ARCHITECTURE OF THE ARGONNE CRAY XC40 KNL SYSTEM ‘THETA’

SCOTT PARKER
Lead, Performance Engineering Team
Argonne Leadership Computing Facility

July 30, 2018

XEON PHI IN THE TOP500

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

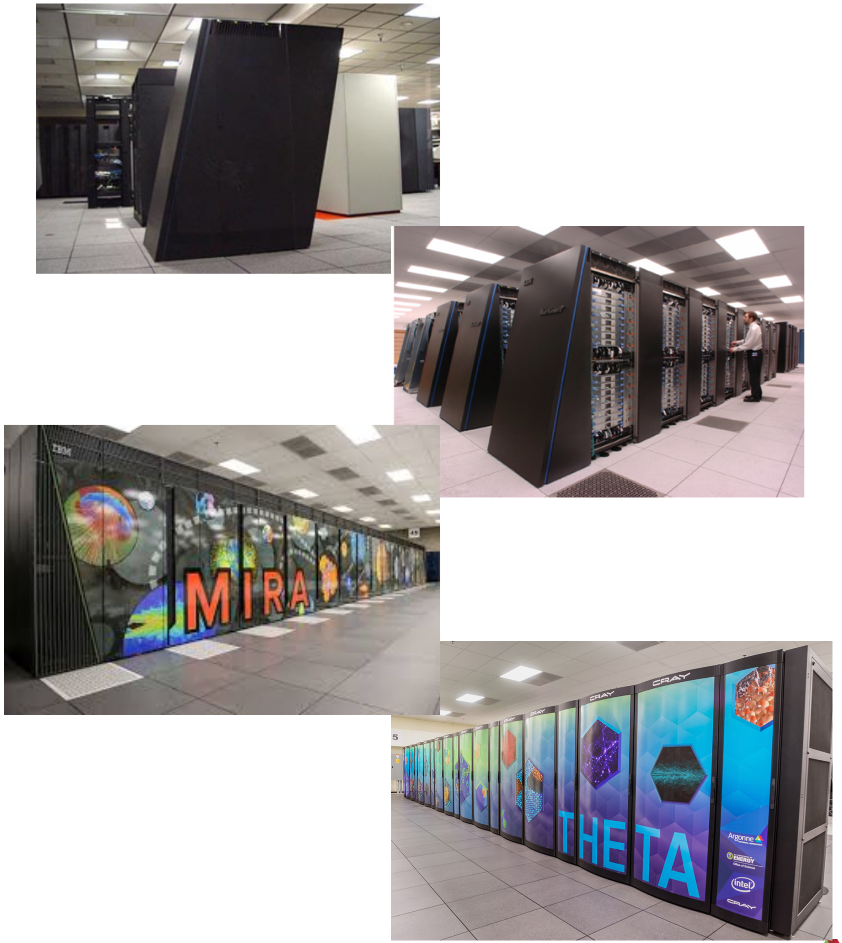
Rank	Facility	Architecture	Linpack (PF)	Peak (PF)
9	Los Alamos/Sandia	Trinity - Cray XC40, Intel Xeon Phi 7250	14	44
10	Berkeley - NERSC	Cori - Cray XC40, Intel Xeon Phi 7250	14	28
11	Korea Institute of Science and Technology Inf.	Nurion - Cray CS500, Intel Xeon Phi 7250	14	26
12	Joint Center for Advanced 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

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

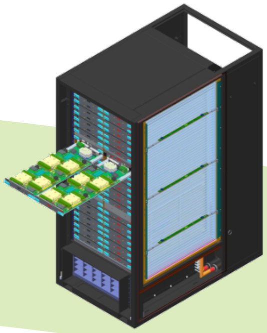
ARGONNE HPC TIMELINE

- 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 (12 PF) Cray XC40 with Xeon Phi (KNL)
- 2021:
  - One Exaflop Aurora Cray/Intel Xeon Phi to be delivered

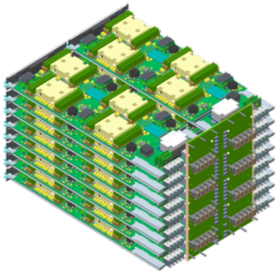


THETA SYSTEM OVERVIEW

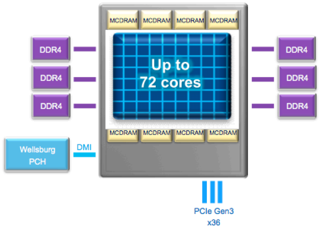
System: 24Cabinets
4,392 Nodes, 1152 Switches
12 groups, Dragonfly
11.7 PF Peak
68.6 TB MCDRAM, 823.5 TB DRAM

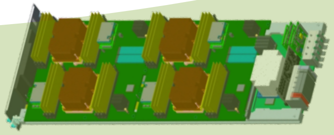


Cabinet: 3 Chassis
510.72 TF
3TB MCDRAM, 36TB DRAM



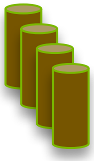
Chassis: 16 Blades
64 Nodes, 16 Switches
170.24 TF
1TB MCDRAM, 12TB DRAM





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 Features

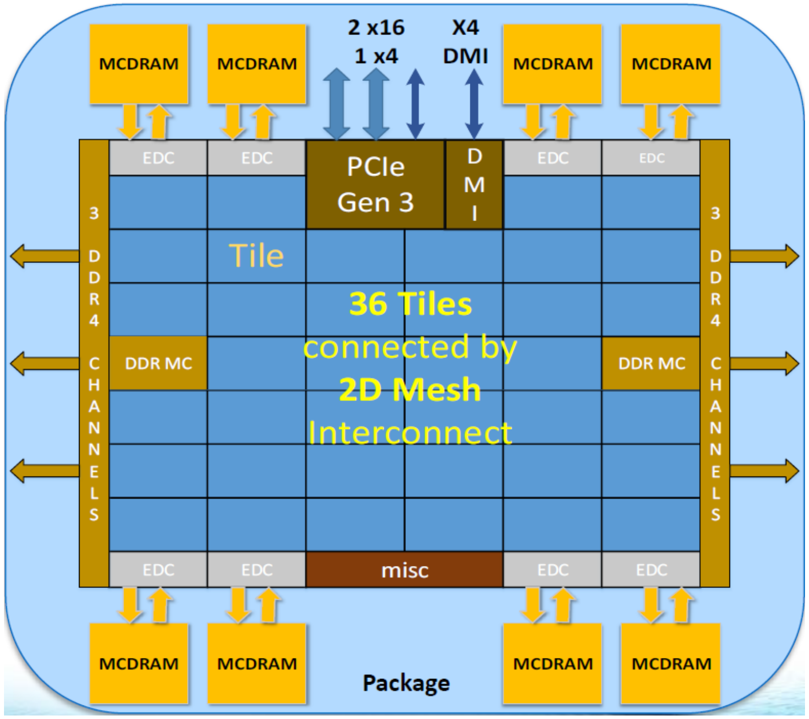
Improvement	Impact
Self Booting	No PCIe bottleneck
Binary Compatible with Xeon	Runs legacy code, no recompile
New Core Architecture (Atom based)	~3x higher performance than KNC
Improved Vector Density	3+ TFlops (DP) Peak per chip
New AVX-512 ISA	New 512 bit vector ISA with Masks
Gather/Scatter Engine	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

