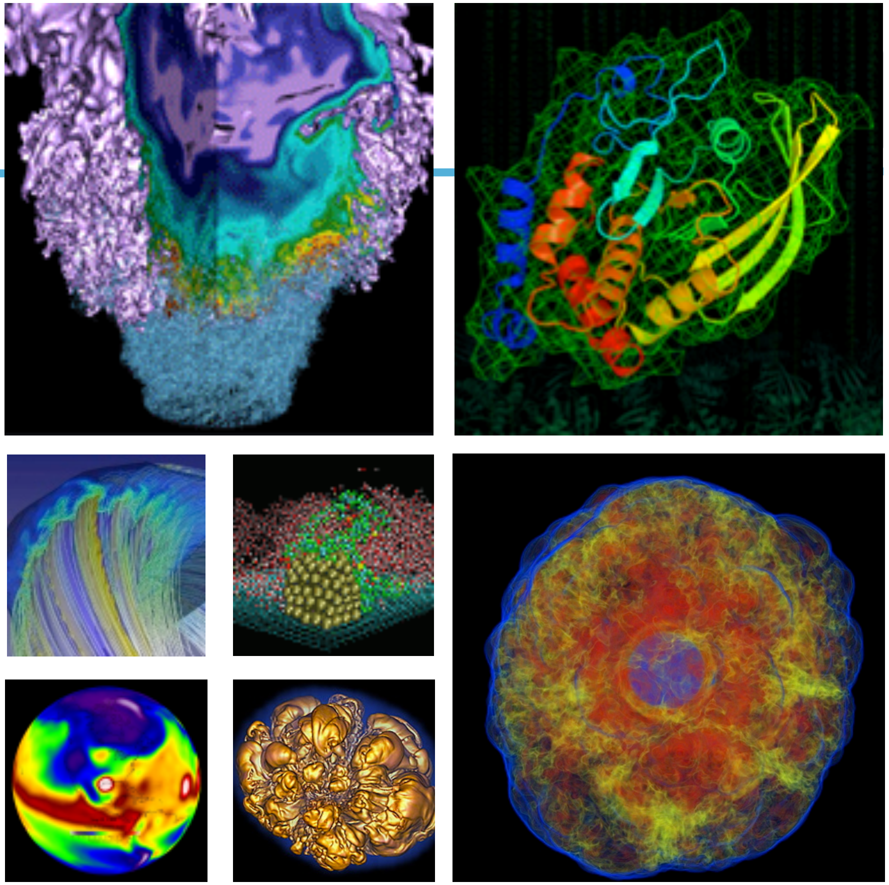
Edison Overview



Richard Gerber

Acting NERSC User Services Group Lead

October 10, 2013

Edison Addresses NERSC's Workload Needs

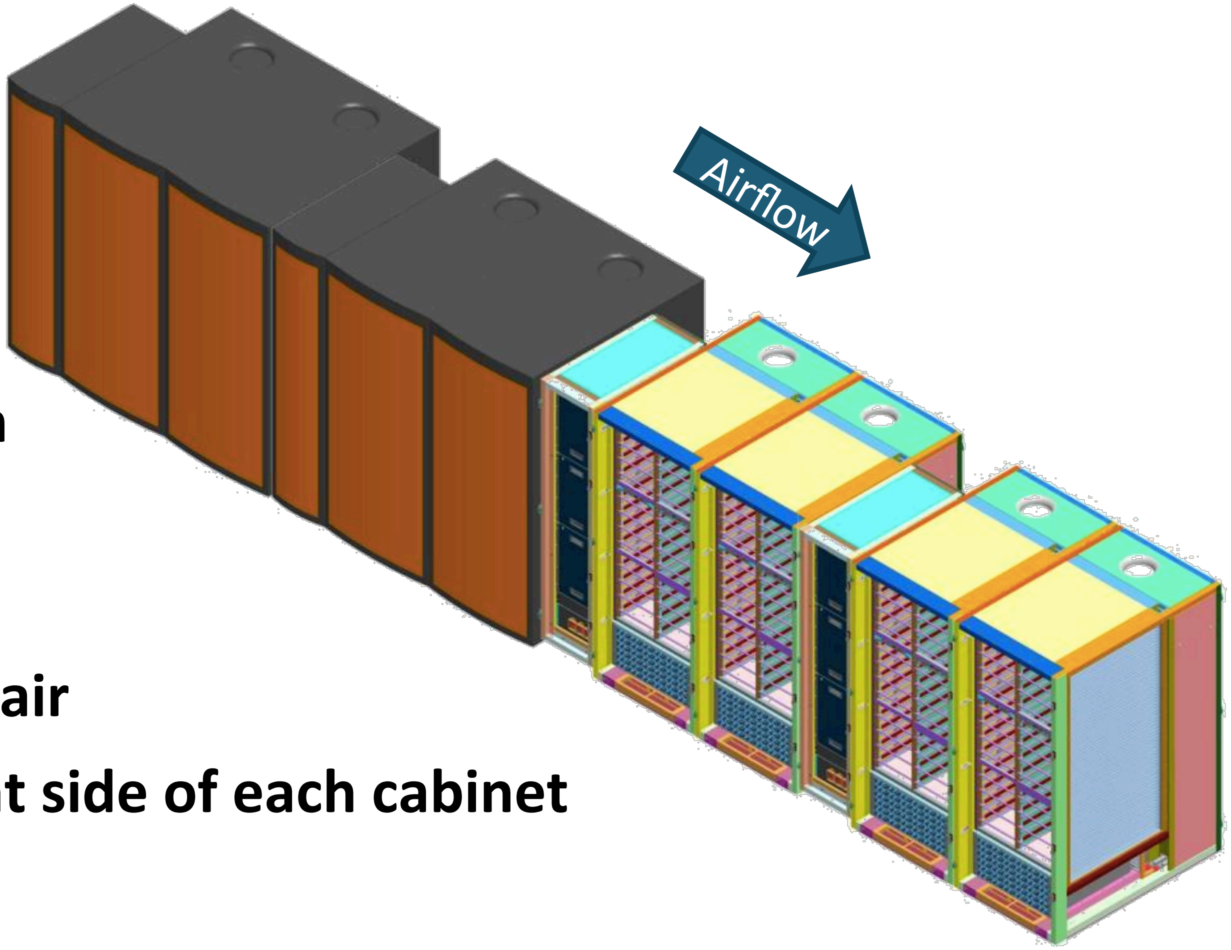
Characteristic	Description	Comment
Processor	Intel Ivy Bridge 2.6 GHz	Fast, cutting-edge, commodity processor Performance for High Throughput Apps
Node	Dual-socket, 64 GB 1866 MHz memory	Large memory per node Excellent memory bandwidth Performance for High Throughput Apps
Interconnect	Cray Aries, dragonfly topology	Excellent latency & bandwidth Excellent scaling Adaptive routing eases congestion Performance at Scale
Storage	6.48 PB 140 GB/sec I/O bandwidth, 3 file systems	Large, dedicated data storage High bandwidth; better metadata Data & I/O Improvements

