

THETA, AND THE FUTURE OF ACCELERATOR PROGRAMMING AT ARGONNE



SCOTT PARKER

Lead, Performance Engineering Team
Argonne Leadership Computing Facility

July 29, 2019

ARGONNE HPC TIMELINE

- 2005:
 - Argonne accepts 1 rack (1024 nodes) of Blue Gene/L (5.6 TF)
- 2008:
 - ALCF accepts 40 racks (160k cores) of Blue Gene/P (557 TF)
- 2012:
 - 48 racks of Mira Blue Gene/Q (10 PF) in production at ALCF
- 2016:
 - ALCF accepts Theta (12 PF) Cray XC40 with Xeon Phi (KNL)
- 2021:
 - One Exaflop Aurora Intel/Cray GPU machine to be delivered



PERFORMANCE FROM PARALLELISM

- Parallelism across nodes (using MPI, etc.)
- Parallelism across sockets within a node
- Parallelism across cores within each socket
- Parallelism across pipelines within each core (i.e. instruction-level parallelism)
- Parallelism across vector lanes within each pipeline (i.e. SIMD)
- Using instructions that perform multiple operations simultaneously (e.g. FMA)

THETA

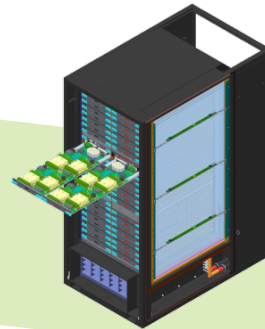
- **System:**
 - Cray XC40 system
 - 4,392 compute nodes/ 281,088 cores
 - 11.7 PetaFlops peak performance
 - Accepted Fall 2016
- **Processor:**
 - Intel Xeon Phi, 2nd Generation (Knights Landing) 7230
 - 64 Cores
 - 1.3 GHz base / 1.1 GHz AVX / 1.4-1.5 GHz Turbo
- **Memory:**
 - 892 TB of total system memory
 - 16 GB MCDRAM per node
 - 192 GB DDR4-2400 per node
- **Network:**
 - Cray Aries interconnect
 - Dragonfly network topology
- **Filesystems:**
 - Project directories: 10 PB Lustre file system
 - Home directories: GPFS



THETA SYSTEM OVERVIEW

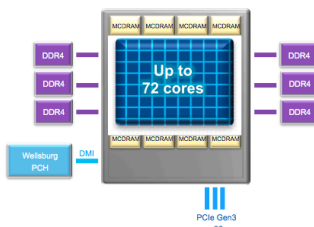
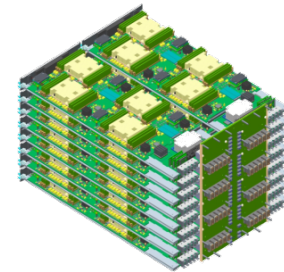


System: 24 Cabinets
4,392 Nodes, 1152 Switches
12 groups, Dragonfly
11.7 PF Peak
68.6 TB MCDRAM, 823.5 TB DRAM

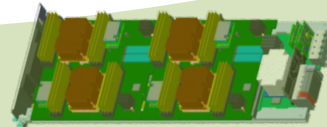


Cabinet: 3 Chassis
184 Nodes
510.72 TF
3TB MCDRAM, 36TB DRAM

Chassis: 16 Blades
64 Nodes, 16 Switches
170.24 TF
1TB MCDRAM, 12TB DRAM



Node: KNL Socket
2.66 TF
16GB MCDRAM, 192 GB DDR4 (6 channels)

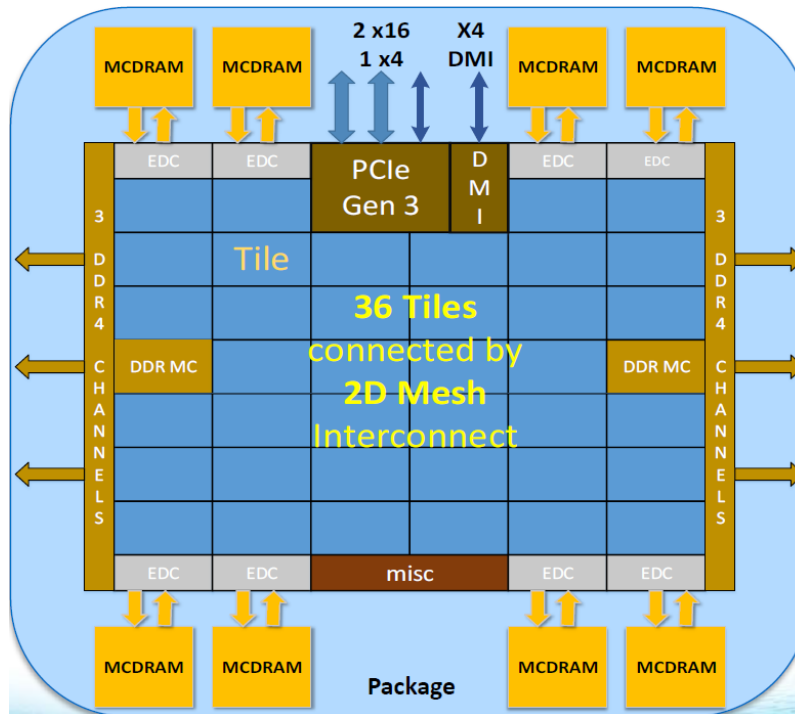


Compute Blade:
4 Nodes/Blade + Aries switch
10.64 TF
64GB MCDRAM, 768GB DRAM
128GB SSD



Sonexion Storage
4 Cabinets
Lustre file system
10 PB usable
210 GB/s

KNIGHTS LANDING PROCESSOR



- Chip**
- 683 mm²
 - 14 nm process
 - 8 Billion transistors

- Up to 72 Cores**
- 36 tiles
 - 2 cores per tile
 - Up to 3 TF per node

- 2D Mesh Interconnect**
- Tiles connected by 2D mesh

- On Package Memory**
- 16 GB MCDRAM
 - 8 Stacks
 - 485 GB/s bandwidth

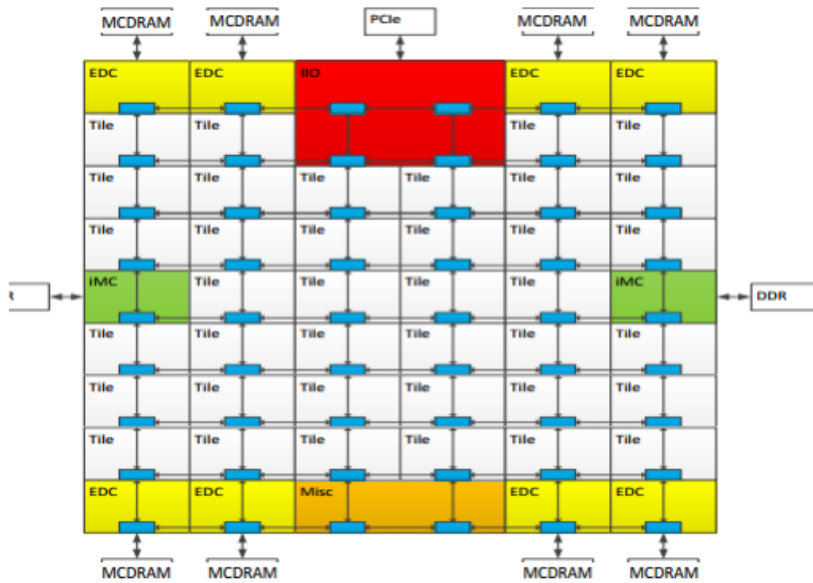
- 6 DDR4 memory channels**
- 2 controllers
 - up to 384 GB external DDR4
 - 90 GB/s bandwidth

- On Socket Networking**
- Omni-Path NIC on package
 - Connected by PCIe

Knights Landing Features

Feature	Impact
Self Booting	No PCIe bottleneck
Binary Compatible with Xeon	Runs legacy code, no recompile
Atom Based Core Architecture	~3x higher performance than KNC
High. Vector Density	3+ TFlops (DP) Peak per chip
AVX-512 ISA	New 512 bit vector ISA with Masks
Gather/Scatter Engine	Hardware support for gather/scatter
MCDRAM + DDR memory	High bandwidth MCDRAM, large capacity DDR
2D mesh. on-die interconnect	High bandwidth connection between cores and memory
Integrated Omni-path Fabric	Better scalability at lower cost

KNL Mesh Interconnect



- 2D mesh interconnect connects
 - Tiles (CHA)
 - MCDRAM controllers
 - DDR controllers
 - Off chip I/O (PCIe, DMI)
- YX routing:
 - Go in Y → turn → Go in X
 - Messages arbitrate on injection and on turn
- Cache coherent
 - Uses MESIF protocol
- Clustering mode allow traffic localization
 - All-to-all, Quadrant, Sub-NUMA

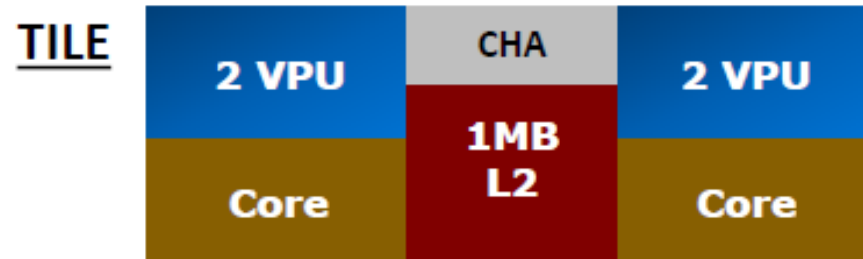
OPENMP OVERHEADS

EPCC OpenMP Benchmarks

Threads	Barrier (µs)	Reduction (µs)	Parallel For (µs)
1	0.1	0.7	0.6
2	0.4	1.3	1.3
4	0.8	1.9	1.9
8	1.5	2.7	2.5
16	1.8	5.9	2.9
32	2.8	7.7	4.0
64	3.9	10.4	5.6
128	5.3	13.7	7.3
256	7.8	19.4	10.5

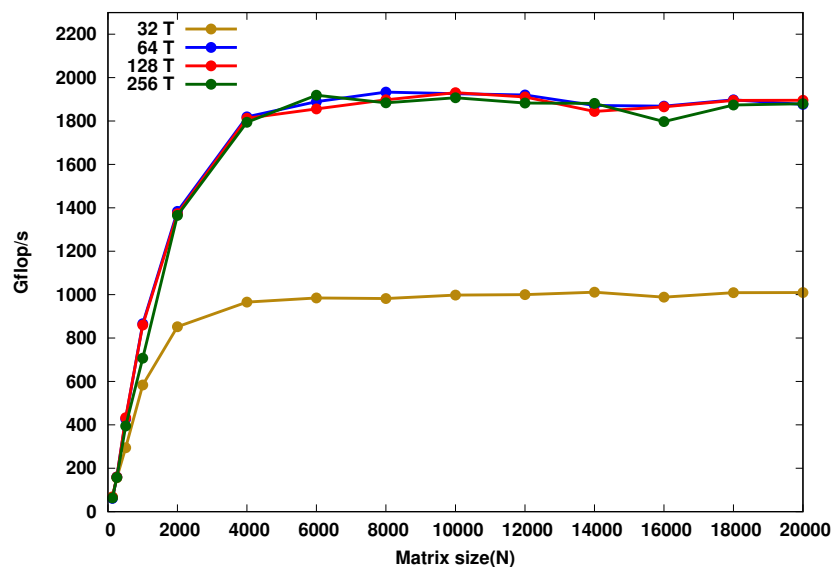
- OpenMP costs related to cost of memory access
 - KNL has no shared last level cache
- Operations can take between 130 – 25,000 cycles
- Cost of operations increases with thread count
 - Scales as $\sim C \cdot \text{threads}^{1/2}$

KNL TILE



- Two CPUs
- 2 vector units (VPUs) per core
- 1 MB Shared L2 cache
 - Coherent across all tiles (32-36 MB total)
 - 16 Way
 - 1 line read and $\frac{1}{2}$ line write per cycle
- Caching/Home agent
 - Distributed tag directory, keeps L2s coherent
 - Implements MESIF cache coherence protocol
 - Interface to mesh

