Computational Science at the Argonne Leadership Computing Facility

Nichols A. Romero
Assistant Computational Scientist
Argonne Leadership Computing Facility
Argonne National Laboratory

Mission need for Leadership Computing

Leadership computing capability is required for scientists to tackle the highest-resolution, multi-scale/multi-physics simulations of greatest interest and impact to both science and the nation.

For scientific grand challenges, the Leadership Computing Facilities provide capability computing that is 10-100X greater than other computational centers. LCF focus is on big jobs that use a substantial fraction of the systems resources.

Leadership Computing research is mission critical to inform policy decisions and advance innovation in far reaching topics such as:

- energy assurance
- ecological sustainability
- scientific discovery
- global security

"We will respond to the threat of climate change, knowing that the failure to do so would betray our children and future generations." – President Obama 1/21/2013

What is the Leadership Computing Facility?

- Two centers/two architectures to address diverse and growing computational needs of the scientific community.
- Highly competitive user allocation programs (INCITE, ALCC).
- Projects receive computational resources typically 10-100x greater than generally available.
- LCF centers partner with users to enable science & engineering breakthroughs.

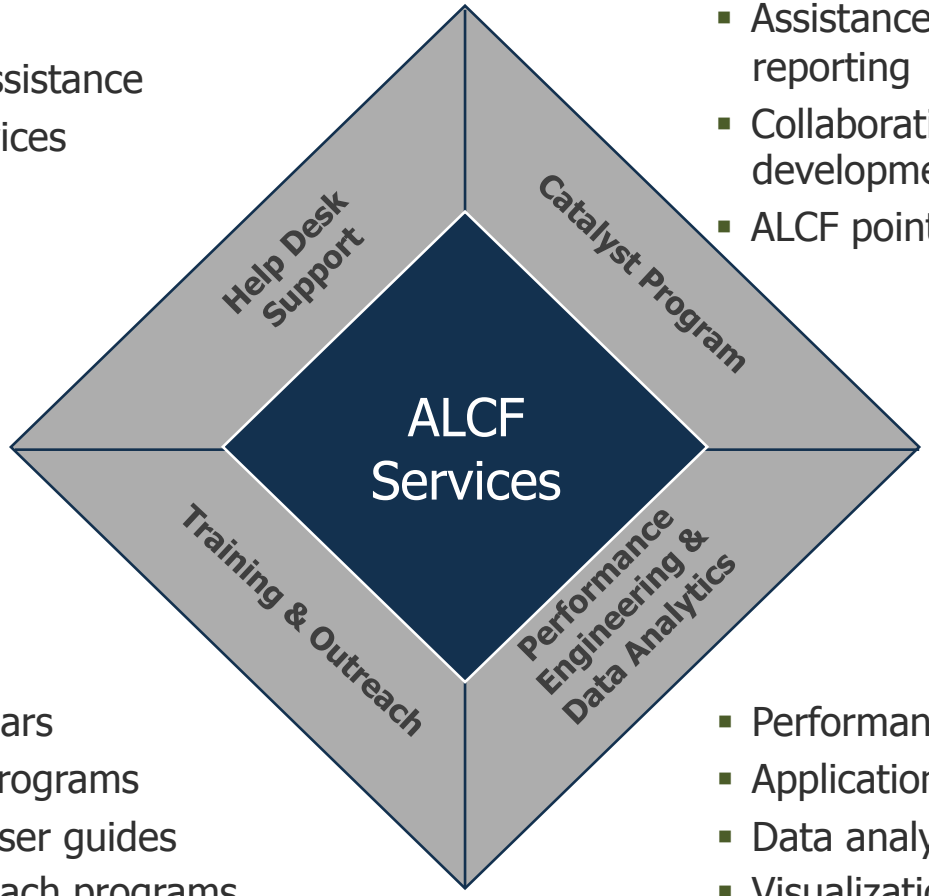
OAK RIDGE National Laboratory

Argonne NATIONAL LABORATORY

Integrated Services Model

- Startup assistance
- User administration assistance
- Job management services
- Technical support

- Science project management
- Assistance with proposals, planning, reporting
- Collaboration on algorithms and development
- ALCF point of coordination



- Workshops and seminars
- Customized training programs
- On-line content and user guides
- Educational and outreach programs
- Reporting and promotion