PERFORMANCE FROM PARALLELISM

Xeon Phi systems achieve performance from parallelism:

- Parallelism across nodes (using MPI, etc.)
- Parallelism across sockets within a node [Not applicable to the KNL]
- 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)

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

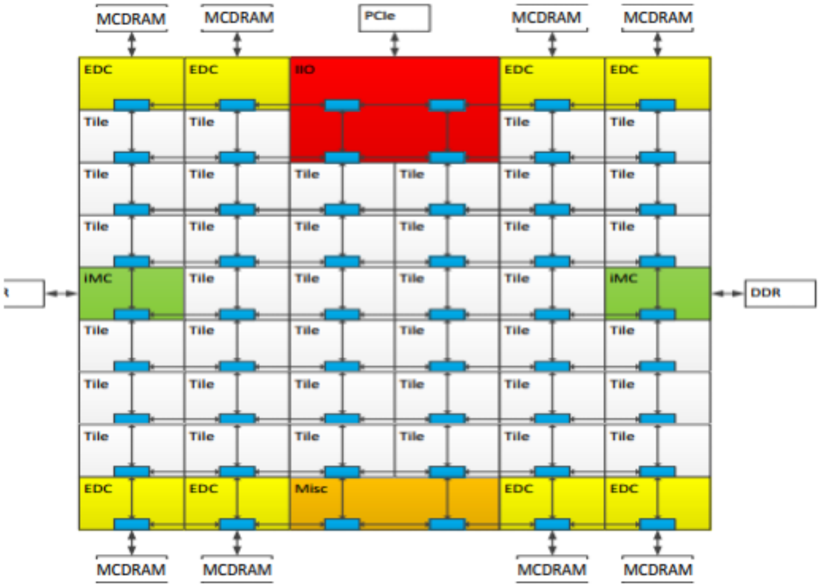
6 DDR4 memory channels
- 2 controllers
- up to 384 GB external DDR4
- 90 GB/s bandwidth

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

KNIGHTS LANDING VARIANTS

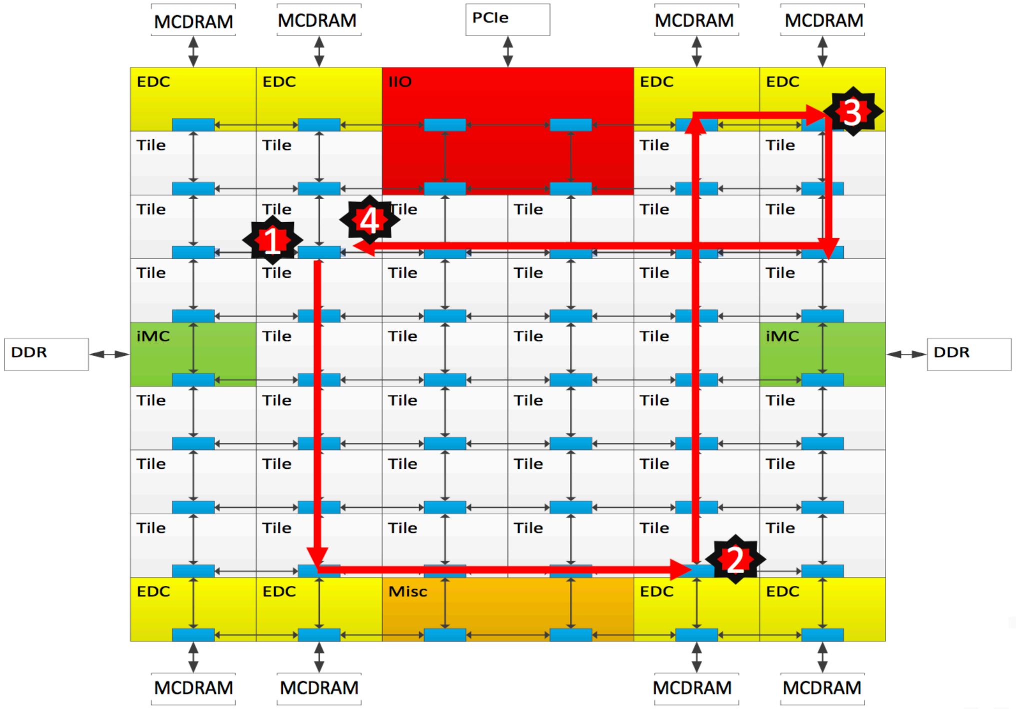
SKU	Cores	TDP Freq (GHz)	AVX Freq (GHz)	Peak Flops (TFlops)	MCDRAM (GB)	DDR Speed	TDP (Watts)
7210	64	1.3	1.1	2.66	16	2133	215
7230	64	1.3	1.1	2.66	16	2400	215
7250	68	1.4	1.2	3.05	16	2400	215
7290	72	1.5	1.3	3.46	16	2400	245