DGEMM PERFORMANCE ON THETA

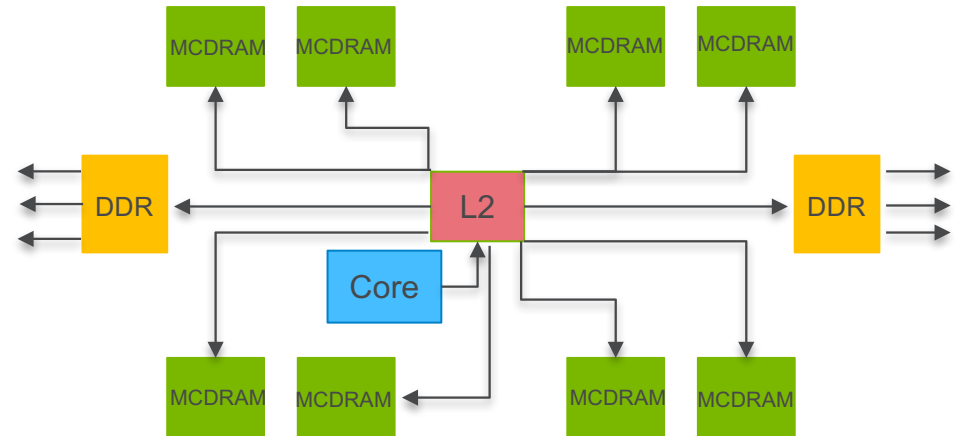


MKL DGEMM Performance

- Peak FLOP rate per node on Theta: 2252.8 Gflops
 - 64 cores
 - 2 Vector pipelines, 8 Wide Vectors, FMA instruction (2 flops)
 - AVX frequency 1.1 GHz
- MKL DGEMM:
 - Peak flop rate: 1945.67 Gflops
 - 86.3% of peak
- Thread scaling:
 - Linear scaling with cores
 - More than 1 hyperthread per core does not increase performance
- Floating point performance is limited by AVX frequency
 - AVX vector frequency is lower than TDP frequency (1.3 GHz)
 - Frequency drops for sustained series of AVX512 instructions

MEMORY

- **Two memory types**
 - In Package Memory (IPM)
 - 16 GB MCDRAM
 - ~485 GB/s bandwidth
 - Off Package Memory (DDR)
 - Up to 384 GB
 - ~90 GB/s bandwidth
- **One address space**
 - Minor NUMA effects
 - Sub-NUMA clustering mode creates four NUMA domains



MEMORY MODES - IPM AND DDR

SELECTED AT NODE BOOT TIME

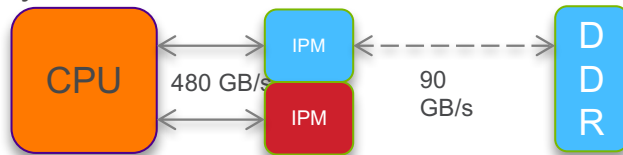
Cache



Flat



Hybrid



- **Memory configurations**

- **Cached:**

- DDR fully cached by IPM
- No code modification required
- Less addressable memory
- Bandwidth and latency worse than flat mode

- **Flat:**

- Data location completely user managed
- Better bandwidth and latency
- More addressable memory

- **Hybrid:**

- $\frac{1}{4}$, $\frac{1}{2}$ IPM used as cache rest is flat

- **Managing memory:**

- jemalloc & memkind libraries
- numctl command
- Pragmas for static memory allocations

STREAM TRIAD BENCHMARK PERFORMANCE

- Measuring and reporting STREAM bandwidth is made more complex due to having MCDRAM and DDR
- Memory bandwidth depends on
 - Mode: flat or cache
 - Physical memory: mcdram or ddr
 - Store type: non-temporal streaming vs regular
- Peak STREAM Triad bandwidth occurs in Flat mode with streaming stores:
 - from MCDRAM, 485 GB/s
 - from DDR, 88 GB/s

Case	GB/s with SS	GB/s w/o SS
Flat, MCDRAM	485	346
Flat, DDR	88	66
Cache, MCDRAM	352	344
Cache, DDR	59	67

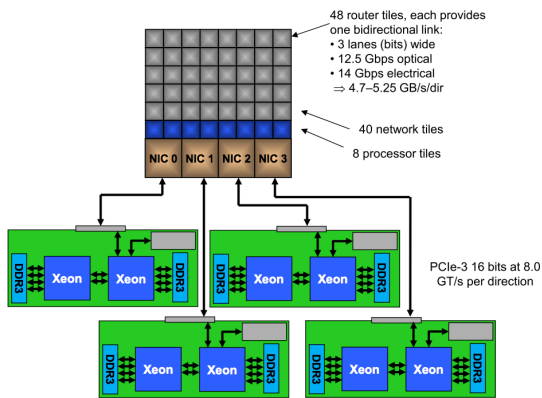
MEMORY LATENCY

	Cycles	Nano seconds
L1 Cache	4	3.1
L2 Cache	20	15.4
MCDRAM	220	170
DDR	180	138

ARIES DRAGONFLY NETWORK

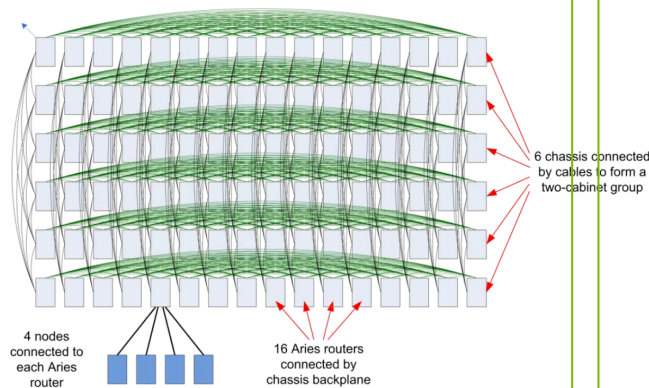
Aries Router:

- 4 Nodes connect to an Aries router
- 4 NIC's connected via PCIe
- 40 Network tiles/links
- 4.7-5.25 GB/s/dir per link



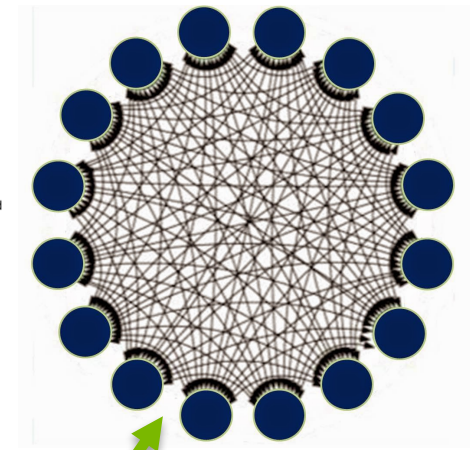
Connections within a group:

- 2 Local all-to-all dimensions
 - 16 all-to-all horizontal
 - 6 all-to-all vertical
- 384 nodes in local group



Connectivity between groups:

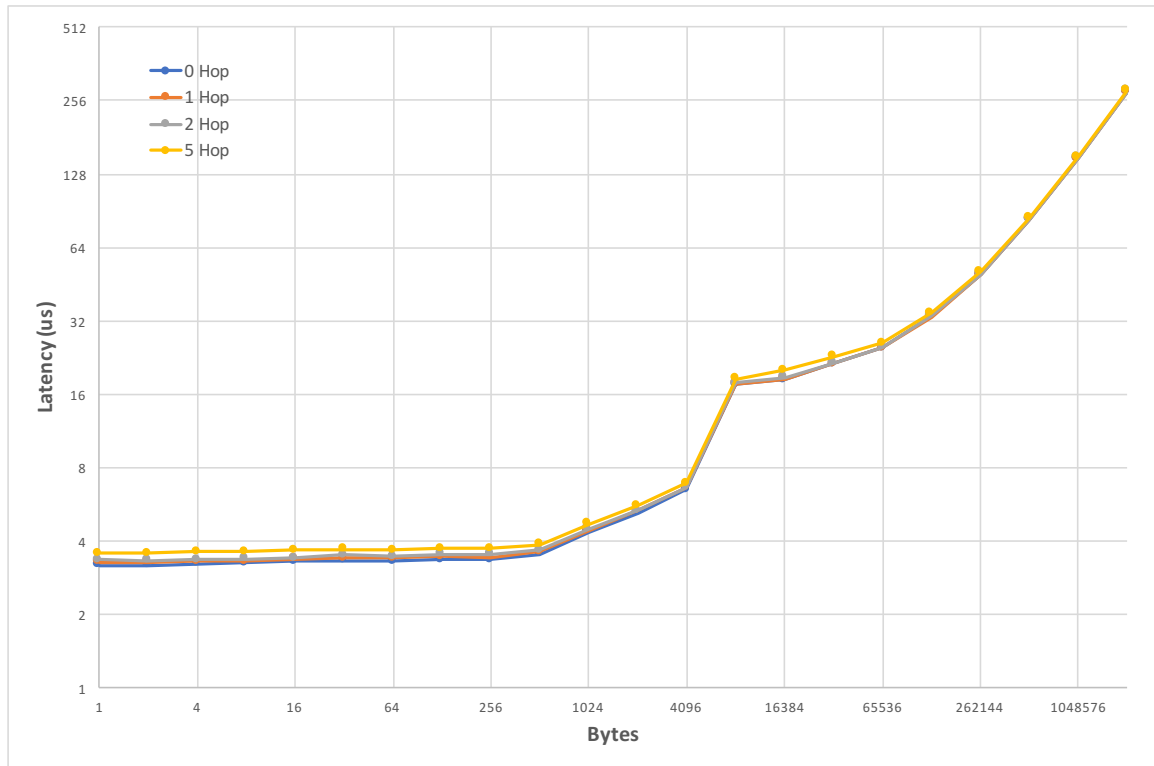
- Each group connected to every other group
- Restricted bandwidth between groups



Theta has 12 groups with 12 links between each group

MPI SEND AND RECEIVE LATENCY

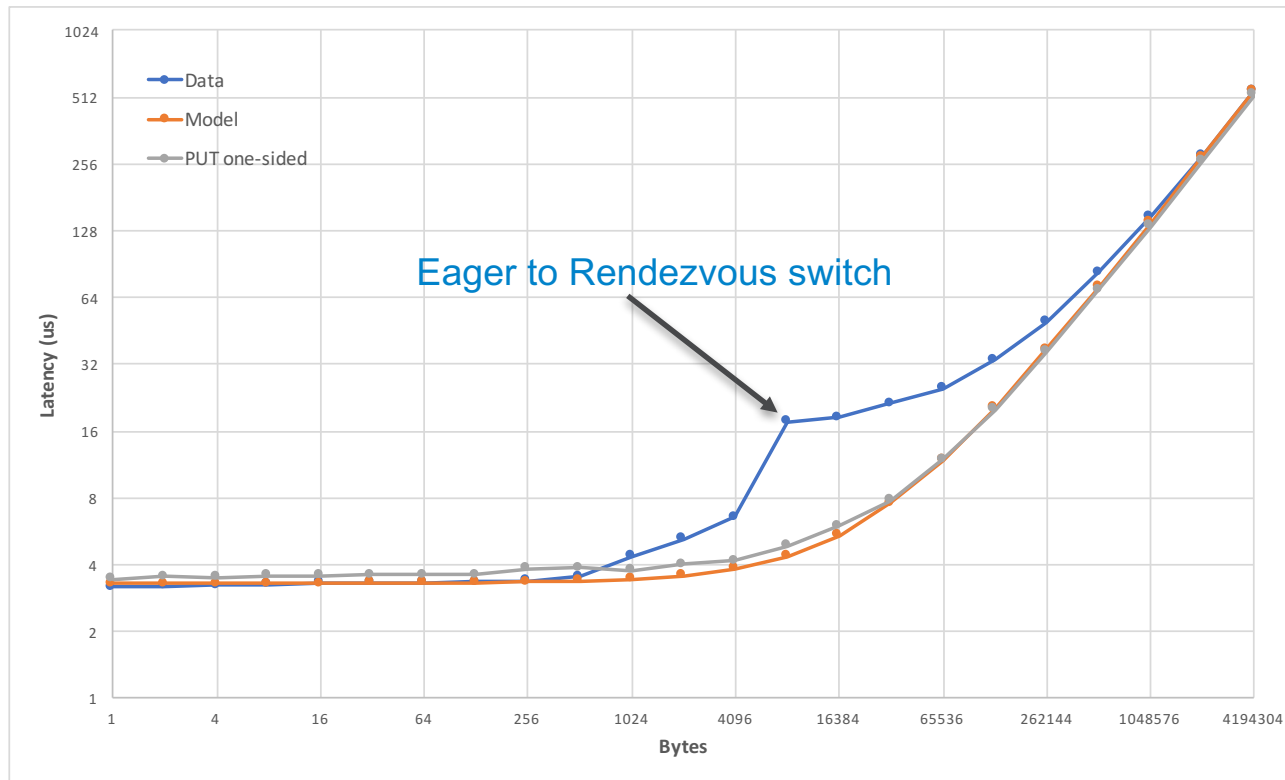
OSU PtoP MPI Latency on Theta



- Latency tested for pairs placed different distances or hops apart
 - 0 – on same Aries
 - 1 – same row/col
 - 2 – same groups
 - 5 – between groups
- Hop count does not strongly influence latency

