Argonne NATIONAL LABORATORY

High-Performance Computing at Argonne National Laboratory

Marc Snir

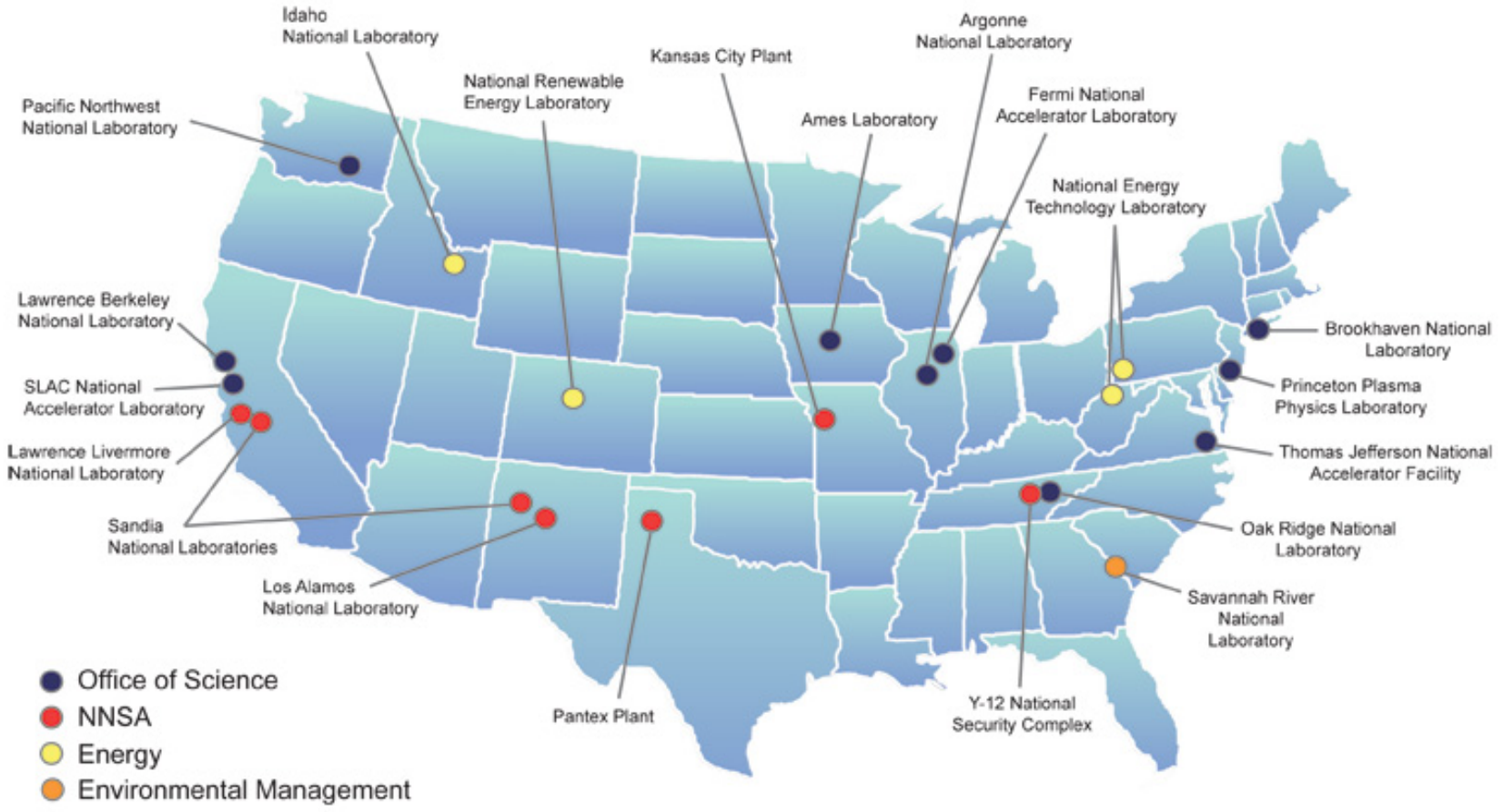
Director, Mathematics and Computer Science Division

Argonne National Laboratory

Professor, Dept. of Computer Science, UIUC

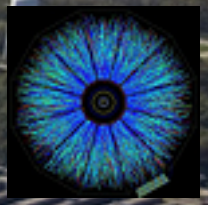
U.S. DEPARTMENT OF ENERGY

Argonne: Vital part of DOE National Laboratory System

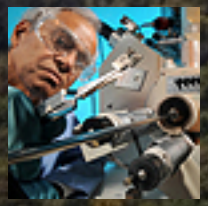


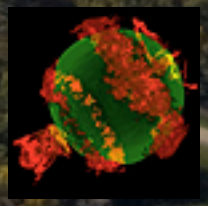
About Argonne

- $675M operating budget
- 3,200 employees
- 1,450 scientists and engineers
- 750 Ph.D.s

