Blue Gene/Q and Knights Landing Many Core Architectures
Argonne Training Program on Extreme Scale Computing

Scott Parker
Argonne Leadership Computing Facility
8/01/2016

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 (8.5 PF) Cray XC40 with Xeon Phi (KNL)
- **2018:**
  - Aurora (180+ PF) Cray/Intel Xeon Phi (KNH) to be delivered

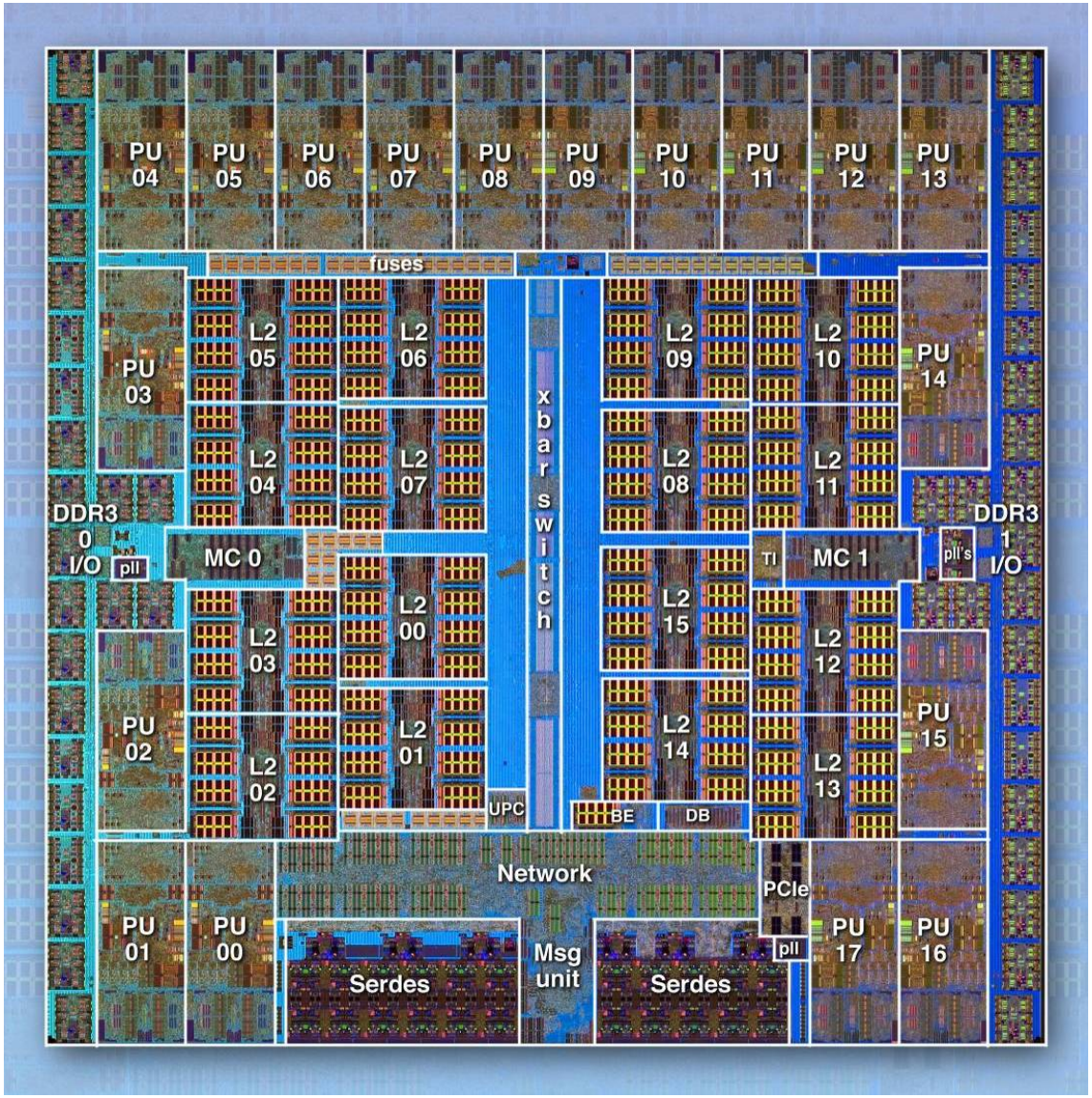
Current ALCF Systems

- ***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 (#6 Top 500)
- ***Cetus & Vesta (T&D)* - BG/Q systems**
  - 4K & 2k nodes / 64k & 32k cores
  - 64 TB & 32 TB of memory
  - 820 TF & 410 TF peak flop rate
- ***Cooley*– x86 & Nvidia system**
  - 126 nodes / 1512x86 cores/ 126 nVidia Tesla K80 GPUs
  - 47 TB x86 memory / 3 TB GPU memory
  - Peak flop rate: 220 TF
- **Storage**
  - Over 30 PB capacity, 240 GB/s bw (GPFS)

