Argonne NATIONAL LABORATORY

EARLY EVALUATION OF THE CRAY XC40 SYSTEM “THETA”

SUDHEER CHUNDURI, SCOTT PARKER, KEVIN HARMS, VITALI MOROZOV, CHRIS KNIGHT, KALYAN KUMARAN

Performance Engineering Group
Argonne Leadership Computing Facility

May 9, 2017

OUTLINE

- Theta System Description
- Variability
- Benchmark Performance
  - Computational
  - Memory
  - OpenMP
  - MPI
  - Power
- Application Performance
  - Nekbone
  - LAMMPS
  - MILC

THETA

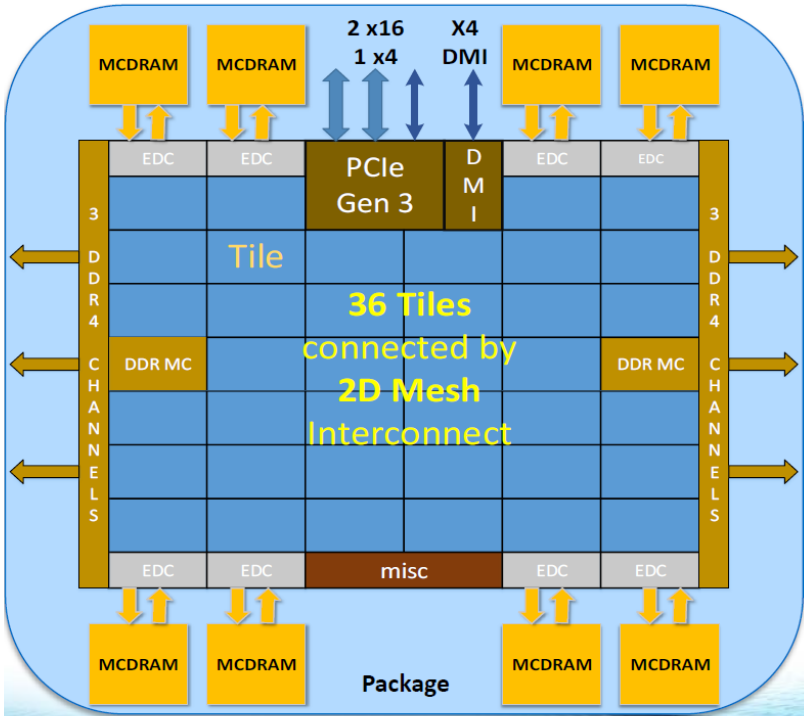
- System:
  - Cray XC40 system
  - 3,624 compute nodes/ 231,936 cores
  - 9.65 PetaFlops peak performance
  - Accepted Fall 2016
- Processor:
  - Intel Xeon Phi, 2nd Generation (Knights Landing) 7230
  - 64 Cores
  - 1.3 GHz base / 1.1 GHz AVX / 1.4-1.5 GHz Turbo
- Memory:
  - 16 GB MCDRAM per node
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

Argonne NATIONAL LABORATORY

VARIABILITY ON THETA

- Variability between runs is frequently 15% or greater, can be up to 100%
- Identified 4 causes of potential variability
  - Core level variability due to OS noise
    - Impact on applications: minimal
    - Available mitigations: Use core specialization, exclude tile 0 & 32
  - Tile level variability due to shared resource contention on tile (L2)
    - Impact on applications: yes
    - Available mitigations: run using only 1 core per tile
  - Memory mode variability due to cache mode page conflicts
    - Impact on applications: yes
    - Available mitigations: run in flat mode
    - Potential mitigations: improved zone sort
  - Network variability due to shared network resources
    - Impact on applications: yes
    - Available mitigations: run without other jobs present on system
    - Potential mitigations: compact job placement

