Argonne NATIONAL LABORATORY

ALCF SYSTEM ARCHITECTURES, HARDWARE

SCOTT PARKER
Lead, Performance Tools & Programming Models
Argonne Leadership Computing Facility

May 2, 2017

ALCF SYSTEMS

Mira – IBM BG/Q

- 49,152 nodes
- 786,432 cores
- 786 TB RAM
- 10 PF

Cetus – IBM BG/Q

- 4,096 nodes
- 65,536 cores
- 64 TB RAM
- 836 TF

Vesta – IBM BG/Q

- 2,048 nodes
- 32,768 cores
- 32 TB RAM
- 419 TF

Cooley - Cray/NVIDIA

- 126 nodes (Haswell)
- 1512 cores
- 126 Tesla K80
- 48 TB RAM (3 TB GPU)

Theta - Cray XC40

- 3,624 nodes (KNL)
- 231,936 cores
- 736 TB RAM
- 10 PF

Storage

HOME: 1.44 PB raw capacity

SCRATCH:

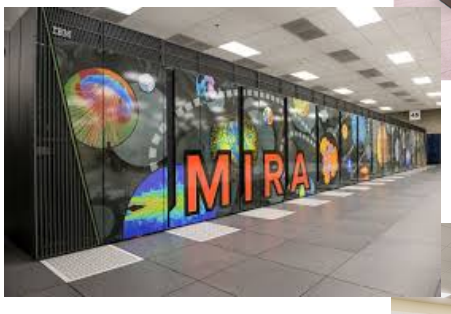
- mira-fs0 - 26.88 PB raw, 19 PB usable; 240 GB/s sustained
- mira-fs1 - 10 PB raw, 7 PB usable; 90 GB/s sustained
- mira-fs2 (ESS) - 14 PB raw, 7.6 PB usable; 400 GB/s sustained (not in production yet)
- theta-fs0 – 10 PB raw, 8.9 useable, 240 GB/s sustained

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

Argonne NATIONAL LABORATORY

ARGONNE HPC TIMELINE

- 2004:
  - Blue Gene/L introduced
  - LLNL 90-600 TF system #1 on Top 500 for 3.5 years
- 2005:
  - Argonne accepts 1 rack (1024 nodes) of Blue Gene/L (5.6 TF)
- 2006:
  - Argonne Leadership Computing Facility (ALCF) created
  - ANL working with IBM on next generation Blue Gene
- 2008:
  - ALCF accepts 40 racks (160k cores) of Blue Gene/P (557 TF)
- 2009:
  - ALCF approved for 10 petaflop system to be delivered in 2012
  - ANL working with IBM on next generation Blue Gene
- 2012:
  - 48 racks of Mira Blue Gene/Q (10 PF) in production at ALCF
- 2014:
  - ALCF CORAL contract awarded to Intel/Cray
  - Development partnership for Theta and Aurora begins
- 2016:
  - ALCF accepts Theta (10 PF) Cray XC40 with Xeon Phi (KNL)
- 2018:
  - Aurora (180+ PF) Cray/Intel Xeon Phi (KNH) to be delivered



Argonne NATIONAL LABORATORY

BLUE GENE/Q ARCHITECTURE

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)