MPI SEND AND RECEIVE MODEL

OSU PtoP MPI Latency on Theta



Simple (Hockney) model:

$$T = \alpha + \beta \cdot n$$

$$n = \text{bytes}$$

$$\alpha = 3.3$$

$$\beta = 0.0013$$

Model fits well for low and high byte counts

Eager to rendezvous protocol switch believed to be producing “bump” in latency

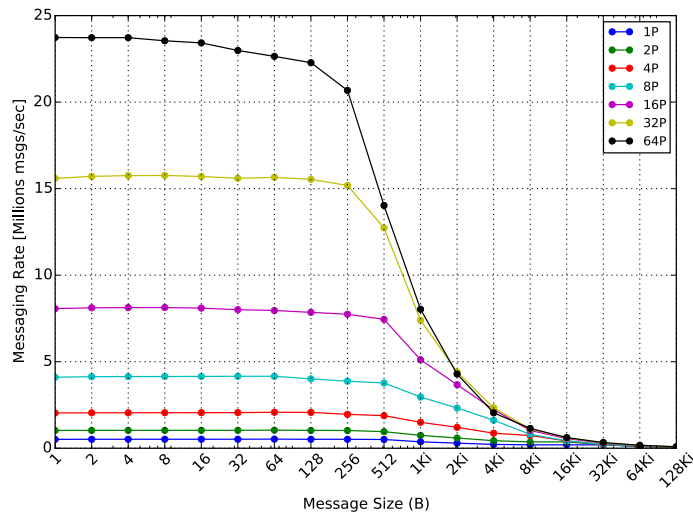
One sided PUT latency results lack “bump” and are close to the model

MPI BANDWIDTH AND MESSAGING RATE

OSU PtoP MPI Multiple Bandwidth / Message Rate Test on Theta

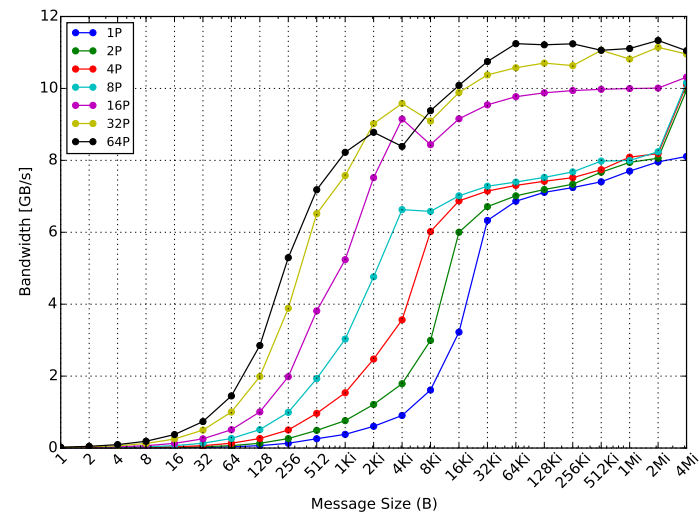
Messaging Rate:

- Maximum rate of 23.7 MMPS
 - At 64 ranks per node, 1 byte, window size 128
- Increases generally proportional to core count for small message sizes

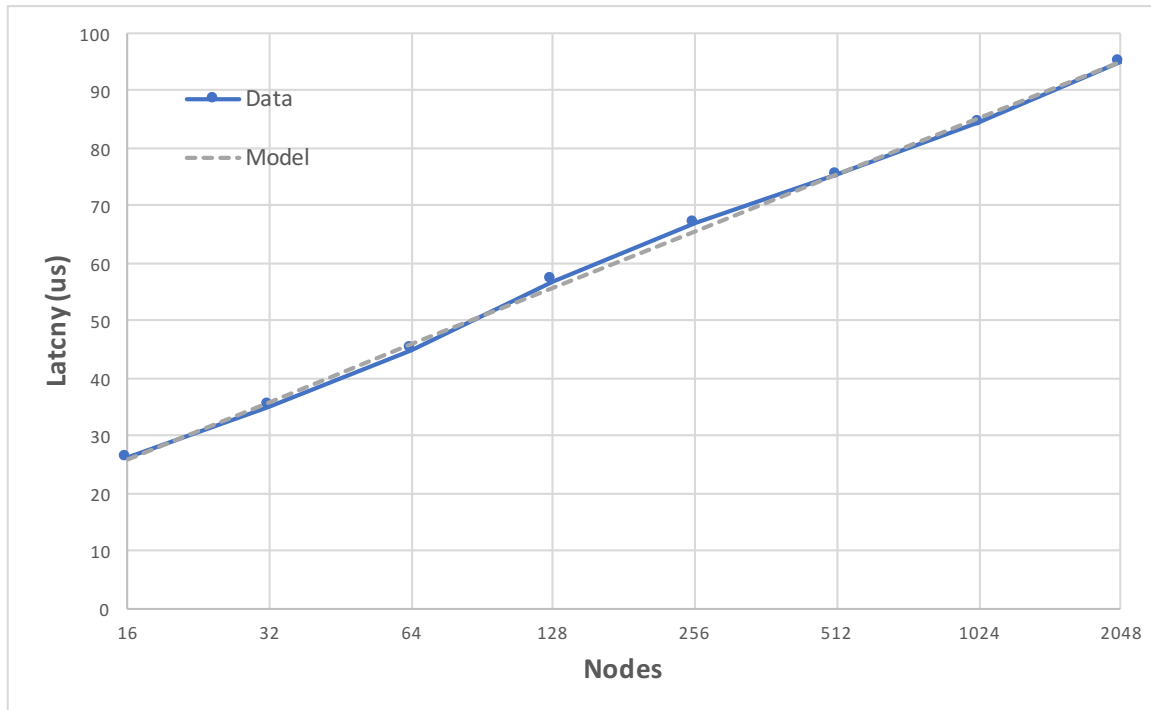


Bandwidth:

- Peak sustained bandwidth of 11.4 GB/s to nearest neighbor
- 1 rank capable of 8 GB/s
- For smaller messages more ranks improve aggregate off node bandwidth



MPI BARRIER MODEL



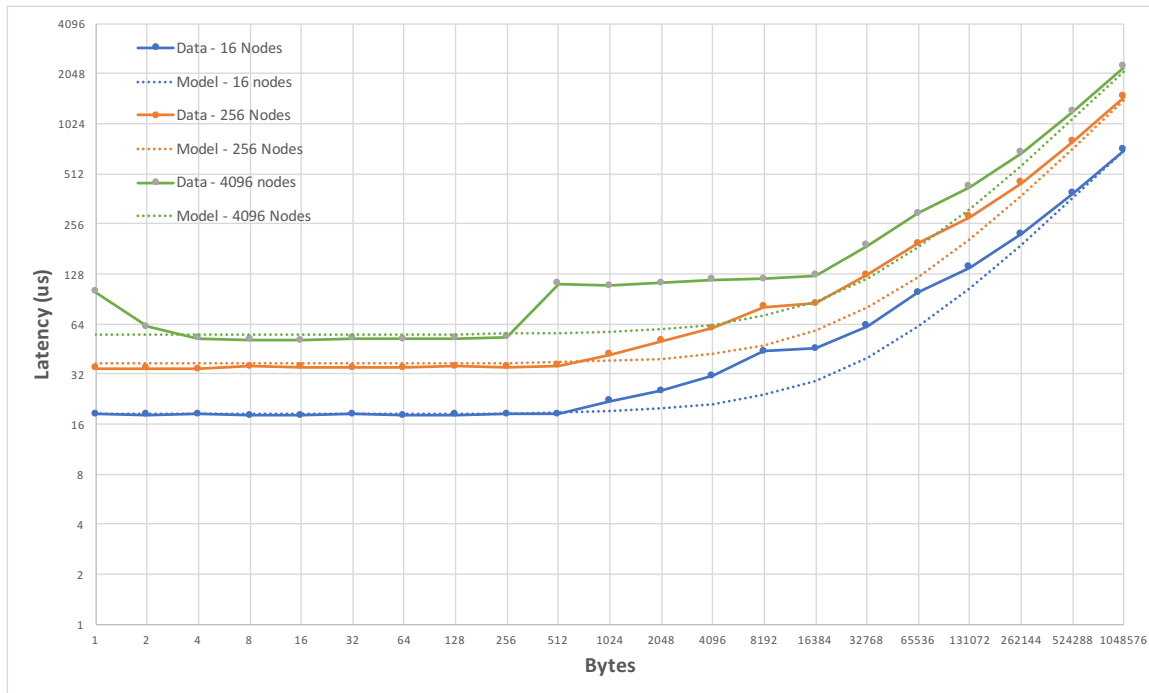
$$T = \alpha + \beta \cdot \log_2(p)$$

$$p = \text{nodes}$$

$$\alpha = -13.5$$

$$\beta = 9.87$$

MPI BROADCAST MODEL



$$T = (\alpha + \beta \cdot n) \text{Log}_2(p)$$

$n = \text{bytes}$

$p = \text{nodes}$

$\alpha = 4.6$

$\beta = 0.0016$

Good fit at low and high byte ranges.

Errors centered around point of protocol switch

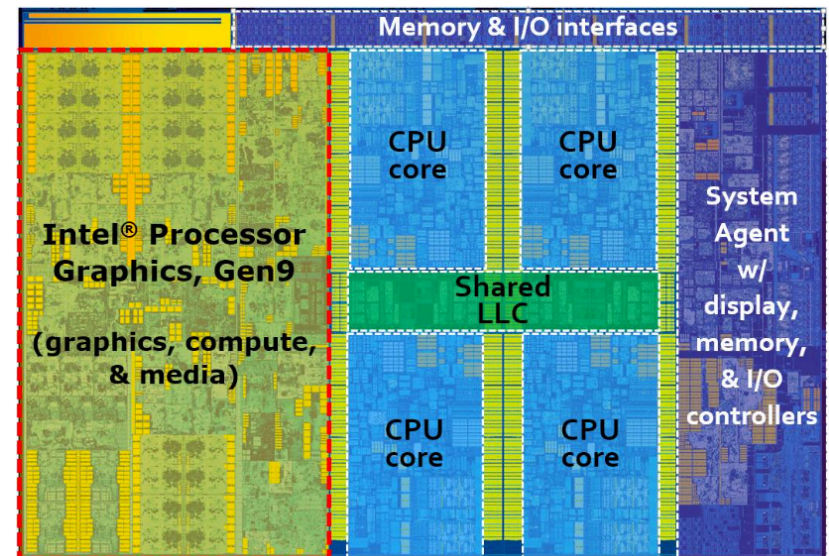
AURORA



- 1+ ExaFlop system
- Arriving at Argonne in 2021
- Intel Xeon processors + Intel X^e GPUs
- Greater than 10 PB of total memory
- Cray Slingshot network