- Performance engineering
- Application tuning
- Data analytics
- Visualization
- Data management services

The Past - Intrepid

- ***Intrepid* - Blue Gene/P system**
  - 40,960 nodes / 163,840 cores
  - 80 TB memory
  - Peak flop rate: 0.56 PF
  - Linpack flop rate: 0.45 PF
- ***Challenger & Surveyor (T&D)* – BG/P systems**
  - 1k & 1k nodes / 4096 & 4096 cores
  - 2 TB & 2 TB of memory
  - 27.8 TF & 27.8 TF peak flop rate
- ***Eureka* – NVidia S-4 cluster**
  - Primary use: Visualization and data analysis
  - 100 nodes / 800 2.0 GHz Xeon cores
  - 3.2 TB memory
  - 200 NVIDIA FX5600 GPUs
  - Peak flop rate: 100 TF
- **Storage – Data Direct Networks (DDN) storage arrays**
  - 6+ PB capability, 80 GB/s bandwidth (GPFS and PVFS)
  - 14+ PB of archival storage, 10,000 volume tape archive (HPSS)

The Present - Mira

- ***Mira* – BG/Q system**
  - 49,152 nodes / 786,432 cores
  - 768 TB of memory
  - Peak flop rate: 10 PF
  - Linpack flop rate: 8.1 PF
- ***Cetus & Vesta (T&D)* - BG/Q systems**
  - 1K & 2k nodes / 32k & 64k cores
  - 16 TB & 32 TB of memory
  - 210 TF & 419 TF peak flop rate
- ***Tukey* – Nvidia system**
  - 100 nodes / 1600 x86 cores/ 200 M2070 GPUs
  - 6.4 TB x86 memory / 1.2 TB GPU memory
  - Peak flop rate: 220 TF
- **Storage**
  - Scratch: 28.8 PB raw capacity, 240 GB/s bw (GPFS)
  - Home: 1.8 PB raw capacity, 45 GB/s bw (GPFS)
  - Storage upgrade planned in 2015

Evolution from P to Q

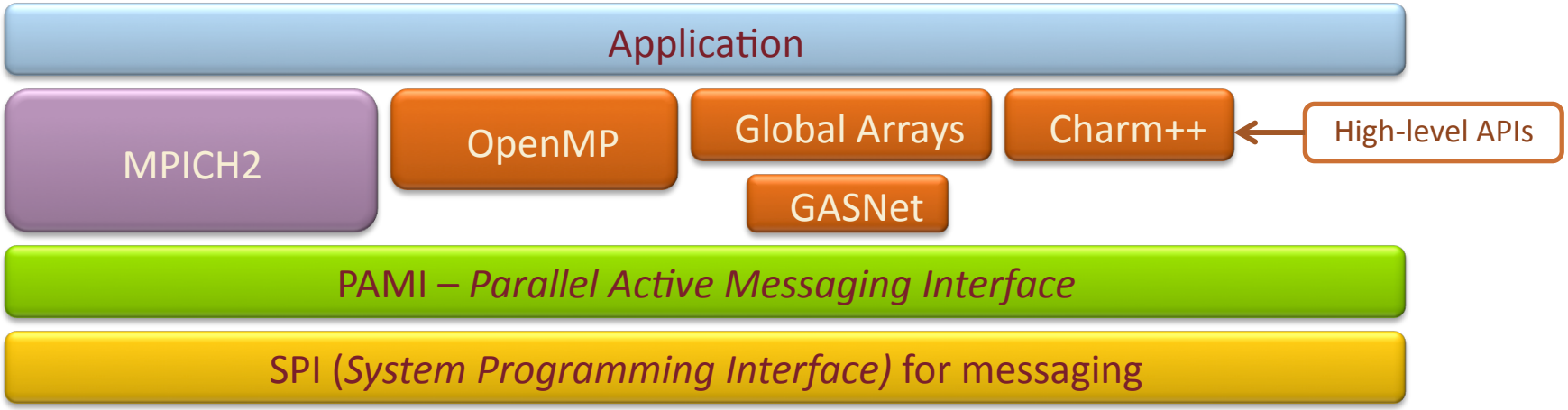
Design Parameters	BG/P	BG/Q	Difference
Cores / Node	4	16	4x
Hardware Threads	1	4	4x
Concurrency / Rack	4,096	65,536	16x
Clock Speed (GHz)	0.85	1.6	1.9x
Flop / Clock / Core	4	8	2x
Flop / Node (GF)	13.6	204.8	15x
RAM / core (GB)	0.5	1	2x
Mem. BW/Node (GB/sec)	13.6	42.6	3x
Latency (MPI zero-length, nearest-neighbor node)	2.6 μs	2.2 μs	~15% less
Bisection BW (32 racks)	1.39TB/s	13.1TB/s	9.42x
Network	3D Torus + Collectives	5D Torus	Smaller diameter
GFlops/Watt	0.77	2.10	3x
Instruction Set	32 bit PowerPC + DH	64 bit PowerPC + QPX	New vector instructions
Programming Models	MPI + OpenMP	MPI + OpenMP	
Cooling	Air	Water	

