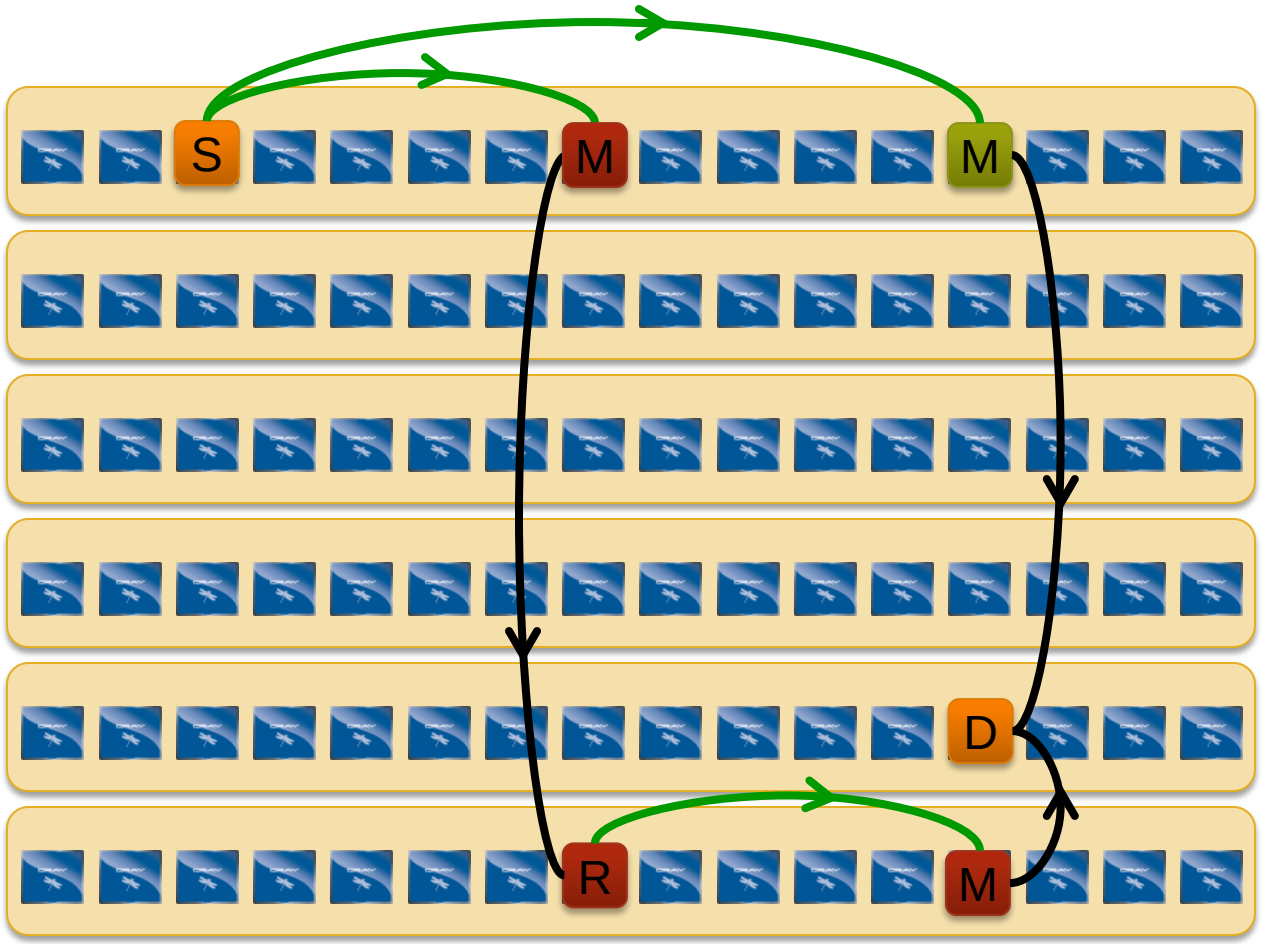


4 nodes connect to a single Aries

Cray XC30 Rank-2 Cabling

- Cray XC30 two-cabinet group
  - 768 Sockets
  - 96 Aries Chips