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)
- **Massive parallelism:**
  - Multi/Many core (L - 2, P – 4, Q – 16)
  - Many aggregate cores (L – 208k, P – 288k, Q – 1.5M)
- **Fast communication network(s)**
  - Low latency, high bandwidth, torus network (L & P – 3D, Q – 5D)
- **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

BlueGene/Q Compute Chip



It's big
- 360 mm² Cu-45 technology (SOI)
- 1.5 B transistors
- 205 GF per node

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

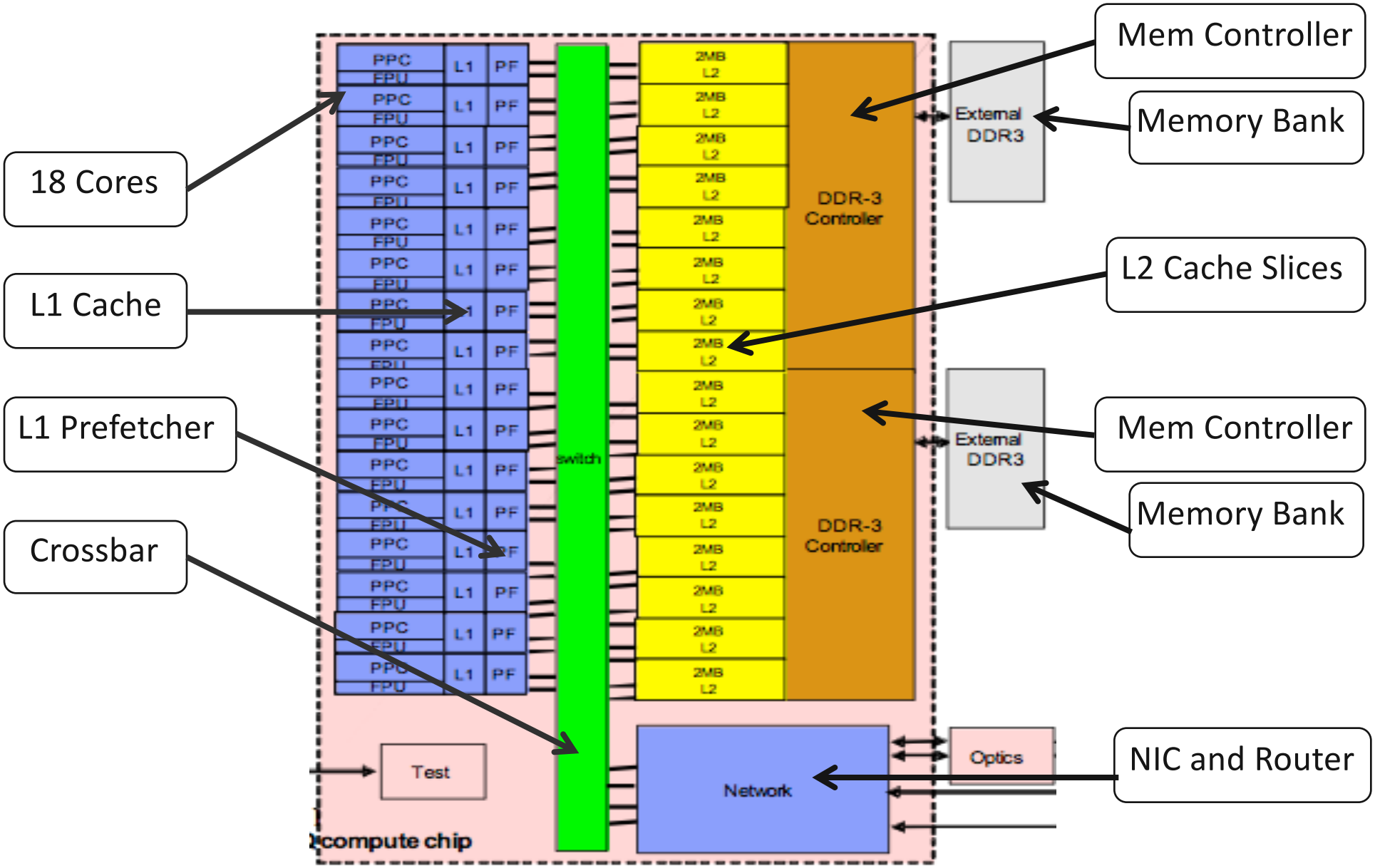
Crossbar switch
- Each core connected to shared L2
- Aggregate read rate of 409.6 GB/s

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, collective operations
- 11 network ports