Node/node comparison details

	BG/P	BG/Q	RATIO
Cores / Node	4	16	4x
Clock Speed (GHz)	0.85	1.6	1.9x
Flop / Clock / Core	4	8	2x
Flops/core	3.4 GF	12.8 GF	3.8x
Flops/node	13.6	204.8	15.1x
Nodes / Rack	1,024	1,024	1x
Flops / Rack	13.9 TF	210 TF	15.1x

Programming and Running on BG/Q

- **MPI**
- **Threads: OpenMP, PTHREADS**
- **QPX intrinsics: vec_ld, vec_add, vec_madd, ….**
- **Topology interfaces**
  - E.g. MPIX_* functions
- **Run modes: combinations of**
  - MPI ranks/node = {1,2,4,…,64}
  - Threads/node = {1,2,4,…,64}

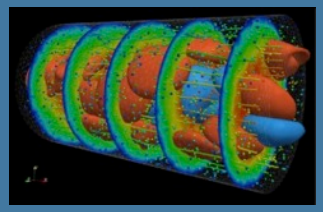


Allocation Programs at the LCFs

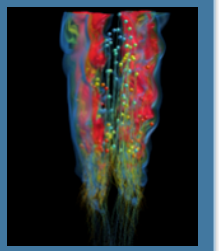
	60% INCITE		30% ALCC		10% Director's Discretionary	
Mission	High-risk, high-payoff science that requires LCF-scale resources*		High-risk, high-payoff science aligned with DOE mission		Strategic LCF goals	
Call	1x/year – (Closes June) ***2014 Call Open Now***		1x/year – (Closes February)		Rolling	
Duration	1-3 years, yearly renewal		1 year		3m,6m,1 year	
Typical Size	30 - 40 projects	50M - 500M core-hours/yr.	5 - 10 projects	10M – 300+M core-hours/yr.	100s of projects	.5M – 10M core-hours
Review Process	Scientific Peer-Review	Computational Readiness	Scientific Peer-Review	Computational Readiness	Strategic impact and feasibility	
Managed By	INCITE management committee (ALCF & OLCF)		DOE Office of Science		LCF management	
Readiness	**High**		**Medium to High**		**Low to High**	
Availability	Open to all scientific researchers and organizations ***Capability > 131,072 cores (16.7% of Mira)***					

Diversity of INCITE science

Simulating a flow of healthy (red) and diseased (blue) blood cells with a Dissipative Particle Dynamics method.
- George Karniadakis, Brown University

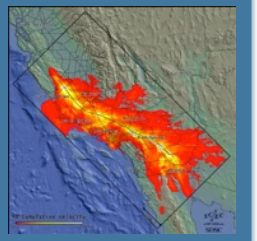


Provide new insights into the dynamics of turbulent combustion processes in internal-combustion engines.
-Jacqueline Chen and Joseph Oefelein, Sandia National Laboratories

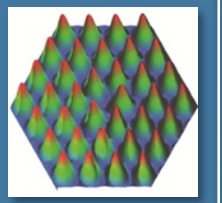


Demonstration of high-fidelity capture of airfoil boundary layer, an example of how this modeling capability can transform product development.
- Umesh Paliath, GE Global Research

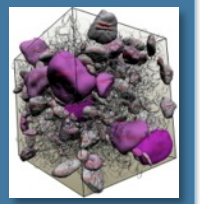
Calculating an improved probabilistic seismic hazard forecast for California.
- Thomas Jordan, University of Southern California