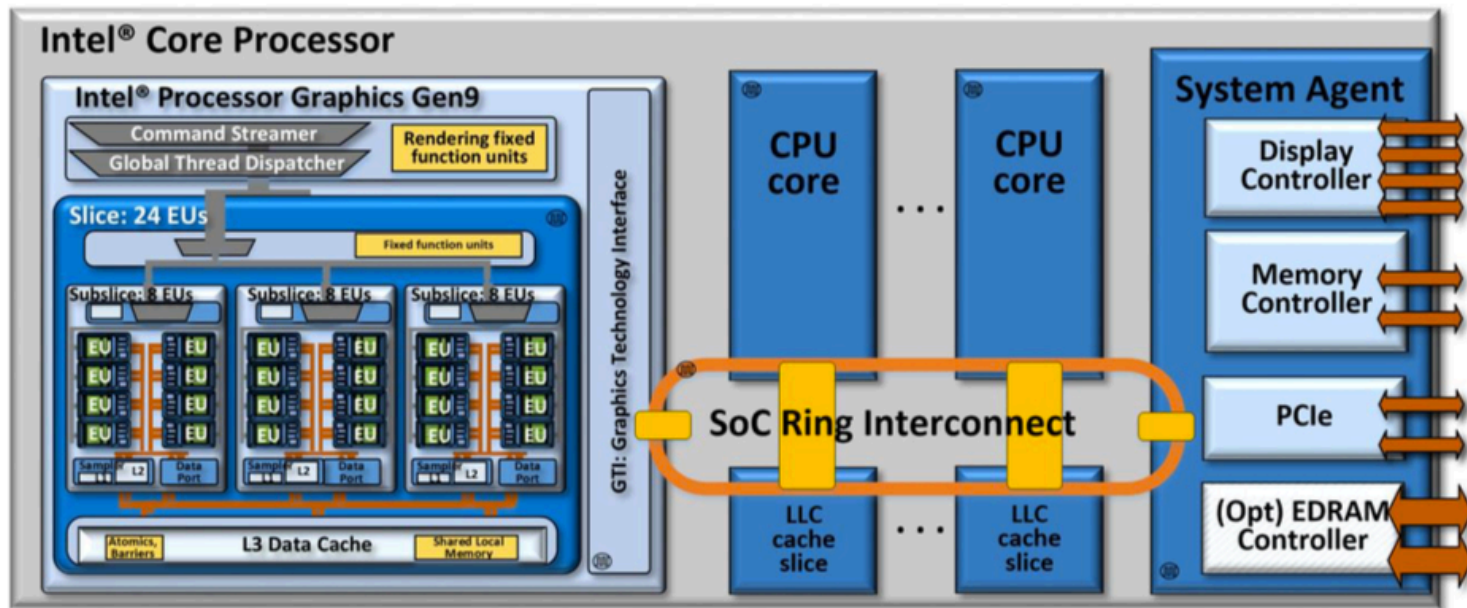
INTEL GPUS

- Intel has been building GPUs integrated with CPUs for over a decade
- Currently released products use the “Gen 9” version
- Soon to be released is “Gen 11”
- After that come the X^e (Gen 12) line of integrated and discrete GPUs



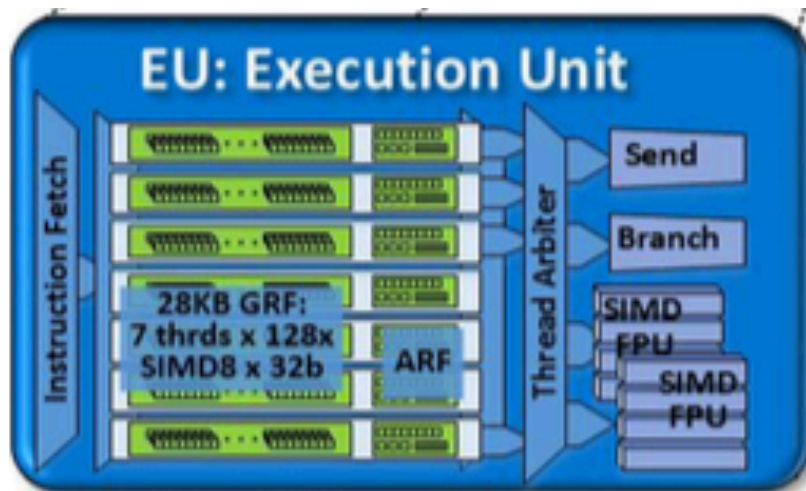
Architecture components layout for an Intel Core i7 processor 6700K for desktop systems.

INTEL INTEGRATED GRAPHICS



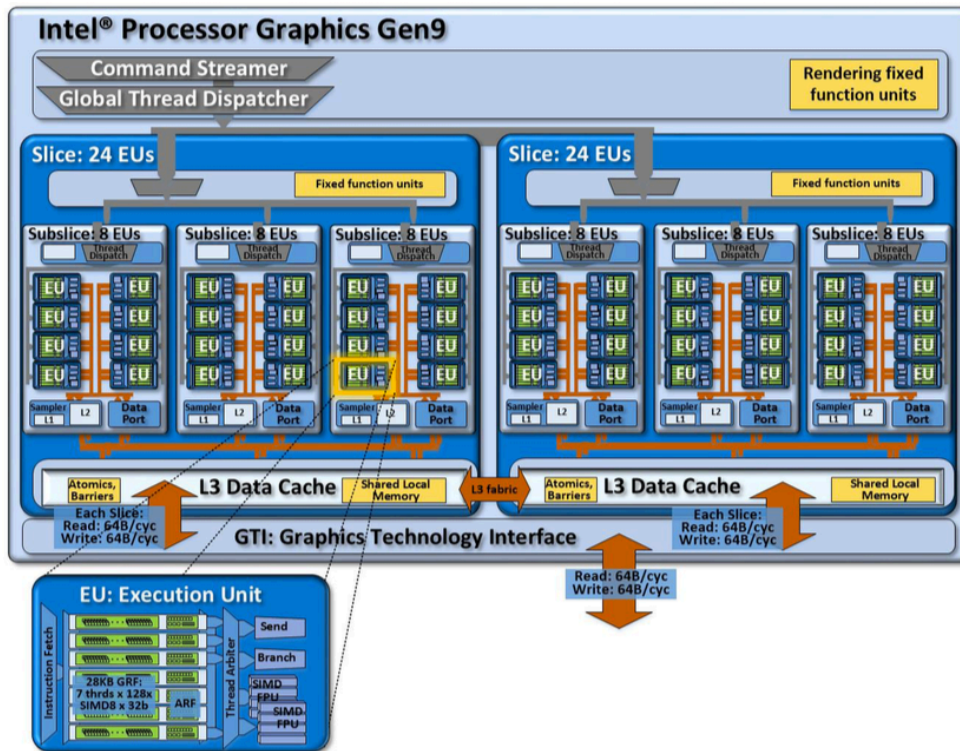
- Cores, GPU, and memory connected by a ring interconnect
- Same memory used by CPU and GPU
- Shared Last Level Cache
- Peak double precision floating point performance 100-300 GF

THE EXECUTION UNIT (EU)



- The EU executes instruction
- Each EU has 7 threads
- Each thread has 128 32 byte registers
- Issues instructions to four processing units:
 - 2 SIMD FPU
 - Branch
 - Send (memory)

SUBSLICES AND SLICES



- A subslice contains 8 EUs
- A slice contains 3 subslices
- Products available with 1, 2, or 3 slices

HETEROGENOUS SYSTEM PROGRAMMING MODELS

- CUDA
- OpenCL
- HIP
- OpenACC
- OpenMP
- SYCL
- Kokkos
- Raja

QUESTIONS?

www.anl.gov

Argonne 
NATIONAL LABORATORY

XEON PHI IN THE TOP500

The KNL Xeon Phi Processor is in 7 of the top 20 systems

Rank	Facility	Architecture	Linpack (PF)	Peak (PF)
9	Los Alamos/Sandia	Trinity - Cray XC40, Intel Xeon Phi 7250	14	44
10	Berkeley - NERSC	Cori - Cray XC40, Intel Xeon Phi 7250	14	28
11	Korea Institute of Science and Technology Inf.	Nurion - Cray CS500, Intel Xeon Phi 7250	14	26
12	Joint Center for Advanced High Performance Computing	Oakforest-PACS - PRIMERGY CX1640 M1, Intel Xeon Phi 7250	14	25
14	Commissariat a l'Energie Atomique	Tera-1000-2 - Bull Sequana X1000, Intel Xeon Phi 7250	12	23
15	Texas Advanced Computing Center	Stampede2 - PowerEdge C6320P/C6420, Intel Xeon Phi 7250	11	18
18	CINECA	Marconi Intel Xeon Phi - CINECA Cluster, Lenovo SD530/S720AP, Intel Xeon Phi 7250	8.4	16
21	Argonne National Laboratory	Theta - Cray XC40, Intel Xeon Phi 7230	6.9	12

KNIGHTS LANDING VARIANTS

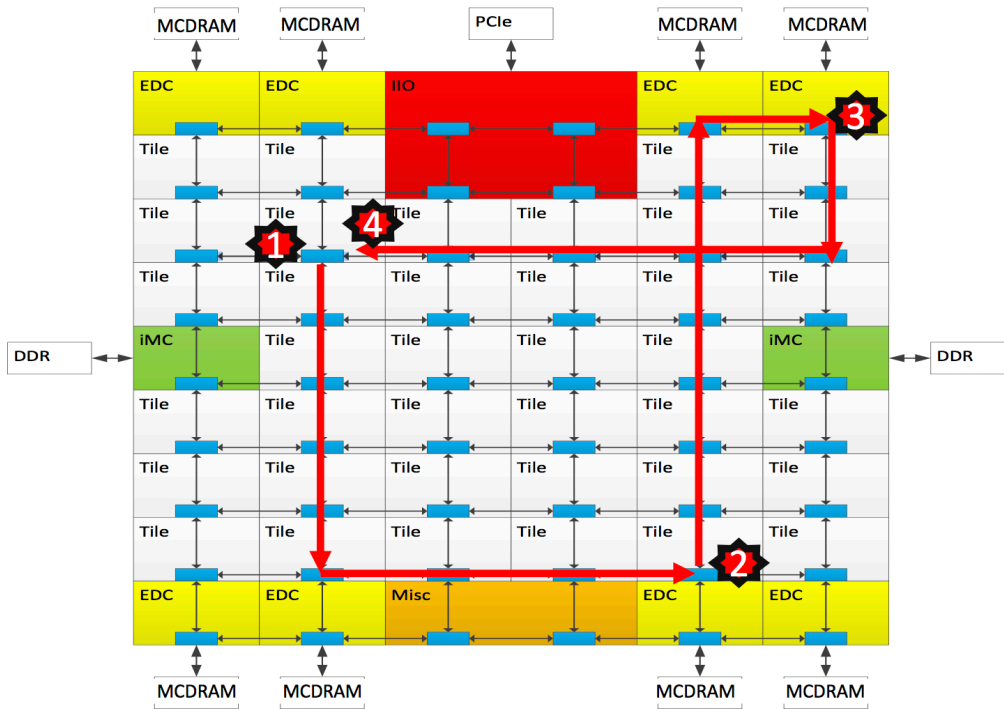
SKU	Cores	TDP Freq (GHz)	AVX Freq (GHz)	Peak Flops (TFlops)	MCDRAM (GB)	DDR Speed	TDP (Watts)
7210	64	1.3	1.1	2.66	16	2133	215
7230	64	1.3	1.1	2.66	16	2400	215
7250	68	1.4	1.2	3.05	16	2400	215
7290	72	1.5	1.3	3.46	16	2400	245

STREAM TRIAD BENCHMARK PERFORMANCE

- Cache mode peak STREAM triad bandwidth is lower
 - Bandwidth is 25% lower than Flat mode
 - Due to an additional read operation on write
- Cache mode bandwidth has considerable variability
 - Observed performance ranges from 225-352 GB/s
 - Due to MCDRAM direct mapped cache conflicts
- Streaming stores (SS) :
 - Streaming stores on KNL by-pass L1 & L2 and write to MCDRAM cache or memory
 - Improve performance in Flat mode by 33% by avoiding a read-for-ownership operation
 - Doesn't improve performance in Cache mode, can lower performance from DDR

Case	GB/s with SS	GB/s w/o SS
Flat, MCDRAM	485	346
Flat, DDR	88	66
Cache, MCDRAM	352	344
Cache, DDR	59	67

Cluster Modes: All-to-All



Address uniformly hashed across all distributed directories

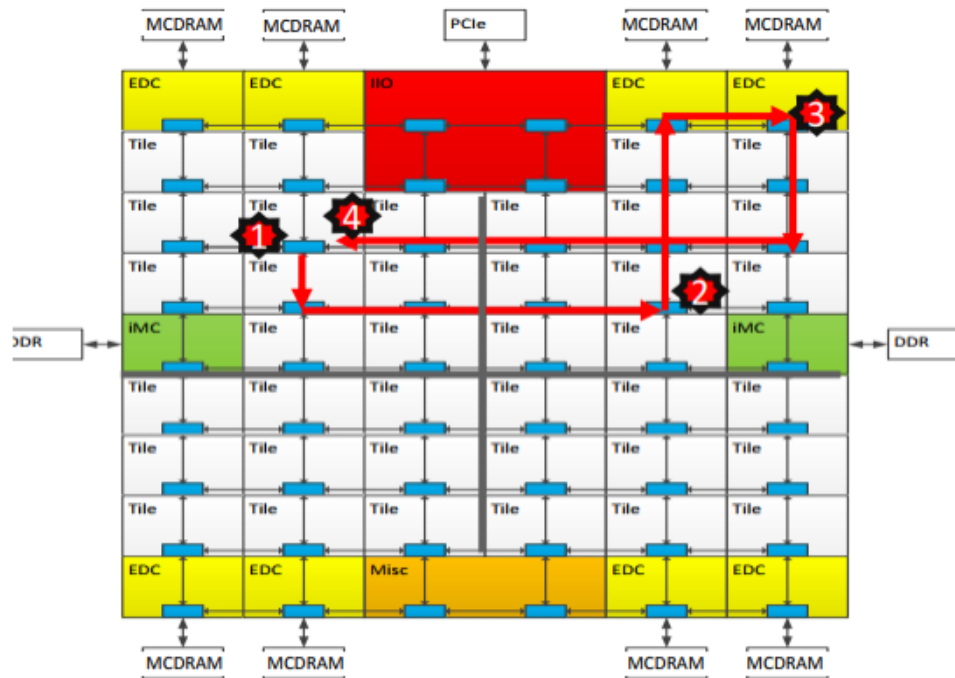
No affinity between Tile, Directory and Memory

Most general mode. Lower performance than other modes.