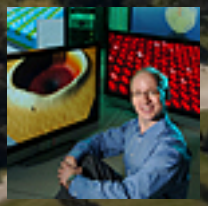














Direct descendent of Enrico Fermi's Metallurgical Laboratory

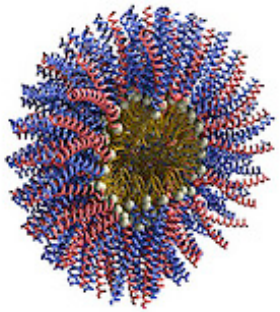
Argonne's mission: To provide science-based solutions to pressing global challenges

Through discovery and transformational science and engineering…

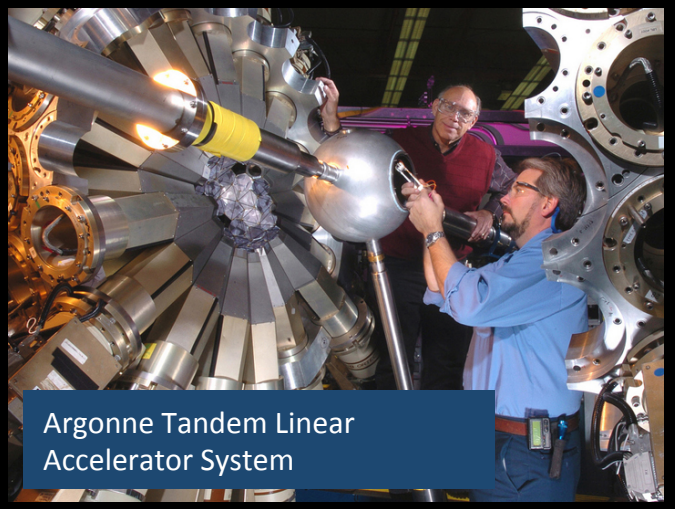
World-leading hard x-ray sciences & sources	Discovery science for energy	Leadership computing and computational ecosystem	Fundamental physics and accelerator capabilities	Materials & systems engineering solutions

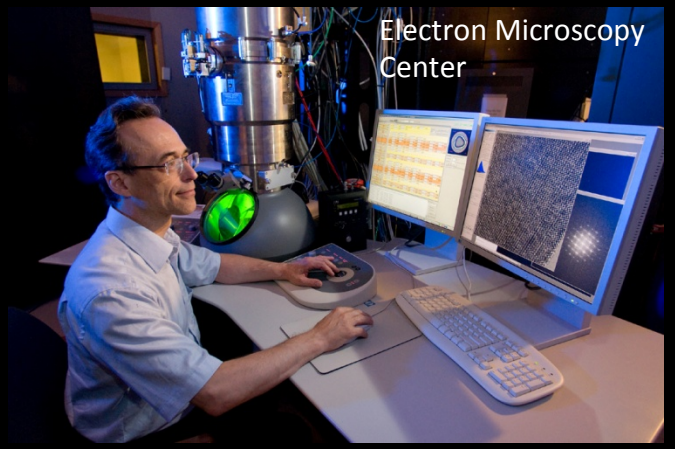
and through use-inspired science and engineering

Energy Storage	Sustainable Transportation	Nuclear Energy	Environmental Genomics	National Security



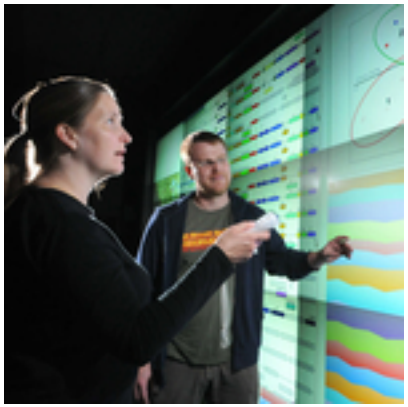
Major Scientific User Facilities





Argonne National Laboratory: 4 Directorates

Computing, Environment & Life Sciences (CELS)



Energy Engineering & Systems Analysis (EESA)

Photon Sciences (PS)

Physical Sciences & Engineering (PSE)