KNIGHTS LANDING PROCESSOR



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

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

2D Mesh Interconnect
- Tiles connected by 2D mesh

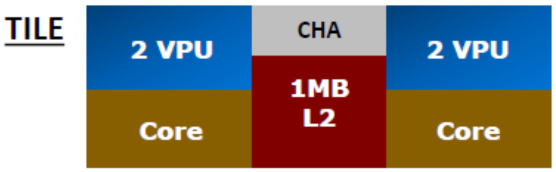
On Package Memory
- 16 GB MCDRAM
- 8 Stacks
- 485 GB/s bandwidth

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

Argonne NATIONAL LABORATORY

KNL TILE AND CORE

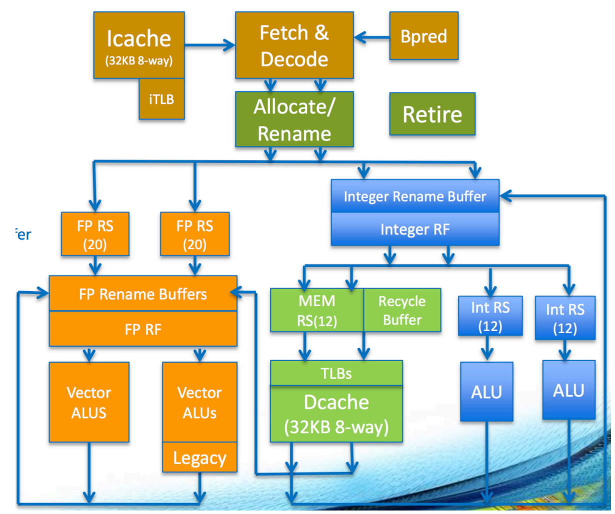


Tile

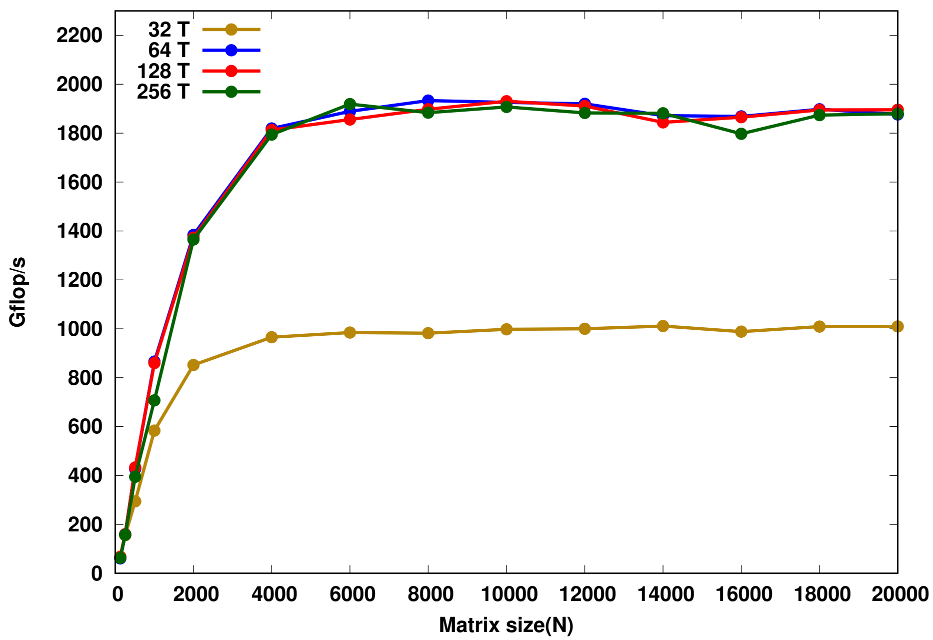
- Two CPUs
- 2 VPUs per core
- Shared 1 MB L2 cache (not global)
- Caching/Home agent
  - Distributed directory, provides coherence