BG/Q Chip, Another View

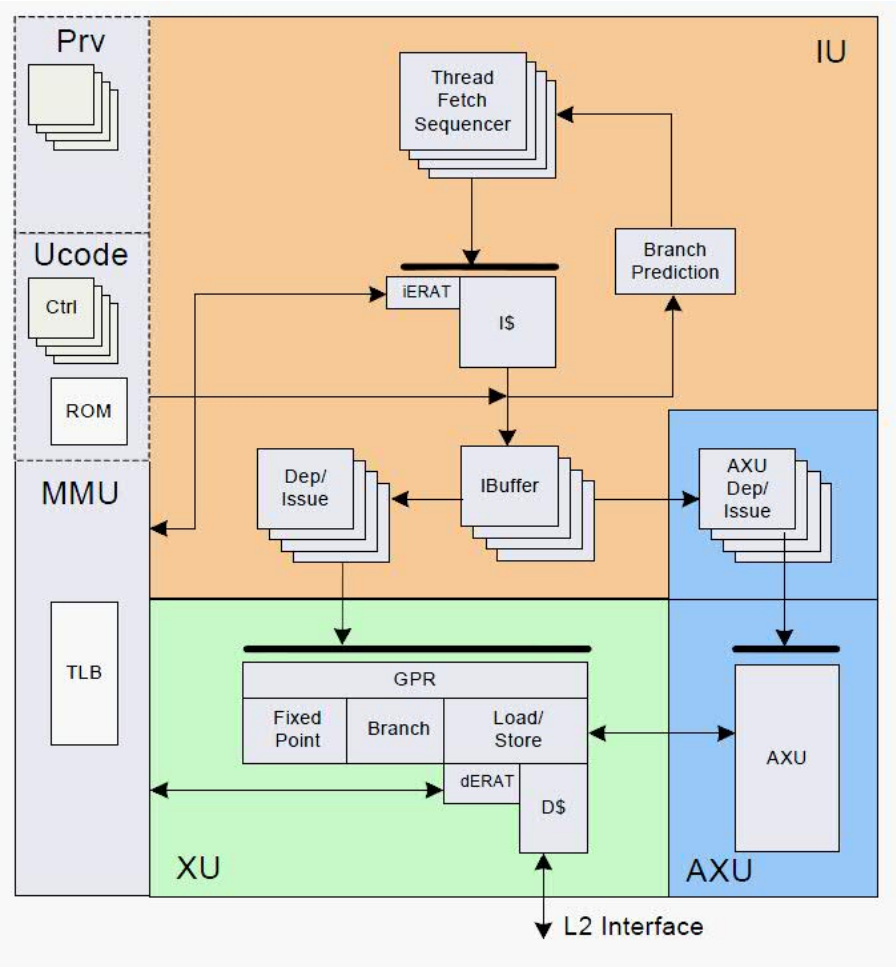


BG/Q Core

- 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)

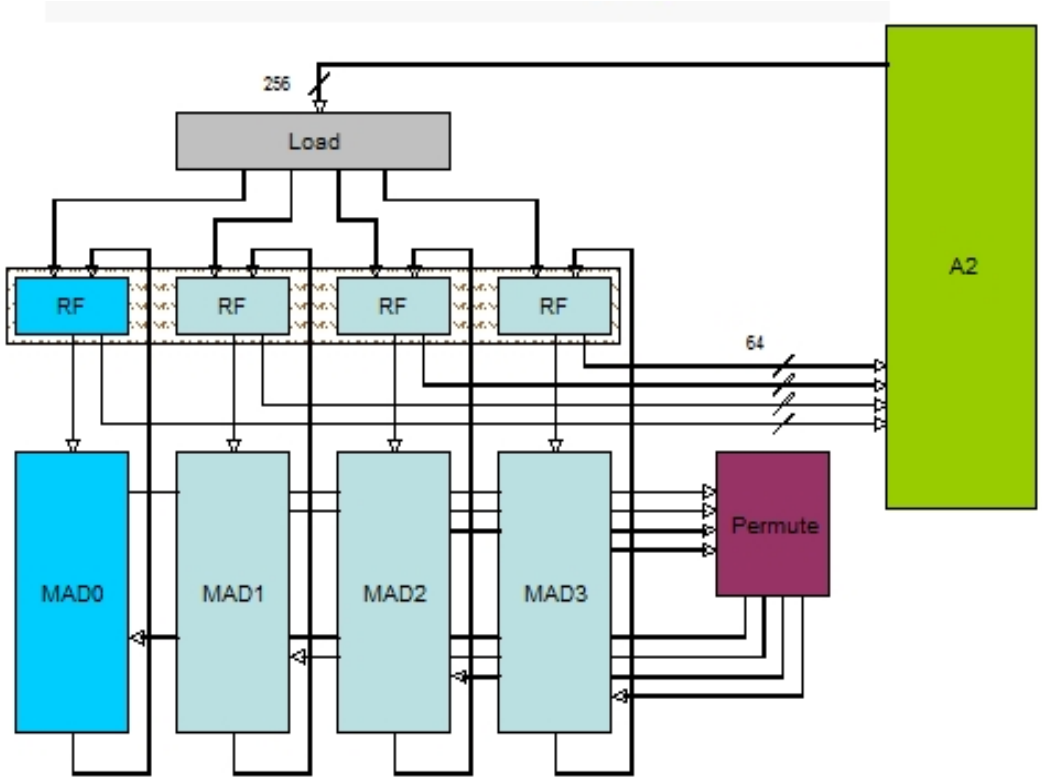


Instruction Issue:

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

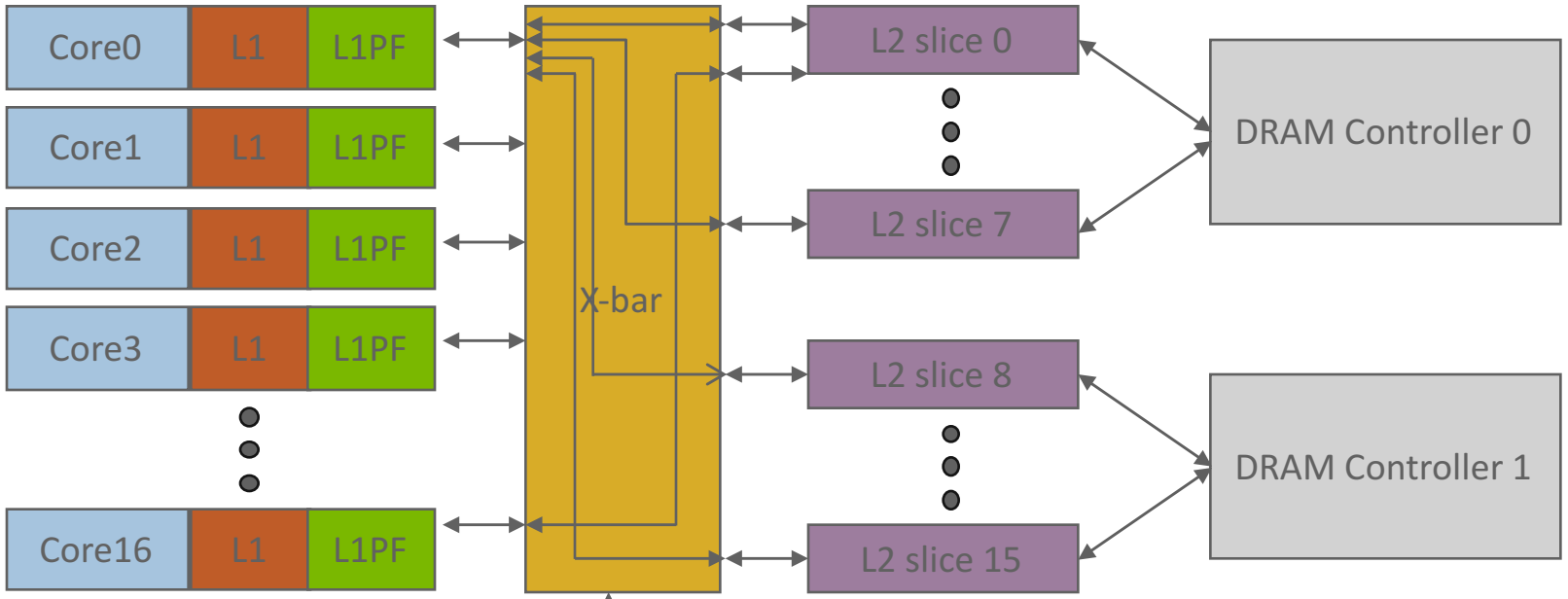


BG/Q Memory Hierarchy

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



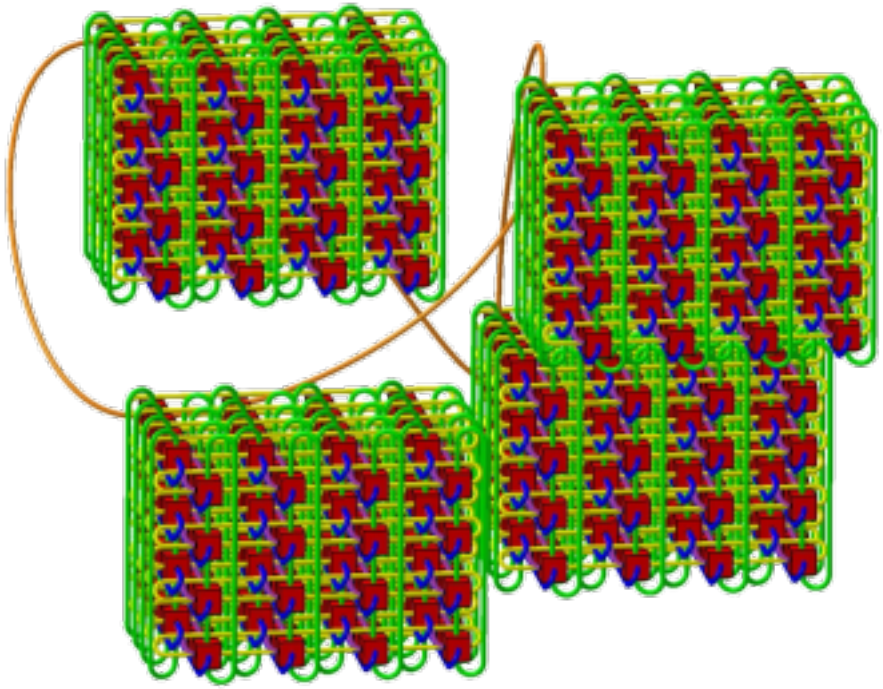
L1 Cache:
- **Data**: 16KB, 8 way assoc., 64 byte line, 6 cycle latency
- **Instruction**: 16KB, 4 way assoc., 3 cycle latency