Vital Statistics

	NERSC-6 (Hopper)	Edison
System	Cray XE-6	**Cray XC30 "Cascade"**
Compute Nodes / Cores	6,384 / 153,216	**5,200 / 124,800**
Processor	2 x AMD "Magny Cours" 2.1GHz, 12 core	2 x Intel Ivy Bridge 2.4GHz, 12 core
Memory	DDR3 1333 MHz	DDR3 1866 MHz
Memory per Node / Core	32 GB / 1.3 GB	**64 GB / 3.2 GB**
Total Memory	217 TB	333 TB
Interconnect	Gemini (Torus)	**Aries (Dragonfly)**
Sustained Performance (SSP)	144 TF	**250 TF**
Peak FLOPS	1.28 PF	2.4 PF
I/O Bandwidth	70 GB/s	>140 GB/s
I/O Capacity	2 PB	**6.48 PB**
File Systems	2	3
Login Nodes	12 x Quad Shanghi/128GB	12 x Quad Sandy Bridge/512GB

System Design

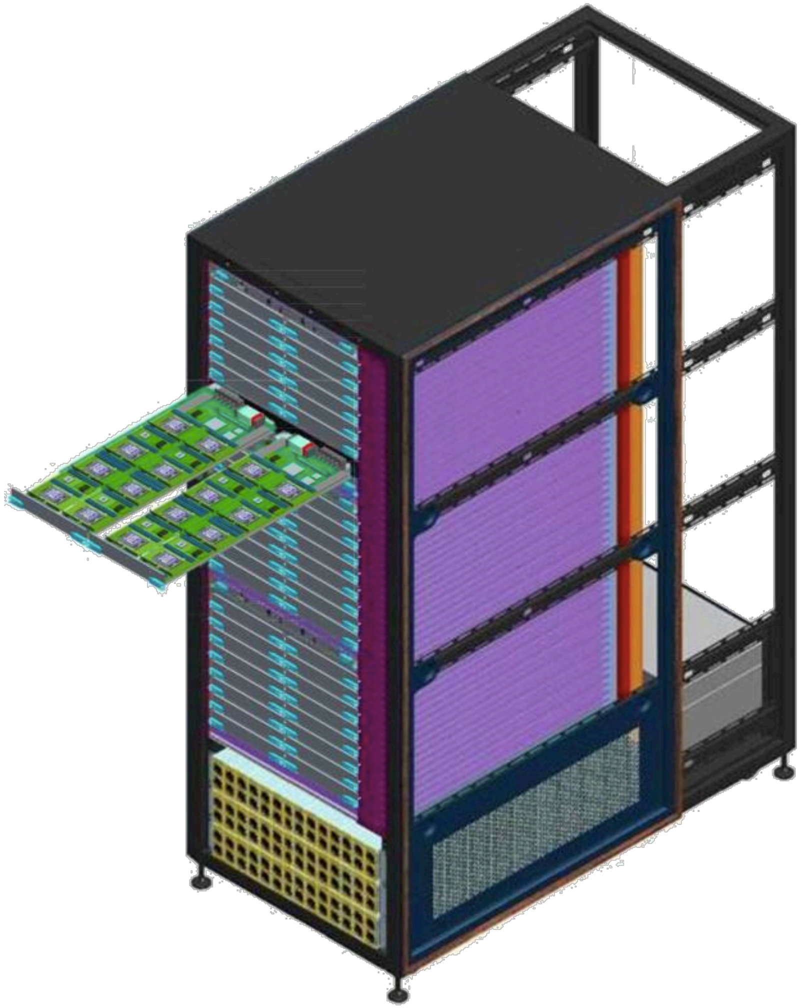
- Primarily water cooled
- One blower assembly for each cabinet pair (group)