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



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

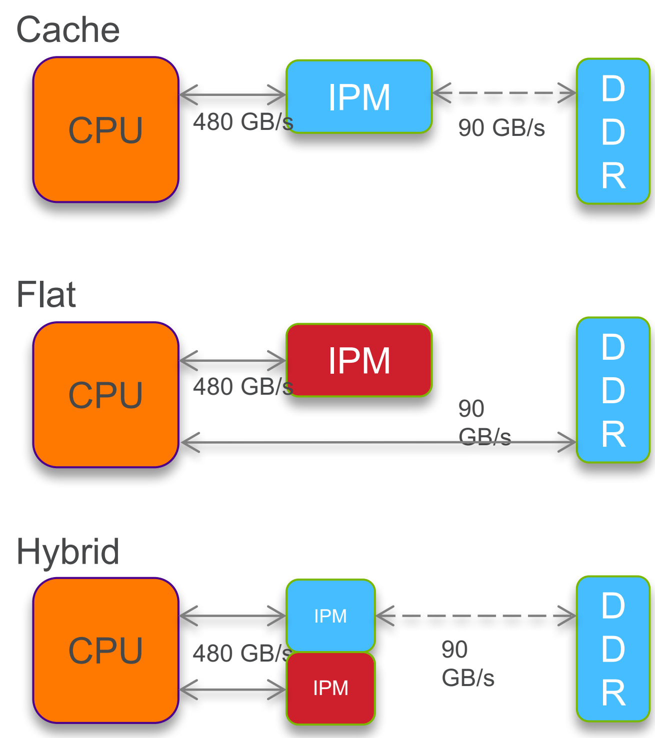
OBSERVATIONS ON FLOATING POINT PERFORMANCE

- Floating point performance is limited by AVX frequency
  - AVX vector frequency (1.1 GHz) 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
- Variability in performance
  - OS noise can produce variability in when timing small kernels even with core specialization
  - L2 cache contention can favor one core leading to differing performance for cores sharing a tile on the same workload
  - Have not observed significant variability caused processor turbo clock rates

KNL MEMORY HIERARCHY AND MODES

- **Two memory types:**
  - In Package Memory (IPM)
    - 16 GB MCDRAM, 8 stacks
  - Off Package Memory (DDR2400)
    - Up to 384 GB, 2 controllers, 6 channels
- **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
- **Cluster modes:**
  - Quadrant, SNC-4, AlltoAll, …
- **Managing memory:**
  - jemalloc & memkind libraries
  - numactl command
  - Pragmas for static memory allocations

MODES, SELECTED AT NODE BOOT TIME



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

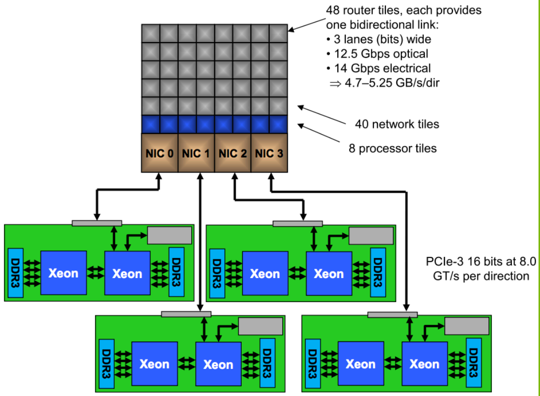
EPCC OpenMP Benchmarks

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

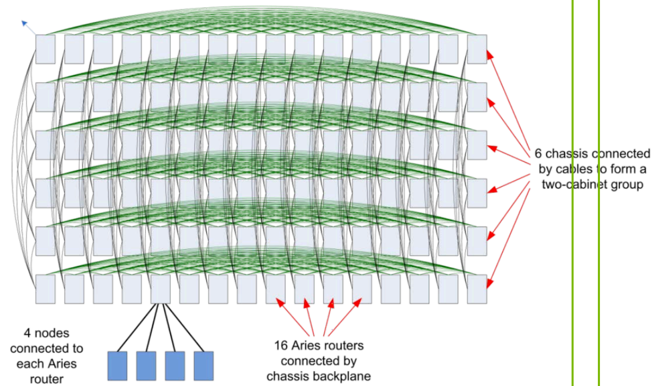
- OpenMP costs related to cost of memory access
  - KNL has no shared last level cache
- Operations can take between 130 – 25,000 cycles
- Cost of operations increases with thread count
  - Scales as ~C*threads$^{1/2}$