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

The Blue Gene/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)
- **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*



Blue Gene/Q Software High-Level Goals & Philosophy

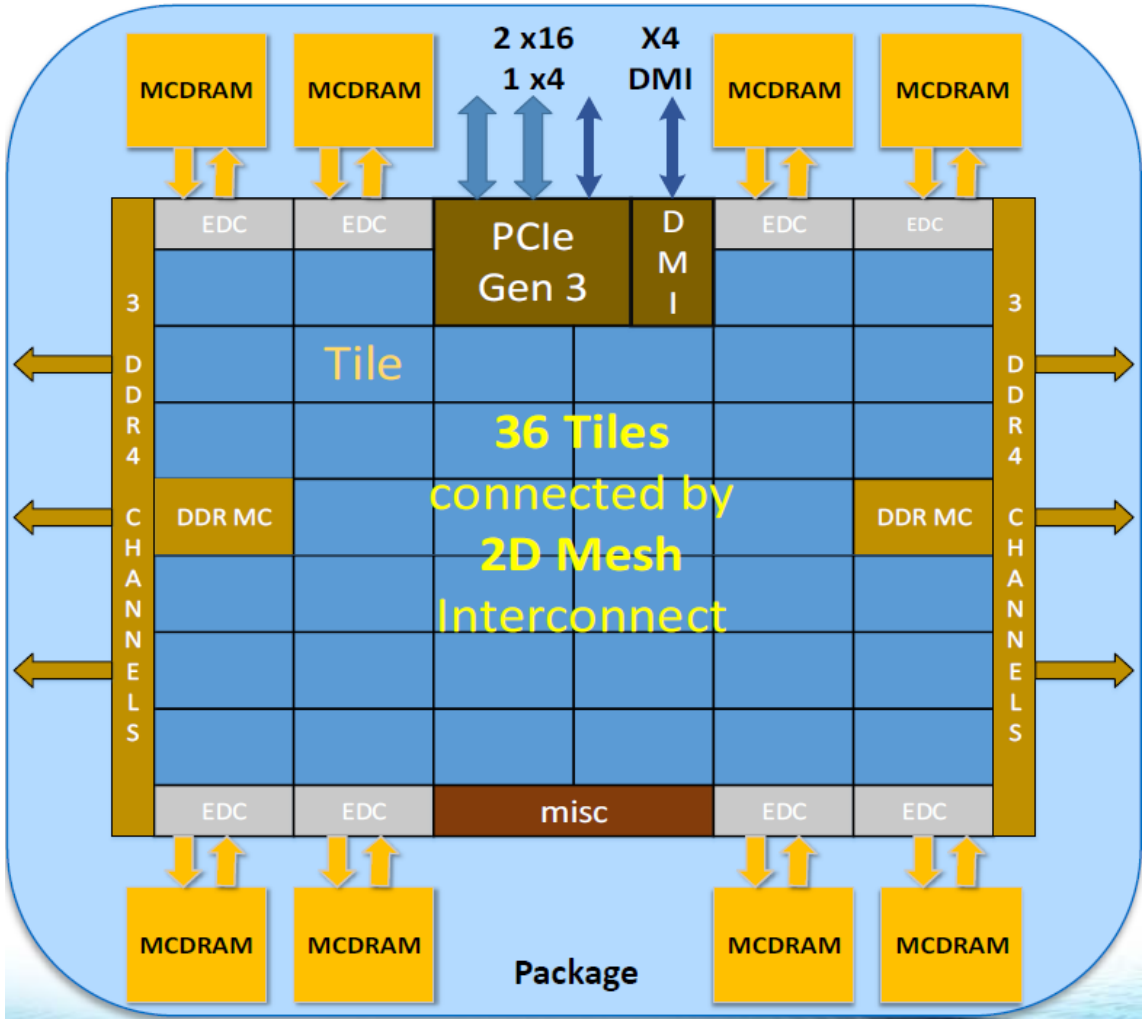
- Facilitate extreme scalability
  - Low noise on compute nodes
  - File I/O offloaded to I/O nodes running full Linux
  - GLIBC environment with a few restrictions for scaling
- Familiar programming modes such as MPI and OpenMP
  - Scalable MPICH2 providing MPI 2.2 with extreme message rate
  - Efficient intermediate (PAMI) and low-level (SPI) message libraries
  - PAMI layer allows easy porting of runtimes like GA/ARMCI, Berkeley UPC, etc
- Standards-based when possible
  - Linux development environment: familiar GNU toolchain with glibc, pthreads
  - XL Compilers C, C++, Fortran with OpenMP 3.1
  - Debuggers: Totalview
  - Tools: HPC Toolkit, TAU, PAPI, Valgrind
- 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
- Flexible and fast job control – with high availability
  - Noise-free partitioned networks
  - Integrated HPC, HTC, MPMD, and sub-block jobs
- Facilitate new programming models