- ≤75F water; ≤74F air
- Water coil on right side of each cabinet

Cabinet Design

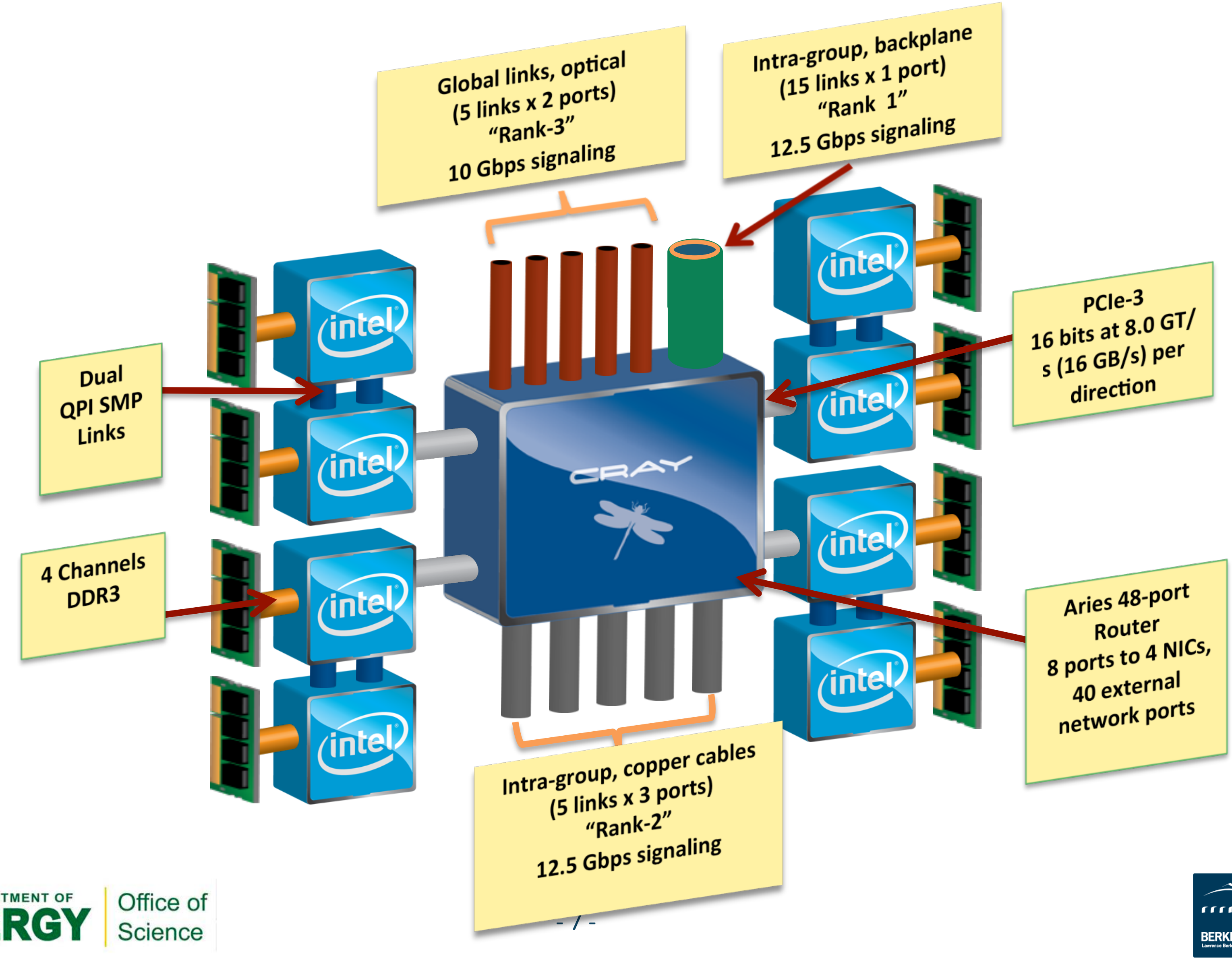
- **3 chassis / cabinet**
- **Up to 16 blades/chassis**
  - Up to 8 I/O blades
- **4 Nodes/compute blade**
  - 2 sockets/node
- **2 single socket nodes/ service/IO blade**



Edison Compute Node

- Intel Xeon Processor E5-2695 v2
- 2.4 GHz (3.2 GHz max turbo)
- 12 cores / 24 Threads (Hyperthreading)
- Intel AVX extensions
- 22 nm lithography
- 8 Flops / cycle max => 230 Gflops/socket
- 2 sockets per node => 460 Gflops/node
- Intel QPI Speed 16 GB/sec x 2
- 64 GB 1866 MHz memory/node
- ~100 GB/sec memory bandwidth