ARIES DRAGONFLY NETWORK

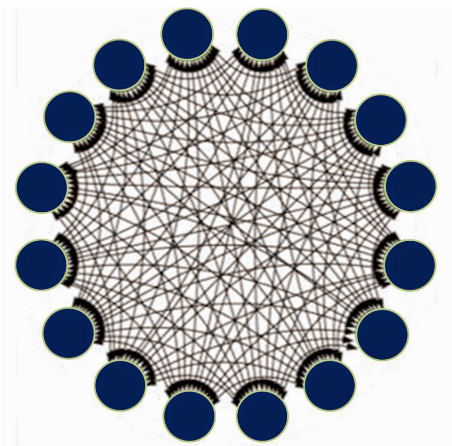
Aries Router:
- 4 Nodes connect to an Aries
- 4 NIC's connected via PCIe
- 40 Network tiles/links
- 4.7-5.25 GB/s/dir per link



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



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



MPI BANDWIDTH AND MESSAGING RATE

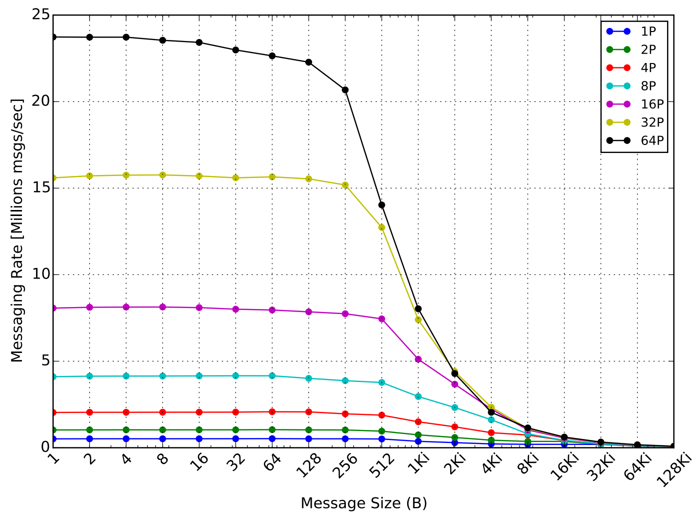
OSU PtoP MPI Multiple Bandwidth / Message Rate Test on Theta

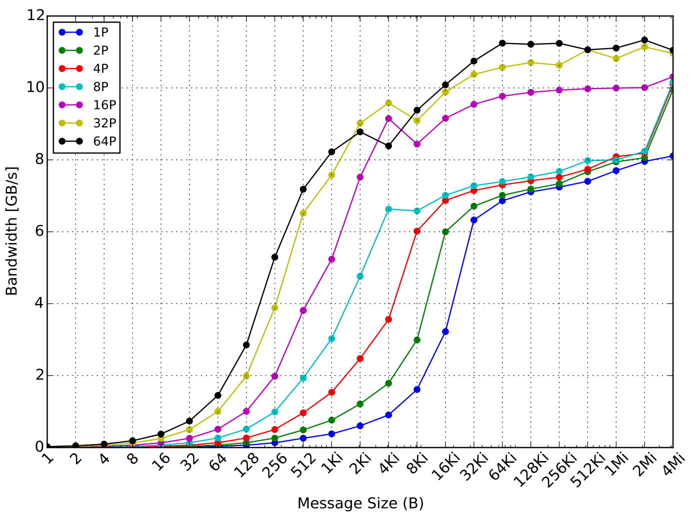
Messaging Rate:

- 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)