Future ALCF Systems

- **Theta**
  - Arriving in 2016
  - Intel Xeon Phi, 2nd Generation (Knights Landing)
  - 3,240 compute nodes: 207,360 cores
  - 660 TB of memory
  - 8.5 PetaFlops peak performance
  - Cray Aries interconnect
  - Dragonfly network topology
  - 10 PB Lustre file system

- **Aurora**
  - Arriving in 2018
  - Intel Xeon Phi, 3rd Generation (Knights Hill)
  - Over 50,000 compute nodes
  - Greater than 7 PB of persistent memory
  - 180+ PetaFlops peak performance
  - Intel Omni-Path interconnect
  - Dragonfly network topology
  - Over 150 PB Lustre file system

Knights Landing Processor



It's bigger
- 683 mm²
- 14 nm process
- 8 Billion transistors

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

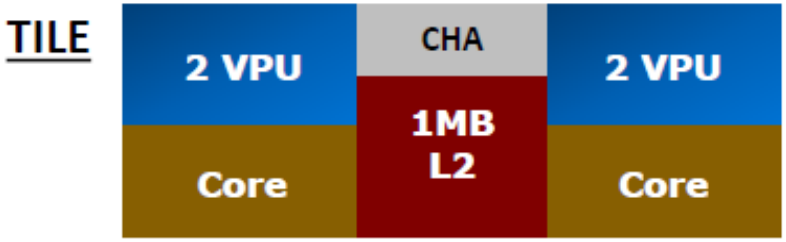
2D Mesh Interconnect
- Tiles connected by 2D mesh

On Package Memory
- 16 GB MCDRAM
- 8 Stacks
- ~450 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