Typical Read L2 miss

1. L2 miss encountered
2. Send request to the distributed directory
3. Miss in the directory. Forward to memory
4. Memory sends the data to the requester

Cluster Modes: Quadrant



Chip divided into four virtual Quadrants

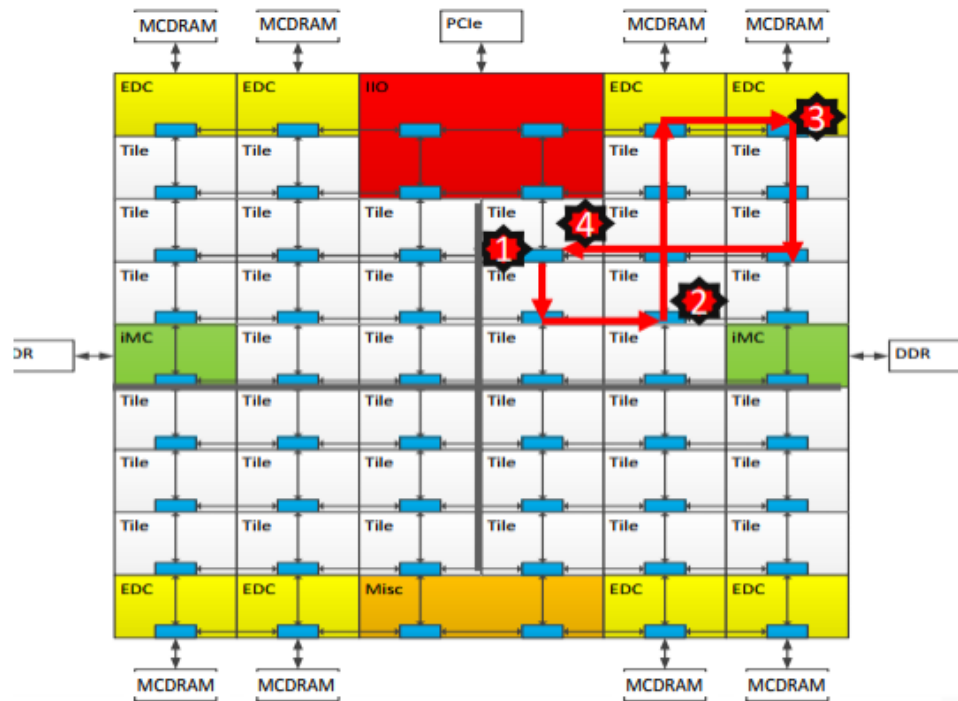
Address hashed to a Directory in the same quadrant as the Memory

Affinity between the Directory and Memory

Lower latency and higher BW than all-to-all. SW Transparent.

1) L2 miss, 2) Directory access, 3) Memory access, 4) Data return

Cluster Modes: Sub-NUMA Clustering



Each Quadrant (Cluster) exposed as a separate NUMA domain to OS.

Looks analogous to 4-Socket Xeon

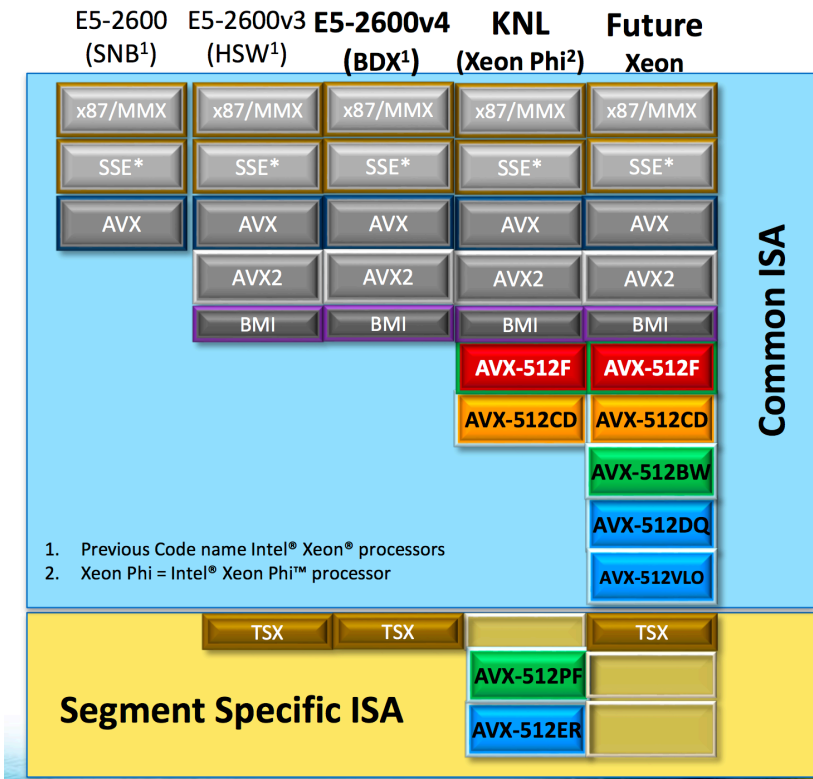
Affinity between Tile, Directory and Memory

Local communication. Lowest latency of all modes.

SW needs to NUMA optimize to get benefit.

1) L2 miss, 2) Directory access, 3) Memory access, 4) Data return

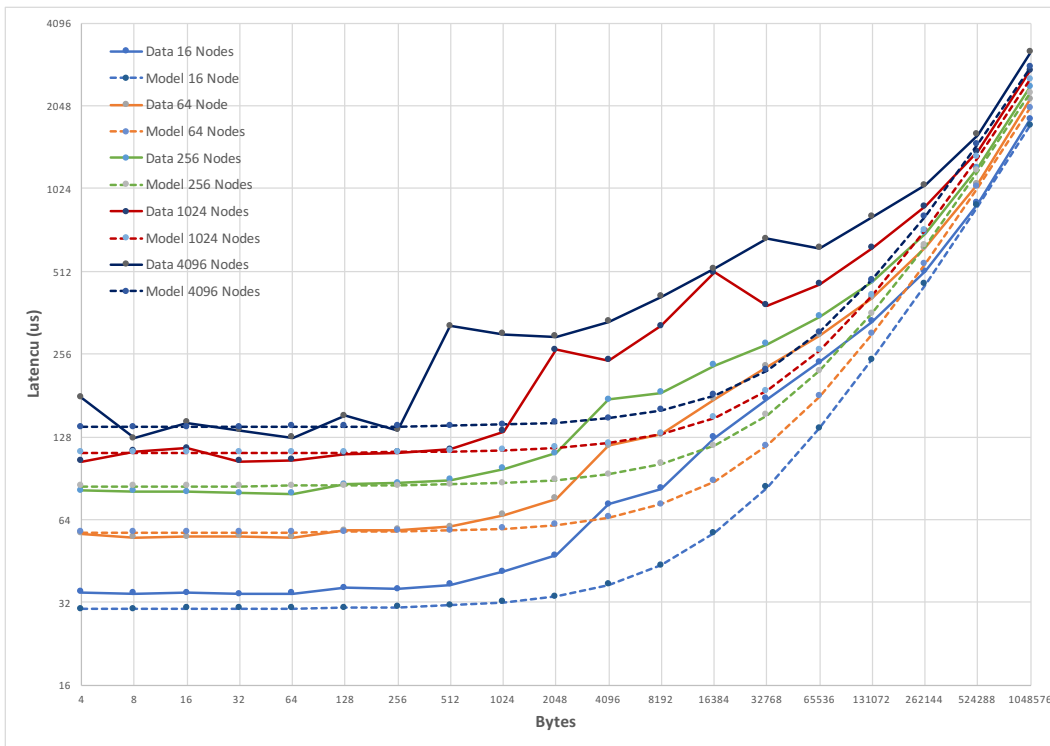
Knights Landing Instruction Set



- KNL implements x86 legacy instructions
 - Don't need to recompile
- KNL introduces AVX-512 instruction
 - 512F – foundation
 - 512 bit FP and integer vectors
 - 32 registers and 8 mask register
 - Gather/scatter
 - 512CD – conflict detection
 - 512PF – gather/scatter prefetch
 - 512ER – reciprocal and sqrt estimates
- KNL does not have
 - TSX – transactional memory
 - 512BW – byte/word (8/16 bit)
 - 512DQ – dword/quad-word (32/64b)
 - 512VLO – vector length orthogonality

MPI ALLREDUCE MODEL

OSU MPI Allreduce Benchmark



$$T = \gamma + \delta n + (\alpha + \beta n) \log_2(p)$$

$n = \text{bytes}$

$p = \text{nodes}$

$\gamma = -24$

$\delta = 0.0012$

$\alpha = 13.6$

$\beta = 0.00012$

Good fit at low and high byte ranges.

Errors centered around point of protocol switch

MOST FREQUENTLY CALLED COLLECTIVE ROUTINES

Approximate relative call frequency from ALCF applications workload

Routine	Relative Call Frequency
Allreduce	5000
Bcast	2500
Barrier	500
Alltoall	500
Alltoallv	250
Reduce	75
Allgatherv	25
Everything else	<1

POWER EFFICIENCY






- Theta #7 on Green500 (Nov. 2016)
- For high compute intensity, 1 thread per core was most efficient
 - Avoids contention with shared resources
- MCDRAM is a 4x improvement over DDR4 in power efficiency

Threads per Core	Time (s)	Power (W)	Efficiency (GF/W)
1	110.0	284.6	4.39
2	118.6	285.4	4.06
4	140.3	295.0	3.32

Memory Type	Bandwidth GB/s	Power (W)	Efficiency (GB/s/W)
MCDRAM	449.5	270.5	1.66
DDR4	87.1	224.4	0.39

BLUE GENE/Q ARCHITECTURE

ALCF SYSTEMS

				
Mira – IBM BG/Q	Cetus – IBM BG/Q	Vesta – IBM BG/Q	Cooley - Cray/NVIDIA	Theta - Cray XC40
<ul style="list-style-type: none"> – 49,152 nodes – 786,432 cores – 786 TB RAM – 10 PF 	<ul style="list-style-type: none"> – 4,096 nodes – 65,536 cores – 64 TB RAM – 836 TF 	<ul style="list-style-type: none"> – 2,048 nodes – 32,768 cores – 32 TB RAM – 419 TF 	<ul style="list-style-type: none"> – 126 nodes (Haswell) – 1512 cores – 126 Tesla K80 – 48 TB RAM (3 TB GPU) 	<ul style="list-style-type: none"> – 3,624 nodes (KNL) – 231,936 cores – 736 TB RAM – 10 PF

Storage

HOME: 1.44 PB raw capacity

SCRATCH:

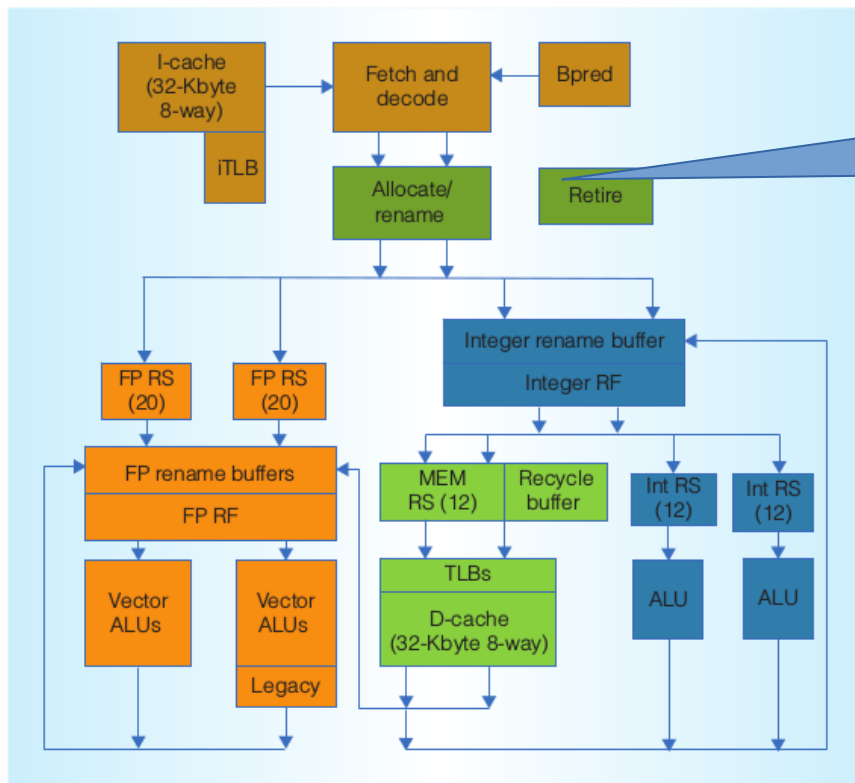
- mira-fs0 - 26.88 PB raw, 19 PB usable; 240 GB/s sustained
- mira-fs1 - 10 PB raw, 7 PB usable; 90 GB/s sustained
- mira-fs2 (ESS) - 14 PB raw, 7.6 PB usable; 400 GB/s sustained (not in production yet)
- theta-fs0 – 10 PB raw, 8.9 useable, 240 GB/s sustained

TAPE: 21.25 PB of raw archival storage [17 PB in use]

COMPARISON OF THETA (KNL) TO MIRA (BG/Q)

- More local parallelism
 - 64 (KNL) vs 16 (BG/Q)
 - 4 hardware threads on both
- Significantly fewer nodes, 48K -> 3.6K
- Clock speed drops, 1.6 GHz -> 1.1 GHz
- Increased vector length
 - 8 wide vectors (KNL) vs 4 wide vectors (BG/Q)
- Increased node performance
 - 2.4 TF (KNL) vs 0.2 TF (BG/Q)
- Instruction issue
 - Out-of-order (KNL) vs in-order (BG/Q)
 - 2 wide instruction issue on both
 - 2 floating point instructions per cycle (KNL) vs 1 per cycle (BG/Q)
- Memory Hierarchy
 - MCDRAM & DDR (KNL) vs uniform 16 GB DDR (BG/Q)
- Different network topology
 - 5D torus vs Dragonfly
- NIC connectivity
 - PCIe (Aries, Omni-Path) vs direct crossbar connection (BG/Q)

KNL Pipeline



Fetch/decode 16 bytes per cycle (i.e. two instructions per cycle)
Careful: AVX-512 instructions can be up to 12 bytes each if they have non-compressed displacements!

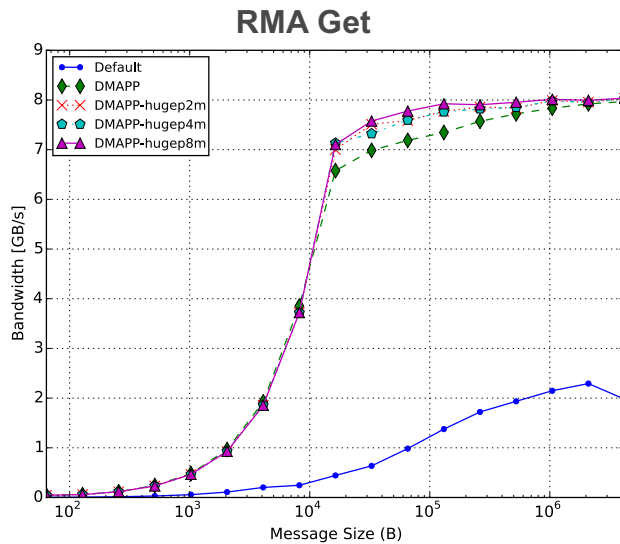
2 FP/vector operations,
2 memory operations,
and 2 scalar integer operations per cycle!

MPI ONE SIDED (RMA)

OSU One Sided MPI Get Bandwidth and Bi-Directional Put Bandwidth

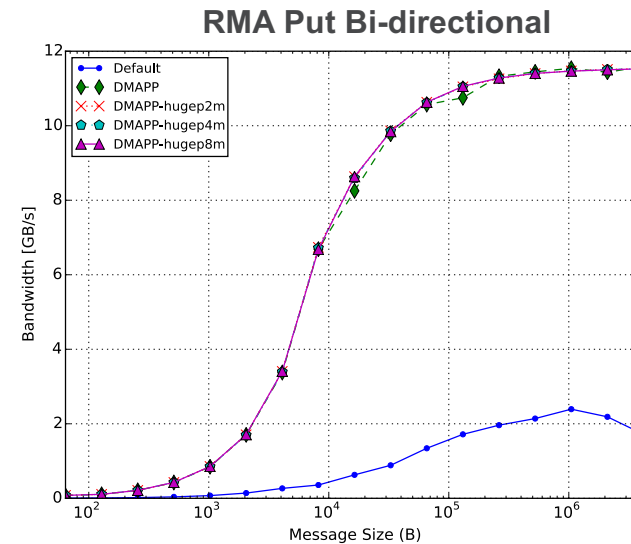
RMA Get

- 2 GB/s using default configuration (uGNI)
- 8 GB/s using RMA over DMAPP
- Huge pages also help.



RMA Put

- 2 GB/s using default configuration (uGNI)
- 11.6 GB/s peak bi-directional bandwidth over DMAPP
- No significant benefit from huge pages



THETA FILESYSTEMS

- Home (GPFS)
 - Home directories (/home) currently live in /gpfs/theta-fs1/home
- Projects (Lustre)
 - /lus/theta-fs0
 - 10 PB raw, 8.9 PB useable space
 - 240 GB/s sustained
 - Project directories (/projects) currently live in /lus/theta-fs0/projects
 - With large I/O, be sure to consider **stripe width**
- SSD
 - Theta compute nodes contain a single SSD with a raw capacity of 128 GB
 - A local volume is presented to the user as an ext3 system on top of an LVM volume
 - Userspace applications can access the SSD via standard POSIX APIs
 - The final capacity available to the end user is still TBD
- **NOTE**
 - No backups at this time
 - No quotas at this time

STREAM TRIAD BENCHMARK PERFORMANCE

- Peak STREAM Triad bandwidth occurs in Flat mode:
 - from MCDRAM, 485 GB/s
 - from DDR, 88 GB/s
- Cache mode bandwidth is 25% lower than Flat mode
 - Due to an additional cache check read operation
- Cache mode bandwidth has considerable variability
 - Observed performance ranges from 225-352 GB/s
 - Due to MCDRAM direct mapped cache page conflicts
- Streaming stores (SS) :
 - Improve performance in Flat mode by 33% by avoiding a read-for-ownership operation
 - Can lower performance from DDR in Cache mode
- Maximum measured single core bandwidth is 14 GB/s
 - Need to use ~half the cores on a node to saturate MCDRAM bandwidth in Flat mode

Case	GB/s with SS	GB/s w/o SS
Flat, MCDRAM	485	346
Flat, DDR	88	66
Cache, MCDRAM	352	344
Cache, DDR	59	67

Memory Modes

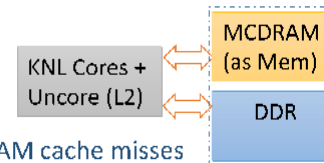
MCDRAM as Cache

- Upside
 - No software modifications required
 - Bandwidth benefit (over DDR)
- Downside
 - Higher latency for DDR access
 - i.e., for cache misses
 - Misses limited by DDR BW
 - All memory is transferred as:
 - DDR -> MCDRAM -> L2
 - Less addressable memory



MCDRAM as Flat Mode

- Upside
 - Maximum BW
 - Lower latency
 - i.e., no MCDRAM cache misses
 - Maximum addressable memory
 - Isolation of MCDRAM for high-performance application use only
- Downside
 - Software modifications (or interposer library) required
 - to use DDR and MCDRAM in the same app
 - Which data structures should go where?
 - MCDRAM is a finite resource and tracking it adds complexity



A BRIEF HISTORY OF THE BLUE GENE

- In 1999 IBM began a \$100 million research project to explore a novel massively parallel architecture
- Initial target was protein folding applications
- Design evolved out of the Cyclops64 and QCDOC architectures
- First Blue Gene/L prototype appeared at #73 on the Top500 on 11/2003
- Blue Gene/L system took #1 on Top500 on 11/2004 (16 Racks at LLNL)
- In 2007 the 2nd generation Blue Gene/P was introduced
- In 2012 the 3rd generation Blue Gene/Q was introduced
- Since being released 14 years ago, on the Top500 list:
 - A Blue Gene was #1 on half of the lists
 - On average 3 of the top 10 machines have been Blue Gene's
- The Blue Gene/Q:
 - Currently #4 on the Top500 (LLNL, 96 racks, 20PF)
 - Also holds #9 (ANL), #19 (Juelich), #21 (LLNL- Vulcan)

BLUE GENE DNA AND THE EVOLUTION OF MANY CORE

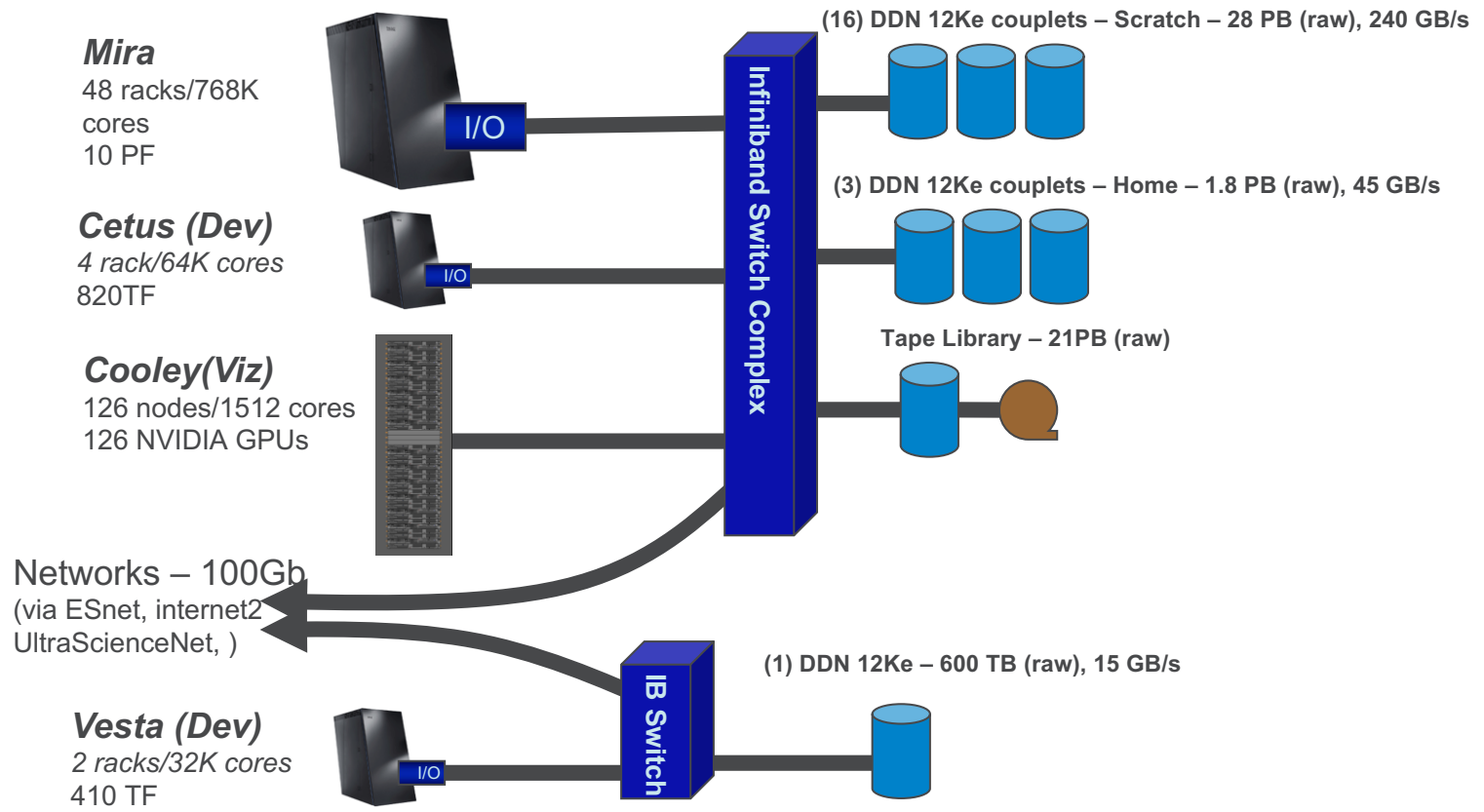
- Leadership computing power
 - Leading architecture since introduction, #1 half Top500 lists over last 10 years
 - On average over the last 12 years 3 of the top 10 machine on Top 500 have been Blue Genes
- Low speed, low power
 - Embedded PowerPC core with custom SIMD floating point extensions
 - Low frequency (L – 700 MHz, P – 850 MHz, Q – 1.6 GH) (KNL – 1.1 GHz)
- Massive parallelism:
 - Multi/Many core (L - 2, P – 4, Q – 16) (KNL - 68)
 - Many aggregate cores (L – 208k, P – 288k, Q – 1.5M) (KNL – 650k)
- Fast communication network(s)
 - Low latency, high bandwidth, network (L & P – 3D Torus, Q – 5D Torus) (KNL - Dragonfly)
- Balance:
 - Processor, network, and memory speeds are well balanced
- Minimal system overhead
 - Simple lightweight OS (CNK) minimizes noise
- Standard Programming Models
 - Fortran, C, C++, & Python languages supported
 - Provides MPI, OpenMP, and Pthreads parallel programming models
- System on a Chip (SoC) & Custom designed Application Specific Integrated Circuit (ASIC)
 - All node components on one chip, except for memory
 - Reduces system complexity and power, improves price / performance
- High Reliability:
 - Sophisticated RAS (reliability, availability, and serviceability)
- Dense packaging
 - 1024 nodes per rack for Blue Gene

