ARCHITECTURE OF THE ARGONNE CRAY XC40 KNL SYSTEM 'THETA'

SCOTT PARKER
Lead, Performance Engineering Team
Argonne Leadership Computing Facility

July 31, 2017

ALCF SYSTEMS

Mira – IBM BG/Q	Cetus – IBM BG/Q	Vesta – IBM BG/Q	Cooley - Cray/NVIDIA	Theta - Cray XC40
– 49,152 nodes	– 4,096 nodes	– 2,048 nodes	– 126 nodes (Haswell)	– 3,624 nodes (KNL)
– 786,432 cores	– 65,536 cores	– 32,768 cores	– 1512 cores	– 231,936 cores
– 786 TB RAM	– 64 TB RAM	– 32 TB RAM	– 126 Tesla K80	– 736 TB RAM
– 10 PF	– 836 TF	– 419 TF	– 48 TB RAM (3 TB GPU)	– 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)

- **2019:**
 - Aurora Cray/Intel Xeon Phi to be delivered

THETA

- System:
  - Cray XC40 system
  - 3,624 compute nodes/ 231,936 cores
  - ~10 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:
  - 736 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
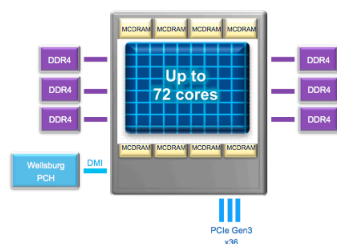 - Home directories: GPFS

THETA SYSTEM OVERVIEW

System: 20 Cabinets
3264 Nodes, 960 Switches
10 groups, Dragonfly 7.2 TB/s Bi-Sec
9.65 PF Peak
56.6 TB MCDRAM, 679.5 TB DRAM
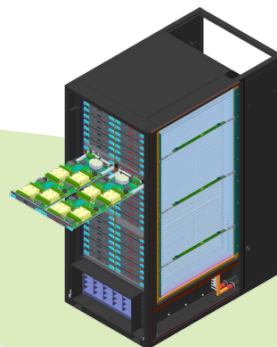
Cabinet: 3 Chassis
510.72 TF
3TB MCDRAM, 36TB DRAM

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

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

Compute Blade:
4 Nodes/Blade + Aries switch
10.64 TF
64GB MCDRAM, 768GB DRAM
128GB SSD
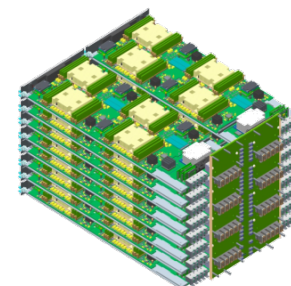
Node: KNL Socket
2.66 TF
16GB MCDRAM, 192 GB DDR4 (6 channels)

DDR4
DDR4
DDR4
MCDRAM MCDRAM MCDRAM MCDRAM
Up to
72 cores
MCDRAM MCDRAM MCDRAM MCDRAM
DDR4
DDR4
DDR4
Wellsburg
PCH
DMI
PCIe Gen3
x36

CRAY
Aries

Argonne
NATIONAL LABORATORY

Knights Landing Improvements

Improvement	Impact
Self Booting	No PCIe bottleneck
Binary Compatible with Xeon	Runs legacy code, no recompile
New Core Architecture (Atom based)	~3x higher performance than KNC
Improved Vector Density	3+ TFlops (DP) Peak per chip
New 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
New on-die interconnect: 2D mesh	High bandwidth connection between cores and memory
Integrated Omni-path Fabric	Better scalability at lower cost

Argonne
NATIONAL LABORATORY

KNIGHTS LANDING PROCESSOR
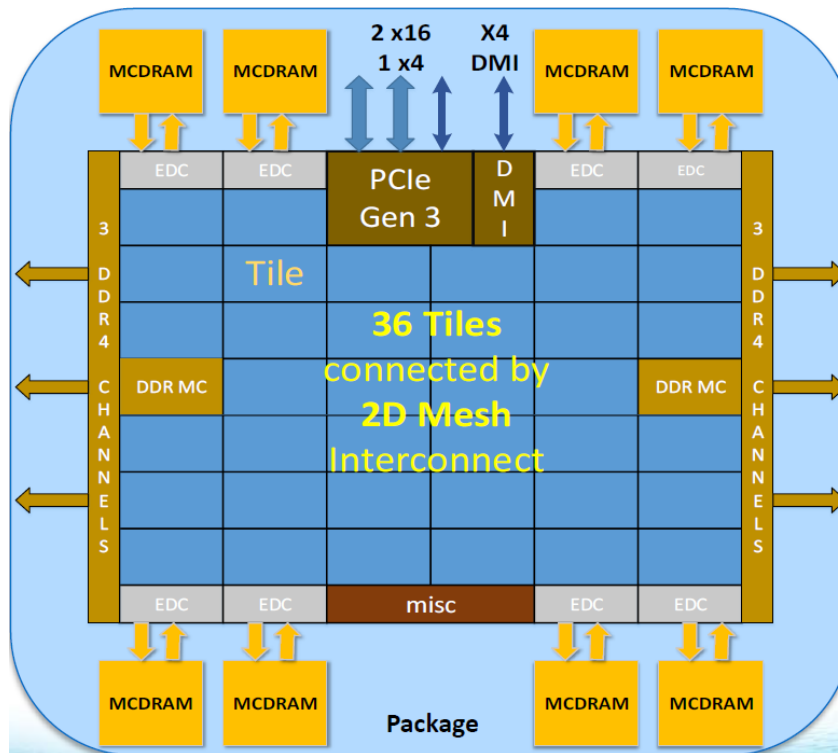
Chip
- 683 mm²
- 14 nm process
- 8 Billion transistors