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

Cluster Modes: All-to-All



Address uniformly hashed across all distributed directories

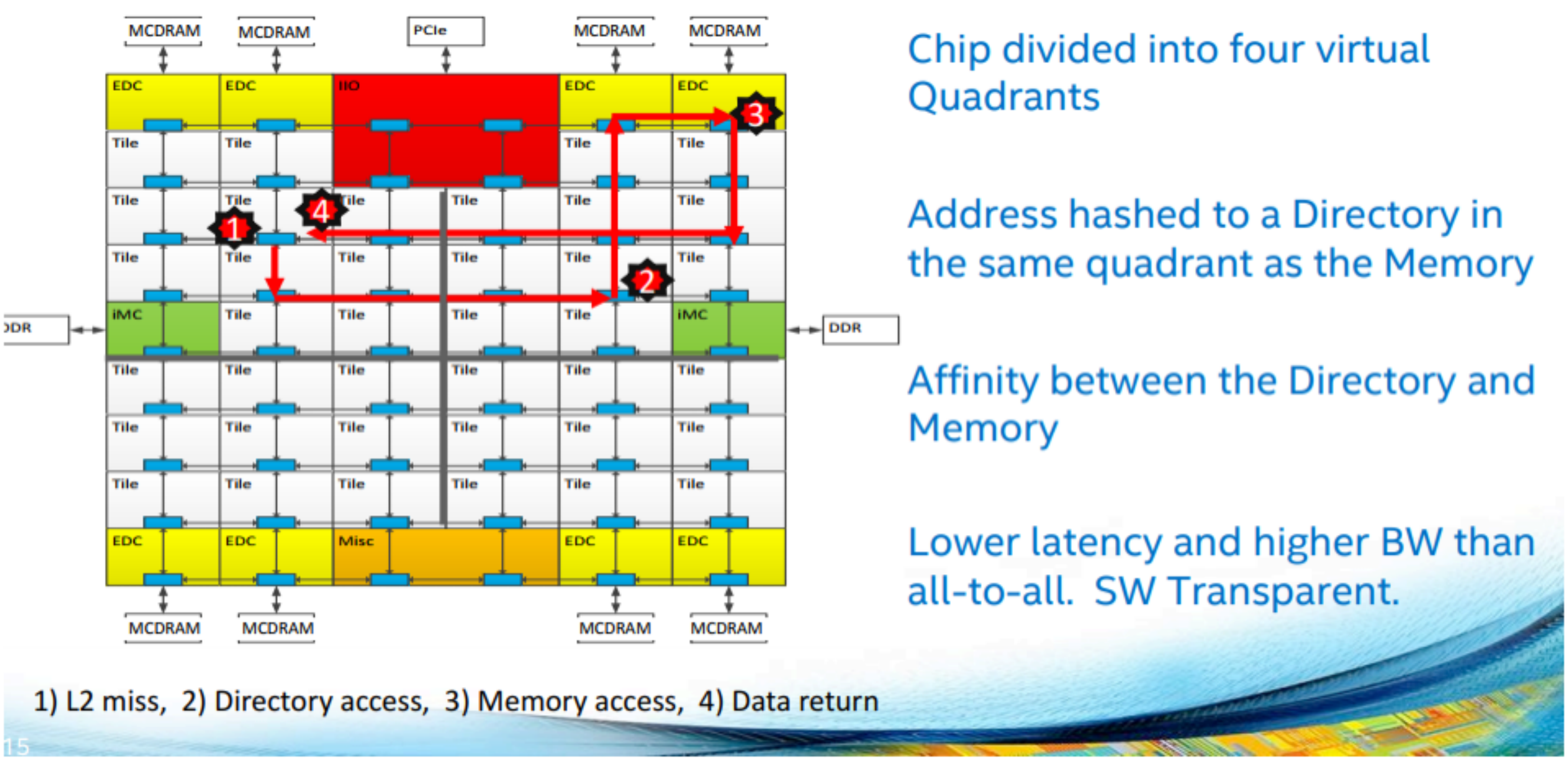
No affinity between Tile, Directory and Memory

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

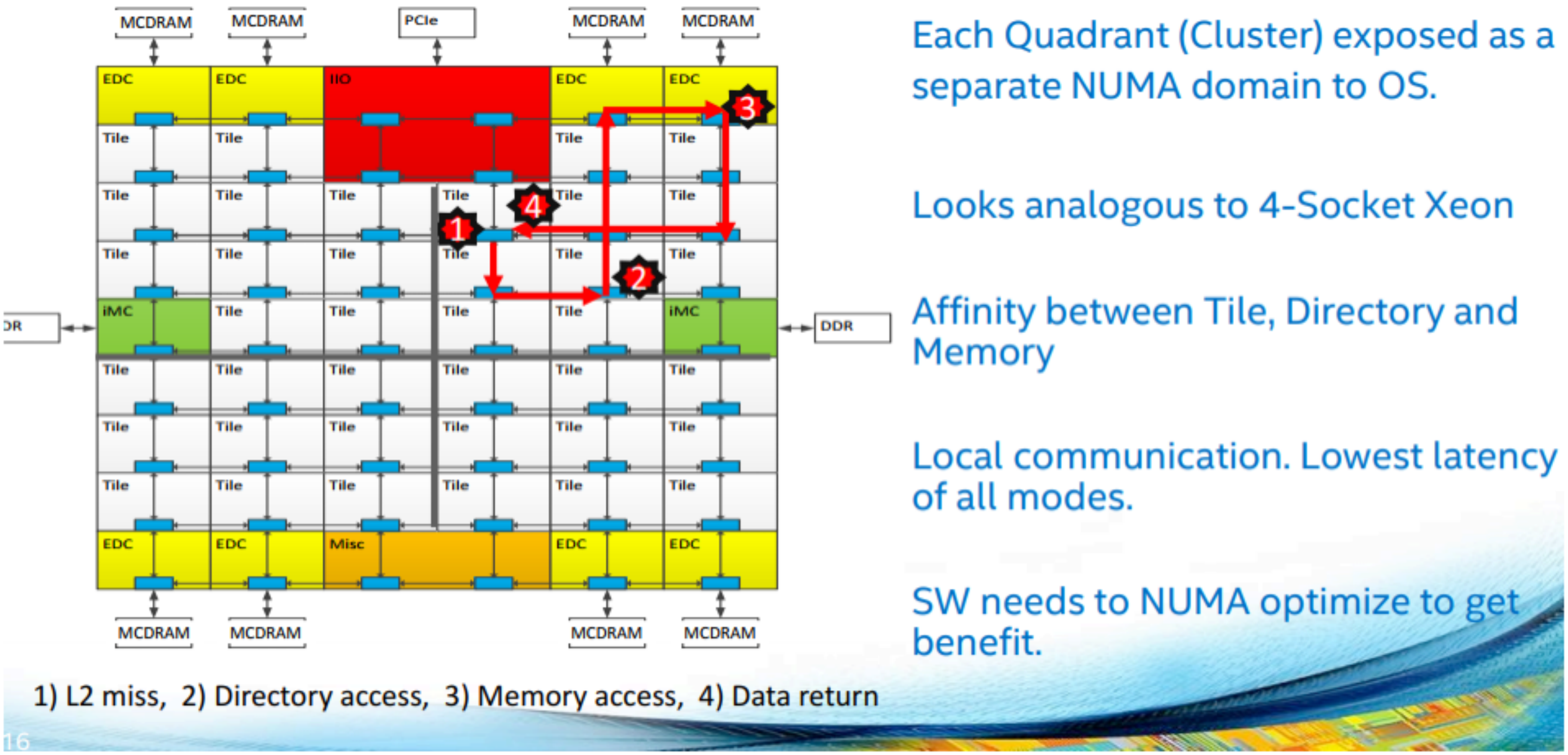
Address hashed to a Directory in the same quadrant as the Memory