ALCF BG/Q SYSTEMS

- *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
 - 4K & 2k nodes / 64k & 32k cores
 - 64 TB & 32 TB of memory
 - 820TF & 410TF peak flop rate
- Storage
 - Scratch: 28.8 PB raw capacity, 240 GB/s bw (GPFS)
 - Home: 1.8 PB raw capacity, 45 GB/s bw (GPFS)



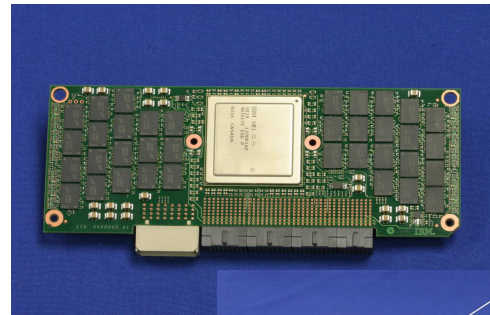
ALCF BG/Q SYSTEMS



BLUE GENE/Q COMPONENTS

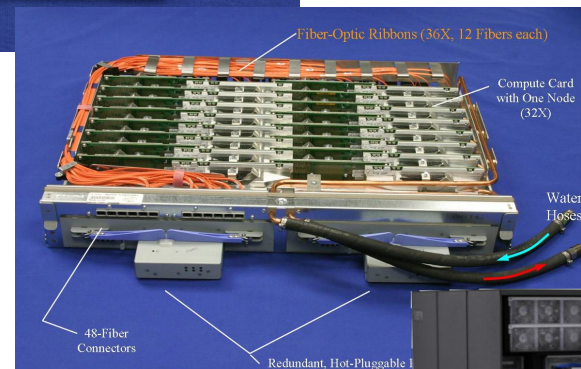
Compute Node:

- **Processor:**
 - 18 cores (205 GF)
 - Memory Controller
 - Network Interface
- **Memory:**
 - 16 GB DDR3
 - 72 SDRAMs, soldered
- **Network connectors**



Node Card Assembly or Tray

- 32 Compute Nodes (6.4 TF)
- Electrical network
- Fiber optic modules and link chips
- Water cooling lines
- Power supplies

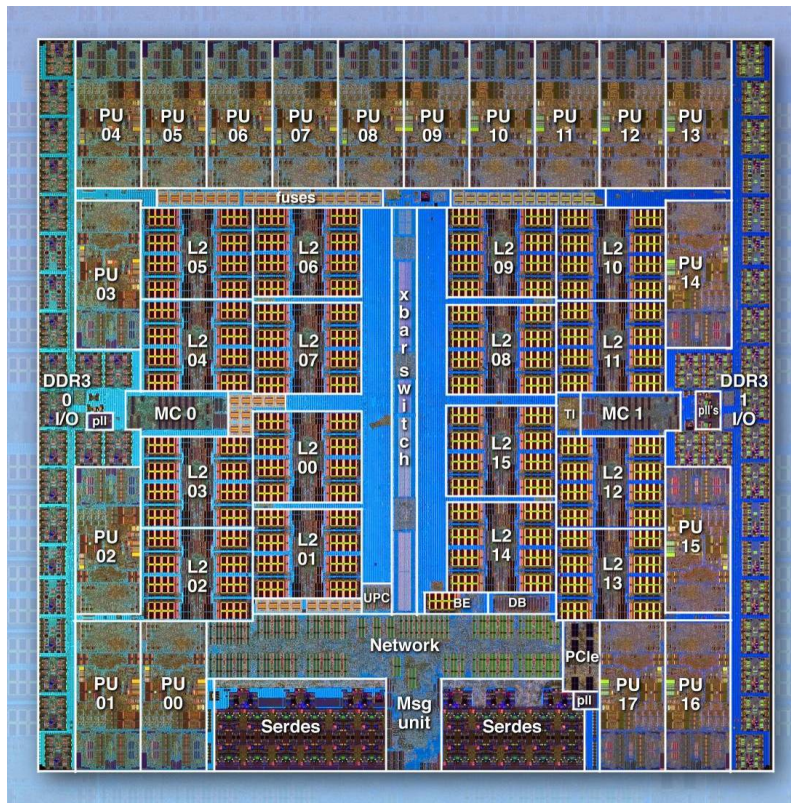


Rack

- 32 Node Trays (1024 nodes) (205 TF)
- 5D Torus Network (4x4x4x8x2)
- 8 IO nodes
- Power Supplies



BLUEGENE/Q COMPUTE CHIP



Chip

- 360 mm² Cu-45 technology (SOI)
- 1.5 B transistors

18 Cores

- 16 compute cores
- 17th core for system functions (OS, RAS)
- plus 1 redundant processor
- L1 I/D cache = 16kB/16kB

Crossbar switch

- Each core connected to shared L2
- Aggregate read rate of 409.6 GB/s

Central shared L2 cache

- 32 MB eDRAM
- 16 slices

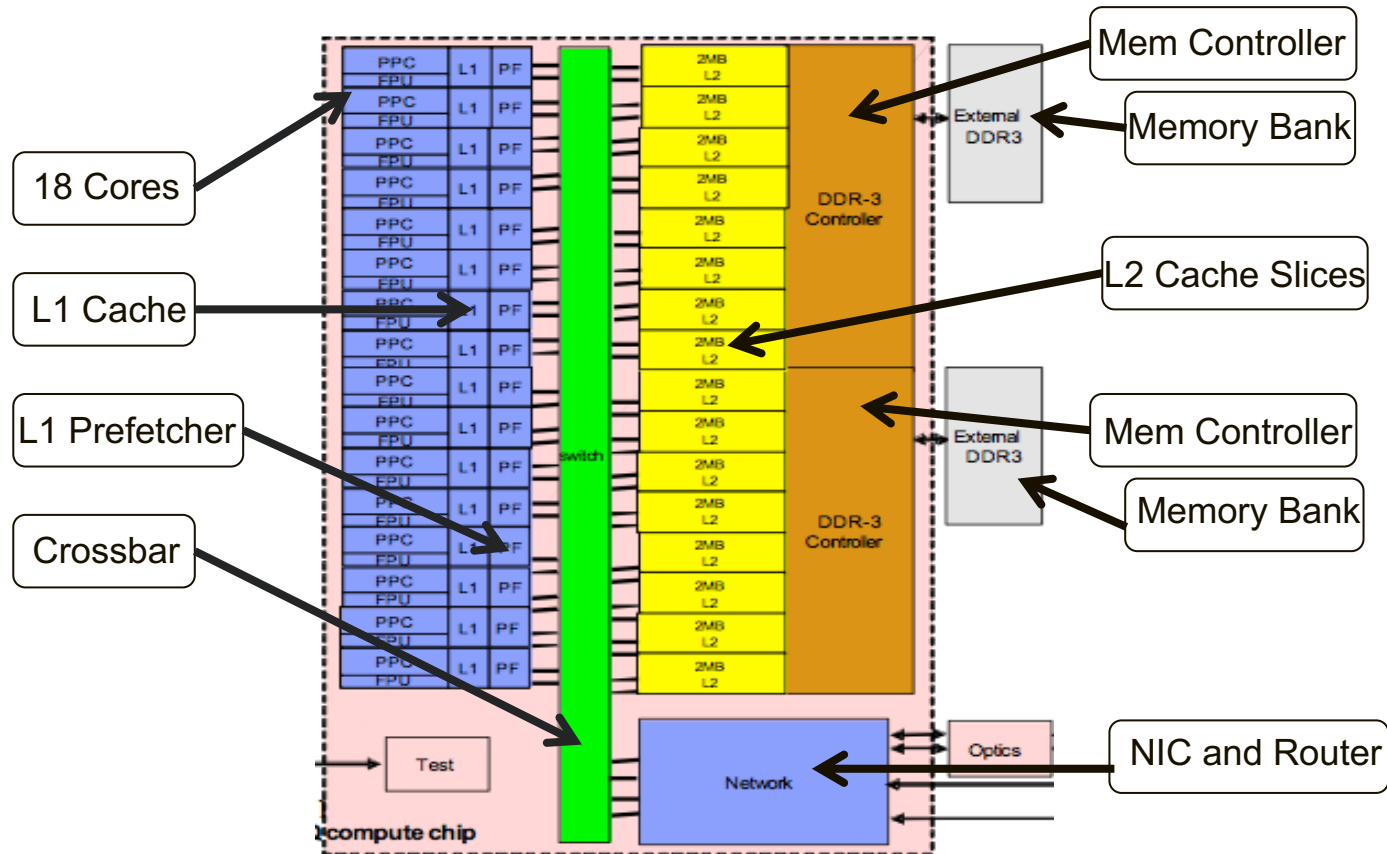
Dual memory controller

- 16 GB external DDR3 memory
- 42.6 GB/s bandwidth

On Chip Networking

- Router logic integrated into BQC chip
- DMA, remote put/get, collective operations
- 11 network ports

BG/Q CHIP, ANOTHER VIEW

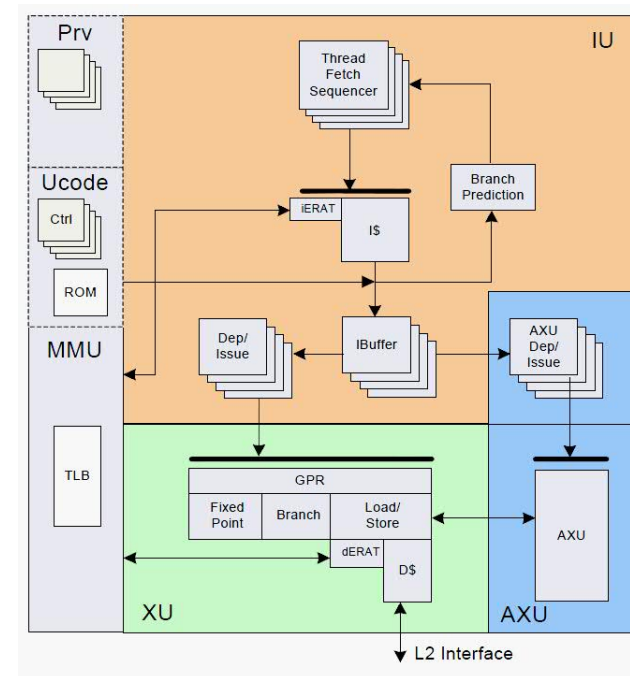


BG/Q Core

- Full PowerPC compliant 64-bit CPU, PowerISA v.206
 - *Plus QPX floating point vector instructions*
- Runs at 1.6 GHz
- In-order execution
- 4-way Simultaneous Multi-Threading
- Registers: 32 64-bit integer, 32 256-bit floating point