Modeling charge carriers in metals and semiconductors to understand the nature of these ubiquitous electronic devices.
- Richard Needs, University of Cambridge, UK



High-fidelity simulation of complex suspension flow for practical rheometry.
- William George, National Institute of Standards and Technology



Other INCITE research topics

- Glimpse into dark matter
- Supernovae ignition
- Protein structure
- Creation of biofuels
- Replicating enzyme functions
- Global climate
- Regional earthquakes
- Carbon sequestration
- Turbulent flow
- Propulsor systems
- Membrane channels
- Protein folding
- Chemical catalyst design
- Combustion
- Algorithm development
- Nano-devices
- Batteries
- Solar cells
- Reactor design
- Nuclear structure

Scientific Support is Collaboration

The ALCF is staffed with a team of computational scientists, expert in their domain, scalable algorithms and performance engineering.

- **Provide a "jump-start" in the use of ALCF resources**
- **Align the availability of ALCF resources with the needs of the project team**
- **Collaborate to maximize the value that ALCF can bring to our projects**
- **Connect the needs of the scientific community with future and current hardware**

Two categories of collaboration and contribution to teams using the ALCF:

Tactical/Collaborative

- Short term, fast solutions
  - Compiling, Debugging, System Use
- Targeted problem resolution
  - Resolve a specific hard problem like restructuring I/O
- Long term collaborations
  - In depth work on a code that be over a long period of time
  - Constrained by staff

Strategic

- Training
  - Postdocs, students, community
- Understand HPC needs for different communities
- Plan for future needs
  - Help planning new facilities
  - Advise/Participate in long term code development paths

Acknowledgements

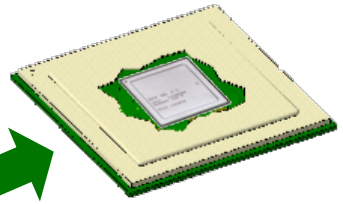
Most of the slides borrowed from Paul Messina, Scott Parker, and Katherine Riley.

Extra Slides on BG/Q hardware

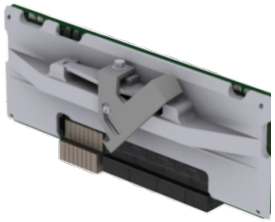
Blue Gene/Q

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



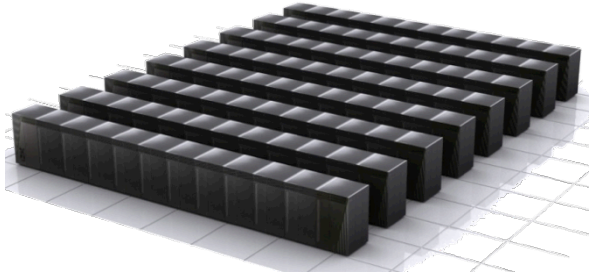
4. Node Card:
32 Compute Cards,
Optical Modules, Link Chips; 5D Torus