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
- 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 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

Argonne
NATIONAL LABORATORY

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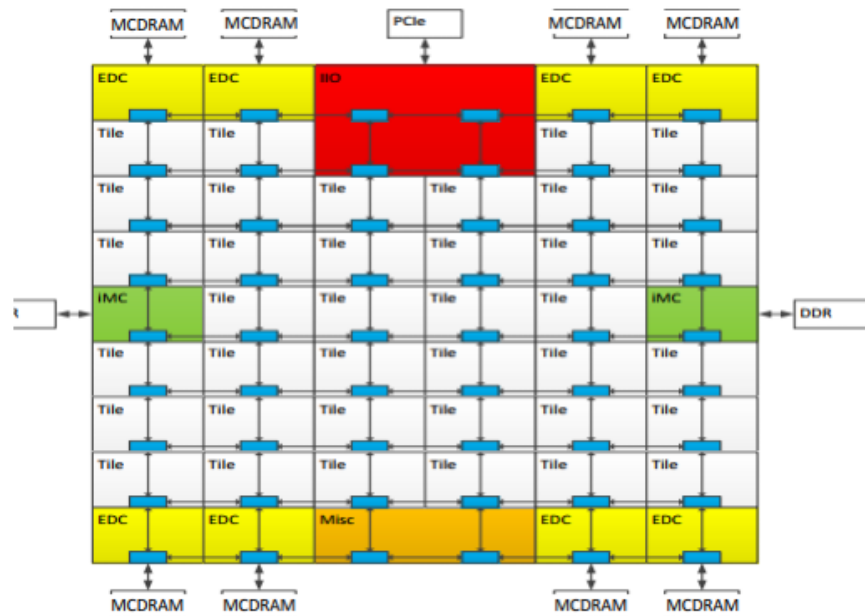
- 2D mesh interconnect connects
  - Tiles (CHA)
  - MCDRAM controllers
  - DDR controllers
  - Off chip I/O (PCIe, DMI)
- YX routing:
  - Go in Y→ turn → Go in X
  - Messages arbitrate on injection and on turn
- Cache coherent
  - Uses MESIF protocol
- Clustering mode allow traffic localization
  - All-to-all,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