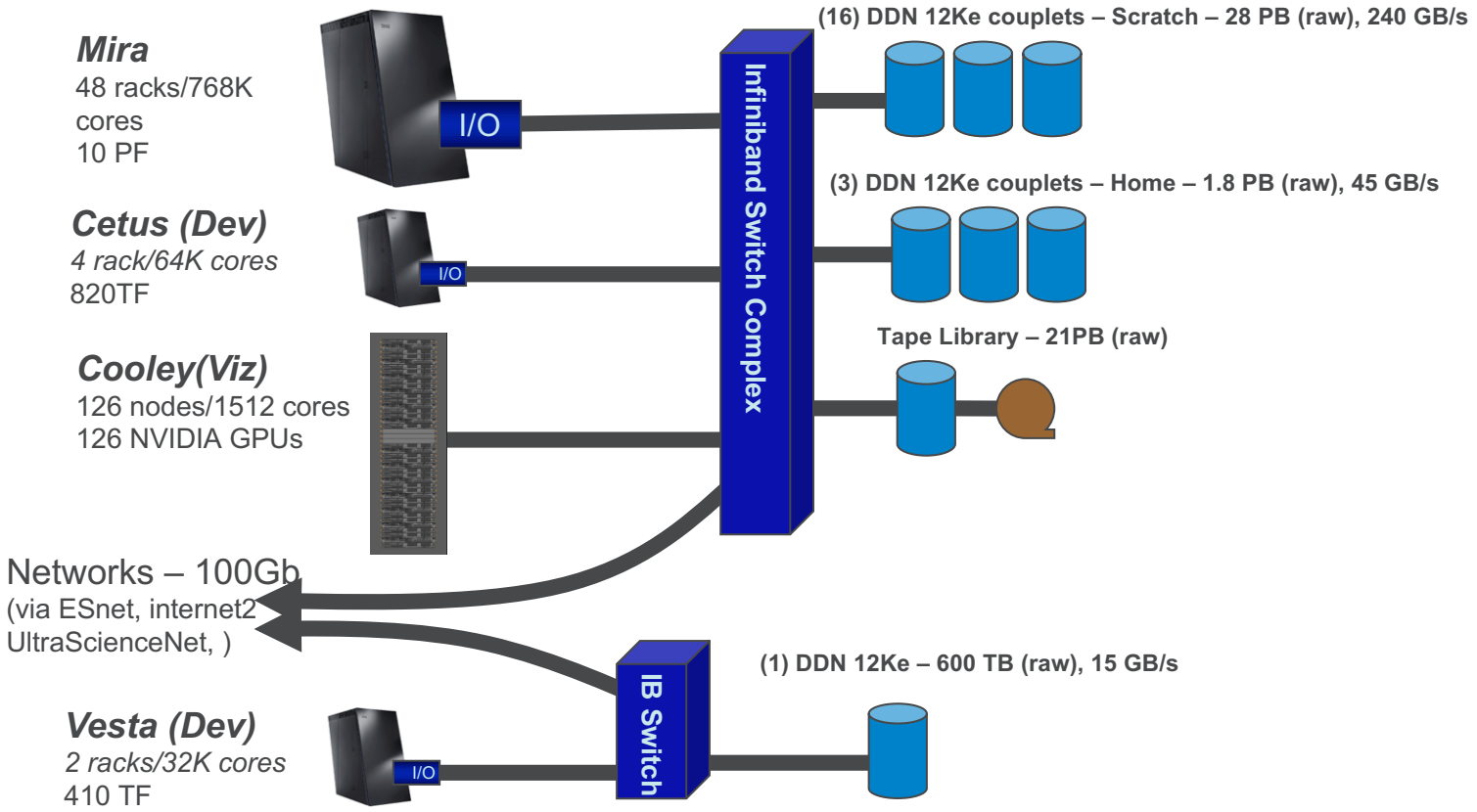
BLUE GENE DNA AND THE EVOLUTION OF MANY CORE

- 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 GHz, 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

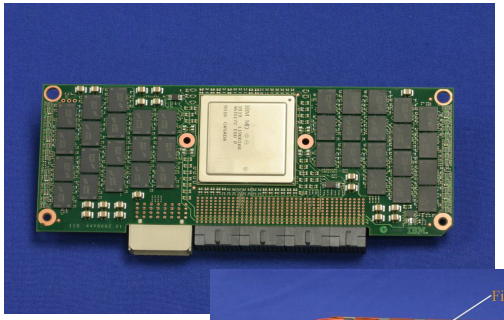


Argonne NATIONAL LABORATORY

BLUE GENE/Q COMPONENTS

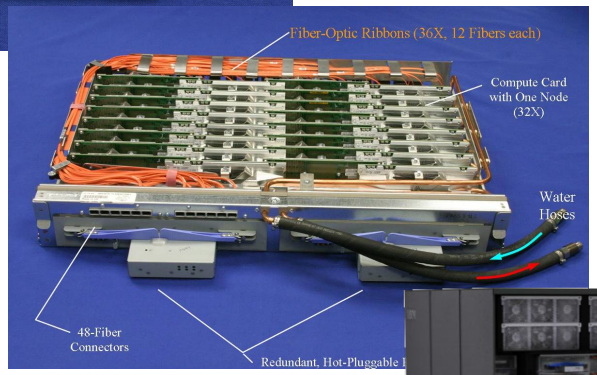
Compute Node:

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



Node Card Assembly or Tray

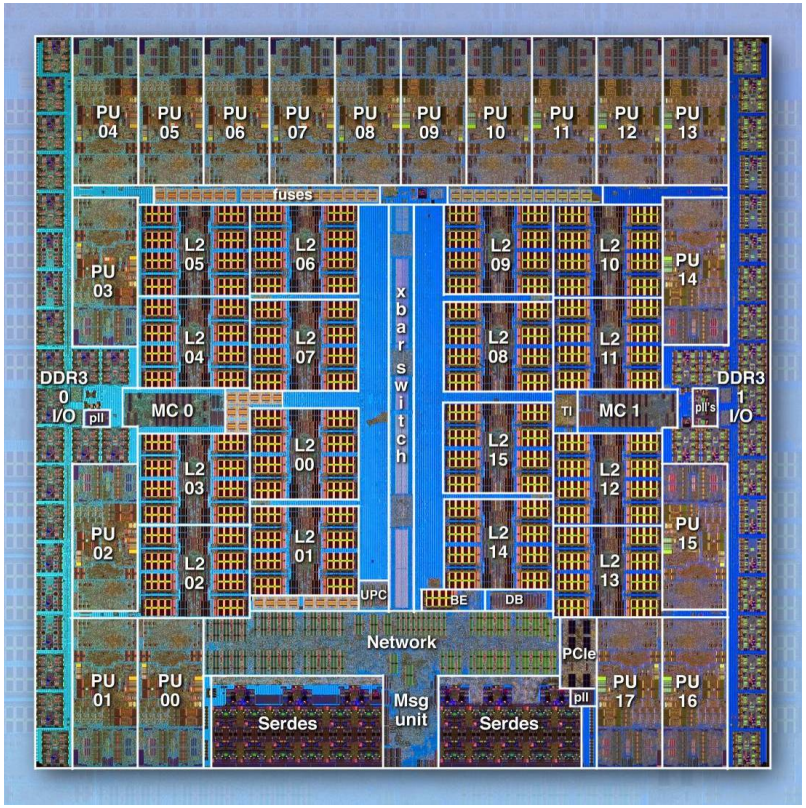
- 32 Compute Nodes (6.4 TF)
- Electrical network
- Fiber optic modules and link chips
- Water cooling lines
- Power supplies



Rack

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

BLUEGENE/Q COMPUTE CHIP



It's big!
- 360 mm² Cu-45 technology (SOI)
- ~ 1.47 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