5a. Midplane:
16 Node Cards

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

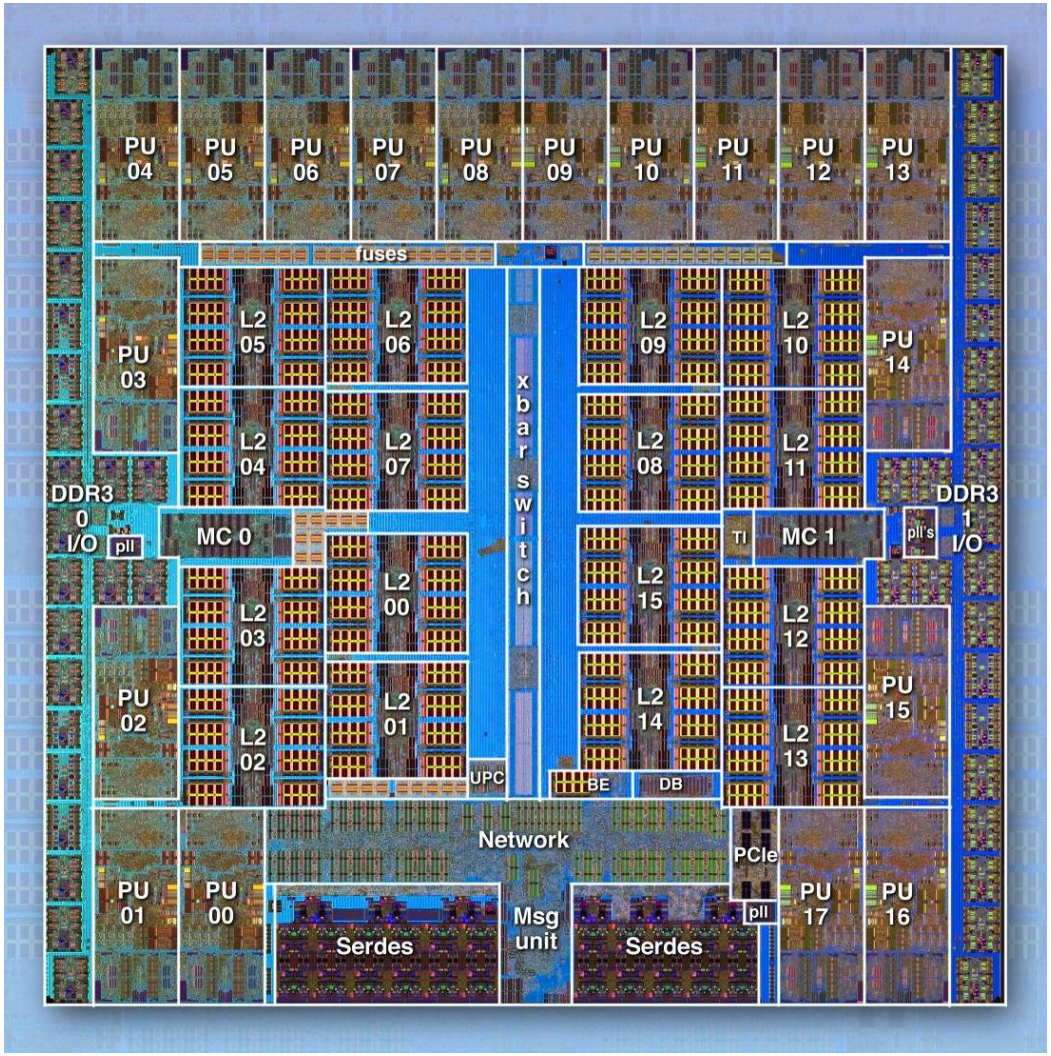
6. Rack: 2 Midplanes

7. System:
96 racks, 20PF/s



BlueGene/Q Compute Chip

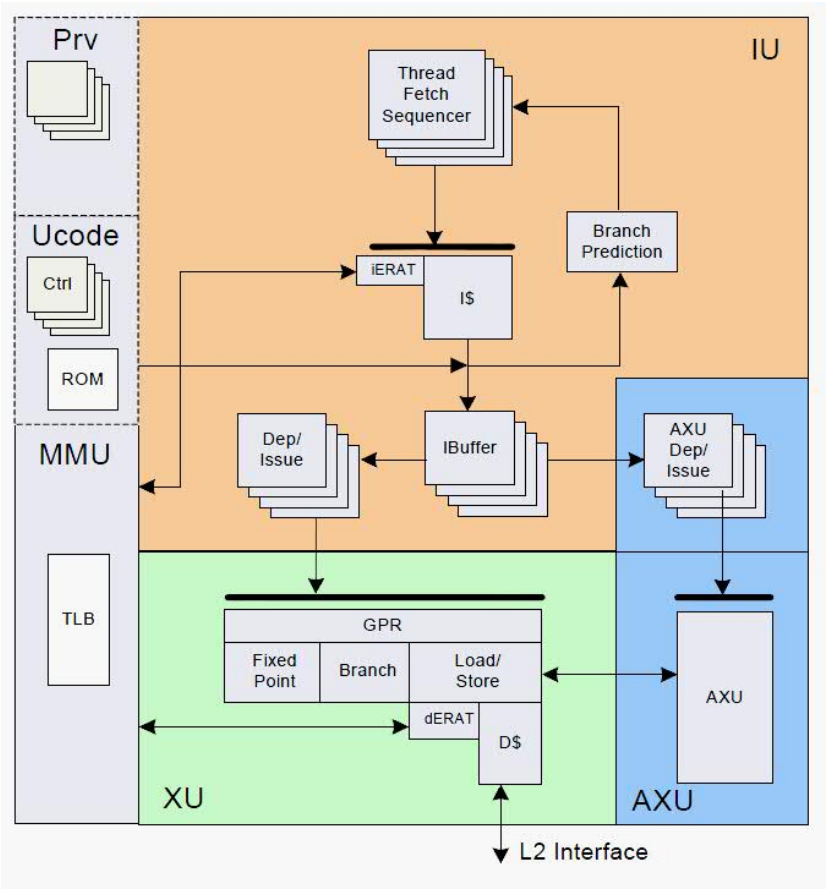
System-on-a-Chip design : integrates processors, memory and networking logic into a single chip