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



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

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

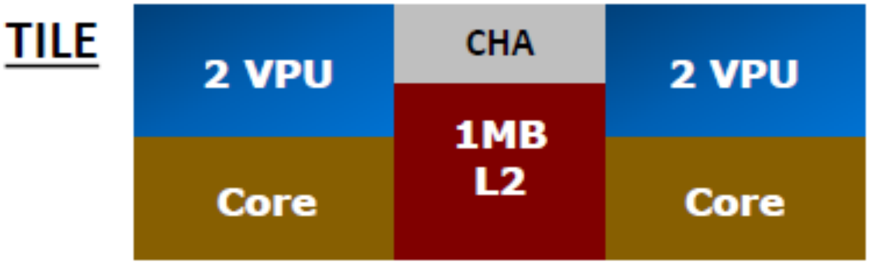
Looks analogous to 4-Socket Xeon

Affinity between Tile, Directory and Memory

Local communication. Lowest latency of all modes.

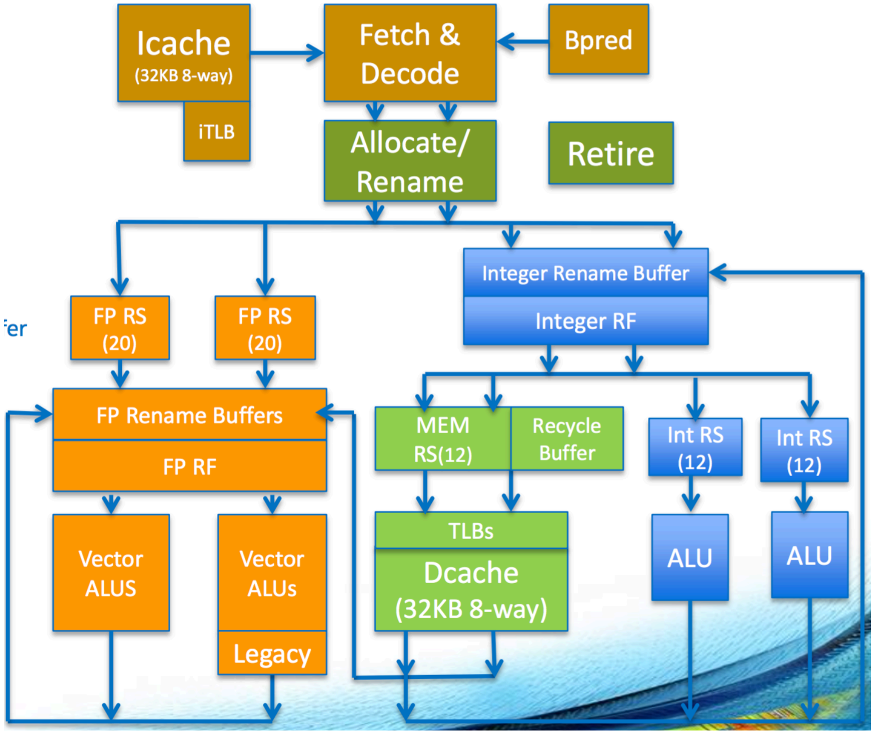
SW needs to NUMA optimize to get benefit.

KNL TILE



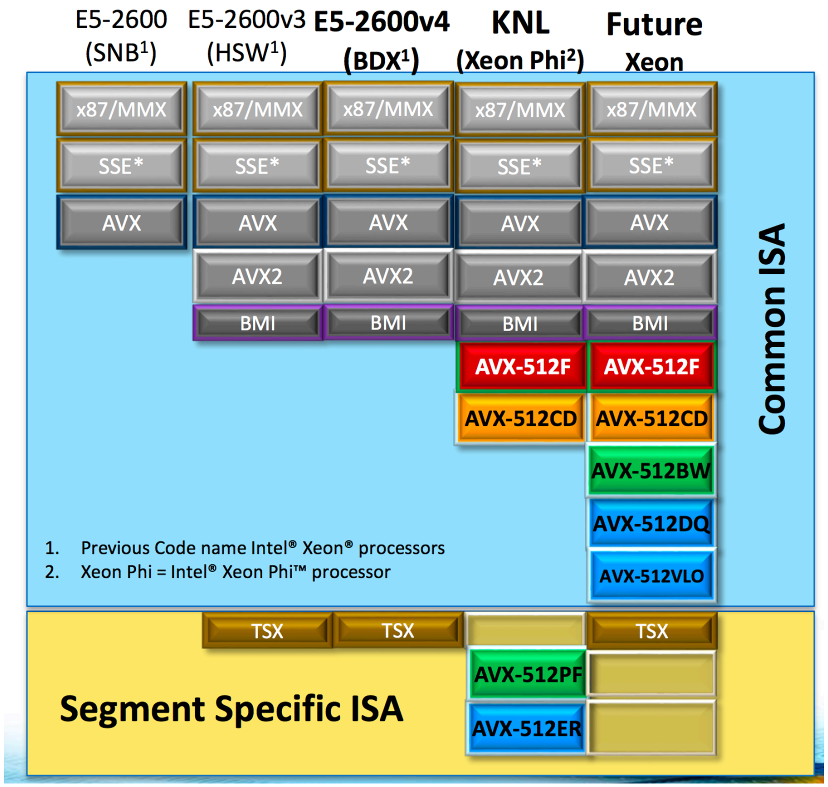
- Two CPUs
- 2 vector units (VPUs) per core
- 1 MB Shared L2 cache
    - Coherent across all tiles (32-36 MB total)
    - 16 Way
    - 1 line read and ½ line write per cycle
- Caching/Home agent
    - Distributed tag directory, keeps L2s coherent
    - Implements MESIF cache coherence protocol
    -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

KNL CORE



- Based on Silvermont (Atom)
- Instruction Issue & Execute:
  - 2 wide decode/rename/retire
  - 6 wide execute
- 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)
- 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 partitioned between threads
  - Caches and TLBs shared