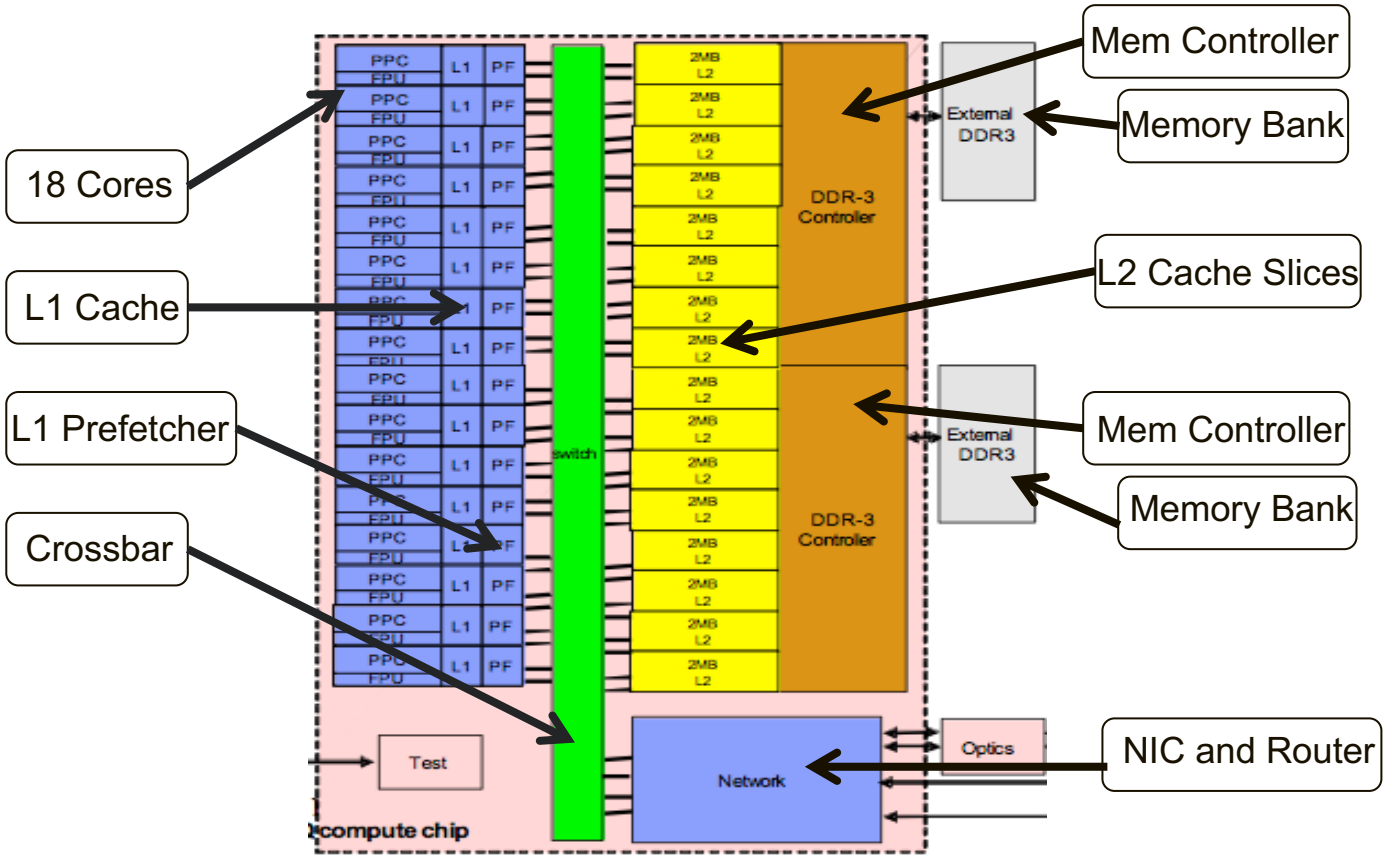
Central shared L2 cache
- 32 MB eDRAM
- 16 slices

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

Argonne NATIONAL LABORATORY

BG/Q CHIP, ANOTHER VIEW



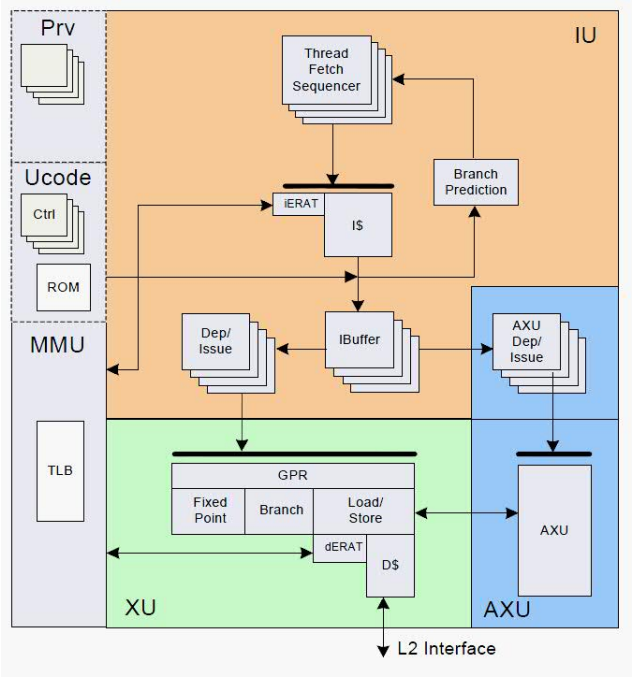
Argonne NATIONAL LABORATORY

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)



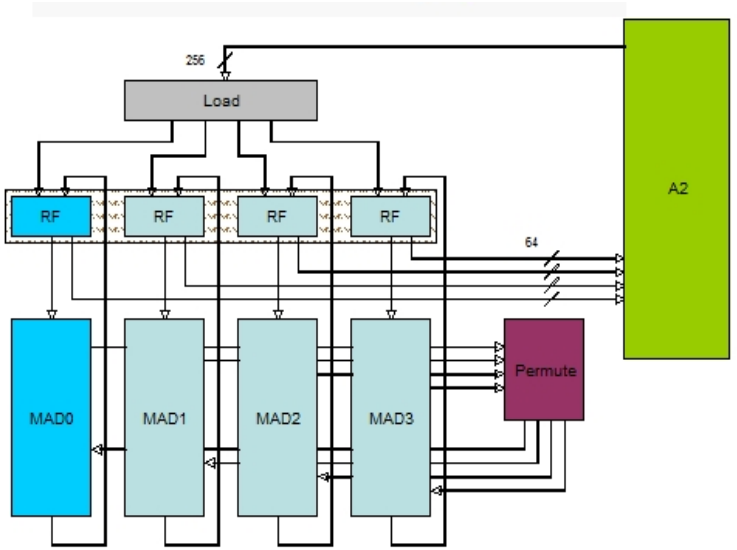
Instruction Issue:

- 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

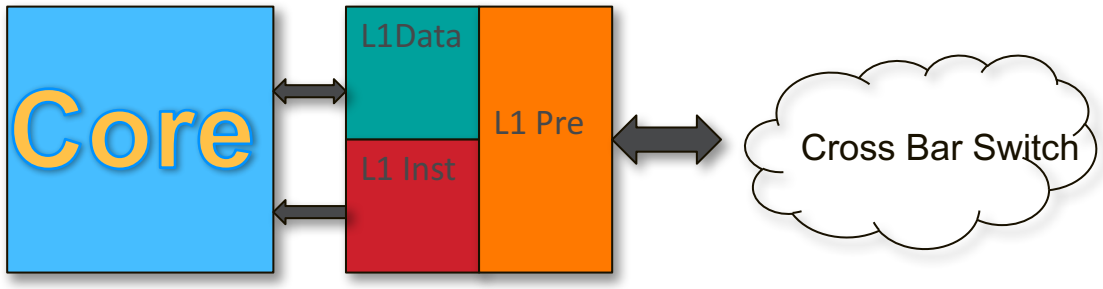
Argonne NATIONAL LABORATORY

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 scalar PowerPC FP instructions
  - 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
- Compiler can generate QPX instructions
- Intrinsic functions mapping to QPX instructions allow easy QPX programming



L1 CACHE & PREFETCHER



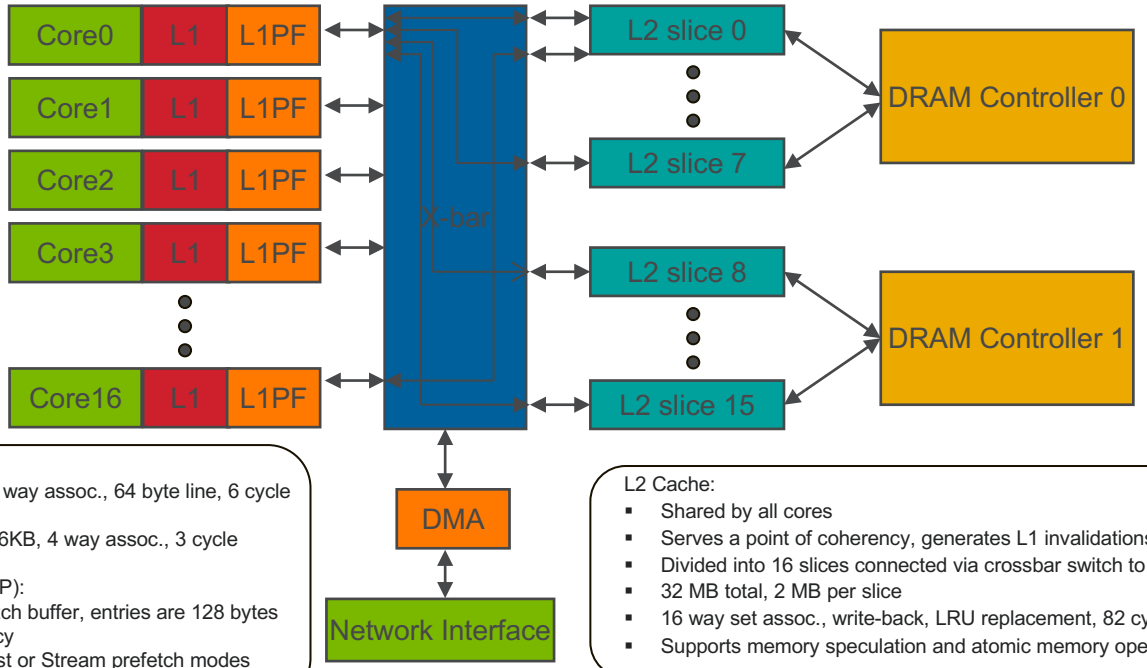
- Each Core has it's own L1 cache and L1 Prefetcher
- L1 Cache:
  - **Data**: 16KB, 8 way set associative, 64 byte line, 6 cycle latency
  - **Instruction**: 16KB, 4 way set associative, 3 cycle latency
- L1 Prefetcher (L1P):
  - 1 prefetch unit for each core
  - 32 entry prefetch buffer, entries are 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 Gb/s
- Peak bandwidth is 42.67 BG/s (excluding ECC)
- Latency > 350 cycles



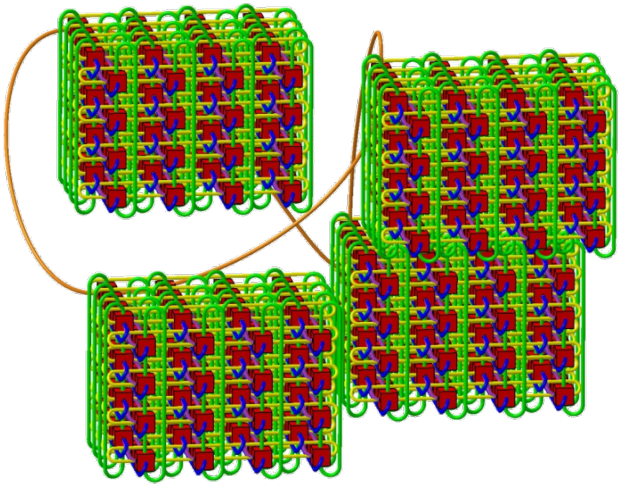
L1 Cache:
- **Data**: 16KB, 8 way assoc., 64 byte line, 6 cycle latency
- **Instruction**: 16KB, 4 way assoc., 3 cycle latency

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

INTER-PROCESSOR COMMUNICATION

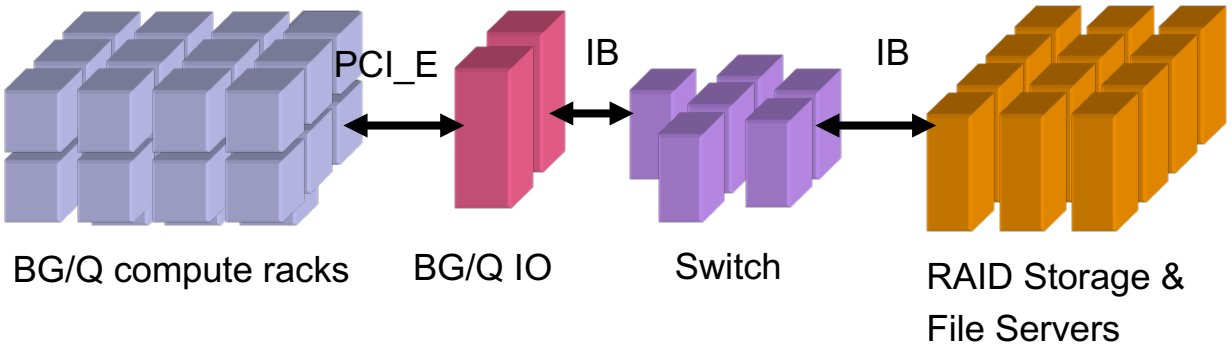
- **5D torus network:**
  - –Achieves high nearest neighbor bandwidth while increasing bisectional bandwidth and reducing hops vs 3D torus
  - –Allows machine to be partitioned into independent sub machines
    - – No impact from concurrently running codes.
  - –Hardware assists for collective & barrier 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 *u*s
  - Barrier hardware latency on 96k nodes ~ 6.3 *u*s