Cray XC30 Routing



Minimal route between any two nodes in a group is just two hops

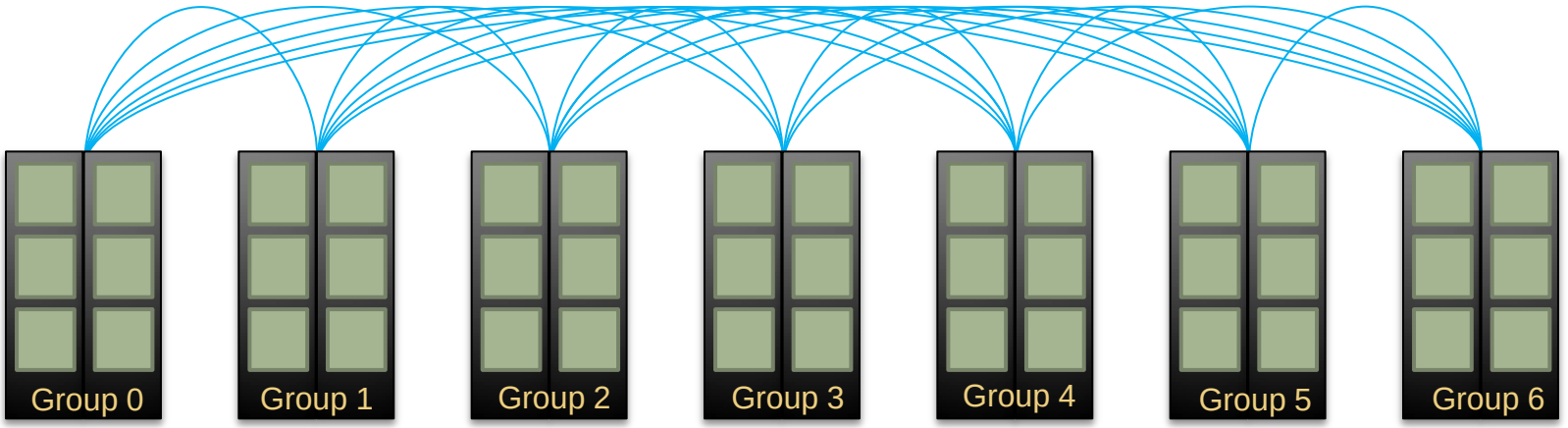
Non-minimal route requires up to four hops.