CELS: 4 Divisions

The Argonne Leadership Computing Facility

Mira: 10 PF BG/Q, ~750,000 cores, 0.75 PB

Biosciences Division

Environmental Science Division

Mathematics and Computer Science Division

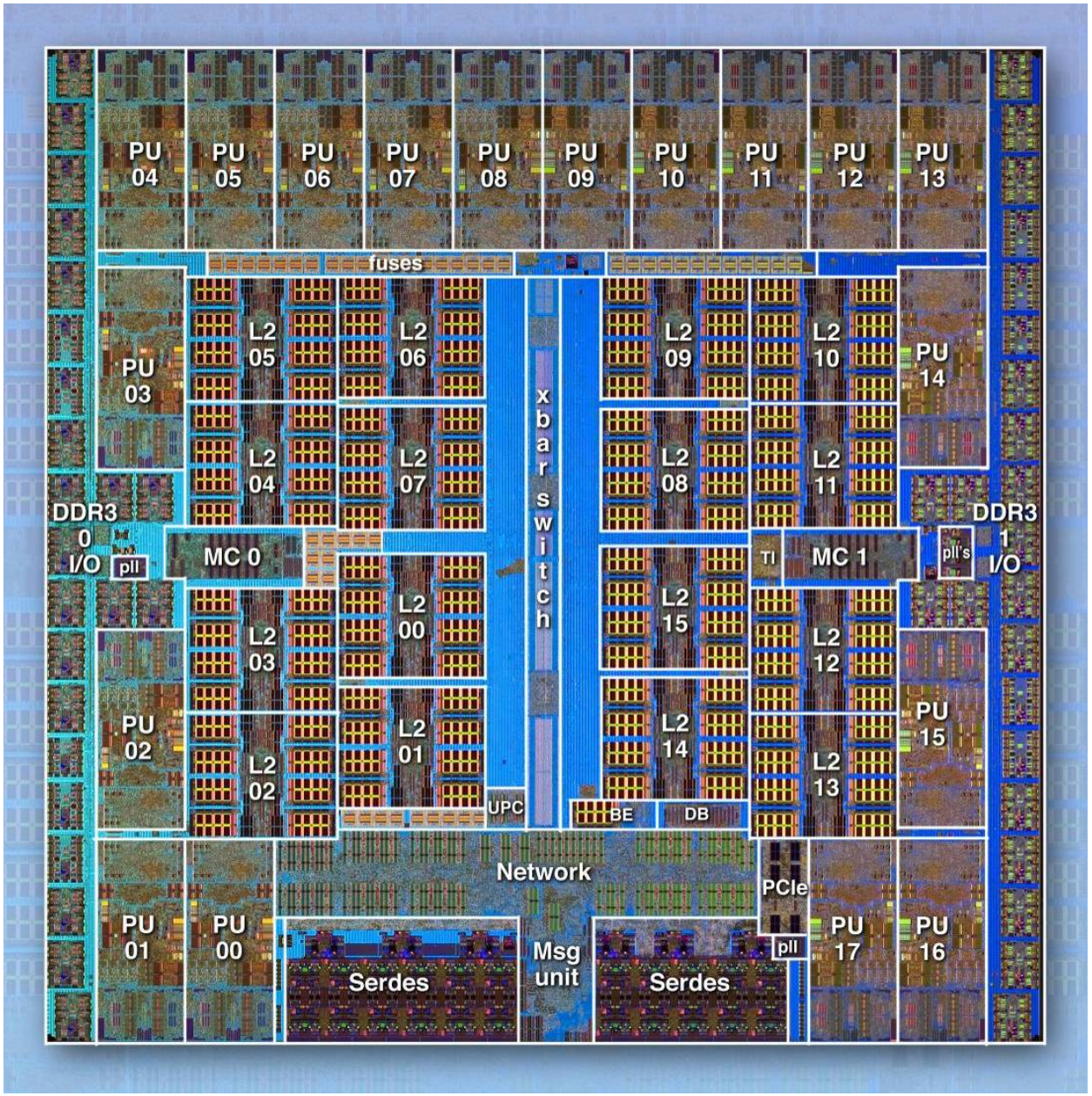
MCS

- **Extreme Computing Group**: Software infrastructure for extreme scale computers
- **Big Data Group**: Software infrastructure for storage, communication and analysis of large & complex data sets
- **Applied Math Group**: Scalable numerical algorithms and scientific libraries
  - Sparse linear algebra (PETSc), PDE solvers (MOAB), Optimization (TAO)
- **Application Group**: Application of advanced computing methods to selected application areas
  - Bioinformatics, climate modeling, nuclear engineering, cosmology
- **Computational Institute** (joint with U. of Chicago): Collaborative environments
  - Grid, cloud
- ~100 people
- Collaborations with France, China, Germany…

Mira - Blue Gene/Q System at ANL

- Computation
  - 48K nodes / 768K cores
  - 786 TB of memory
  - Peak flop rate: 10 PF
- Storage
  - ~35 PB capacity, 240GB/s bandwidth (GPFS)
  - Disk storage upgrade planned in 2015
    - Double capacity and bandwidth

BlueGene/Q Compute chip



System-on-a-Chip design : integrates processors, memory and networking logic into a single chip

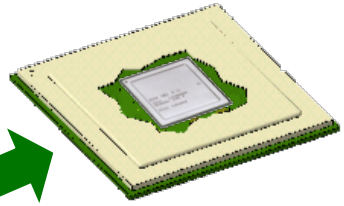
BlueGene/Q Compute chip

- **360 mm² Cu-45 technology (SOI)**
  - ~ 1.47 B transistors
- **16 user + 1 service processors**
  - plus 1 redundant processor
  - all processors are symmetric
  - each 4-way multi-threaded
  - 64 bits PowerISA™
  - 1.6 GHz
  - L1 I/D cache = 16kB/16kB
  - L1 prefetch engines
  - each processor has Quad FPU (4-wide double precision, SIMD)
  - peak performance 204.8 GFLOPS@55W
- **Central shared L2 cache: 32 MB**
  - eDRAM
  - multiversioned cache will support transactional memory, speculative execution.
  - supports atomic ops
- **Dual memory controller**
  - 16 GB external DDR3 memory
  - 42.6 GB/s
  - 2 * 16 byte-wide interface (+ECC)
- **Chip-to-chip networking**
  - Router logic integrated into BQC chip.
- **External IO**
  - PCIe Gen2 interface