Edison Node Layout



Global links, optical
(5 links x 2 ports)
"Rank-3"
10 Gbps signaling

Intra-group, backplane
(15 links x 1 port)
"Rank 1"
12.5 Gbps signaling

Dual
QPI SMP
Links

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

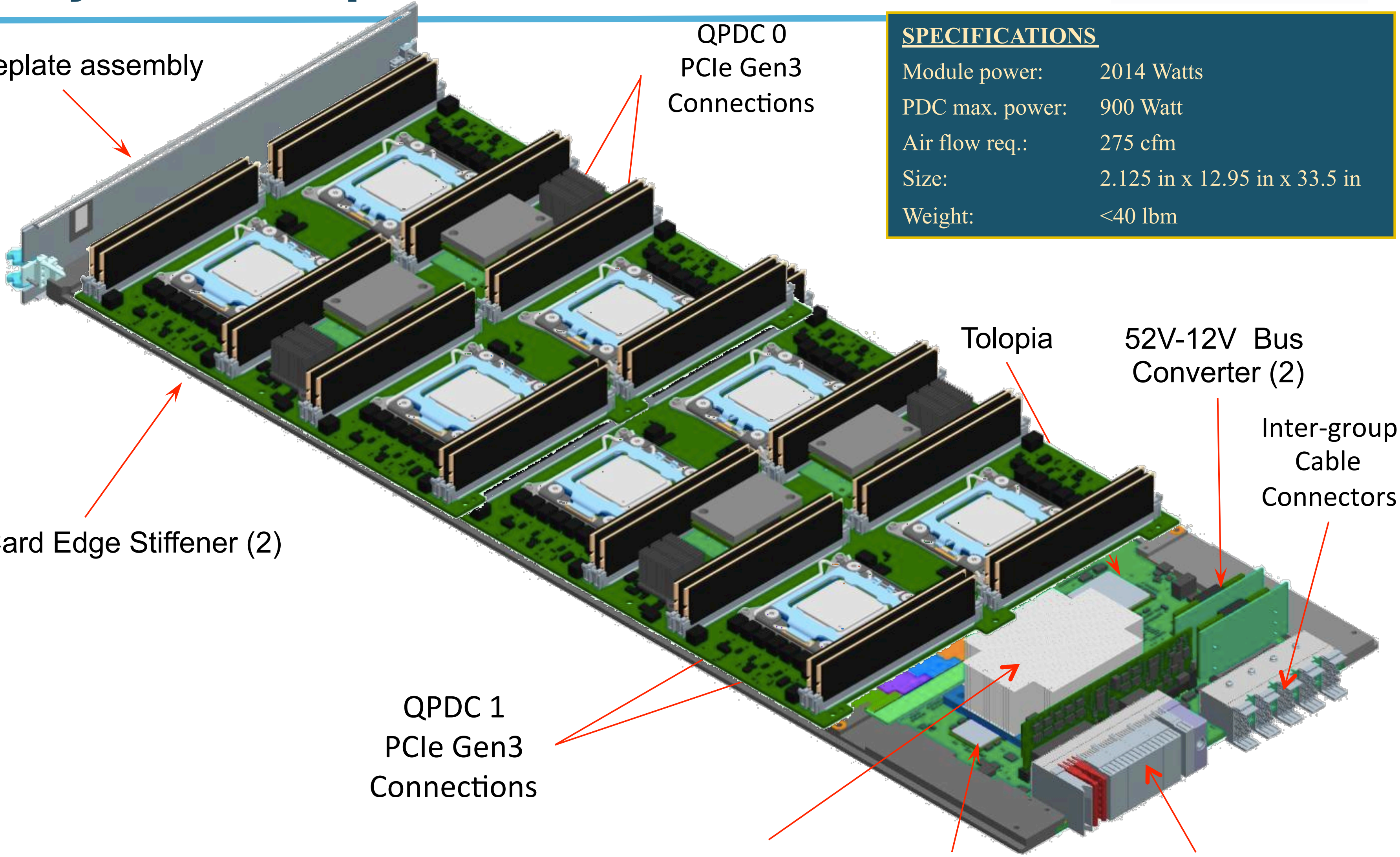
4 Channels
DDR3

Aries 48-port
Router
8 ports to 4 NICs,
40 external
network ports

Intra-group, copper cables
(5 links x 3 ports)
"Rank-2"
12.5 Gbps signaling

Cray XC30 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



Faceplate assembly

QPDC 0 PCIe Gen3 Connections

Tolopia

52V-12V Bus Converter (2)

Inter-group Cable Connectors

Card Edge Stiffener (2)