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 – Rank-3 Network

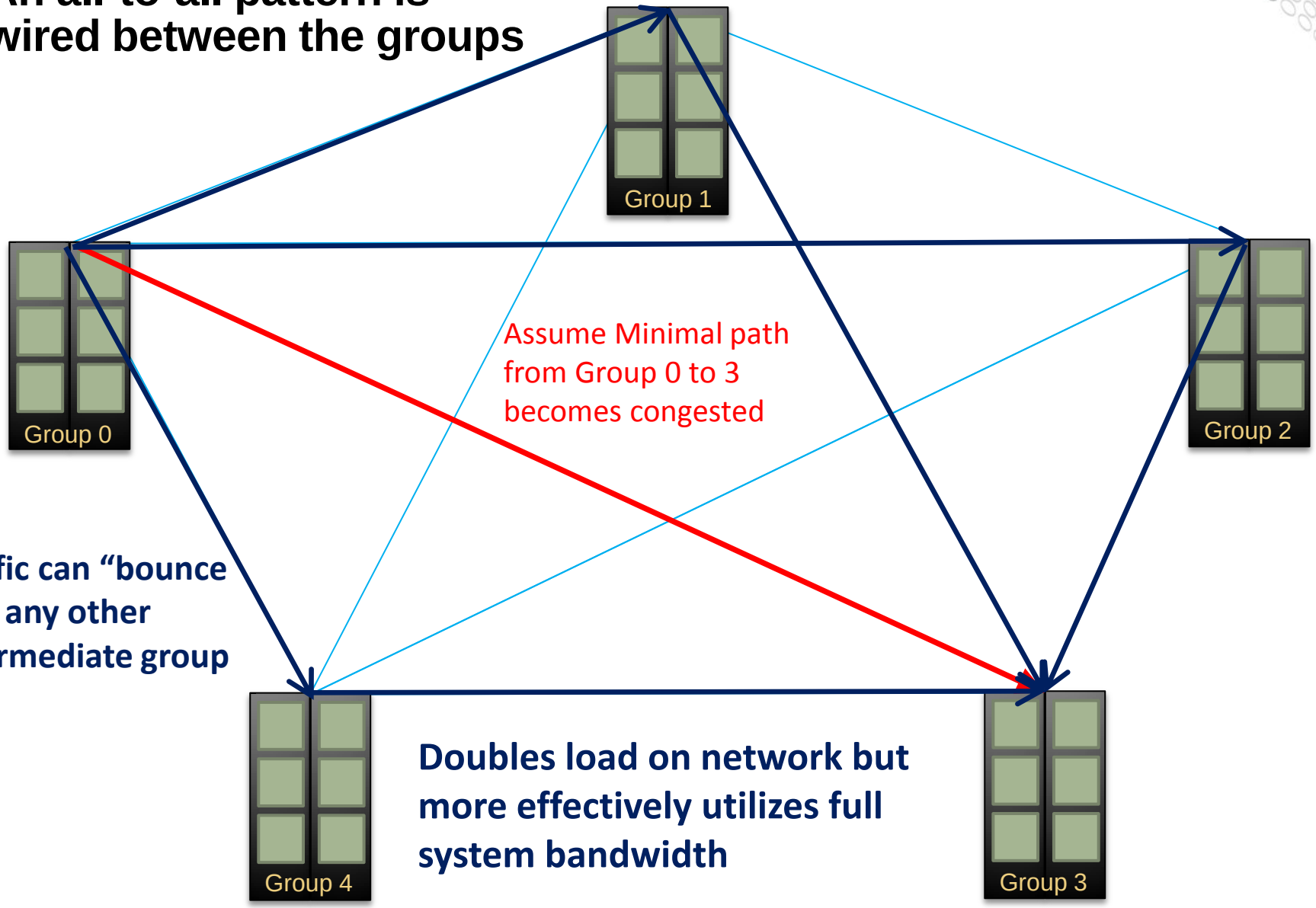
- An all-to-all pattern is wired between the groups using optical cables (blue network)
- The global bandwidth can be tuned by varying the number of optical cables in the group-to-group connections