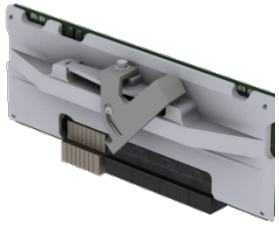
Blue Gene/Q

1) Chip:
16+2 cores

2) Single Chip Module



3) Compute card:
One chip module,
16 GB DDR3 Memory,
Heat Spreader for H_2O Cooling



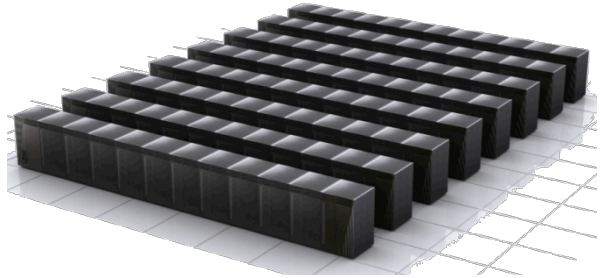
4) Node Card:
32 Compute Cards,
Optical Modules, Link Chips; 5D Torus

5b) IO drawer:
8 IO cards w/16 GB
8 PCIe Gen2 x8 slots
3D I/O torus

5a) Midplane:
16 Node Cards

6. Rack: 2 Midplanes

7) System:
96 (48) racks,
20 (10) PF/s

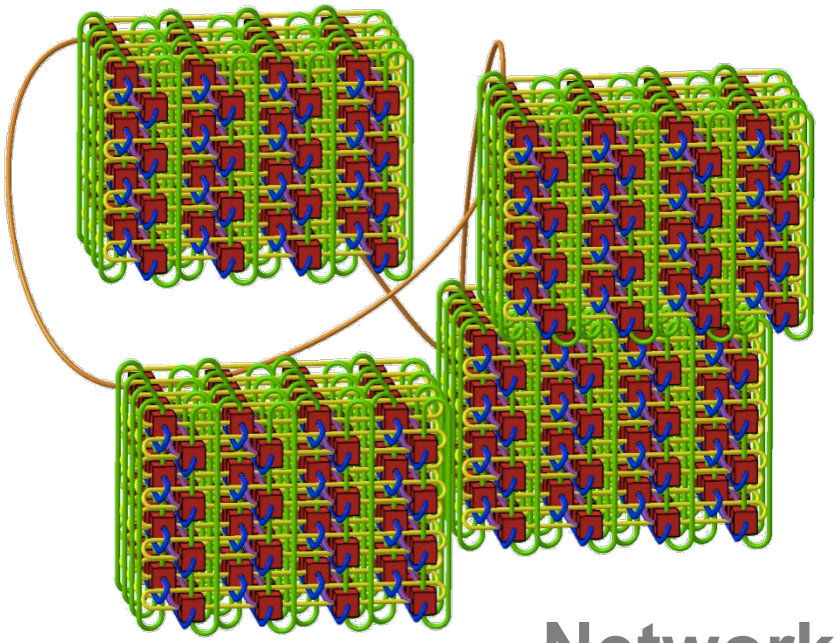


•Sustained single node perf: 10x P, 20x L
• MF/Watt: (6x) P, (10x) L (~2GF/W, Green 500 criteria)
• Software and hardware support for programming models for exploitation of node hardware concurrency

Inter-Processor Communication

- **Integrated 5D torus**
  - Virtual Cut-Through routing
  - Hardware assists for collective & barrier functions
  - FP addition support in network
  - RDMA
    - Integrated on-chip Message Unit
- **2 GB/s raw bandwidth on all 10 links**
  - each direction -- i.e. 4 GB/s bidi
  - 1.8 GB/s user bandwidth
    - protocol overhead
- **Hardware latency**
  - Nearest: 80ns
  - Farthest: 3us
    (96-rack 20PF system, 31 hops)
- **Additional 11th link for communication to IO nodes**
  - BQC chips in separate enclosure
  - IO nodes run Linux, mount file system
  - IO nodes drive PCIe Gen2 x8 (4+4 GB/s)
    ↔ IB/10G Ethernet ↔ file system & world