KNL Tile and Core

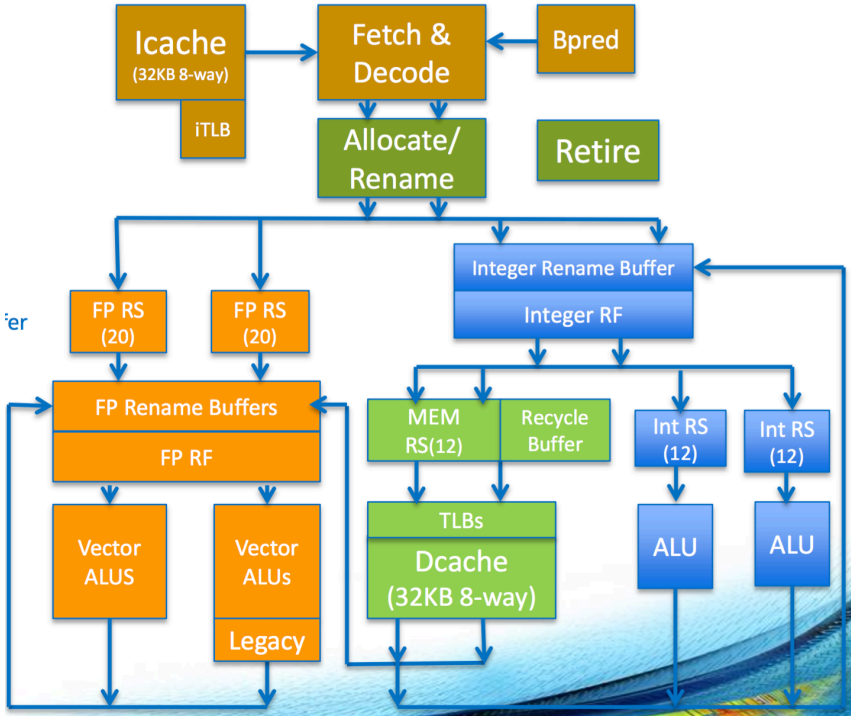


Tile

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

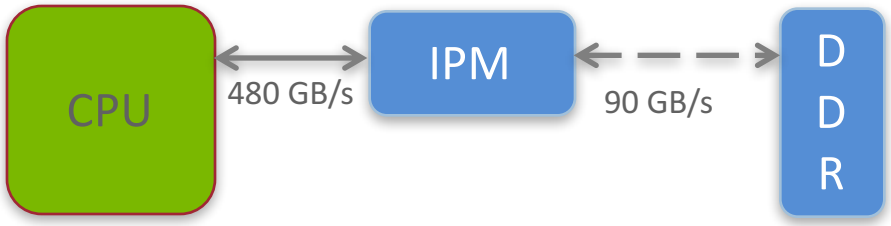
Core

- 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

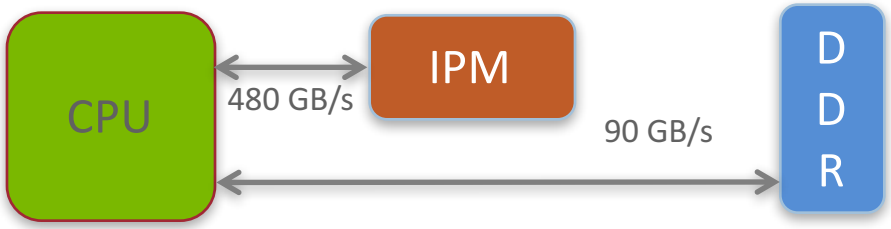


Programming with IPM and DRR

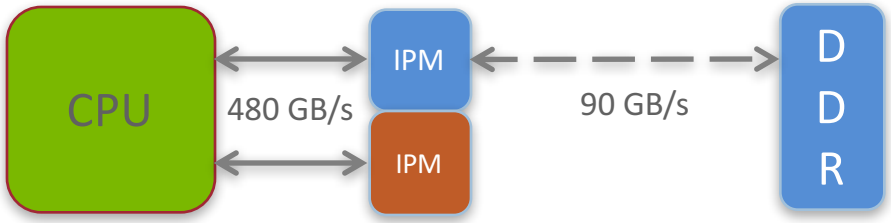
Fully Cached



Direct Mapped



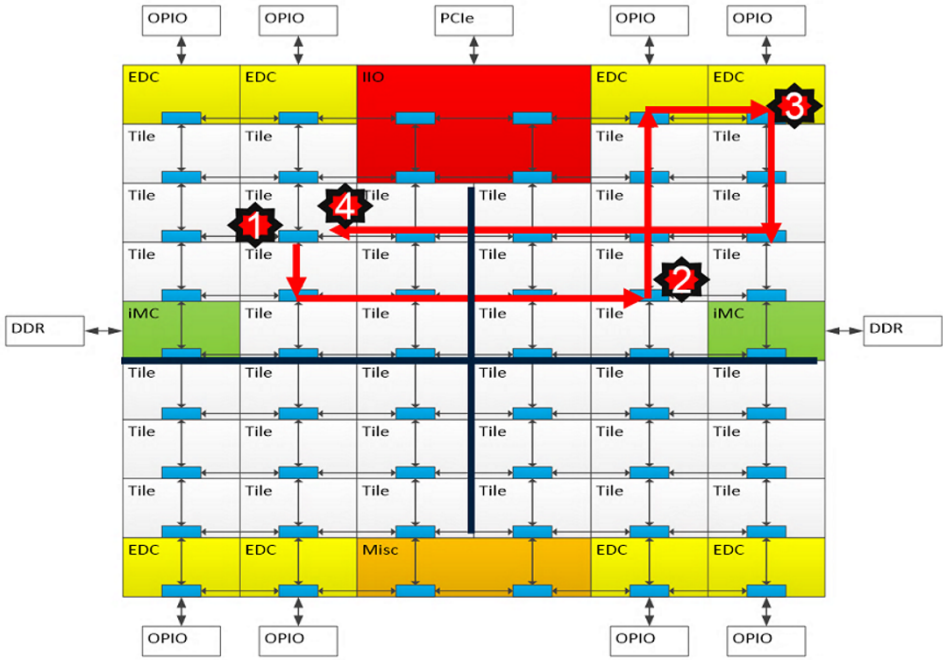
Hybrid



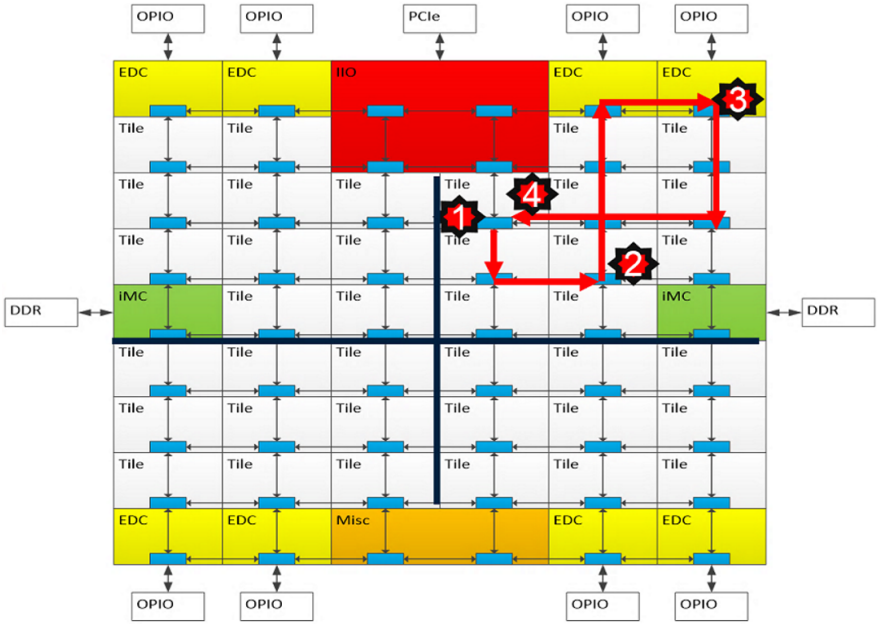
- **Two memory types**
  - In Package Memory (IPM)
    - 16 GB MCDRAM
    - ~480 GB/s bandwidth
  - Off Package Memory (DDR)
    - Up to 384 GB
    - ~90 GB/s bandwidth
- **One address space**
  - Possibly multiple NUMA domains
- **Memory configurations**
  - Cached: DDR fully cached by IPM
  - Direct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