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

THETA ARCHITECTURE

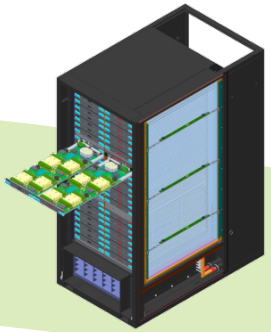
Argonne NATIONAL LABORATORY

THETA

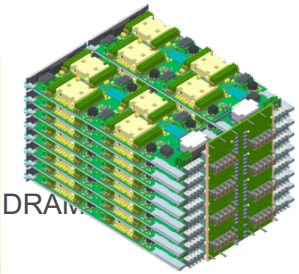
- System:
  - Cray XC40 system
  - 3,624 compute nodes/ 231,936 cores
  - ~10 PetaFlops peak performance
  - Accepted September 2016
  - Serves as a bridge between Mira and Aurora
- Memory:
  - 736 TB of total system memory
  - 16 GB MCDRAM per node
  - 192 GB DDR4-2400 per node
- Processor:
  - Intel Xeon Phi, 2nd Generation (Knights Landing) 7230
  - 64 Cores
  - 1.3 GHz base / 1.1 GHz AVX / 1.4-1.5 GHz Turbo
- Network:
  - Cray Aries interconnect
  - Dragonfly network topology
- Filesystems:
  - Project directories: 10 PB Lustre file system
  - Home directories: GPFS

THETA SYSTEM OVERVIEW

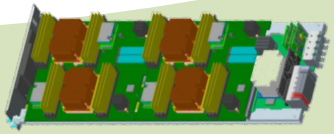
System: 20 Cabinets
3264 Nodes, 960 Switches
Dual-plane, 10 groups, Dragonfly 7.2 TB/s Bi-Sec
9.65 PF Peak
56.6 TB MCDRAM, 679.5 TB DRAM