- **360 mm² Cu-45 technology (SOI)**
  - ~ 1.47 B transistors
- **16 user + 1 service processors**
  - 16 compute cores
  - 17th core for system functions (OS, RAS)
  - plus 1 redundant processor
  - all processors are symmetric
  - L1 I/D cache = 16kB/16kB
  - L1 prefetch engines
- **Crossbar switch**
  - Connects cores via L1P to L2 slices
  - Aggregate read rate of 409.6 GB/s
- **Central shared L2 cache**
  - 32 MB eDRAM
  - 16 slices
- **Dual memory controller**
  - 16 GB external DDR3 memory
  - 42.6 GB/s bandwidth
- **Chip-to-chip networking**
  - Router logic integrated into BQC chip
  - DMA, remote put/get, collective operations
  - 11 network ports
- **External IO**
  - PCIe Gen2 interface

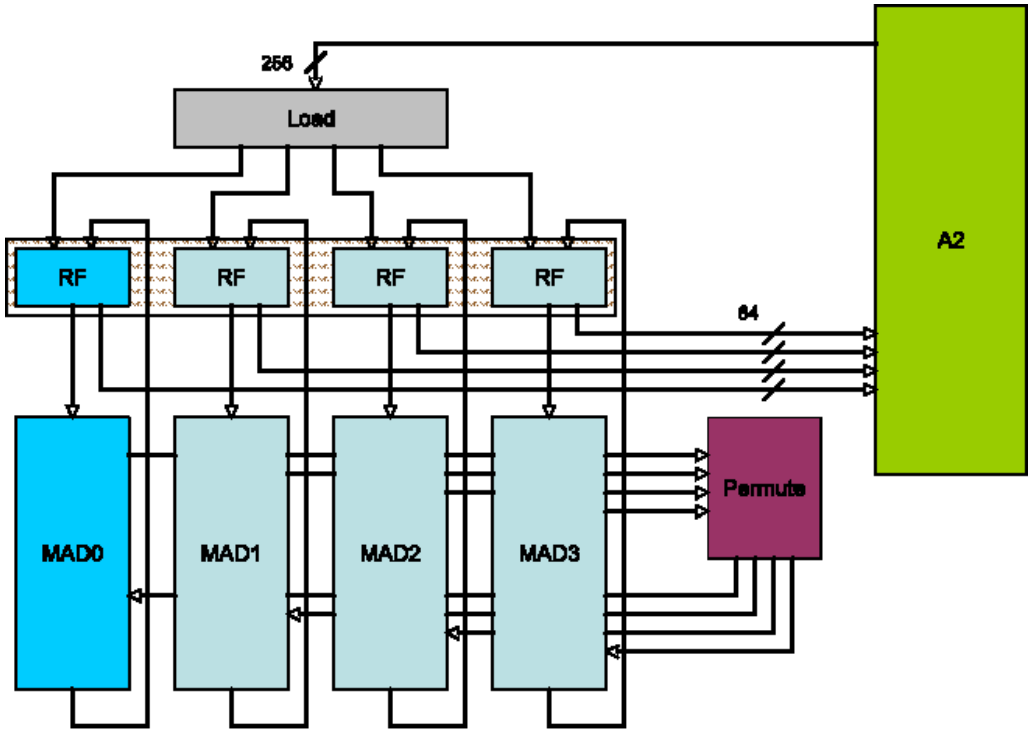
BG/Q Core

- In-order execution
- 4-way Simultaneous Multi-Threading
- Dynamic branch prediction
- 32 64 bit integer registers, 32 256 bit FP registers
- Functional Units:
  - IU – instructions fetch and decode
  - XU – Branch, Integer, Load/Store instructions
  - AXU – Floating point instructions
    - Standard PowerPC instructions
    - QPX 4 wide SIMD
  - MMU – memory management (TLB)
- Instruction Issue:
  - 2-way concurrent issue 1 XU + 1 AXU
  - A given thread may only issue 1 instruction per cycle
  - Two threads may issue 1 instruction each cycle