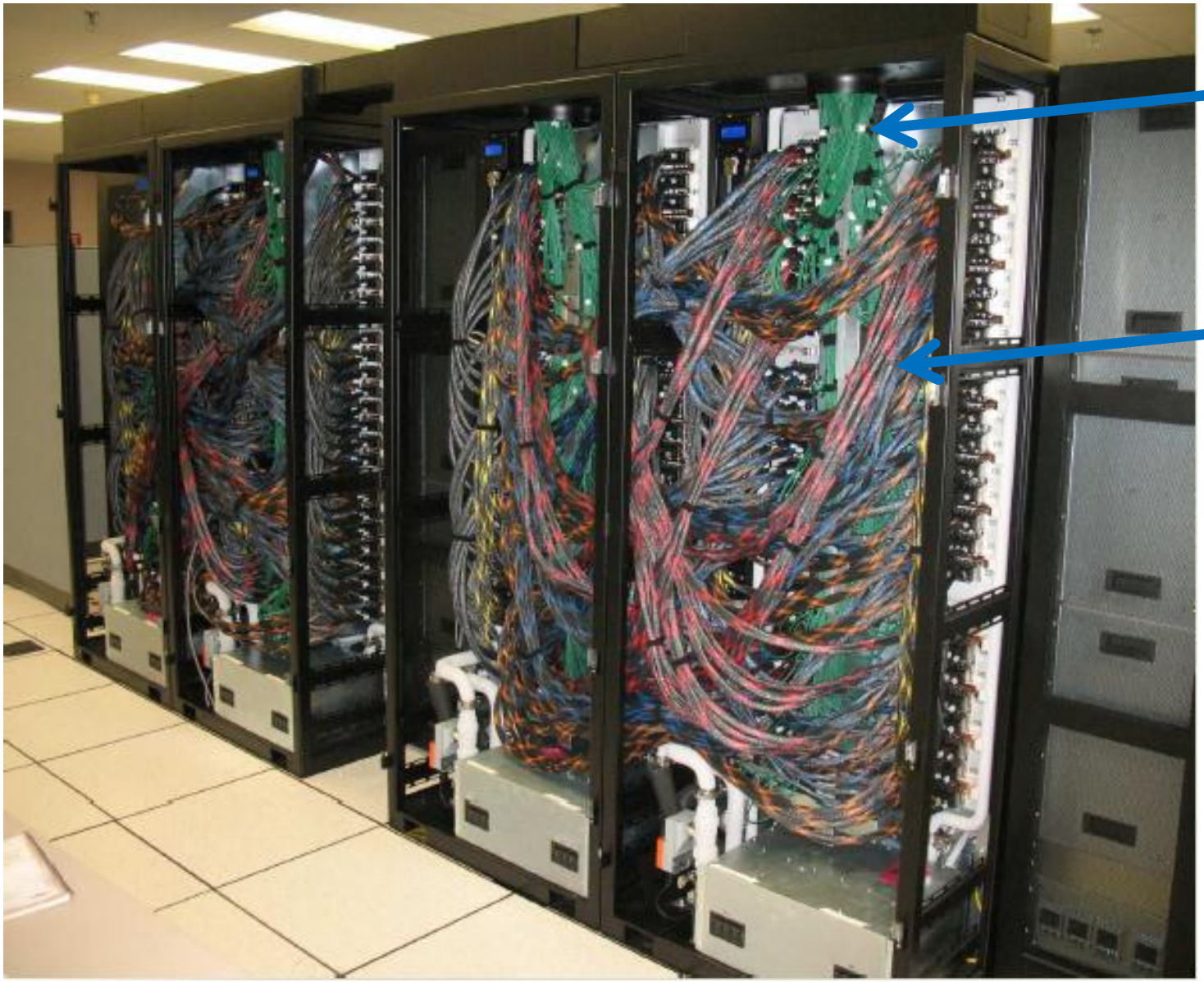
Example: A 7-group system is interconnected with 21 optical "bundles". The "bundles" can be configured between 2 or more cables wide, subject to the group limit.

Adaptive Routing over the Blue Network

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



Copper & Optical Cabling



Cray Software

Cray Programming Environment Distribution Focus on Performance and Productivity

Programming Languages

- Fortran
- C
- C++
- Python

Programming models

- Distributed Memory (Cray MPT)
  - MPI
  - SHMEM
- Shared Memory
  - OpenMP 3.0
  - OpenACC
- PGAS & Global View
  - UPC (CCE)
  - CAF (CCE)
  - Chapel

Compilers

- Cray Compiling Environment (CCE)
- GNU
- 3rd Party Compilers

Tools

- Environment setup
  - Modules
- Debuggers
  - TotalView
  - DDT
  - lgdb
- Debugging Support Tools
  - Abnormal Termination Processing
  - STAT
- Performance Analysis
  - CrayPat
  - Cray Apprentice²
- Scoping Analysis
  - Reveal