Knights Landing Instruction Set

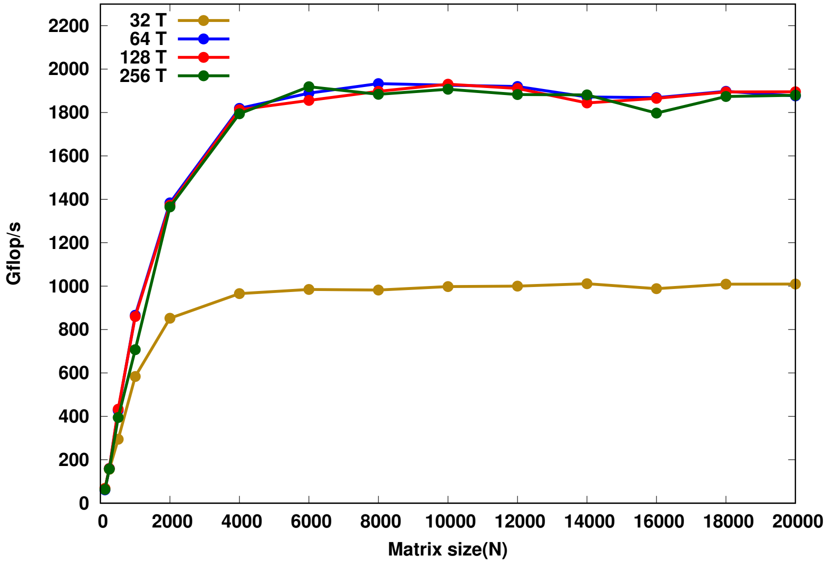


	E5-2600 (SNB[1])	E5-2600v3 (HSW[1])	E5-2600v4 (BDX[1])	KNL (Xeon Phi[2])	Future Xeon
Common ISA	x87/MMX	x87/MMX	x87/MMX	x87/MMX	x87/MMX
	SSE*	SSE*	SSE*	SSE*	SSE*
	AVX	AVX	AVX	AVX	AVX
		AVX2	AVX2	AVX2	AVX2
		BMI	BMI	BMI	BMI
				AVX-512F	AVX-512F
				AVX-512CD	AVX-512CD
					AVX-512BW
					AVX-512DQ
	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

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

- 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

DGEMM PERFORMANCE ON THETA

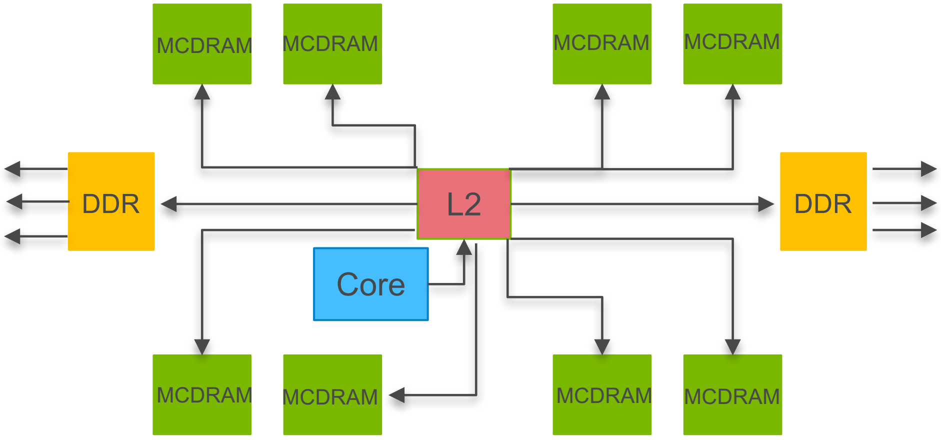


MKL DGEMM Performance

- 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
- 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

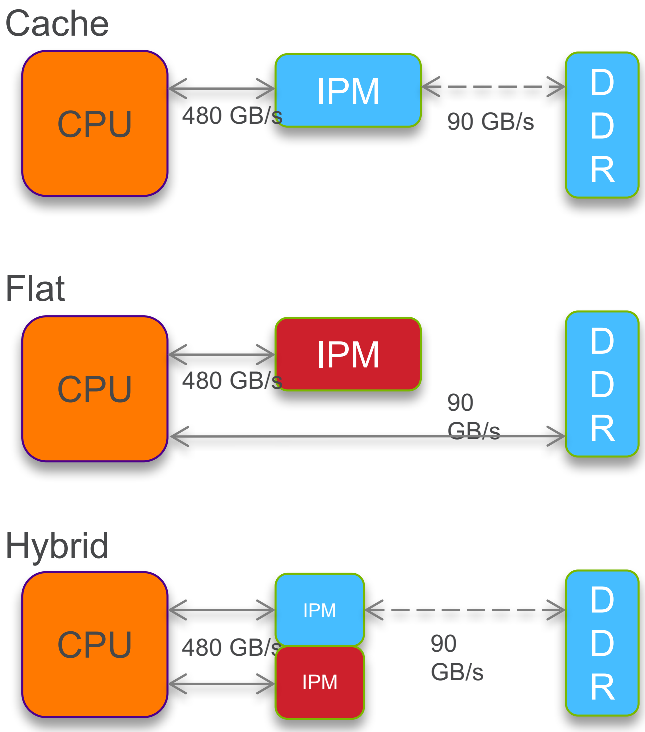
MEMORY

- **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



- **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

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**

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

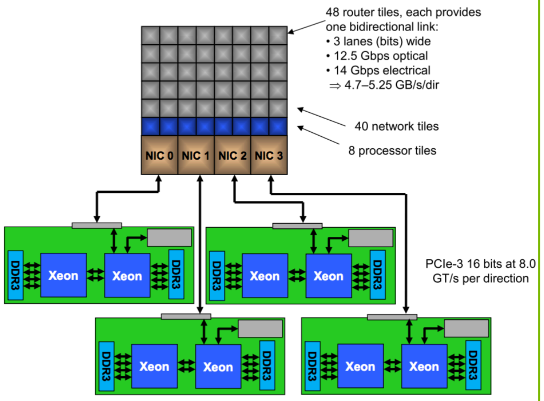
Threads	Barrier (µs)	Reduction (µs)	Parallel For (µs)
1	0.1	0.7	0.6
2	0.4	1.3	1.3
4	0.8	1.9	1.9
8	1.5	2.7	2.5
16	1.8	5.9	2.9
32	2.8	7.7	4.0
64	3.9	10.4	5.6
128	5.3	13.7	7.3
256	7.8	19.4	10.5