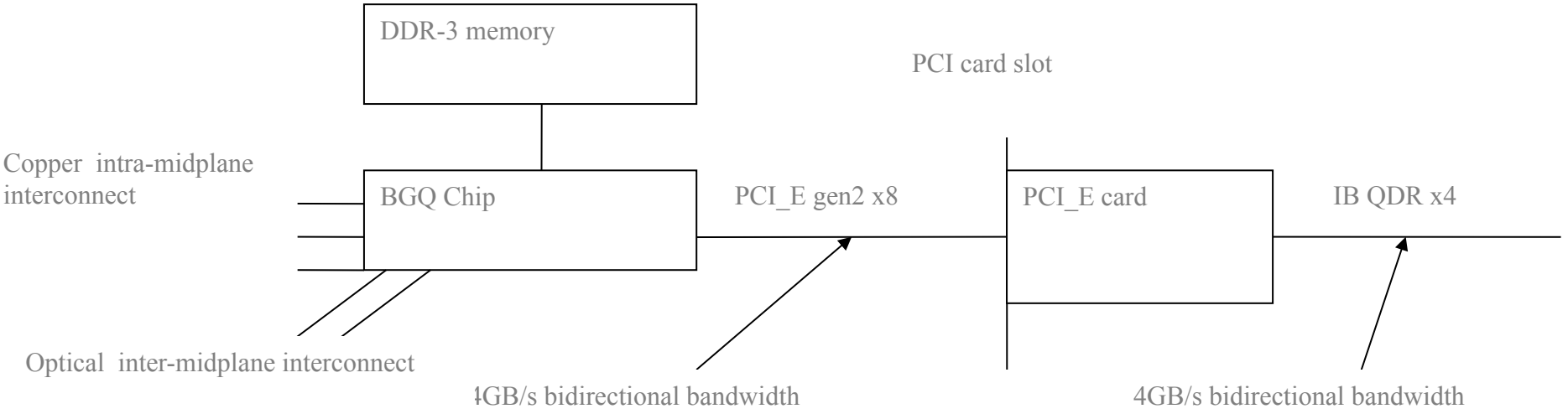
Inter-Processor Communication



Network Performance

- 5D nearest neighbor exchange measured at 1.76 GB/s per link (98% efficiency)
- All-to-all: 97% of peak
- Bisection: > 93% of peak
- Nearest-neighbor: 98% of peak
- Collective: FP reductions at 94.6% of peak

Blue Gene/Q I/O node



Alternatives:

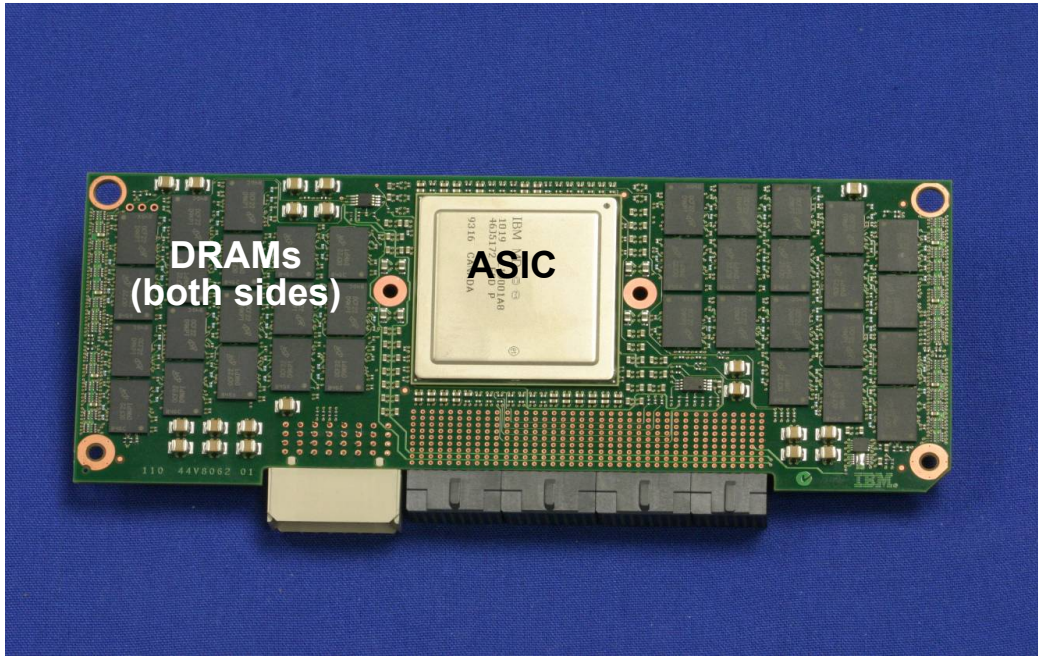
-- PCI_E to IB QDR x4 (shown)

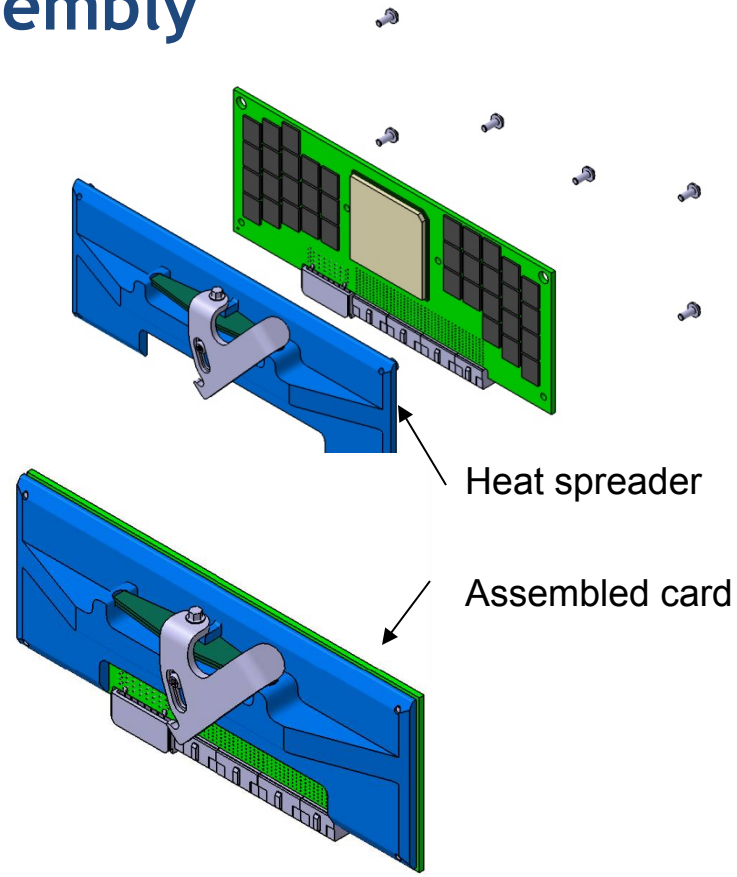
-- PCI_E to (dual) 10 Gb ethernet card (log in nodes)