Argonne
NATIONAL LABORATORY

Cluster Modes: All-to-All
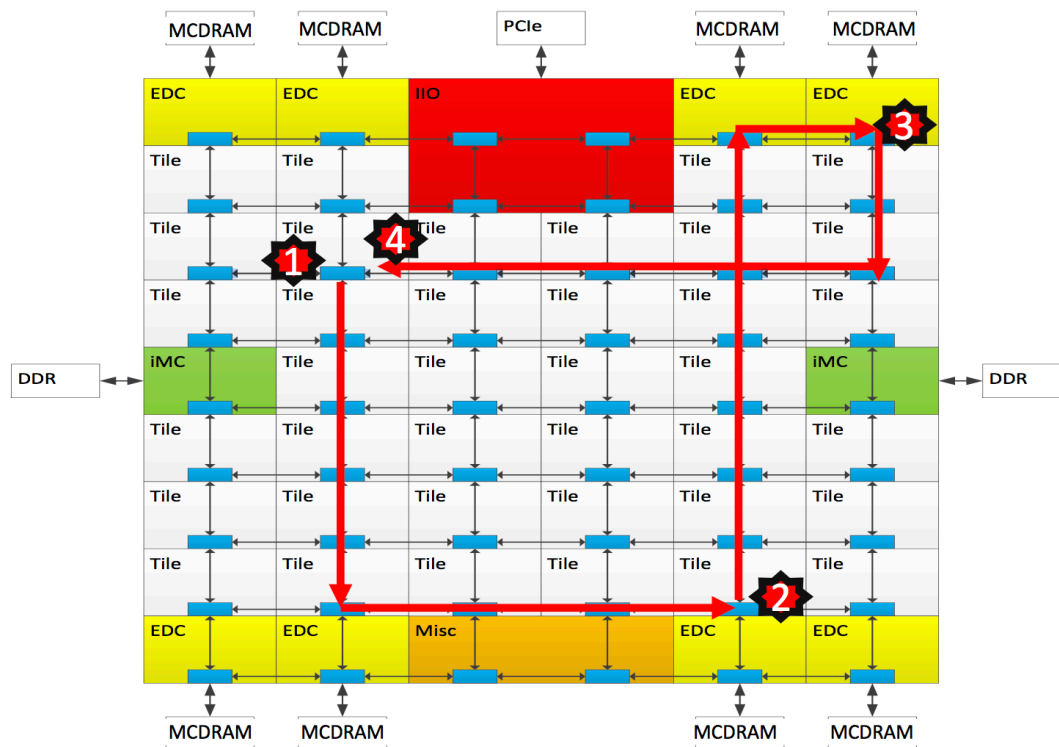
Address uniformly hashed across all distributed directories
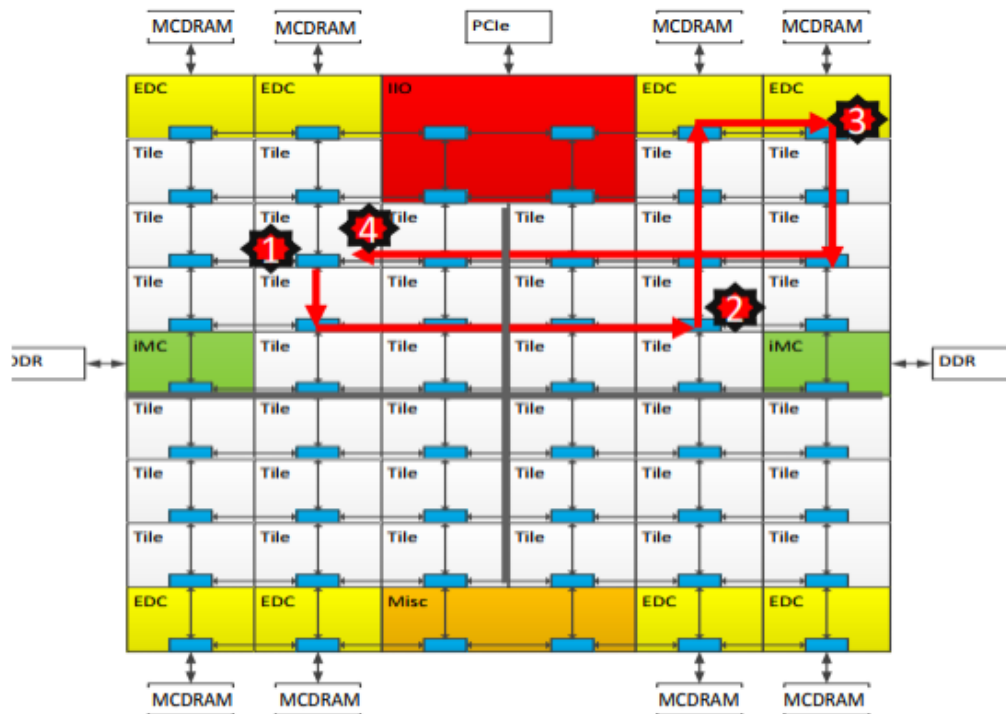
No affinity between Tile, Directory and Memory