Clustering Modes

Quadrant Mode

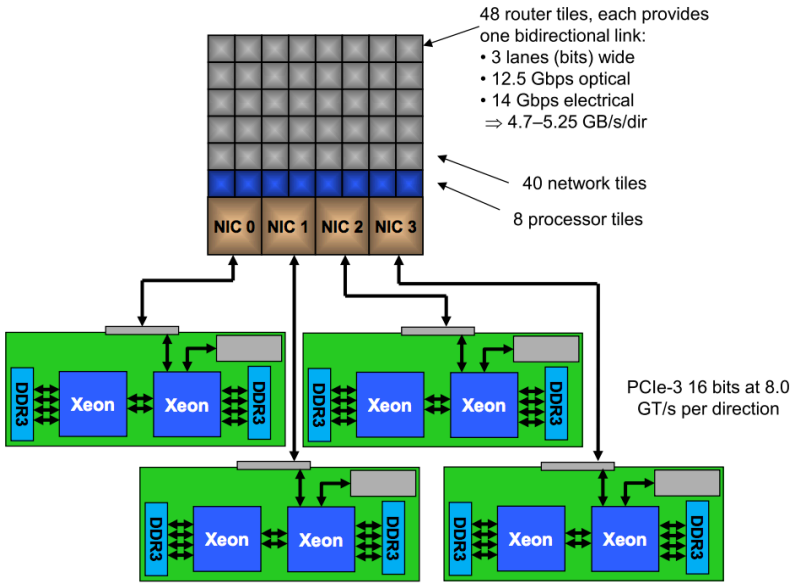


SNC-4 Mode



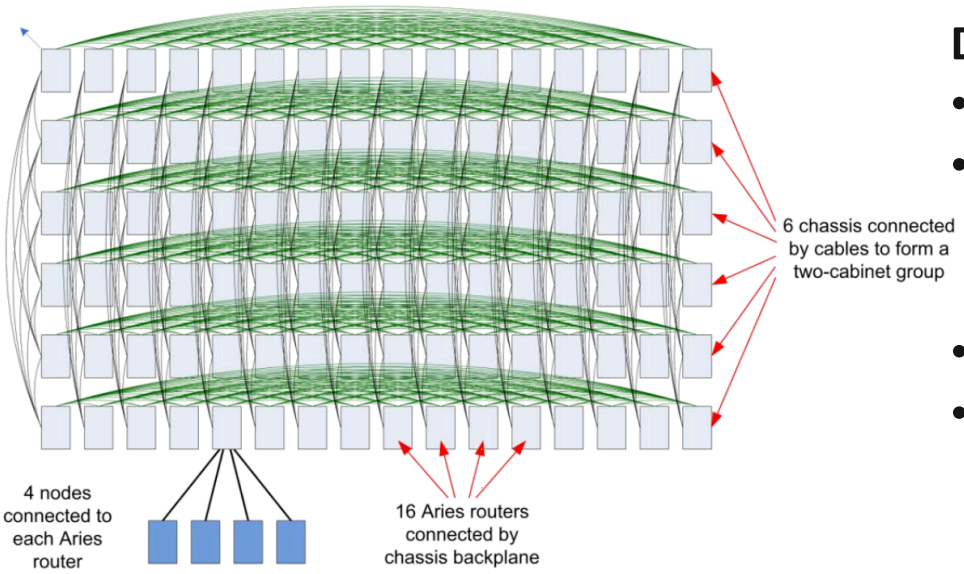
1. Tile has Cache Miss
2. CHA selected
3. Memory Controller
4. Memory Received

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

Considerations In Moving From BG/Q to Xeon Phi (KNL)

- **More local parallelism**
  - 64-72 cores (KNL) vs 16 (BG/Q)
  - 4 hardware threads on both
- **No increase in the number of nodes**
- **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)

Questions?