-- PCI_E to single 10GbE + IB QDR

-- PCI_E to SATA for direct disk attach

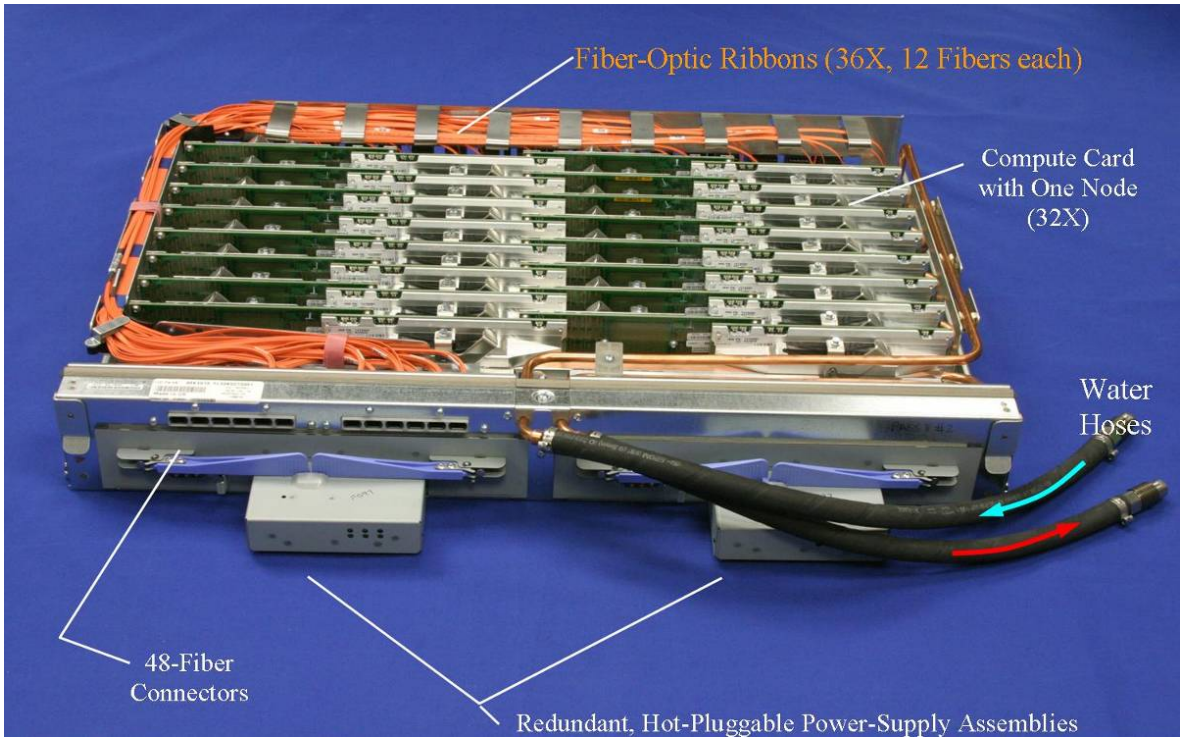
Blue Gene/Q Compute Card Assembly





- Basic field replaceable unit of a Blue Gene/Q system
- Compute Card has 1 BQC chip + 72 SDRAMs (16GB DDR3)
- Two heat sink options: Water-cooled → "Compute Node" / air-cooled → "IO Node"
- Connectors carry power supplies, JTAG etc, and 176 Torus signals (4 and 5 Gbps)