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 requestor

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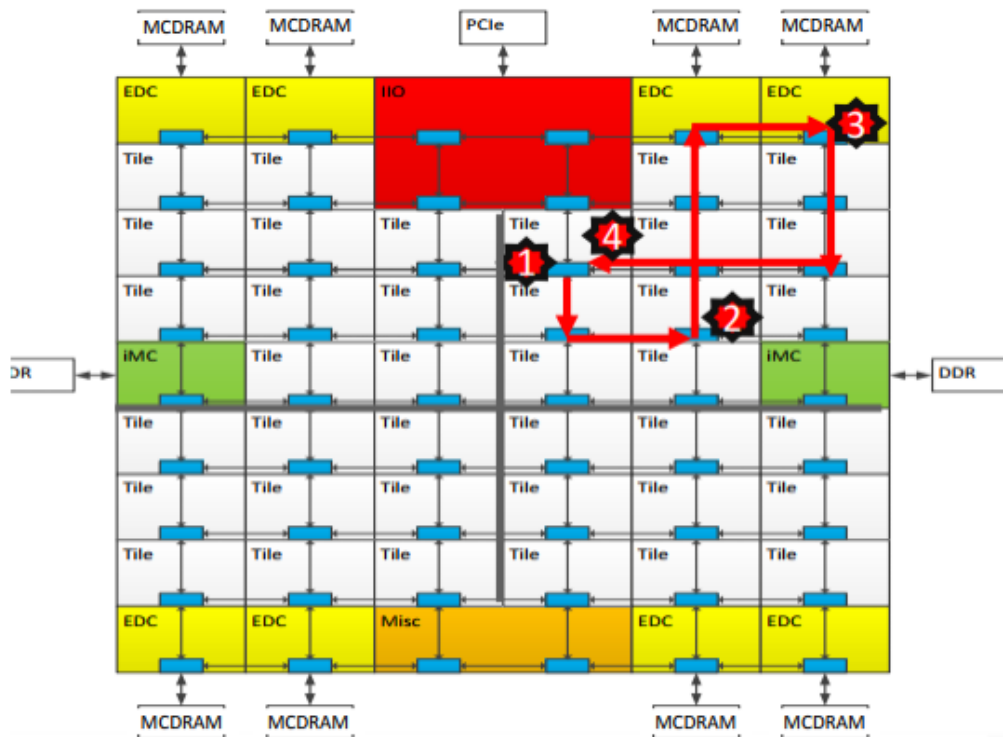
Each Quadrant (Cluster) exposed as a separate NUMA domain to OS.

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

KNL TILE
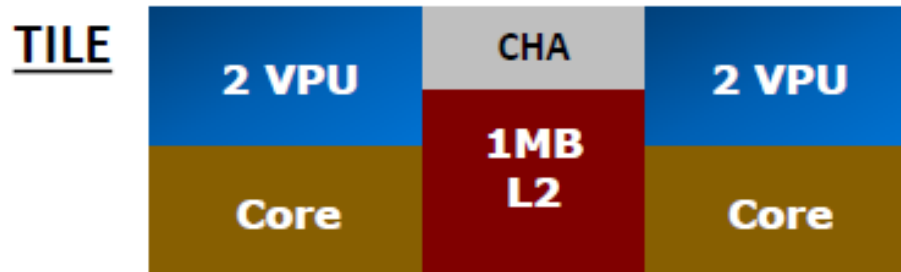
- Two CPUs
- 2 vector units (VPUs) per core
- 1 MB Shared L2 cache
    - Coherent across all tiles (32-36 MB total)
    - 16 Way
    - 1 line read and ½ line write per cycle
- Caching/Home agent
    - Distributed tag directory, keeps L2s coherent
    - Implements MESIF cache coherence protocol
    - Interface to mesh

KNL CORE

- Based on Silvermont (Atom)
- Functional units:
    - 2 Integer ALUs (Out of Order)
    - 2 Memory units (In Order reserve, OoO complete)
    - 2 VPU's with AVX-512 (Out of Order)
- Instruction Issue & Execute:
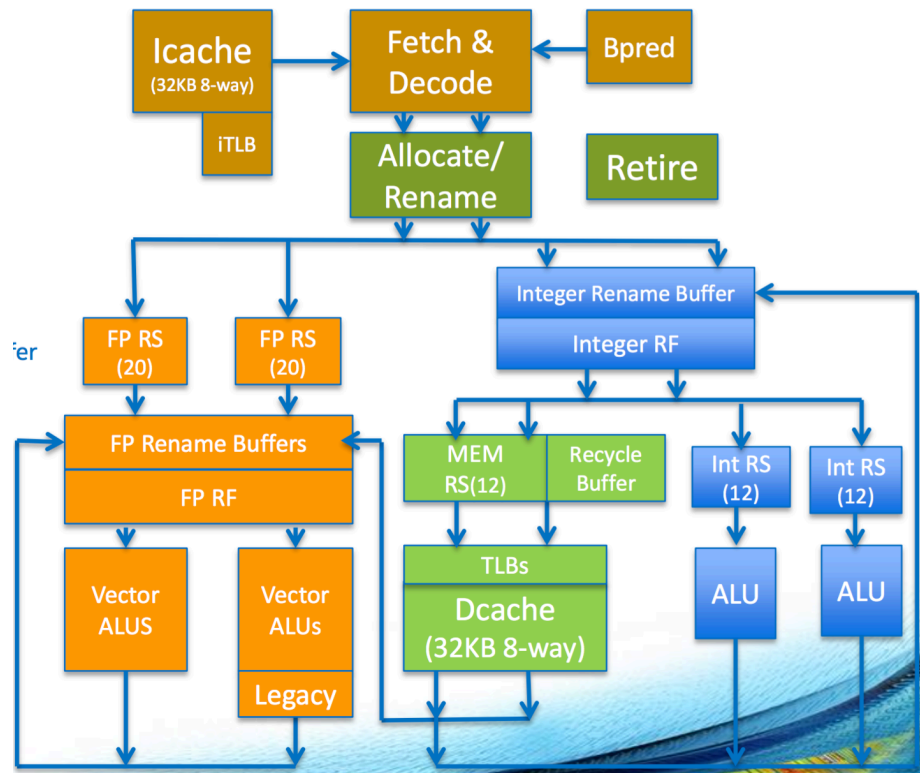   - 2 wide decode/rename/retire
    - 6 wide execute
- L1 data cache
    - 32 KB, 8 way associative
    - 2 64B load ports, 1 64B write port
- 4 Hardware threads per core
    - 1 active thread can use full resources of core
    - ROB, Rename buffer, RD dynamically paritioned between threads
    - Caches and TLBs shared