QPDC 1 PCIe Gen3 Connections

Aries

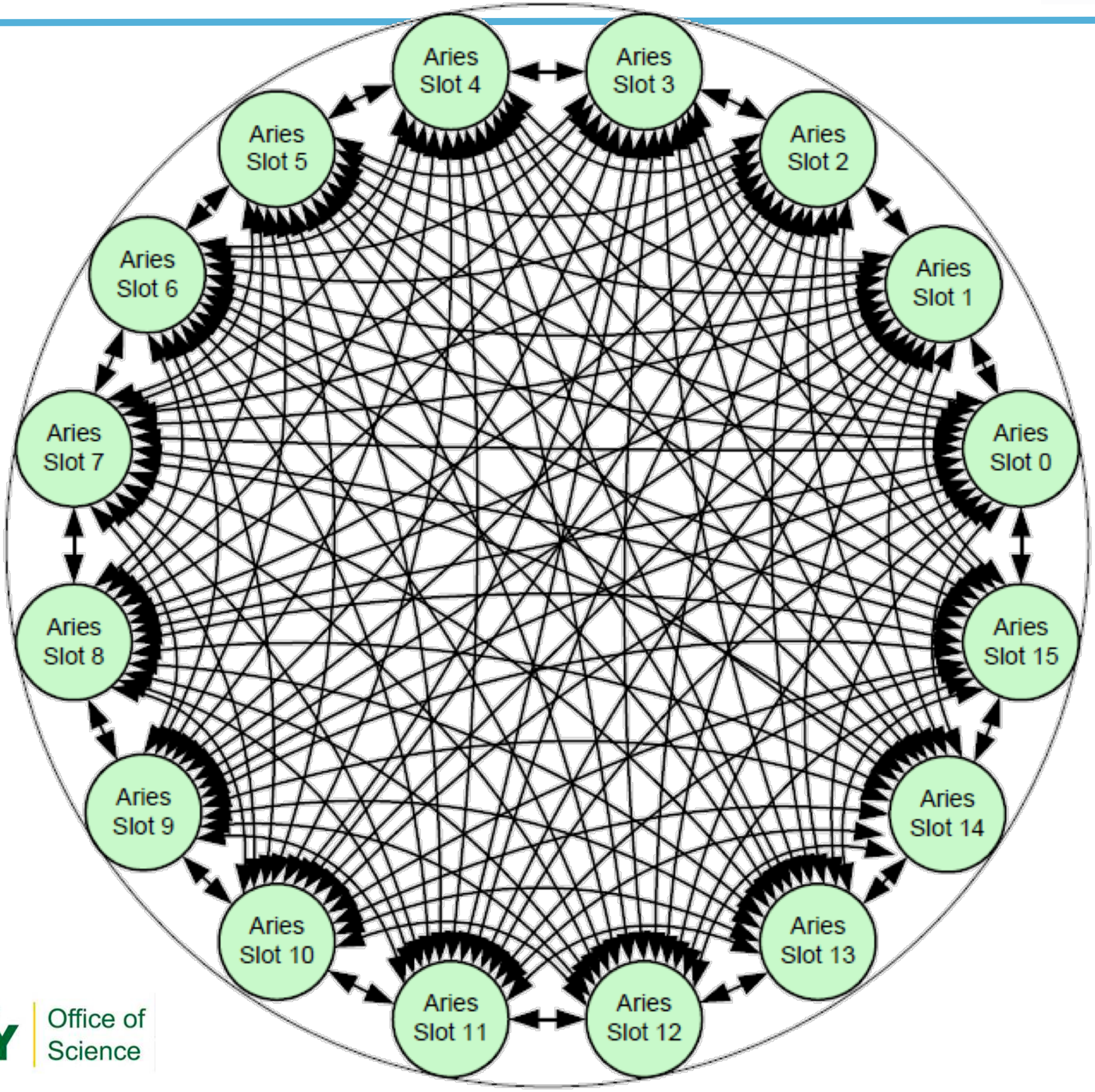
FPGA

Power and Backplane Connector

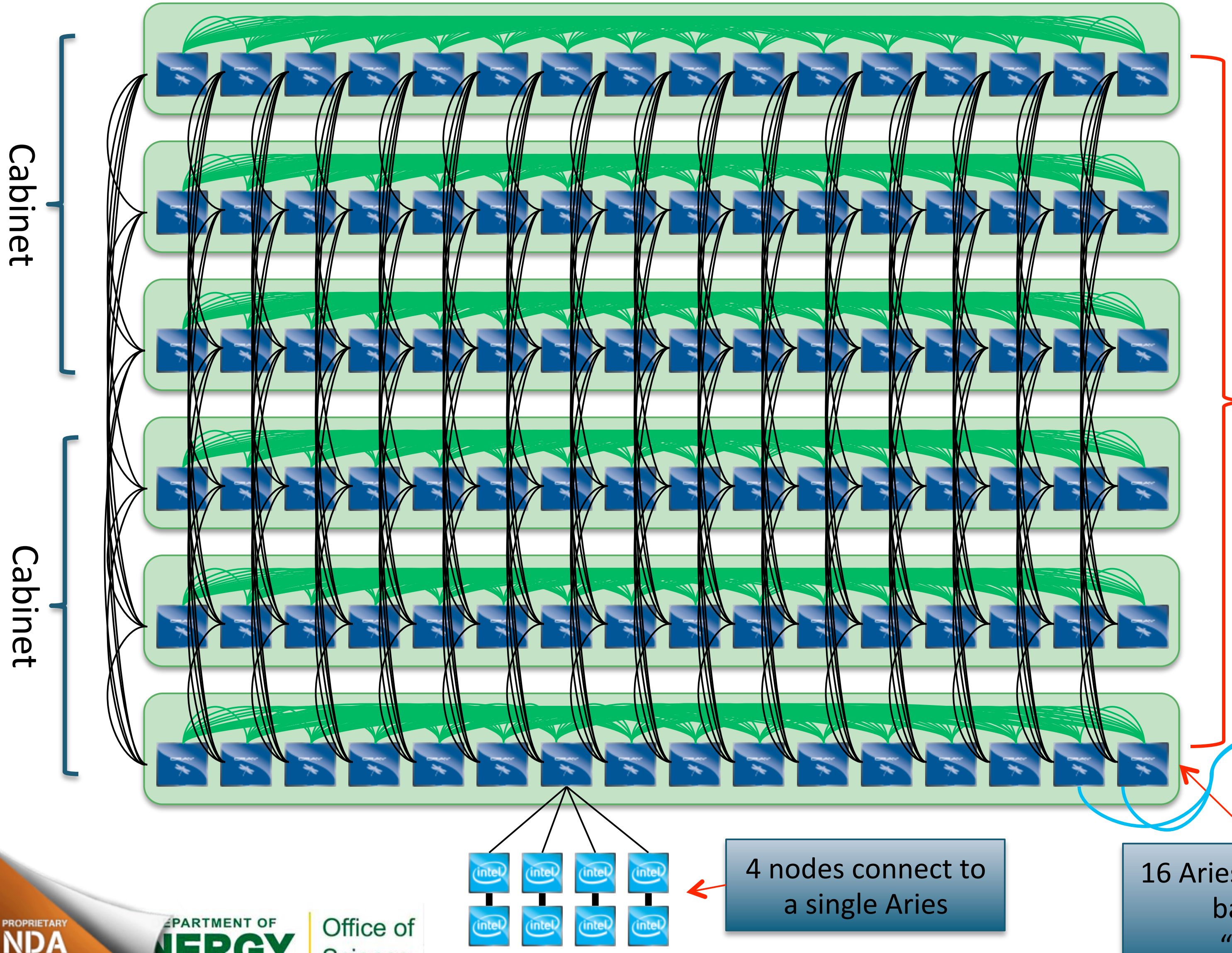
Aries Network

- **“Dragonfly” topology**
- **Configured in 3 ranks:**
  - Rank-1 is chassis level
  - Rank-2 is cabinet level
  - Rank-3 is system level
- **Global bandwidth tuned by number of optical cables (Rank 3)**
  - Edison: 11 TB/sec global bandwidth
- **Within a 2 cabinet group**
  - Minimal routing – 2 hops