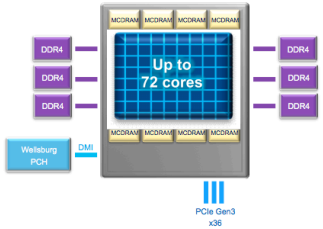


Cabinet: 3 Chassis, 75kW liquid/air cooled
510.72 TF 3TB MCDRAM, 36TB DRAM



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





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

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

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

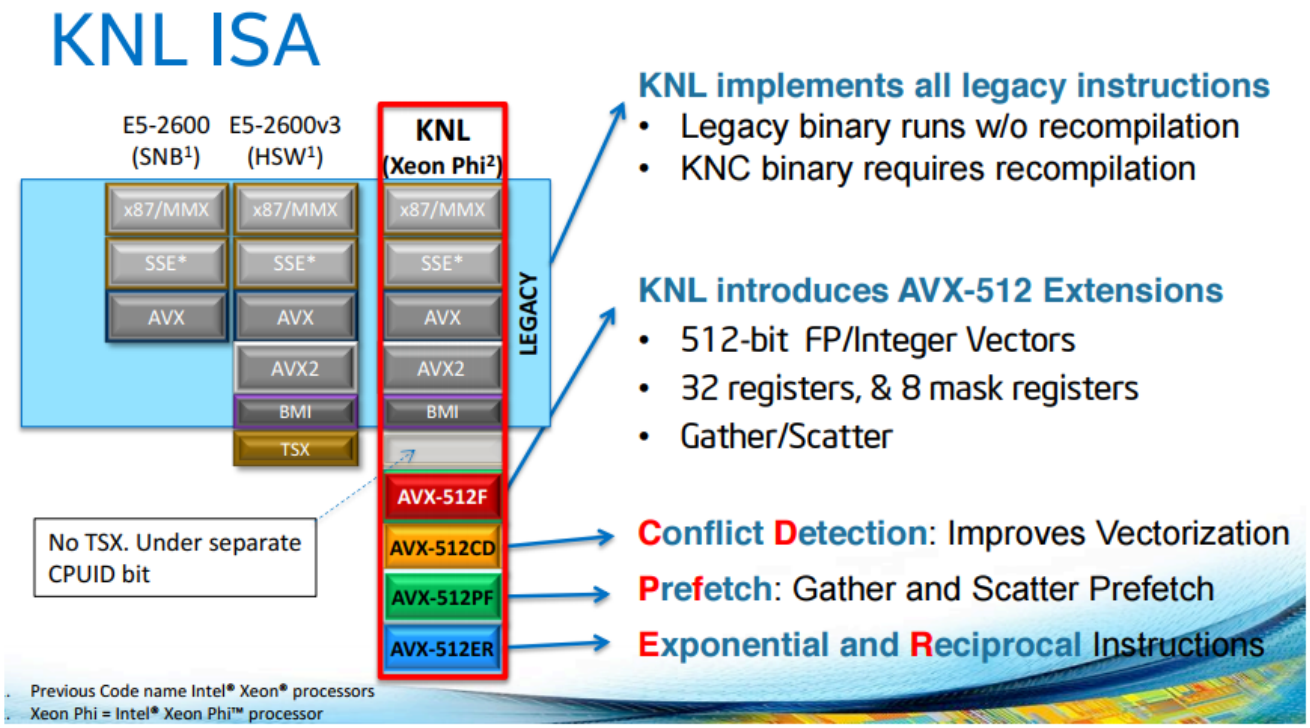
Argonne NATIONAL LABORATORY

Knights Landing Improvements

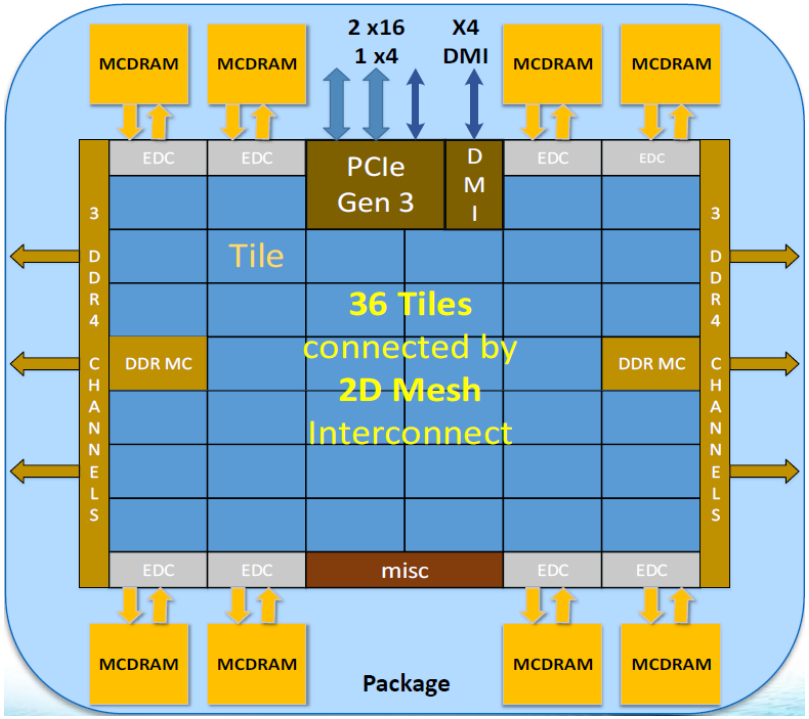
Many Trailblazing Improvements in KNL

Improvements	What/Why
Self Boot Processor	No PCIe bottleneck
Binary Compatibility with Xeon	Runs all legacy software. No recompilation.
New Core: Atom™ based	~3x higher ST performance over KNC
Improved Vector density	3+ TFLOPS (DP) peak per chip
New AVX 512 ISA	New 512-bit Vector ISA with Masks
Scatter/Gather Engine	Hardware support for gather and scatter
New memory technology: MCDRAM + DDR	Large High Bandwidth Memory → MCDRAM Huge bulk memory → DDR
New on-die interconnect: Mesh	High BW connection between cores and memory
Integrated Fabric: Omni-Path	Better scalability to large systems. Lower Cost

Knights Landing Instruction Set



KNIGHTS LANDING PROCESSOR



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

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

2D Mesh Interconnect
- Tiles connected by 2D mesh

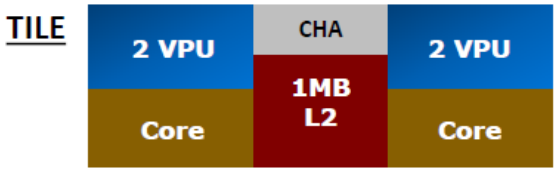
On Package Memory
- 16 GB MCDRAM
- 8 Stacks
- 480 GB/s bandwidth

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

Argonne NATIONAL LABORATORY

KNL TILE AND CORE

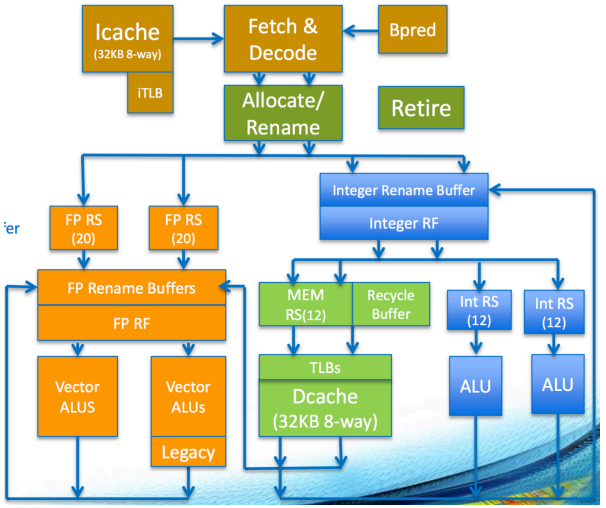


Tile

- Two CPUs
- 2 VPUs per core
- Shared 1 MB L2 cache (not global)
- Caching/Home agent
  - Distributed directory, provides coherence

Core

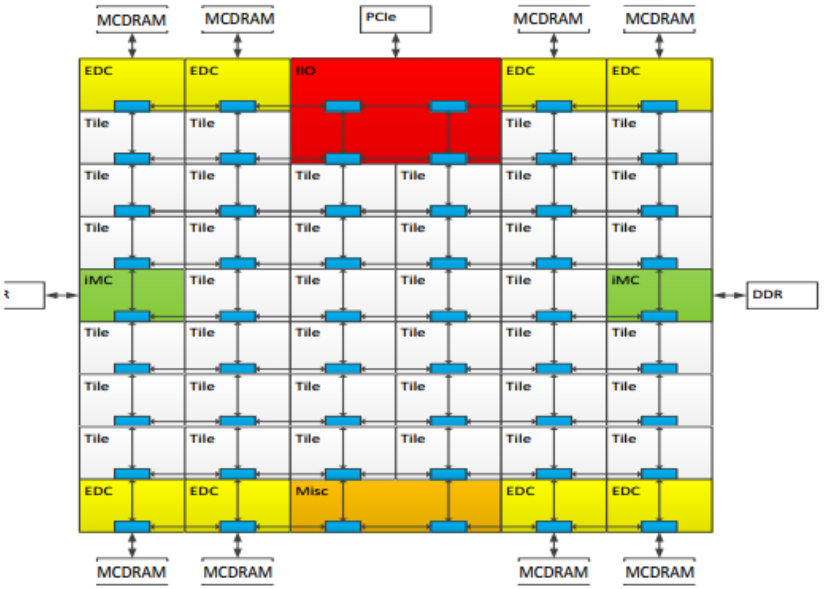
- Based on Silvermont (Atom)
- Functional units:
  - 2 Integer ALUs
  - 2 Memory units
  - 2 VPU's with AVX-512
- Instruction Issue & Exec:
  - 2 wide decode
  - 6 wide execute
  - Out of order
- 4 Hardware threads per core