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)

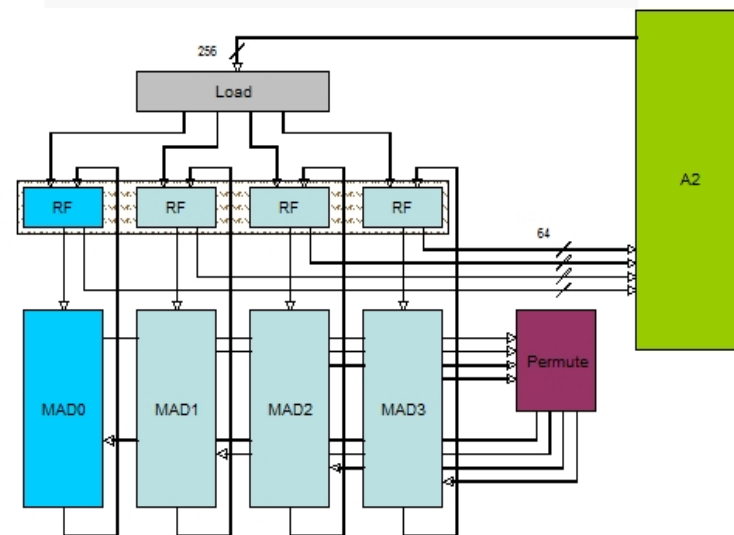


Instruction Issue:

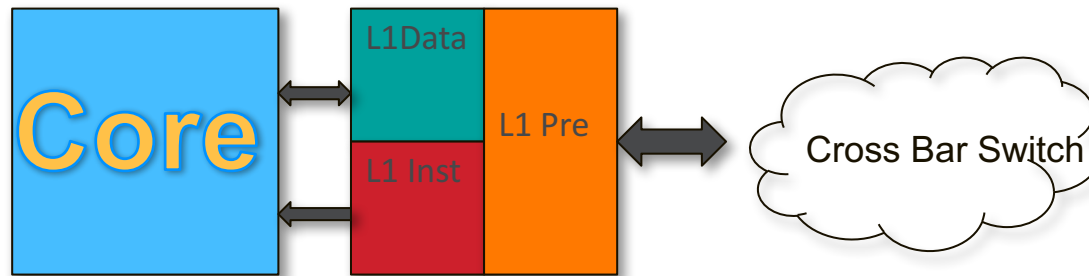
- 2-way concurrent issue if 1 XU + 1 AXU instruction
- A given thread may only issue 1 instruction per cycle
- Two threads may each issue 1 instruction each cycle

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
- Standard 64 bit floating point registers are extended to 256 bits
- FPU operates on:
 - Standard scalar PowerPC FP instructions
 - 4 wide SIMD instructions
 - 2 wide complex arithmetic SIMD arithmetic
- 4 wide FMA (mult-add) instructions allow 8 flops/inst
- Attached to AXU port of A2 core
- A2 issues one instruction/cycle to AXU
- 6 stage pipeline
- Compiler can generate QPX instructions
- Intrinsic functions mapping to QPX instructions allow easy QPX programming



L1 CACHE & PREFETCHER



- Each Core has its own L1 cache and L1 Prefetcher
- L1 Cache:
 - **Data:** 16KB, 8 way set associative, 64 byte line, 6 cycle latency
 - **Instruction:** 16KB, 4 way set associative, 3 cycle latency
- L1 Prefetcher (L1P):
 - 1 prefetch unit for each core
 - 32 entry prefetch buffer, entries are 128 bytes, 24 cycle latency
 - Operates in List or Stream prefetch modes
 - Operates as write-back buffer

BG/Q MEMORY HIERARCHY

Crossbar switch connects:

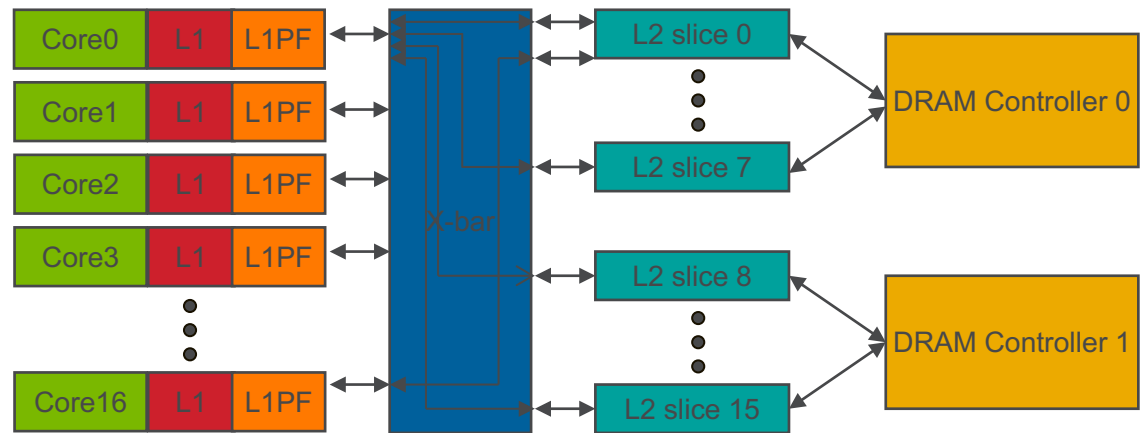
- L1P's, L2 slices, Network, PCIe interface

Aggregate bandwidth across slices:

- Read: 409.6 GB/s, Write: 204.8 GB/s

Memory:

- Two on chip memory controllers
- Each connects to 8 L2 slices via 2 ring buses
- Each controller drives a 16+2 byte DDR-3 channel at 1.33 GT/s
- Peak bandwidth is 42.67 GB/s (excluding ECC)
- Latency > 350 cycles



L1 Cache:

- **Data:** 16KB, 8 way assoc., 64 byte line, 6 cycle latency
- **Instruction:** 16KB, 4 way assoc., 3 cycle latency

L1 Prefetcher (L1P):

- 32 entry prefetch buffer, entries are 128 bytes
- 24 cycle latency
- Operates in List or Stream prefetch modes
- Operates as write-back buffer

L2 Cache:

- Shared by all cores
- Serves a point of coherency, generates L1 invalidations
- Divided into 16 slices connected via crossbar switch to each core
- 32 MB total, 2 MB per slice
- 16 way set assoc., write-back, LRU replacement, 82 cycle latency
- Supports memory speculation and atomic memory operations

THE BG/Q NETWORK

▪ 5D torus network:

- Achieves high nearest neighbor bandwidth while increasing bisectional bandwidth and reducing hops vs 3D torus
- 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 dim always 2)

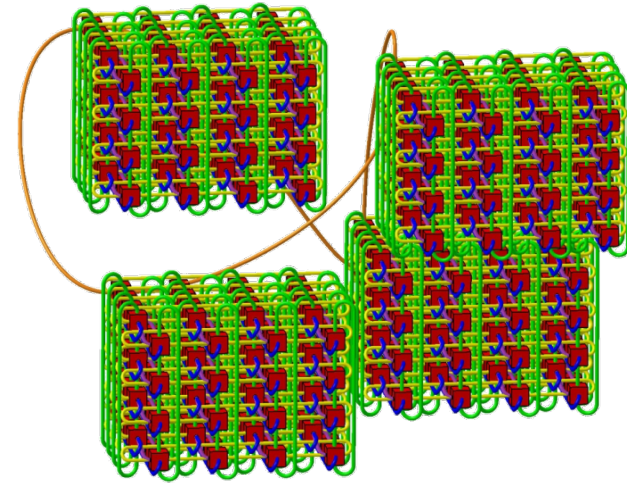
▪ No separate Collectives or Barrier network:

- Single network used for point-to-point, collectives, and barrier operations

▪ Additional 11th link to IO nodes

▪ Two type of network links

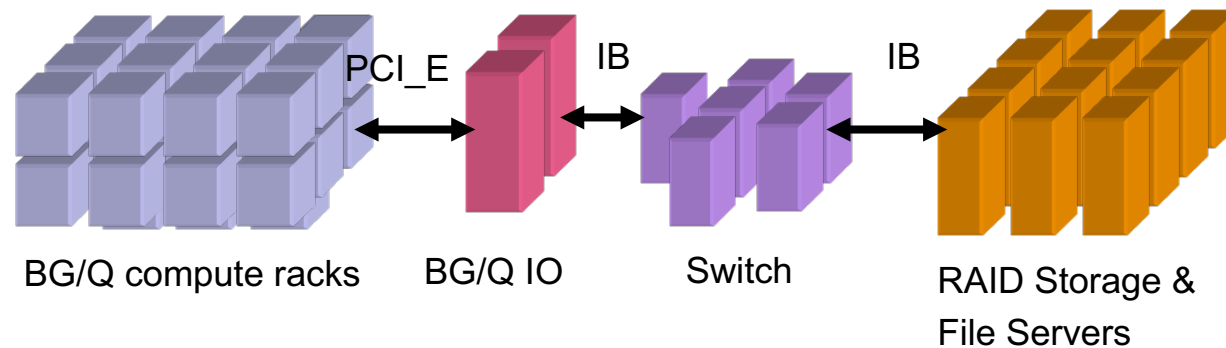
- Optical links between midplanes
- Electrical inside midplane



NETWORK PERFORMANCE

- **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
- **Hardware latency**
 - ~40 ns per hop through network logic
 - Nearest: 80ns
 - Farthest: 3us (96-rack 20PF system, 31 hops)
- **Network Performance**
 - Nearest-neighbor: 98% of peak
 - Bisection: > 93% of peak
 - All-to-all: 97% of peak
 - Collective: FP reductions at 94.6% of peak
 - Allreduce hardware latency on 96k nodes ~ 6.5 *us*
 - Barrier hardware latency on 96k nodes ~ 6.3 *us*

BG/Q IO



IO is sent from Compute Nodes to IO Nodes to storage network

- IO Nodes handle function shipped IO calls to parallel file system client
- IO node hardware is identical to compute node hardware
- IO nodes run Linux and mount file system
- Compute Bridge Nodes use 1 of the 11 network links to link to IO nodes
- IO nodes connect to 2 bridge nodes
- IO nodes are not shared between compute partitions

BLUE GENE/Q SOFTWARE HIGH-LEVEL GOALS & PHILOSOPHY

- Facilitate extreme scalability
 - Extremely low noise on compute nodes running CNK OS
- High reliability: a corollary of scalability
- Familiar programming modes such as MPI and OpenMP
- Standards-based when possible
- Open source where possible
- Facilitate high performance for unique hardware:
 - Quad FPU, DMA unit, List-based prefetcher
 - TM (Transactional Memory), SE (Speculative Execution)
 - Wakeup-Unit, Scalable Atomic Operations
- Optimize MPI and native messaging performance
- Optimize libraries
- Facilitate new programming models

COOLEY

- System:
 - 126 nodes/1512 cores
 - 293 TF
- Processor:
 - Haswell E5-2620v3 processors
 - 2 per node
 - 6 cores per processor
 - 2.4 GHz
- GPUS:
 - 126 NVIDIA Tesla K80 GPUs
- Memory:
 - 384 GB per CPU
 - 2x12 GB per GPU
- Network:
 - FDR Infiniband interconnect



AURORA – COMING 2018

- Over 13X Mira's application performance
- Over 180 PF peak performance
- More than 50,000 nodes with 3rd Generation Intel® Xeon Phi™ processor
 - codename Knights Hill, > 60 cores
- Over 7 PB total system memory
 - High Bandwidth On-Package Memory, Local Memory, and Persistent Memory
- 2nd Generation Intel® Omni-Path Architecture with silicon photonics in a dragonfly topology
- More than 150 PB Lustre file system capacity with > 1 TB/s I/O performance