  - Non-minimal routing – 4 hops

Aries Rank-1 Network



Aries Rank-2 Network



6 backplanes connected with copper wires in a 2-cabinet group: “Rank-2”

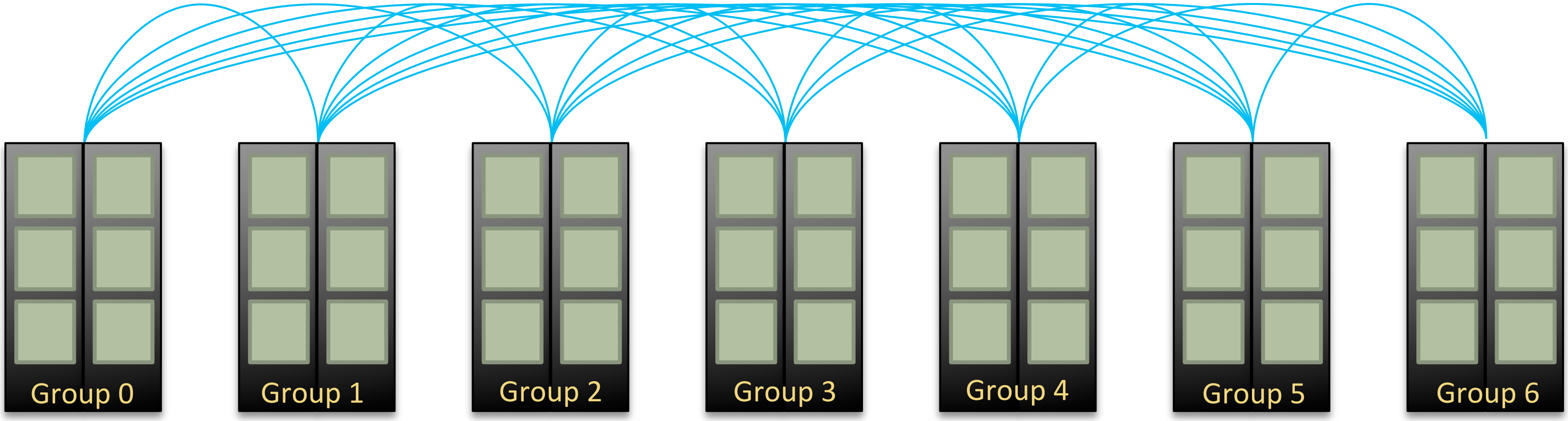
Pairs of Aries connect to a optical fiber to interconnect groups “Rank-3”

4 nodes connect to a single Aries

16 Aries connected in backplane “Rank-1”

Cascade 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



Edison has 546 optical cables in 6-cable bunches at Rank 3.