Optimized Scientific Libraries

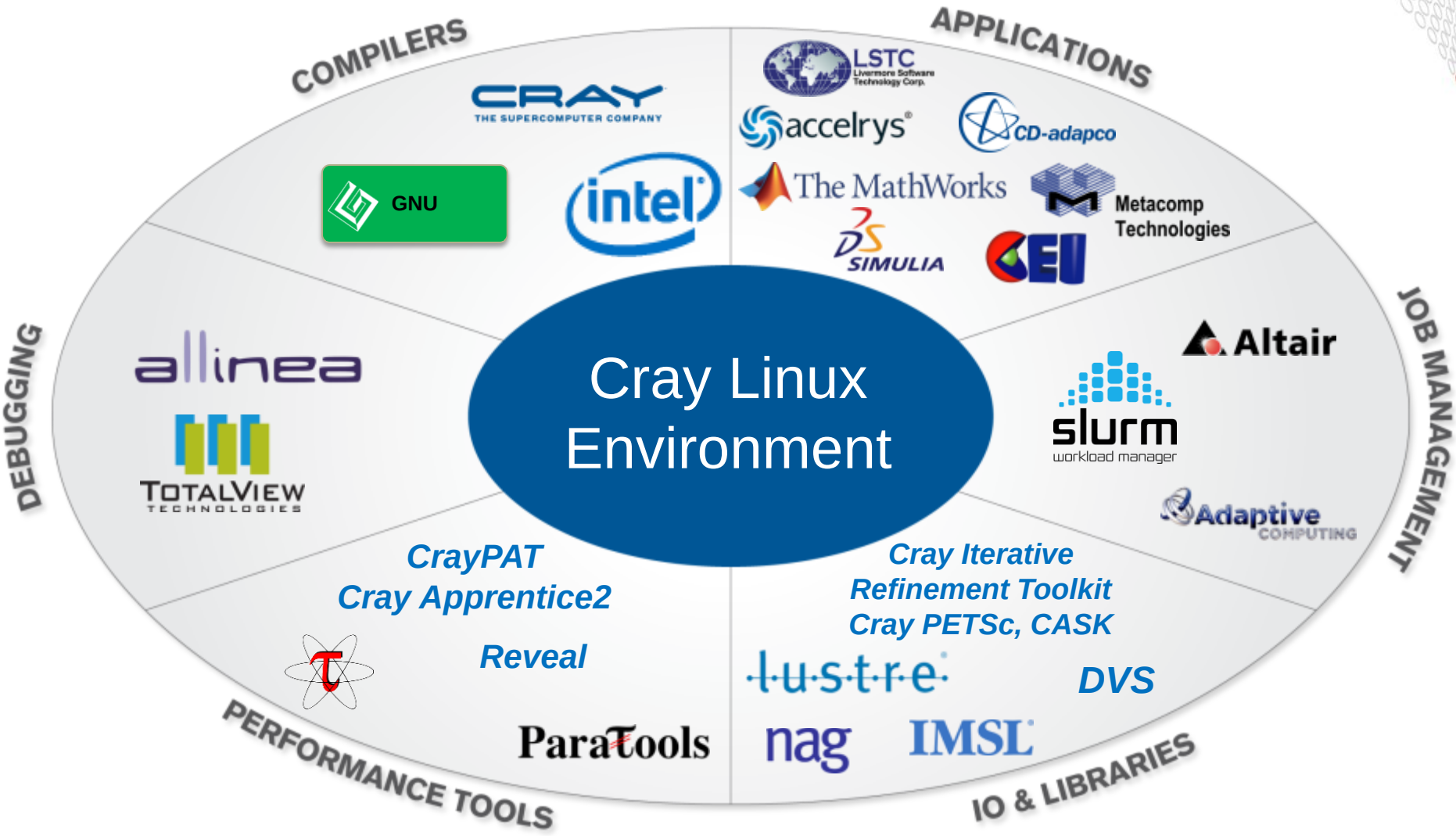
- LAPACK
- ScaLAPACK
- BLAS (libgoto)
- Iterative Refinement Toolkit
- Cray Adaptive FFTs (CRAFFT)
- FFTW
- Cray PETSc (with CASK)
- Cray Trilinos (with CASK)

I/O Libraries

- NetCDF
- HDF5

Cray developed
Licensed ISV SW
3rd party packaging
Cray added value to 3rd party

Cray Software Ecosystem



An Adaptive Linux OS optimized specifically for HPC

CLE
CRAY LINUX ENVIRONMENT

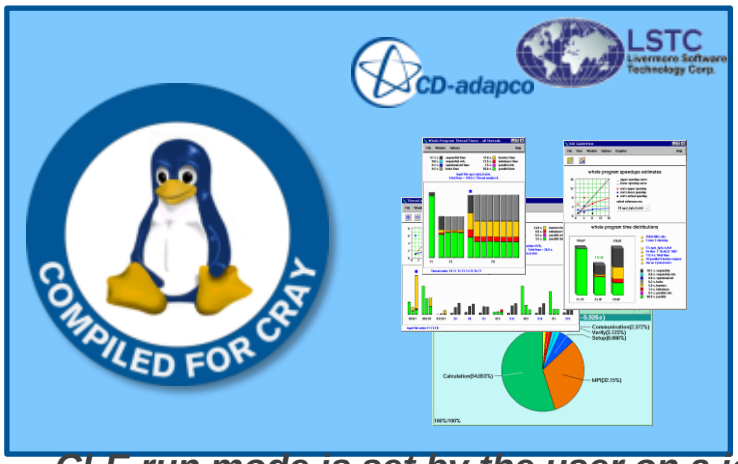
ESM – *Extreme Scalability Mode*

- 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

Linux as the foundation
Solid, robust, feature rich
Familiar, easy-to-use



CLE run mode is set by the user on a job-by-job basis to provide full flexibility