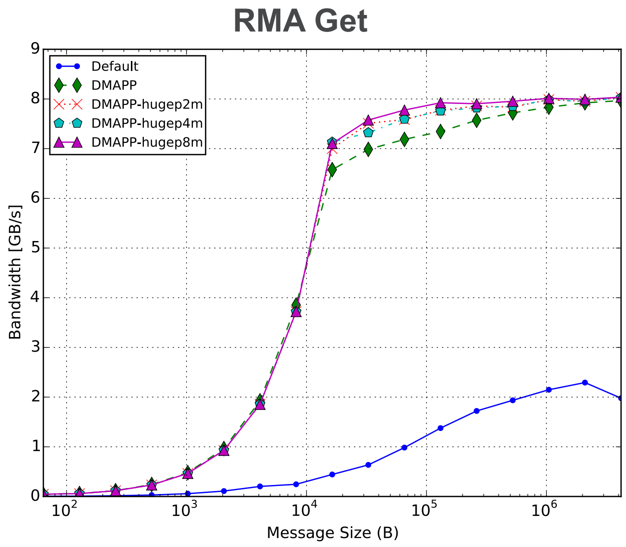
OSU One Sided MPI Get Bandwidth and Bi-Directional Put Bandwidth

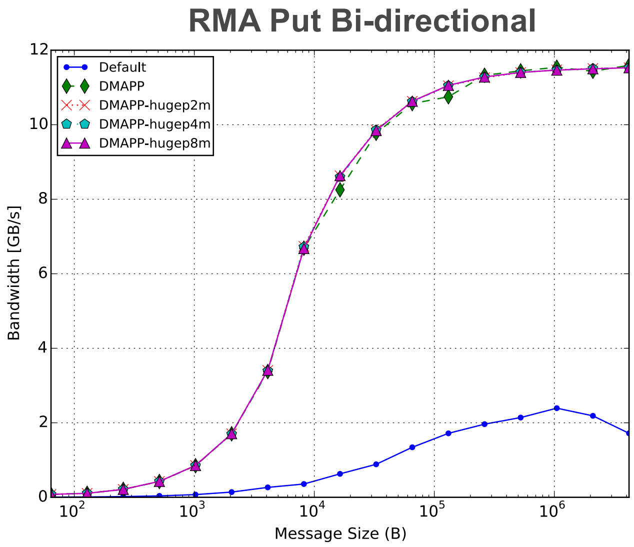
RMA Get

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

RMA Put

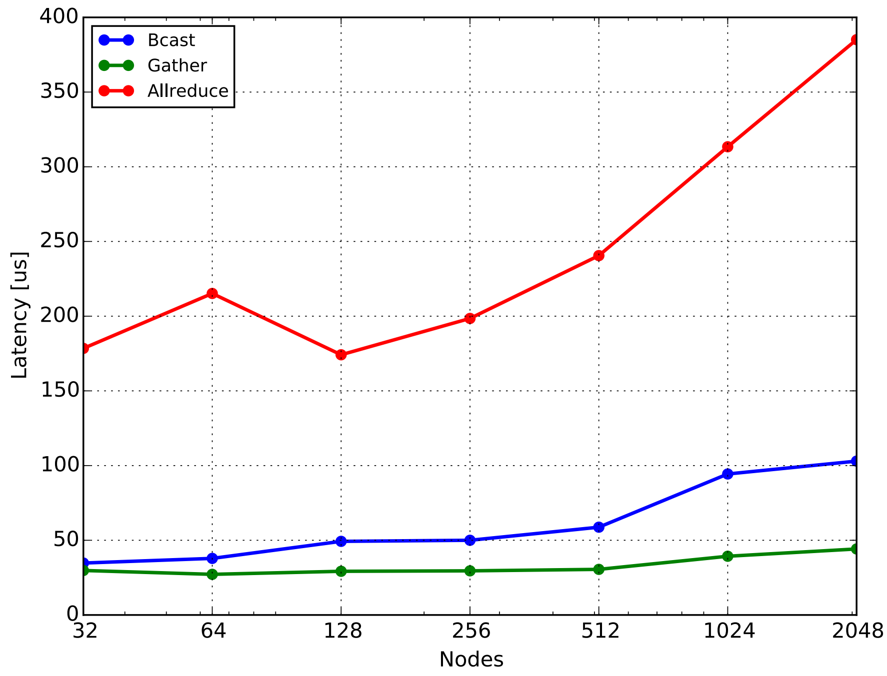
- 2 GB/s using default configuration (uGNI)
- 11.6 GB/s peak bi-directional bandwidth over DMAPP
- No significant benefit from huge pages