Argonne NATIONAL LABORATORY

KNL Mesh Interconnect

KNL Mesh Interconnect



Mesh of Rings

- Every row and column is a (half) ring
- YX routing: Go in Y → Turn → Go in X
- Messages arbitrate at injection and on turn

Cache Coherent Interconnect

- MESIF protocol (F = Forward)
- Distributed directory to filter snoops

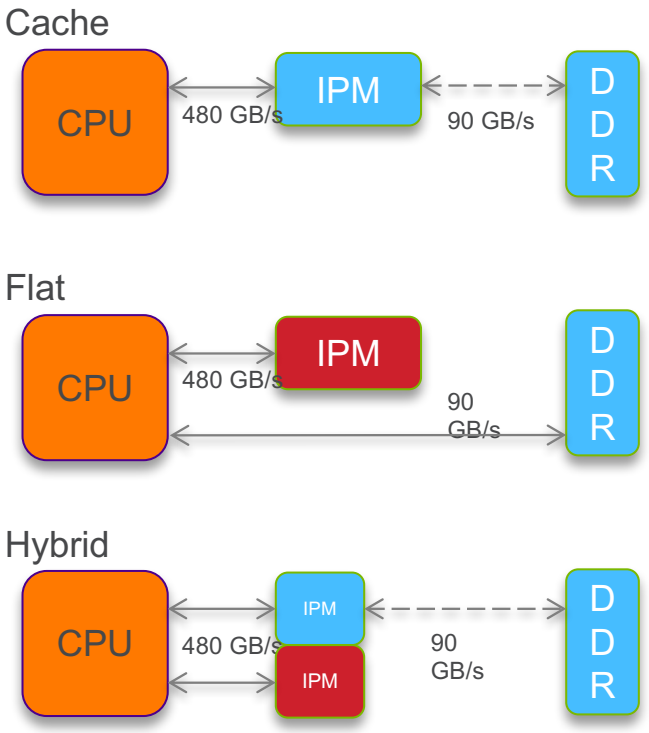
Three Cluster Modes

(1) All-to-All (2) Quadrant (3) Sub-NUMA Clustering

Argonne NATIONAL LABORATORY

MEMORY MODES - IPM AND DDR

SELECTED AT NODE BOOT TIME



- **Two memory types**
  - In Package Memory (IPM)
    - 16 GB MCDRAM
    - ~480 GB/s bandwidth
  - Off Package Memory (DDR)
    - Up to 384 GB
    - ~90 GB/s bandwidth
- **One address space**
  - Possibly multiple NUMA domains
- **Memory configurations**
  - Cached: DDR fully cached by IPM
  - Flat: user managed
  - Hybrid: ¼, ½ IPM used as cache
- **Managing memory:**
  - jemalloc & memkind libraries
  - Pragmas for static memory allocations