Blue Gene/Q Node Card Assembly

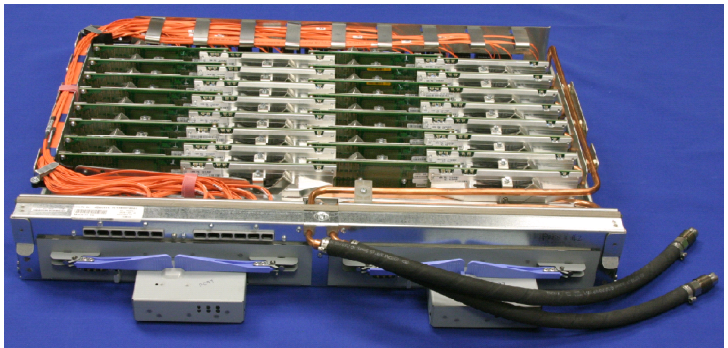


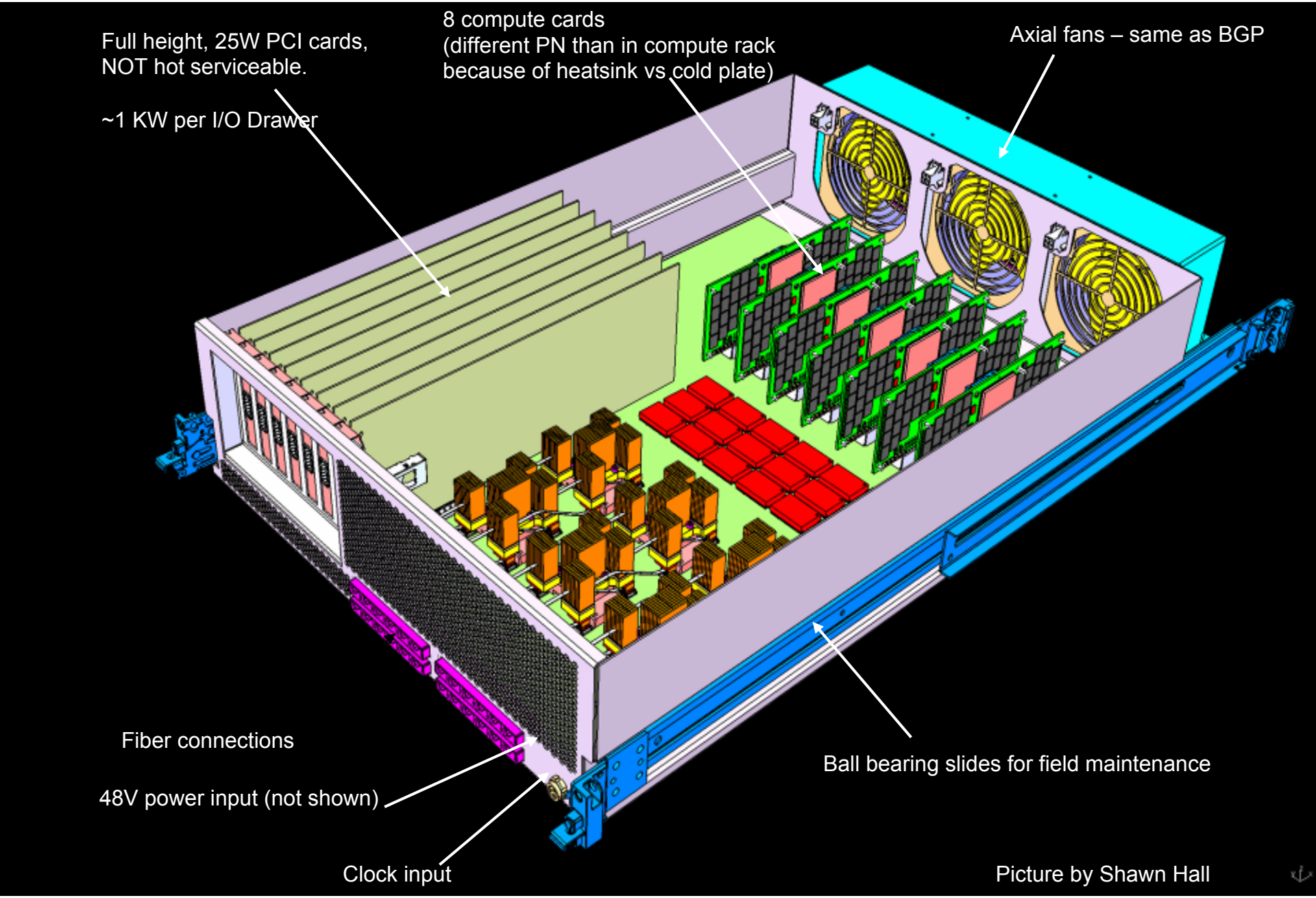
- Power efficient processor chips allow dense packaging
- High bandwidth / low latency electrical interconnect on-board
- 18+18 (Tx+Rx) 12-channel optical fibers @10Gb/s
  - Recombined into 8*48-channel fibers for rack-to-rack (Torus) and 4*12 for Compute-to-IO interconnect
- Compute Node Card assembly is water-cooled (18-25ºC – above dew point)
- Redundant power supplies with distributed back-end ~ 2.5 kW

Packaging and Cooling

Water	18C to 25C
Flow	20 gpm to 30 gpm
Height	2095 mm (82.5 inches)
Width	1219 mm (48 inches)
Depth	1321 mm (52 inches)
Weight	2000 kg (4400 lbs) *(including water)*
	I/O enclosure with 4 drawers 210 kg (480 lbs)

- Water cooled node board
- 32 compute cards, 8 link ASICs drive 4D links using 10Gb/s optical transceivers
- Hot pluggable front-end power supplies




Full height, 25W PCI cards,
NOT hot serviceable.
~1 KW per I/O Drawer
8 compute cards
(different PN than in compute rack
because of heatsink vs cold plate)
Axial fans – same as BGP
Fiber connections
48V power input (not shown)
Clock input
Ball bearing slides for field maintenance
Picture by Shawn Hall