MPI COLLECTIVE PERFORMANCE

OSU MPI Gather, Bcast, and Allreduce Benchmarks



- Node counts from 32 to 2048
- 1 process per node
- 8 KB message sizes
- Plan to review MPI data for performance consistency

Argonne NATIONAL LABORATORY

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

NEKBONE PERFORMANCE ON KNL

- Nekbone mini-app derived from Nek5000 (Spectral Element CFD code)
  - Solves 3D Poisson problem in rectangular geometry
  - Spectral elements and conjugate gradient
  - Contains key kernels, operations, and work from Nek5000
  - Implemented using Fortran 77, C, MPI, and OpenMP
- KNL performance 3.2x Haswell (solve time per element)
  - KNL : 0.38 ms
  - Haswell E5-2699 v3 (dual socket, 36 cores): 1.22 ms
- KNL kernel mix for run on 1024 nodes scaled to 80% parallel efficiency:
  - Streaming kernels – 48% of time (BW limited)
    - Streaming kernels are achieving 70-98% of Stream bandwidth from MCDRAM
  - Matrix multiply – 21% of time (Compute limited)
    - Simple triple loop : ~2.5% of peak
    - Unrolled loops : ~20% of peak
    - LIBXSMM : ~40% of peak
  - Communication – 31% of time (Communication limited)

NEKBONE - THREADS AND RANKS

Identical problem with Nekbone using different number of hyper-threads, threads and ranks

1 Hyper-thread

Ranks	Thds	Solve Time
1	64	3.07
2	32	**3.00**
4	16	3.02
8	8	3.02
16	4	3.05
32	2	3.04
64	1	3.14