- 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 ~C*threads$^{1/2}$

ARIES DRAGONFLY NETWORK

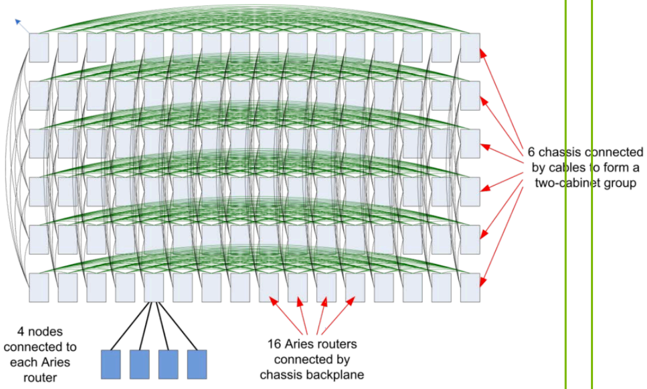
Aries Router:

- 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



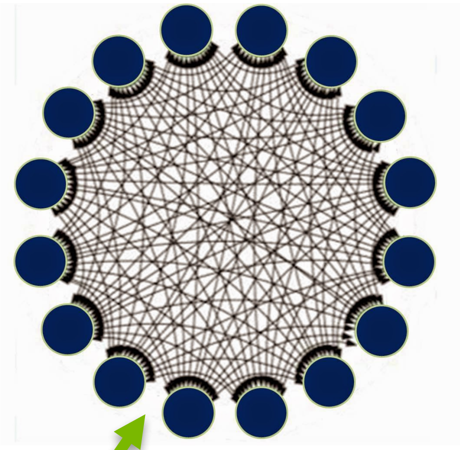
Connections within a group:

- 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:

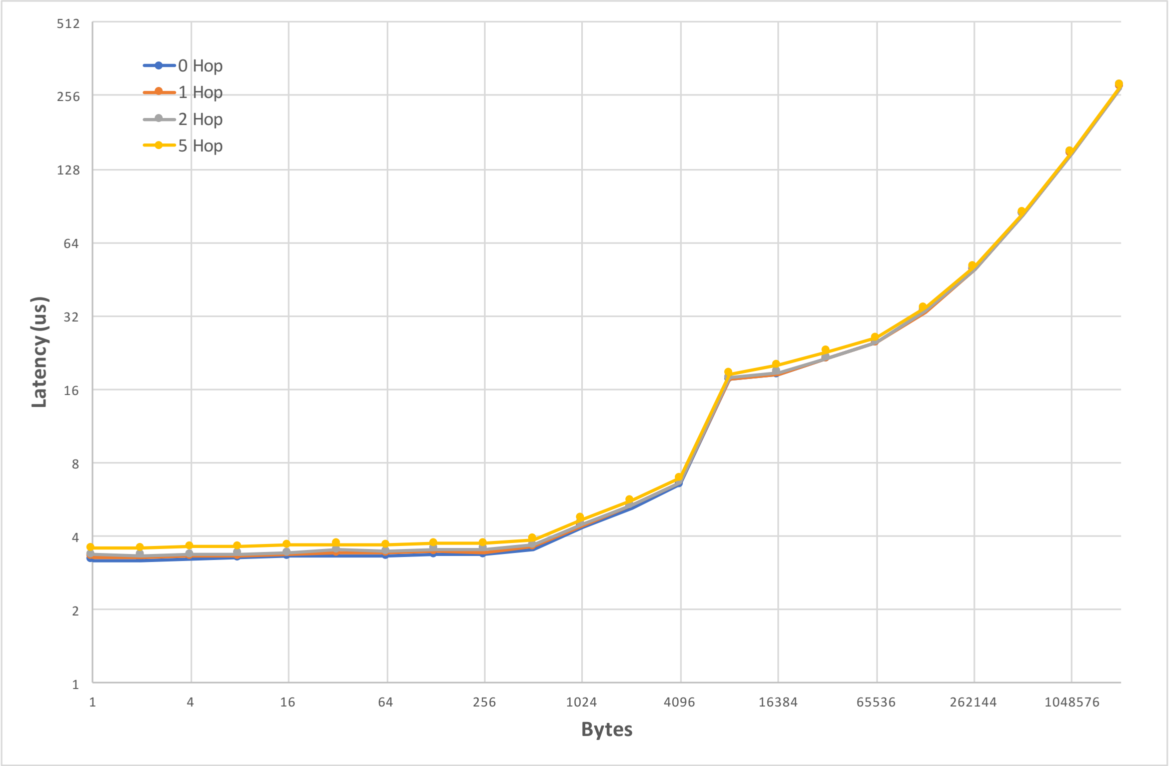
- Each group connected to every other group
- Restricted bandwidth between groups