Knights Landing Instruction Set

E5-2600 (SNB[1])	E5-2600v3 (HSW[1])	E5-2600v4 (BDX[1])	KNL (Xeon Phi[2])	Future Xeon	
x87/MMX	x87/MMX	x87/MMX	x87/MMX	x87/MMX	
SSE*	SSE*	SSE*	SSE*	SSE*	
AVX	AVX	AVX	AVX	AVX	
	AVX2	AVX2	AVX2	AVX2	**Common ISA**
	BMI	BMI	BMI	BMI	
			AVX-512F	AVX-512F	
			AVX-512CD	AVX-512CD	
			AVX-512BW		
			AVX-512DQ		
			AVX-512VLO		

1. Previous Code name Intel® Xeon® processors
2. Xeon Phi = Intel® Xeon Phi™ processor

TSX	TSX		TSX	
		AVX-512PF		
Segment Specific ISA		AVX-512ER		

- 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

15

Argonne
NATIONAL LABORATORY

DGEMM PERFORMANCE ON THETA
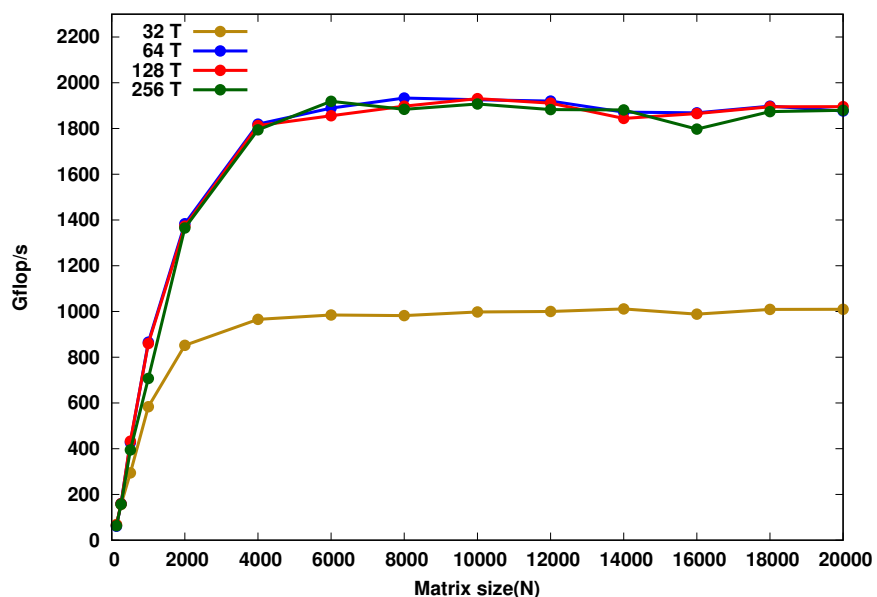
MKL DGEMM Performance

- Peak FLOP rate per node on Theta: 2252.8 GFlops
  - 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
- Performance may be limited by instruction fetch and decode
  - Instruction fetch is limited to 16 bytes
  - Up to 2 instructions may be fetched and decoded per cycle
  - AVX512 instructions with non-compressed displacements can be 12 bytes long limiting fetch to 1 instruction
- Thermal limitations restrict sustained AVX512 performance to around 1.8 instructions per cycle

16

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
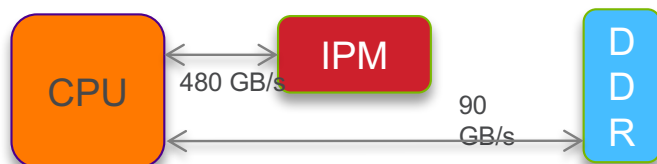 - Sub-NUMA clustering mode creates four NUMA domains