Application and Development Environment

- **Eases adoption by existing users and projects**
  - Easy to port and run production codes
- **Supports production software applications, libraries, and tools needed by the entire NERSC workload**
  - A robust set of programming languages, models
  - A rich set of highly optimized libraries, tools and applications
  - Community and pre-packaged applications
  - Shared-object libraries and socket communication
- **Enables effective application performance at scale, single node (high-throughput computing), and everything in between**

Programming Languages and Compilers supported

Programming languages

- Fortran
- Python, Perl, Shells
- C
- Java
- C++
- Chapel
- UPC

Supported compilers

- Cray Compiling Environment (CCE)
- intel
- GNU

Default compiler: Intel

Supported Programming Models

MPI

Cray SHMEM

OpenMP

POSIX Threads

POSIX Shared Memory

UPC

Coarray Fortran

Chapel

Cray Scientific and Math Libraries

LIBSCI

- LAPACK
- ScaLAPACK
- BLACS
- PBLAS

Third party scientific libraries

- MUMPS
- SuperLU
- ParMETIS
- HYPRE
- Scotch

Trilinos

FFTW

PETSc

DMAPP API for Aries

Intel MKL

MPI-IO Library

IO libraries

- HDF5
- NetCDF
- Parallel-netcdf

Development and Performance Tools

Scalable Debuggers

- DDT
- Totalview

Profiling tools

- CrayPat
- Appentice2
- IPM

Abnormal Termination Processing (ATP)

PAPI

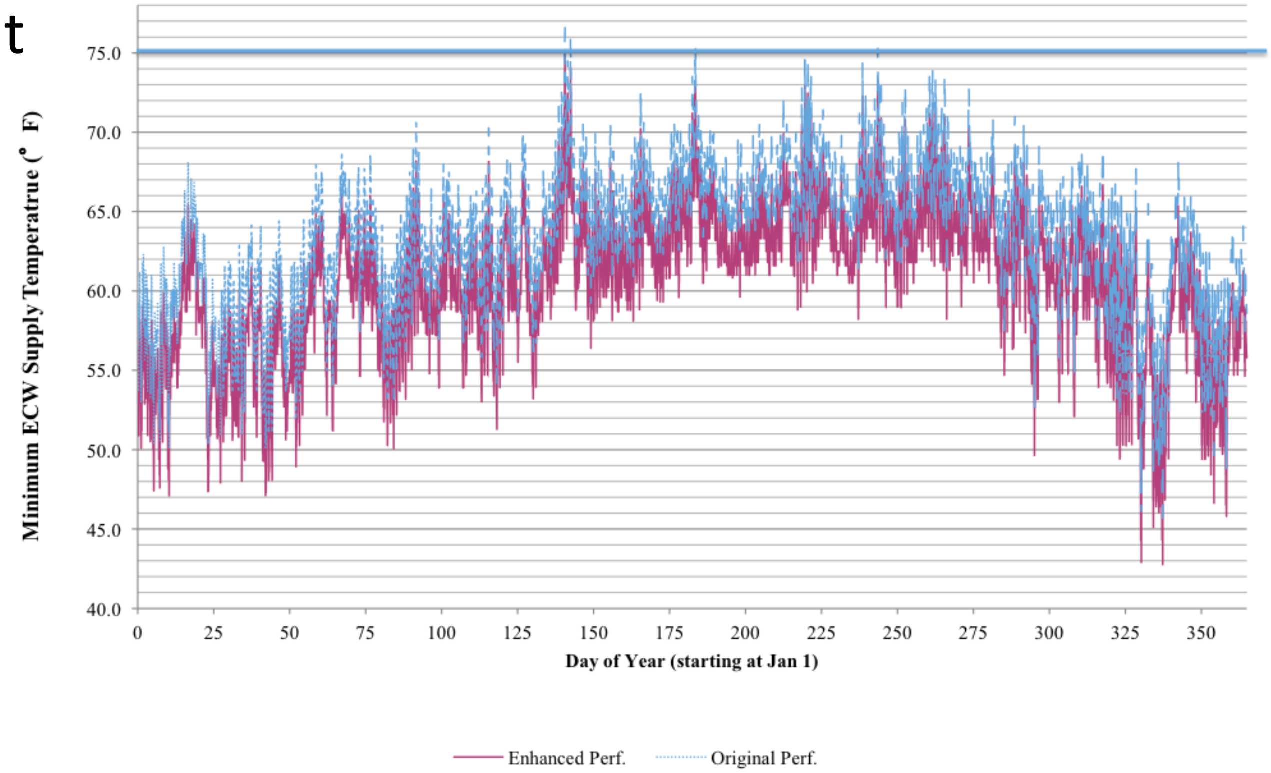
Key Features of Storage

- **3 Lustre v2.2 scratch file systems for spreading user bandwidth needs**
  - Spread users among the 2 file systems to evenly distribute load
  - One file reserved for runs with extreme bandwidth needs (up to 70 GB/s to a single file system)
- **2 x the metadata rates from Hopper in aggregate**
  - Also isolates metadata performance to 1 of 3 file systems