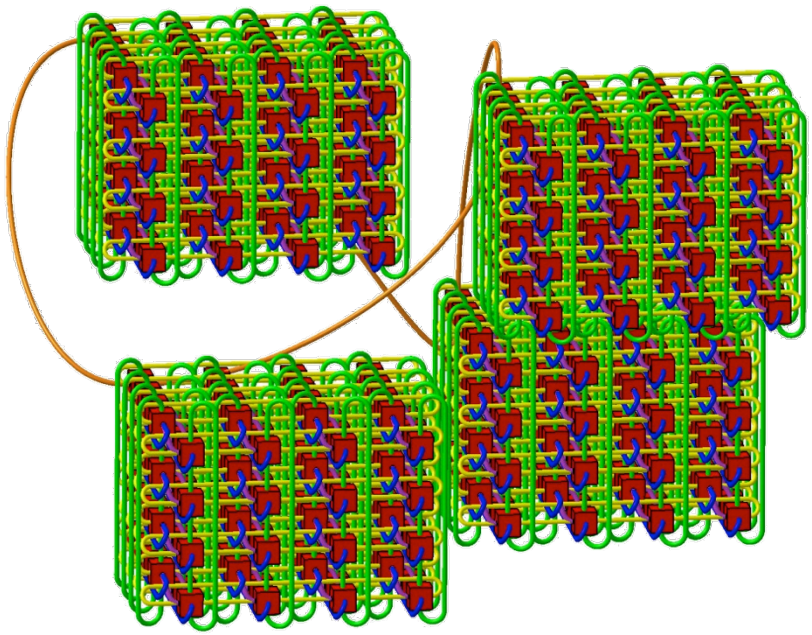


QPX Overview

- Unique 4 wide double precision SIMD instructions extending standard PowerISA with:
  - Full set of arithmetic functions
  - Load/store instructions
  - Permute instructions to reorganize data
- 4 wide FMA instructions allow 8 flops/inst
- FPU operates on:
  - Standard scale PowerPC FP instructions (slot 0)
  - 4 wide SIMD instructions
  - 2 wide complex arithmetic SIMD arithmetic
- Standard 64 bit floating point registers are extended to 256 bits
- Attached to AXU port of A2 core – A2 issues one instruction/cycle to AXU
- 6 stage pipeline
- 32B (256 bits) data path to/from L1 cache
- Compiler can generate QPX instructions
- Intrinsic functions mapping to QPX instructions allow easy QPX programming



Inter-Processor Communication



Network Performance

- All-to-all: 97% of peak
- Bisection: > 93% of peak
- Nearest-neighbor: 98% of peak
- Collective: FP reductions at 94.6% of peak
- On chip per hop latency ~40 ns
- Allreduce hardware latency on 96k nodes ~ 6.5 *us*
- Barrier hardware latency on 96k nodes ~ 6.3 *us*

- **5D torus network:**
  - Virtual cut-through routing with Virtual Channels to separate system and user messages
  - 5D torus achieves high nearest neighbor bandwidth while increasing bisectional bandwidth and reducing hops
  - Allows machine to be partitioned into independent sub machines. No impact from concurrently running codes.
  - Hardware assists for collective & barrier functions over COMM_WORLD and rectangular sub communicators
  - Half rack (midplane) is 4x4x4x4x2 torus
  - Last dimension is always 2
- **No separate Collectives or Barrier network:**
  - Single network used for point-to-point, collectives, and barrier operations
- **Nodes have 10 links with 2 GB/s raw bandwidth each**
  - Bi-directional: send + receive gives 4 GB/s
  - 90% of bandwidth (1.8 GB/s) available to user
- **Additional 11th link for communication to IO nodes**
- **Optical links between midplanes, electrical inside midplane**
- **Hardware latency**
  - Nearest: 80ns
  - Farthest: 3us (96-rack 20PF system, 31 hops)