Theta has 12 groups with 12 links between each group

MPI SEND AND RECEIVE LATENCY

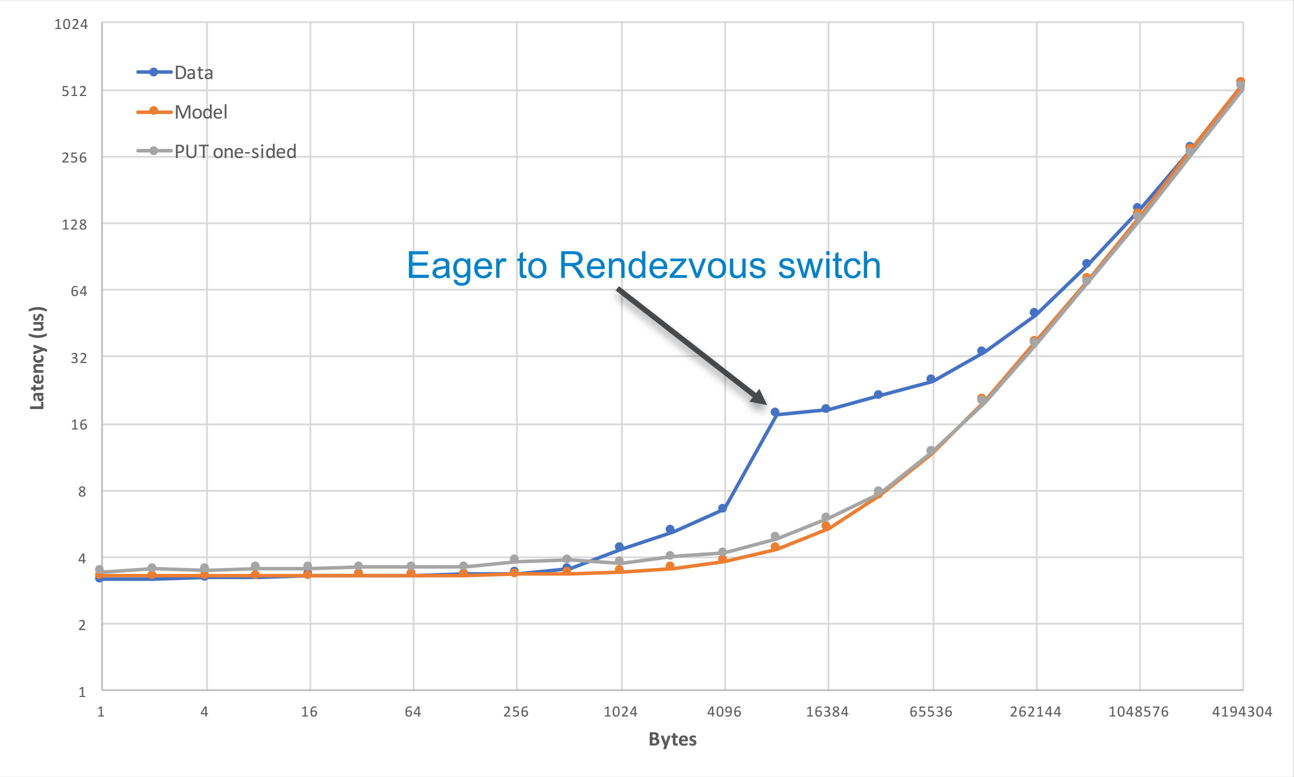
OSU PtoP MPI Latency on Theta



- 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:

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

$$n = bytes$$

$$\propto = 3.3$$

$$\beta = 0.0013$$

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

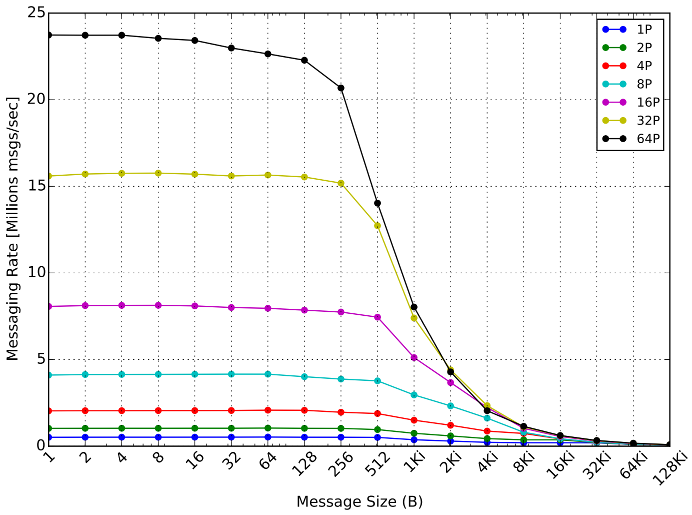
OSU PtoP MPI Multiple Bandwidth / Message Rate Test on Theta

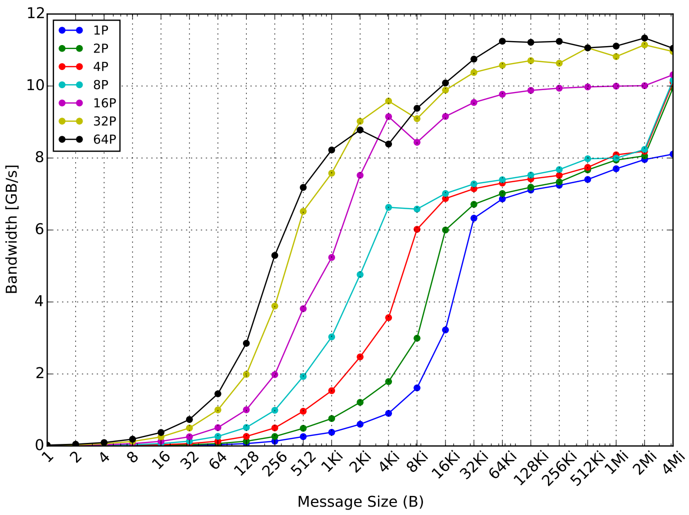
Messaging Rate:

- 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



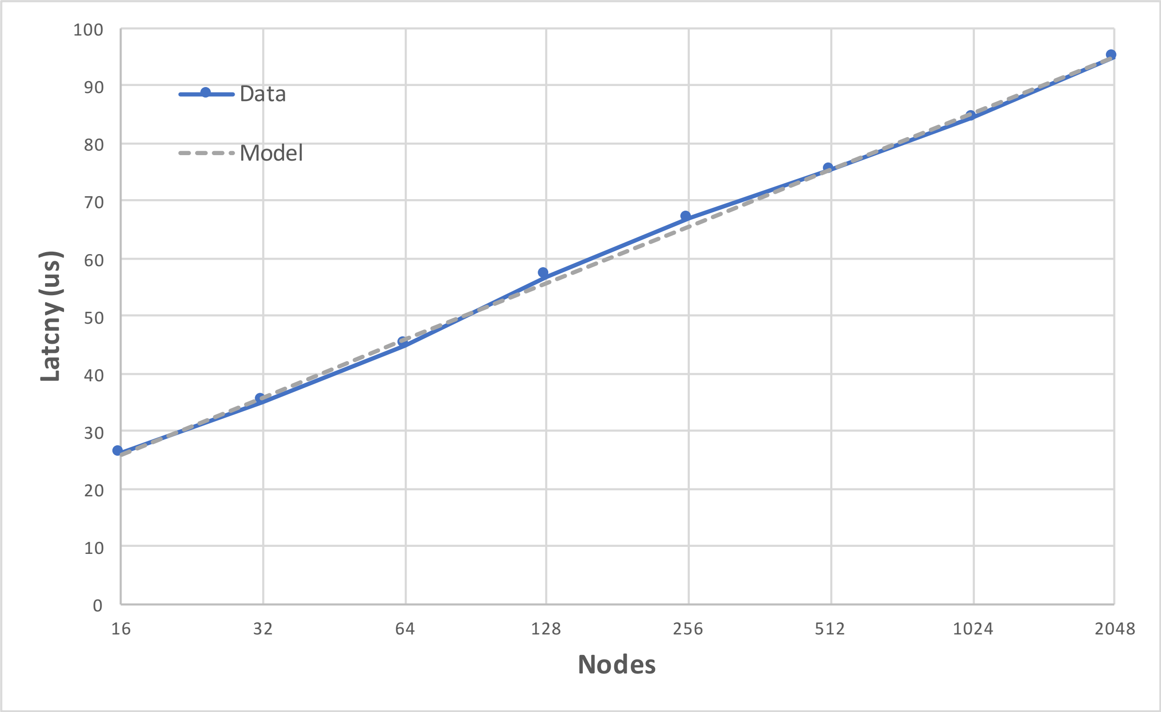


MOST FREQUENTLY CALLED COLLECTIVE ROUTINES

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

MPI BARRIER MODEL



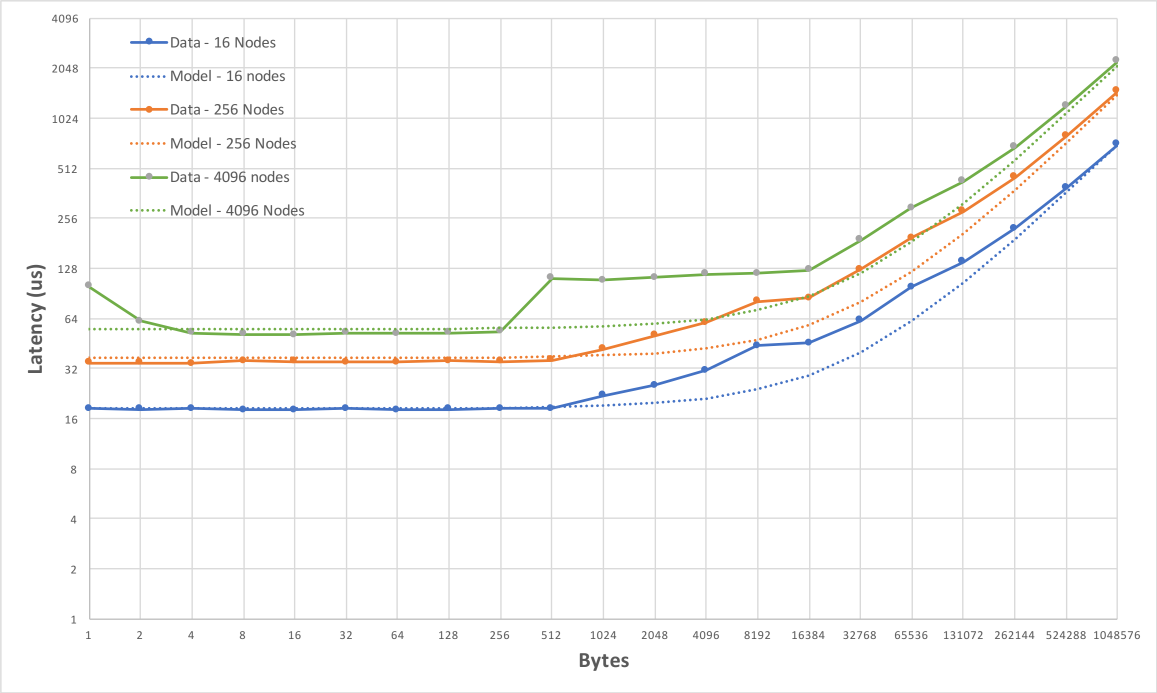
$$T = \alpha + \beta \cdot log_2(p)$$

$$p = nodes$$

$$\propto = -13.5$$

$$\beta = 9.87$$

MPI BROADCAST MODEL



$$T = (\alpha + \beta \cdot n) Log_2(p)$$

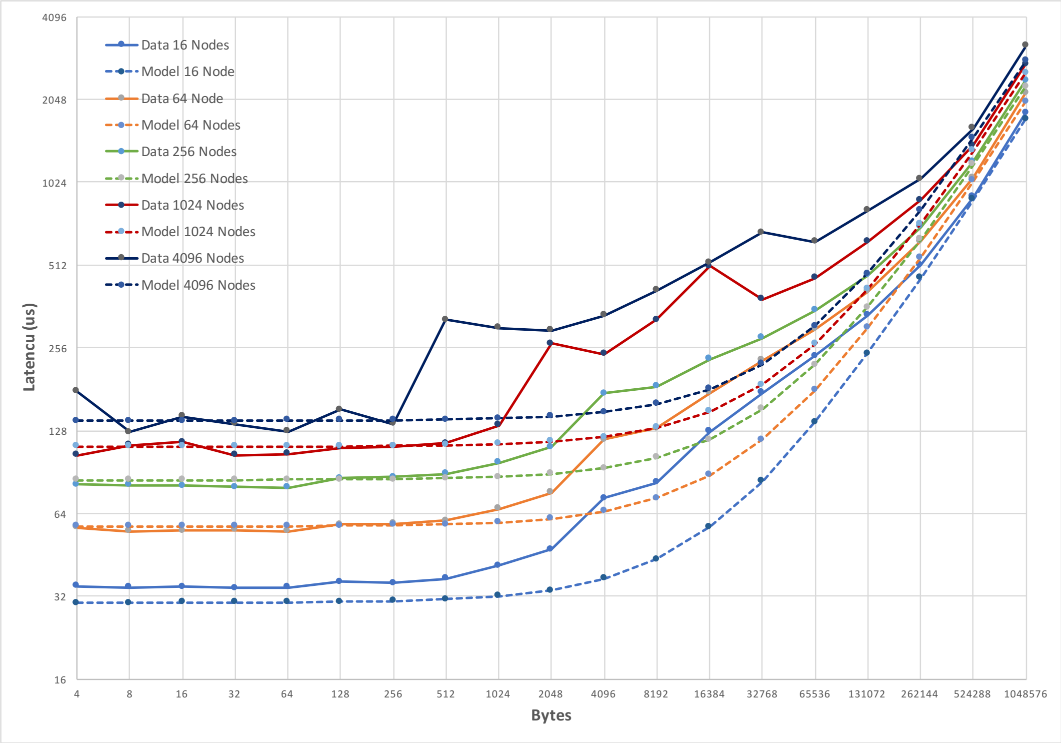
$$n = bytes$$
$$p = nodes$$
$$\propto = 4.6$$
$$\beta = 0.0016$$

Good fit at low and high byte ranges.

Errors centered around point of protocol switch

MPI ALLREDUCE MODEL

OSU MPI Allreduce Benchmark



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

$$n = bytes$$

$$p = nodes$$

$$\gamma = -24$$

$$\delta = 0.0012$$

$$\propto = 13.6$$

$$\beta = 0.00012$$

Good fit at low and high byte ranges.

Errors centered around point of protocol switch

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

QUESTIONS?

www.anl.gov

Argonne NATIONAL LABORATORY