Overview of BG/Q: Another step forward

Design Parameters	BG/P	BG/Q	Improvement
Cores / Node	4	16	4x
Clock Speed (GHz)	0.85	1.6	1.9x
Flop / Clock / Core	4	8	2x
Nodes / Rack	1,024	1,024	--
RAM / core (GB)	0.5	1	2x
Flops / Node (GF)	13.6	204.8	15x
Mem. BW/Node (GB/sec)	13.6	42.6	3x
Latency (MPI zero-length, nearest-neighbor node)	2.6 μs	2.2 μs	~15% less
Bisection BW (32 racks)	1.39TB/s	13.1TB/s	9.42x
Network Interconnect	3D torus	5D torus	Smaller diameter
Concurrency / Rack	4,096	65,536	16x
GFlops/Watt	0.77	2.10	3x

Notes on Mira Science Applications

- Applications cannot be manually tuned; only compiler optimizations are allowed.
- 3 of the applications are threaded – i.e., use both OpenMP and MPI (GTC, GFMC, GFDL).
- The remainder are 100% MPI applications (DNS3D, FLASH, GPAW, LS3DF, MILC, NAMD & NEK 5000).
- For 100% MPI applications, we tested multiple MPI ranks per core (max of 4 ranks per core).
- For MPI + OpenMP applications, we tested 1 MPI rank per core and multiple OpenMP threads per core (max of 4 threads per core)

Comments on using all hardware threads

- Speed up with hardware threads will be limited if the issue rate is already high with 1 thread/core (NEK is an example).
- Speed-up with hardware threads will be limited if the problem is already near the scaling limit at 1 thread/core. Using all threads will require 4x more threads.
- Speed-up can be limited if there is contention for L1-D and L1P resources.
- In some cases using OpenMP or Pthreads instead of MPI might reduce L1 contention.

BG/Q Performance Tools

- Early efforts were initiated to bring widely used performance tools to the BG/Q
- A variety of tools providers are currently working with IBM and Argonne to port and test tools on the Q
- BG/Q provides a hardware & software environment that supports many standard performance tools:
  - Software:
    - Environment similar to 64 bit PowerPC Linux
      - provides standard GNU binutils
    - New performance counter API bgpm
  - Performance Counter Hardware:
    - BG/Q provides 424 64-bit counters in a central node counting unit
    - Counter for all cores, prefetchers, L2 cache, memory, network, message unit, PCIe, DevBus, and CNK events
    - Provides support for hardware threads and counters can be controlled at the core level
    - Countable events include: instruction counts, flop counts, cache events, and many more

- Argonne working with IBM on providing POMP features for tools developers. Also collaborating with LLNL on OpenMP performance interface standard in OpenMP ARB

BG/Q Parallel Debuggers Status

- IBM CDTI (Code Development and Tools Interface)
  - Collaboration of IBM/LLNL/ANL resulted in update v1.7 (August 2011)
  - Refined interface for multiple tool support, breakpoint handling, stepping, and signal handling
- Rogue Wave TotalView
  - Ported to BG/Q (Q32 at IBM) with basic functionality in August 2011
  - Pre-release testing by LLNL December 2011
  - Status
    - Tested working: basic ops (step, breakpoint, stack), QPX instructions, fast conditional breakpoints, job control for C/C++/Fortran with MPI/OMP/threads.
    - Still testing: **scalability**, fast conditional watchpoints, debugging in TM/SE
    - Working on research license on VEAS. Will present at second ESP workshop.
- Allinea DDT
  - Preparation via ANL scalability research contract on BG/P to address I/O node bottlenecks
    - Multiplexed debug daemons – complete and tested (Nov 2011)
    - Multiplexed gdbserver processes – complete and tested for single threading (Dec 2011)
    - BG/P Beta release (March 2012)
  - Status
    - Exploring performance of multiplexed gdbserver with multiple threads/process.
    - BG/Q port commenced on VEAS. Will present at second ESP workshop in April.

BG/Q Libraries

- ESSL available through IBM
- PETSc is being optimized as part of BG/Q Tools ESP project
- Porting and tuning 3rd party libraries (FFTW, BLAS, LAPACK, ScaLAPACK, ParMETIS, …) using compiler optimizations
- Reporting and fixing library bugs found on BG/Q (e.g., ScaLAPACK 2.0.2)
- Collecting actual library usage data; libraries are stamped with a detectable string id
- Collaborating with Robert van de Geijn's group on rewriting Goto-BLAS so that it can be easily ported and tuned to new architectures like BG/Q (BLIS)
- Exploring an optimized FFT library with Spiral Gen


40


50
45

First in *Mira* Queue: Early Science Program

Science Areas
Astrophysics
Biology
CFD/Aerodynamics
Chemistry
Climate
Combustion
Cosmology
Energy
Fusion Plasma
Geophysics
Materials
Nuclear Structure

- Earliest access to BG/Q platform and special assistance porting and tuning code from ALCF
- 16 projects
  - Large target allocations
  - Postdoc
- 2 billion core-hours to burn in a few months, as machine is brought up

7 National Lab PIs
9 University PIs