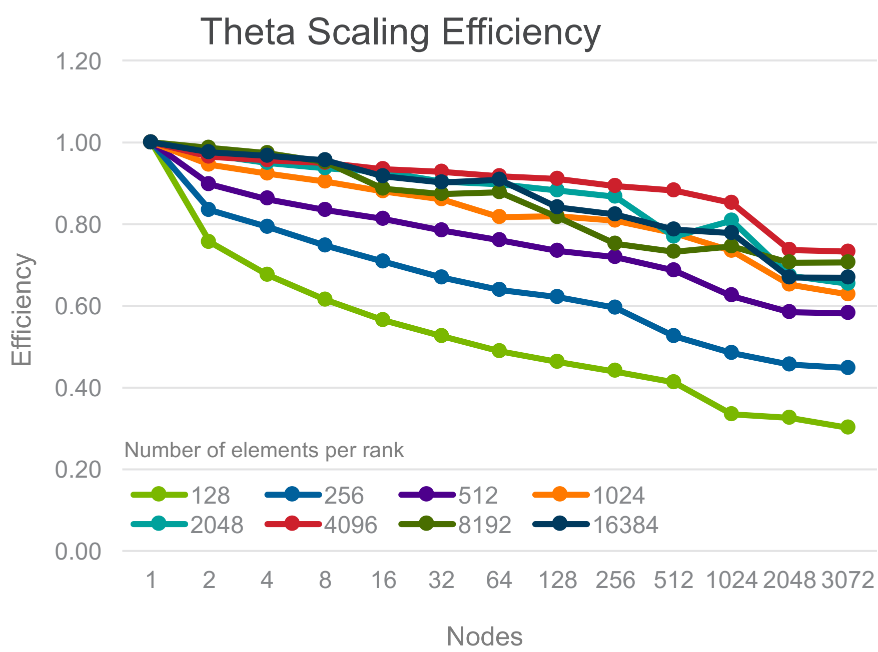
2 Hyper-threads

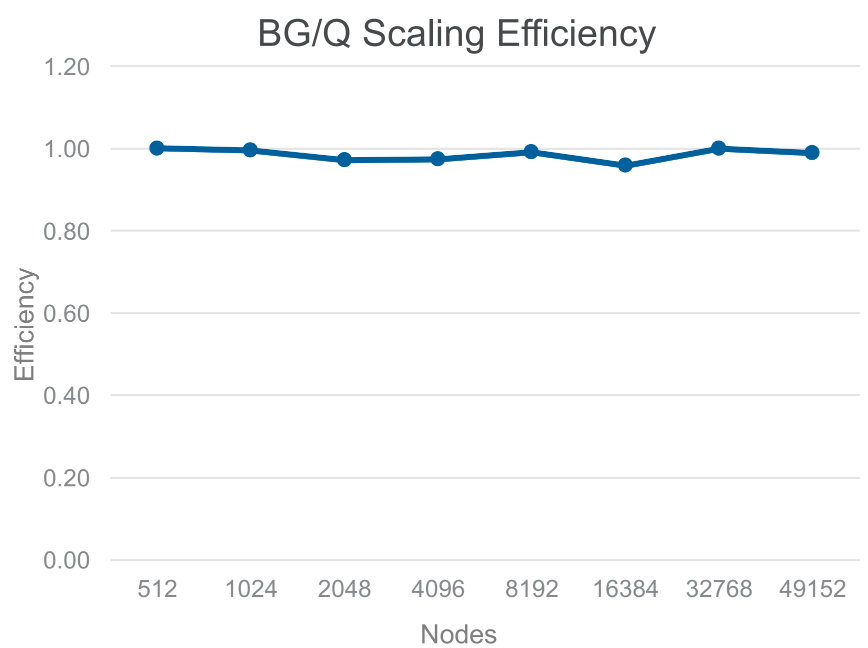
Ranks	Thds	Solve Time
1	128	3.65
2	64	3.66
4	32	3.61
8	16	3.63
16	8	3.67
32	4	3.66
64	2	3.75
128	1	4.10

4 Hyper-threads

Ranks	Thds	Solve Time
1	256	4.72
2	128	4.70
4	64	4.57
8	32	4.59
16	16	4.58
32	8	4.61
64	4	4.69
128	2	5.08
256	1	14.09