Argonne
NATIONAL LABORATORY
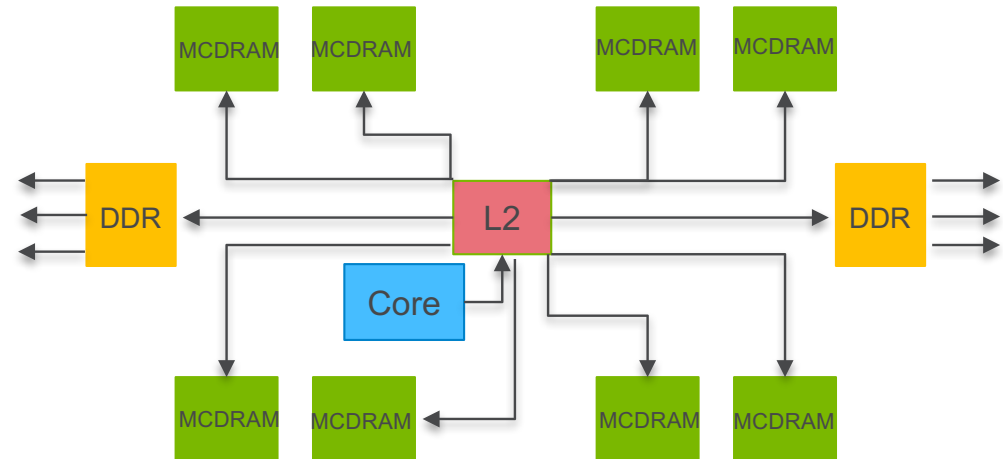
MEMORY MODES - IPM AND DDR
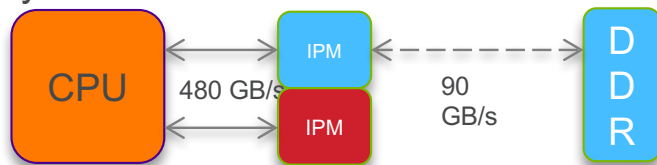## SELECTED AT NODE BOOT TIME

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:
    - ¼, ½ 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
- Observations:
  - No significant performance differences have yet been observed in different cluster modes (Quad, SNC-4, …)
  - Maximum measured single core bandwidth is 14 GB/s. Need about half the cores to saturate MCDRAM bandwidth
  - Core specialization improves memory bandwidth by ~10%

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

Argonne
NATIONAL LABORATORY

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

Argonne
NATIONAL LABORATORY

MEMORY LATENCY

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

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 $\sim C*threads^{1/2}$
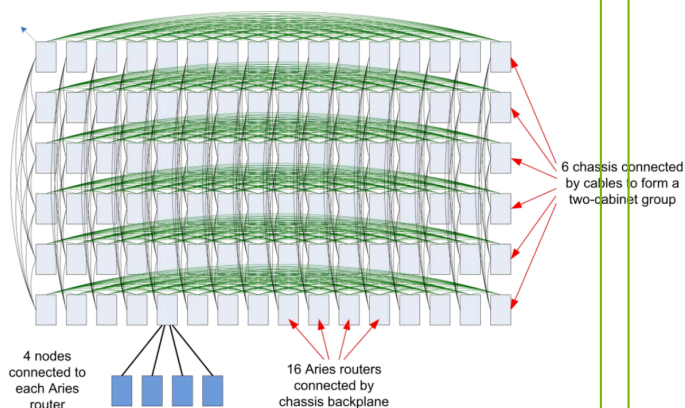
22

Argonne
NATIONAL LABORATORY

ARIES DRAGONFLY NETWORK

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

48 router tiles, each provides one bidirectional link:
- 3 lanes (bits) wide
- 12.5 Gbps optical
- 14 Gbps electrical
$\Rightarrow$ 4.7–5.25 GB/s/dir

40 network tiles

8 processor tiles

NIC 0 NIC 1 NIC 2 NIC 3

PCIe-3 16 bits at 8.0 GT/s per direction
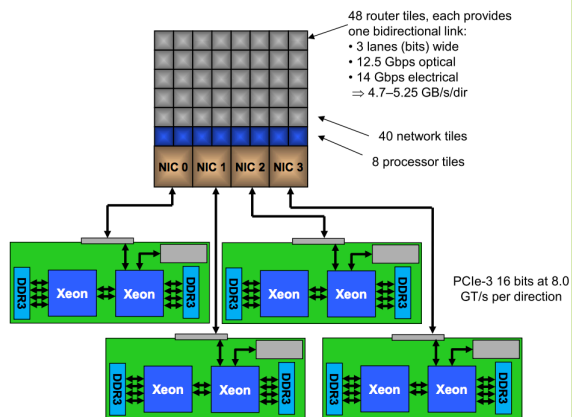
Connections within a group:
- 2 Local all-to-all dimensions
  - 16 all-to-all horizontal
  - 6 all-to-all vertical
- 384 nodes in local group