Energy Efficiency

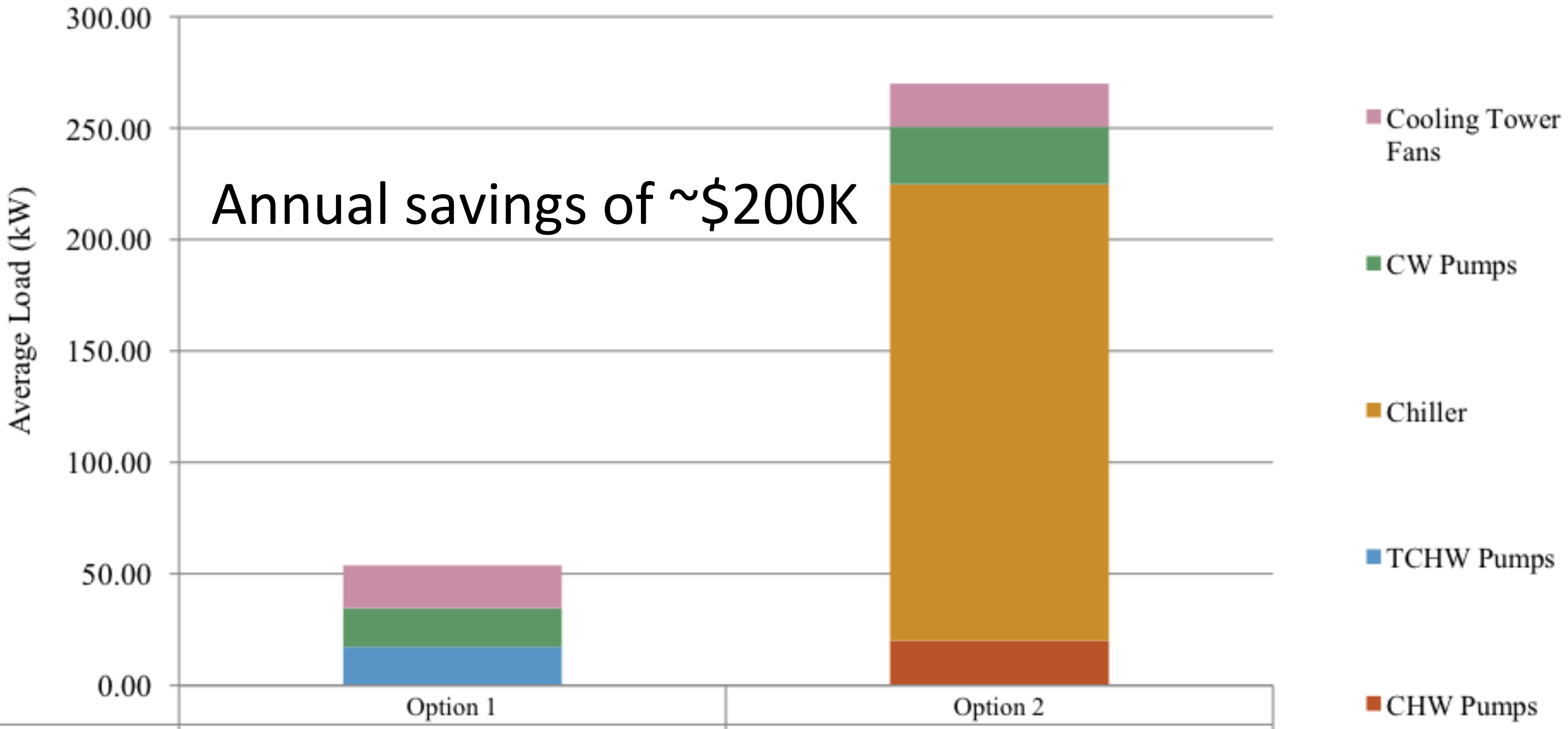
Energy-efficient design and components allows chiller-free cooling 100% of the year.

First DOE PF system to use year-round chiller-free cooling.



The Bay Area climate allows NERSC to used evaporative cooling for Edison.

Power for Cooling with and without chillers



Annual savings of ~$200K

	Option 1	Option 2
🟪 Cooling Tower Fans	19.32	19.32
🟩 CW Pumps	17.38	25.71
🟧 Chiller	0.00	204.98
🟦 TCHW Pumps	17.17	0.00
🟥 CHW Pumps	0.00	19.97

External Nodes

- **esLogin**
  - Quad processor Sandy Bridge
  - 512 GB DDR3 memory
  - 2 dual-port 10GB ethernet
  - 2 dual-port FDR IB HBAs
- **esMS**
  - Management workstation for esLogin nodes
  - Runs Bright Cluster Management software