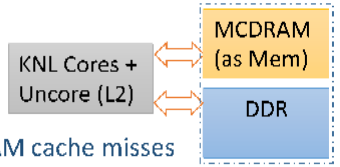
Memory Modes

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

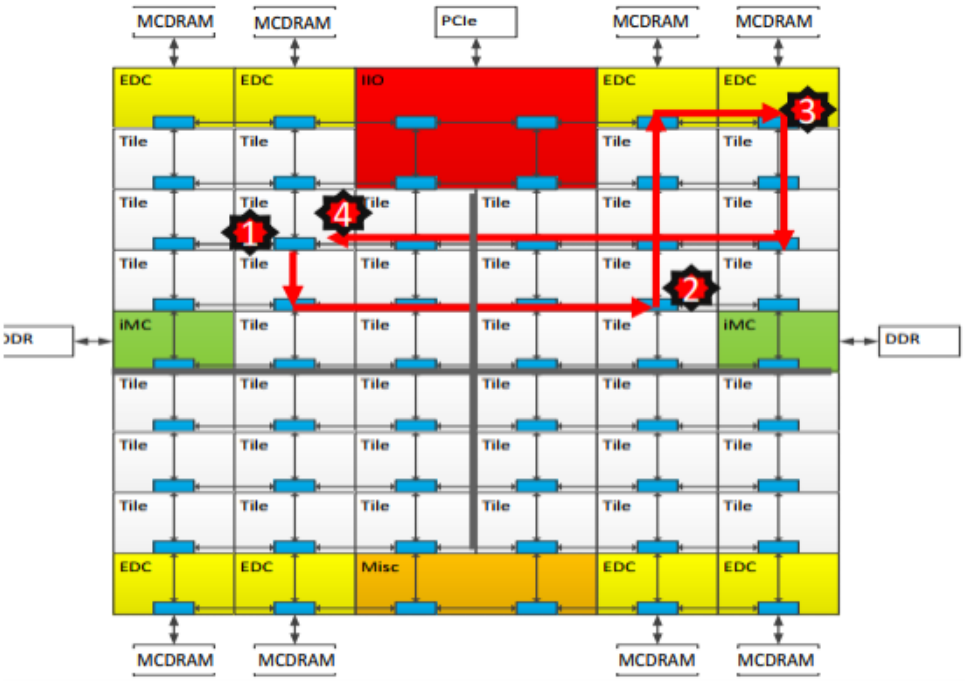


STREAM TRIAD BENCHMARK PERFORMANCE

- 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

Cluster Modes: Quadrant



Chip divided into four virtual Quadrants

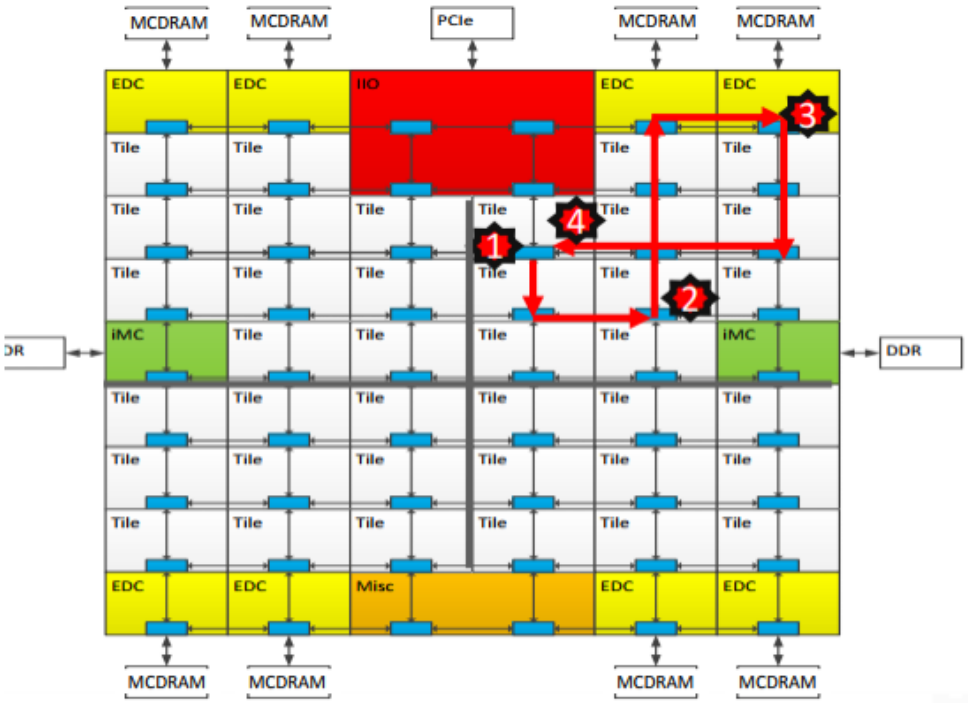
Address hashed to a Directory in the same quadrant as the Memory

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



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

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.

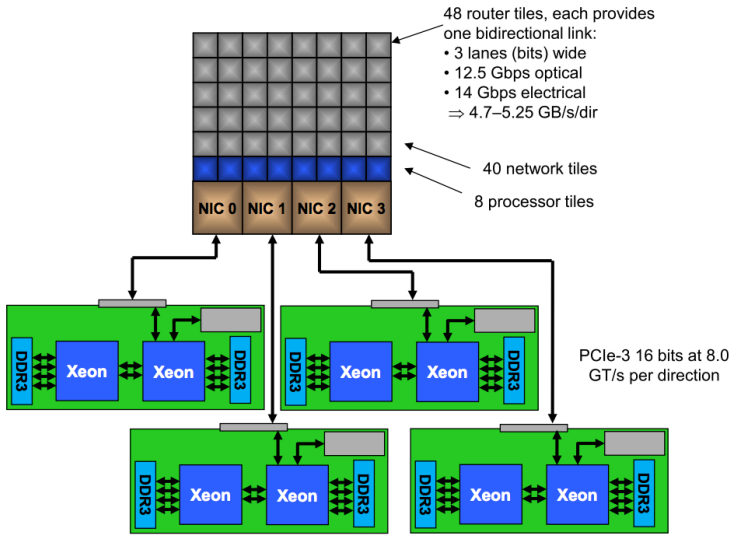
SW needs to NUMA optimize to get benefit.

Argonne NATIONAL LABORATORY

ARIES DRAGONFLY NETWORK

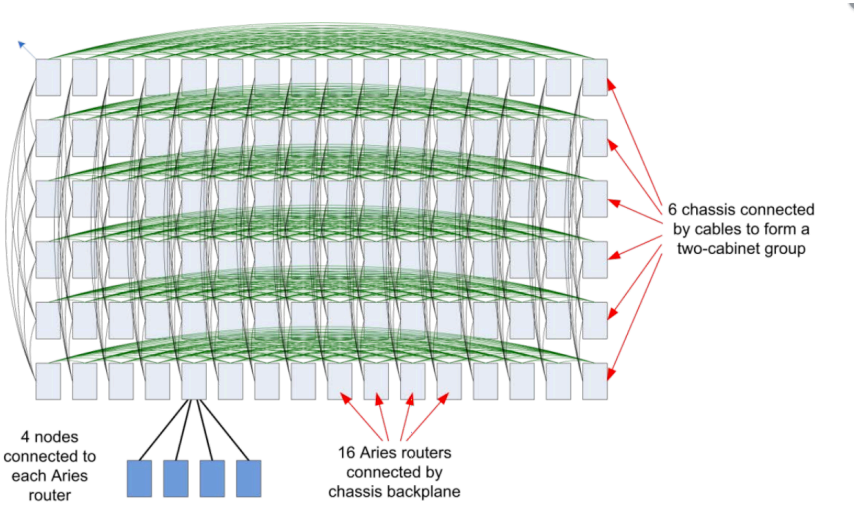
Aries Router:

- 4 NIC's connected via PCIe
- 40 Network tiles/links
- 4.7-5.25 GB/s/dir per link



Dragonfly topology

- 4 nodes connected to an Aries
- 2 Local all-to-all dimensions
  - 16 all-to-all horizontal
  - 6 all-to-all vertical
- 384 nodes in local group
- All-to-all connections between groups



Argonne NATIONAL LABORATORY

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

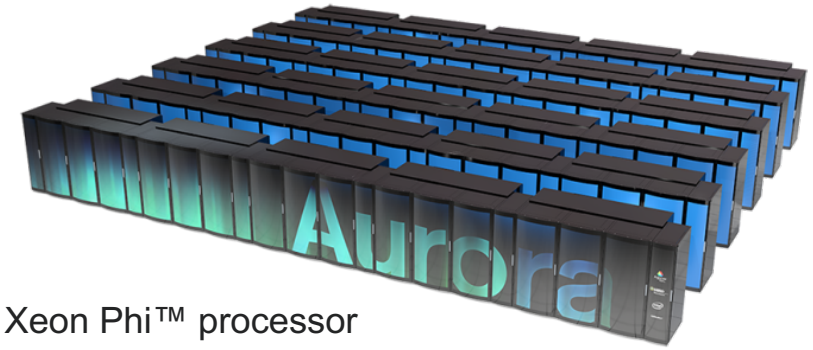
CONSIDERATIONS IN MOVING FROM MIRA TO THETA

- 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)

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 Telsa 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

Argonne NATIONAL LABORATORY

QUESTIONS?