Cray Operating Systems Focus

Performance

- Maximize compute cycles delivered to applications while also providing necessary services
  - Lightweight operating system on compute node
  - Standard Linux environment on service nodes
- Optimize network performance through close interaction with hardware
- GPU infrastructure to support high performance

Stability and Resiliency

- Correct defects which impact stability
- Implement features to increase system and application robustness

Scalability

- Scale to large system sizes without sacrificing stability
- Provide system management tools to manage complicated systems

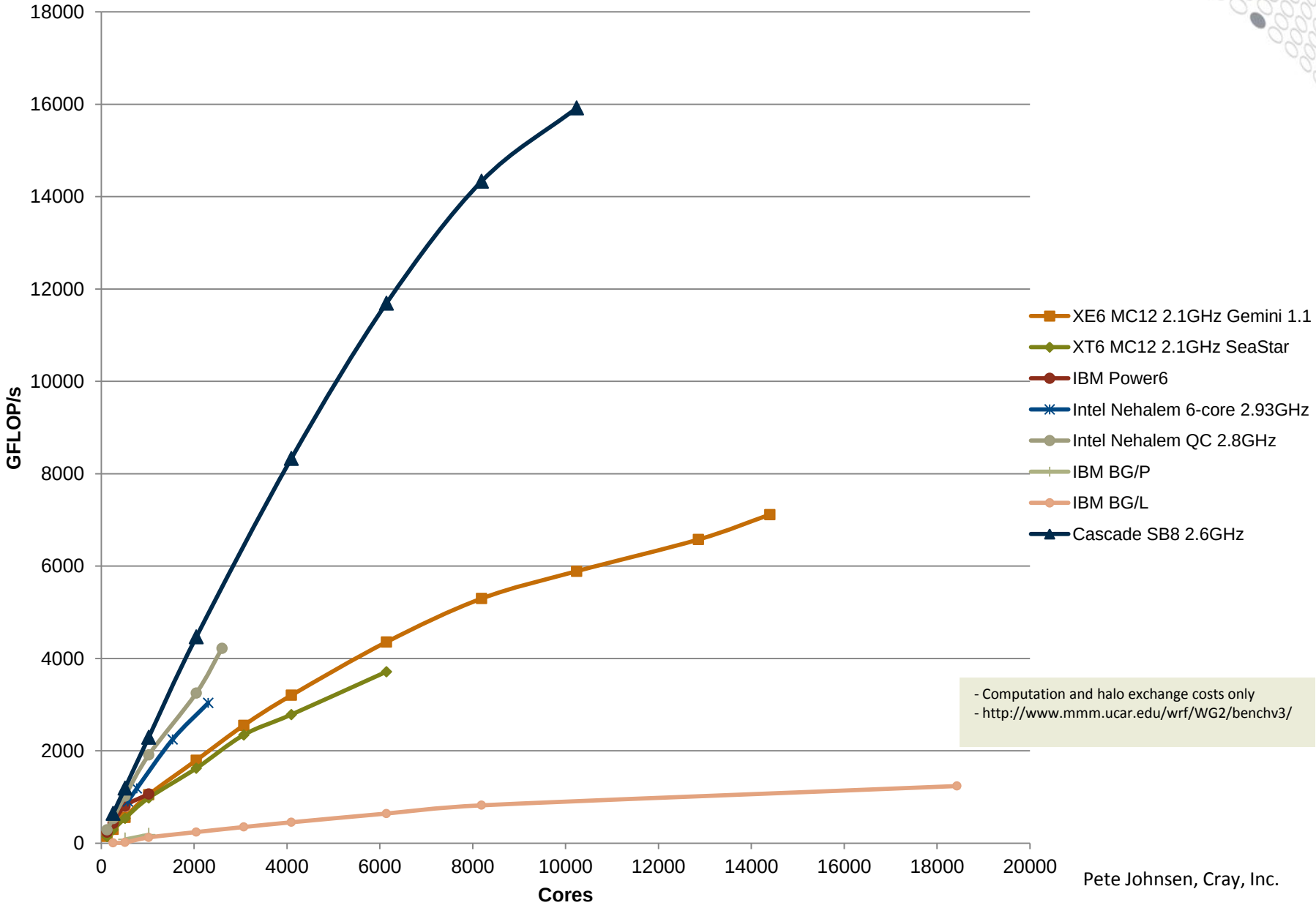
Benchmarks

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%

WRF NewConus 2.5KM Benchmark



- Computation and halo exchange costs only
- http://www.mmm.ucar.edu/wrf/WG2/benchv3/

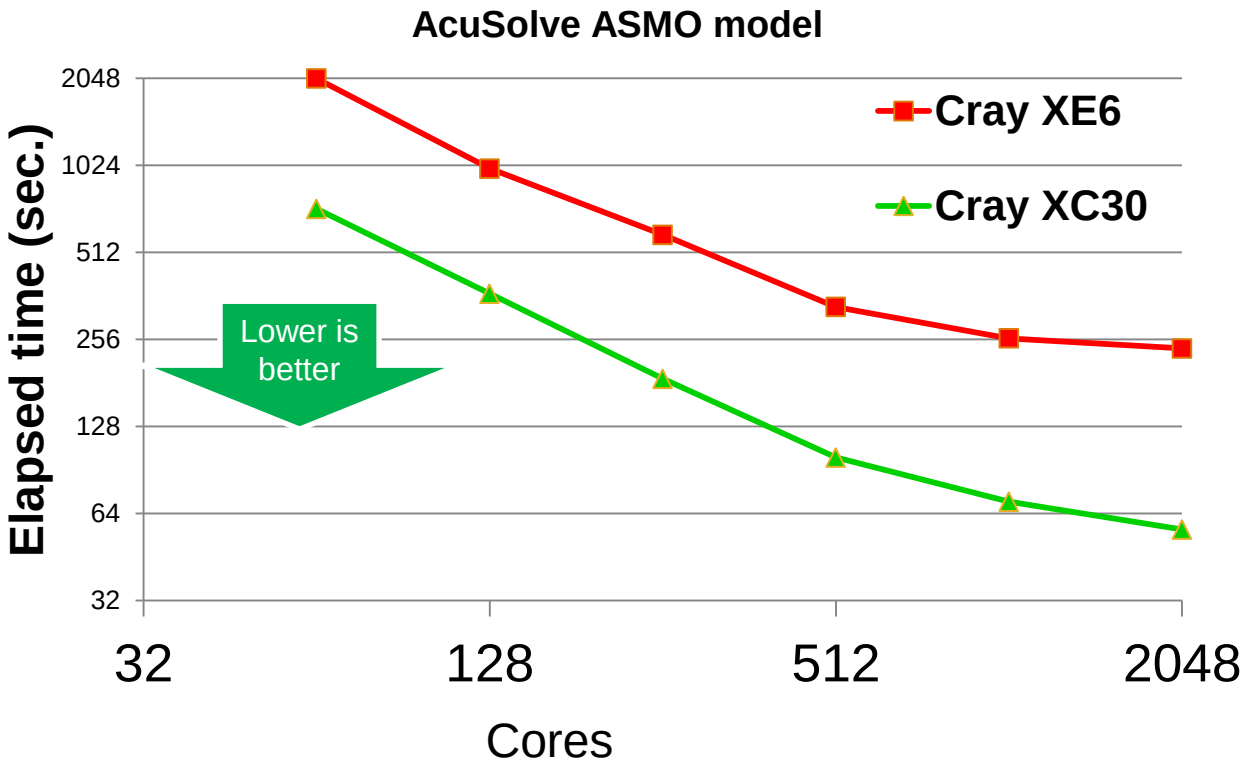
Pete Johnsen, Cray, Inc.