6 chassis connected by cables to form a two-cabinet group

4 nodes connected to each Aries router

16 Aries routers connected by chassis backplane

Connectivity between groups:
- Each group connected to every other group
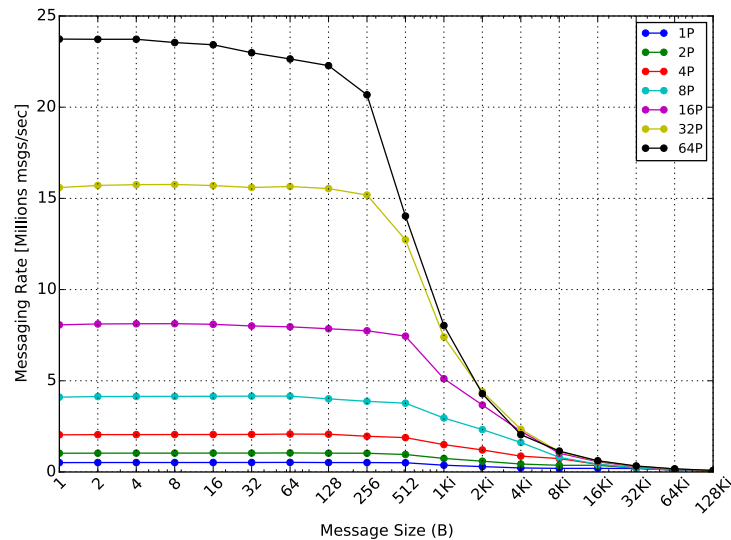- Restricted bandwidth between groups
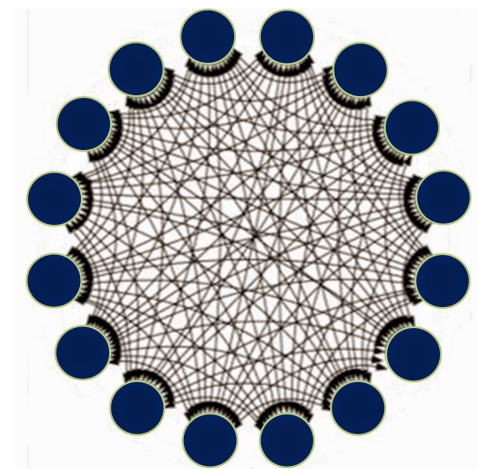
MPI BANDWIDTH AND MESSAGING RATE

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
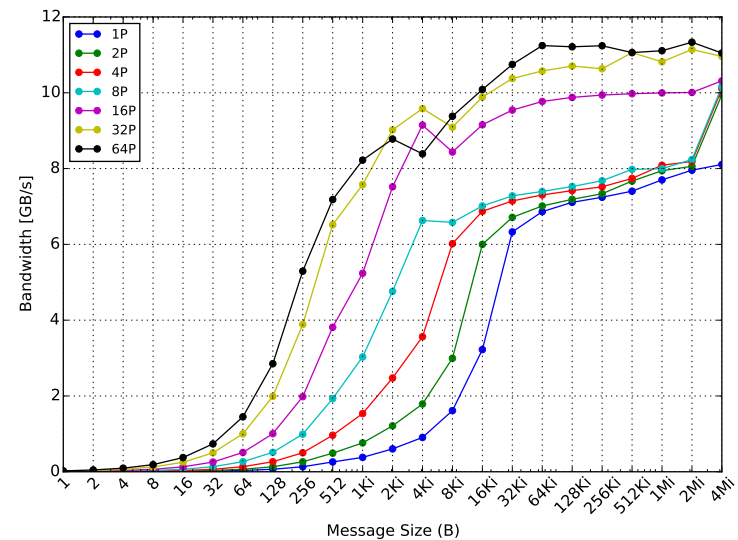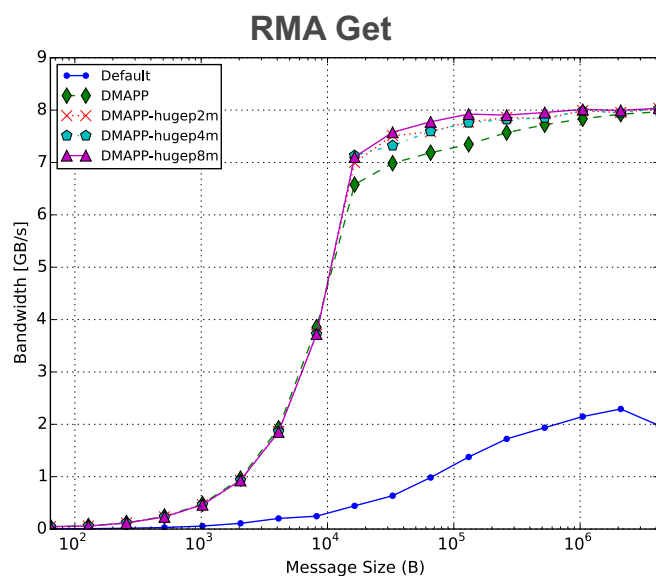
MPI LATENCY