AcuSolve* (CFD) Performance

70M ASMO model

At 16 nodes:

- **Cray XE6: 512 Interlagos cores = 331 sec**
- **Cray XC30: 256 Sandy Bridge cores = 187 sec**

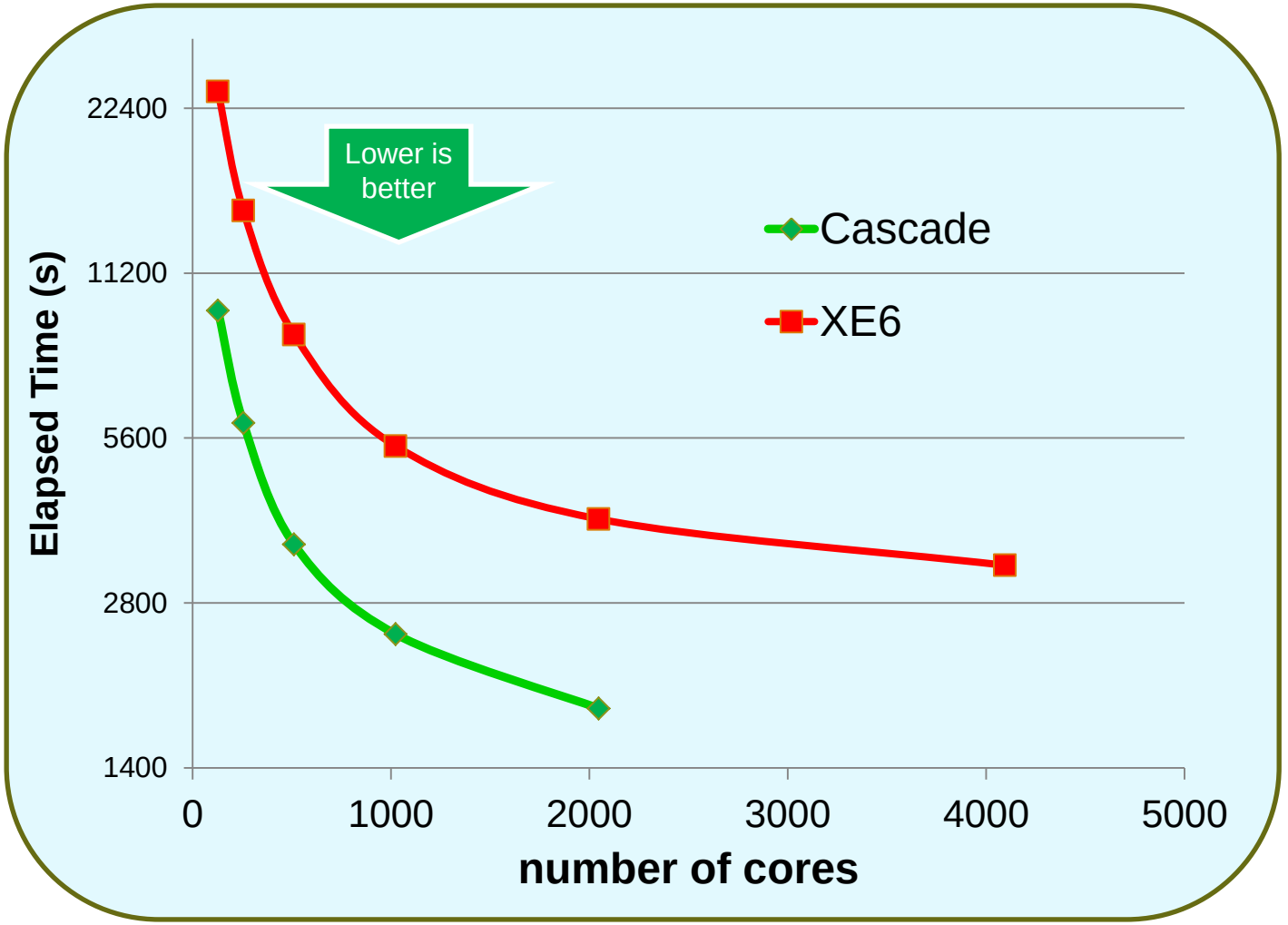


* Pre-release version of AcuSolve from Altair

LS-DYNA benchmark

Two car crash simulation, 2.4M elements, Hybrid parallel

XC30 provides significantly better per core performance



The End