OSU Ping Pong, Put, Get Latency

Benchmark	Zero Bytes (µs)	One Byte (µs)
Ping Pong	3.07	3.22
Put	0.61	2.90
Get	0.61	4.70

MPI ONE SIDED (RMA)

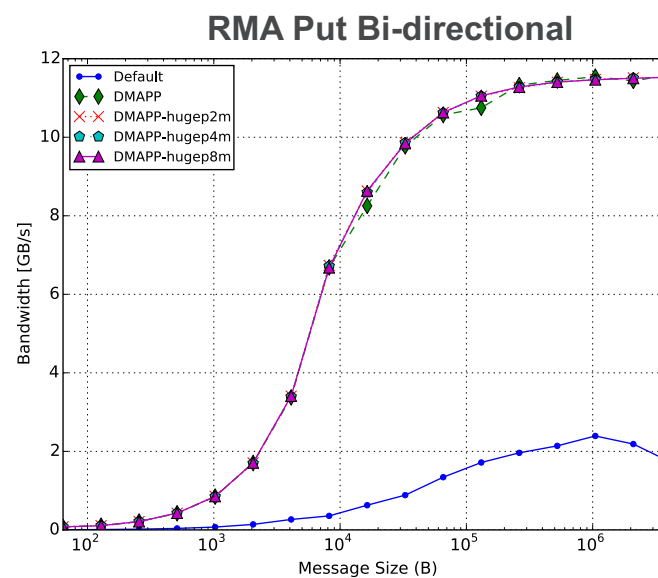## OSU One Sided MPI Get Bandwidth and Bi-Directional Put Bandwidth

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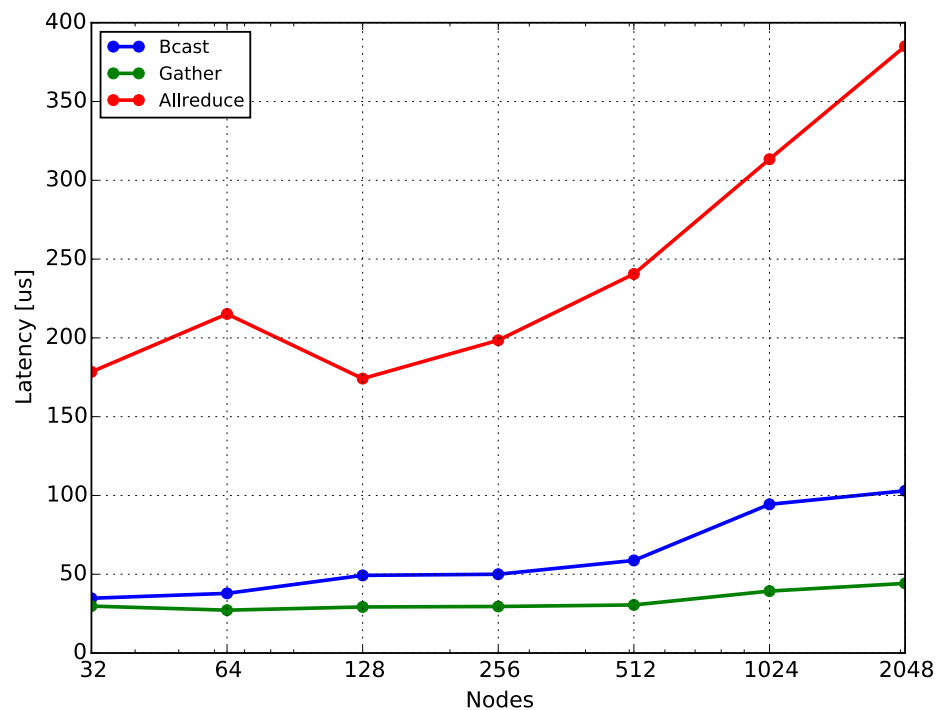- Huge pages also help.

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

RMA Get

RMA Put Bi-directional

Argonne
NATIONAL LABORATORY

MPI COLLECTIVE PERFORMANCE
OSU MPI Gather, Bcast, and Allreduce Benchmarks

- Node counts from 32 to 2048
- 1 process per node
- 8 KB message sizes

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

Argonne
NATIONAL LABORATORY

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)

Argonne
NATIONAL LABORATORY

QUESTIONS?

www.anl.gov

Argonne

NATIONAL LABORATORY