NEKBONE WEAK SCALING

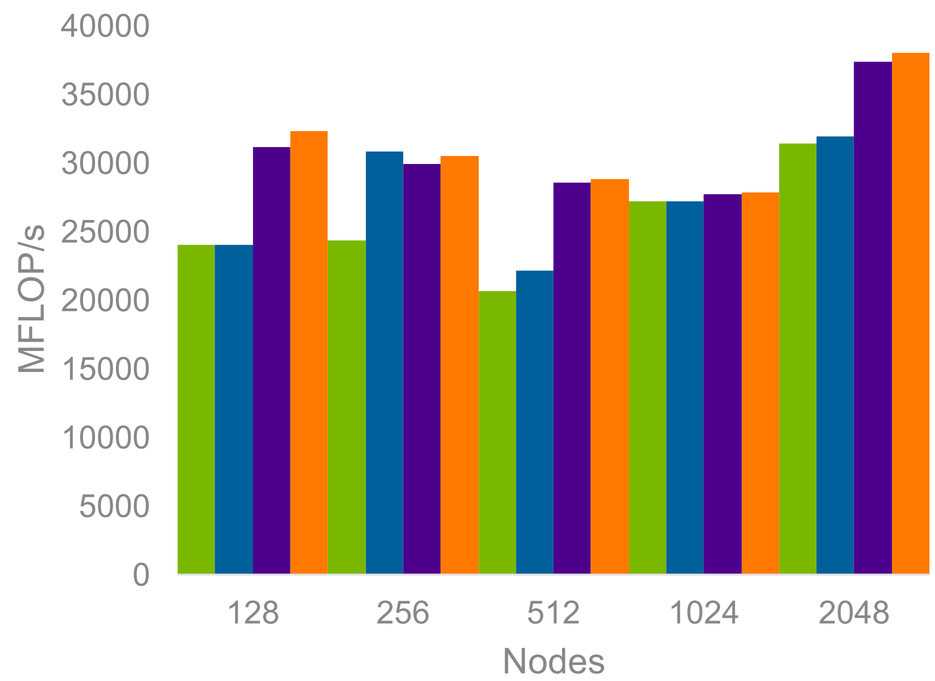




Argonne NATIONAL LABORATORY

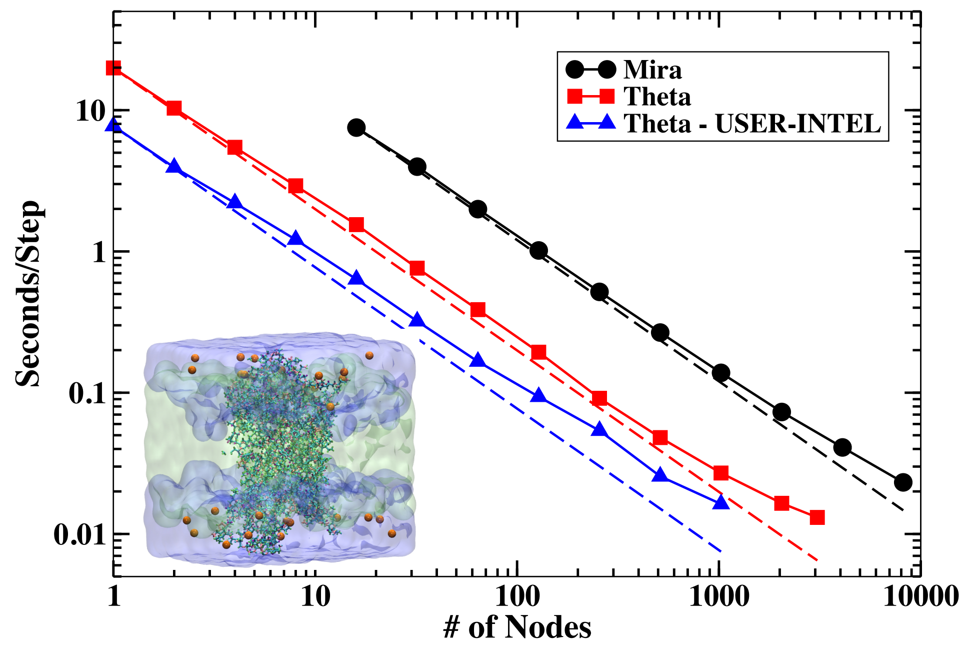
MILC WEAK SCALING

- su3_rhmd_hisq application run in cache mode with other jobs running
- grid_order and core specialization are used but not huge pages
- ~84% difference between lowest and highest performance between different days
- Application is subject to variance from MCDRAM cache mode and MPI traffic from other jobs running



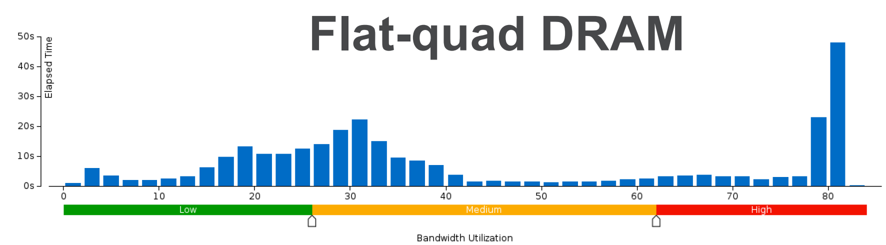
LAMMPS – STRONG SCALING COMPARISON

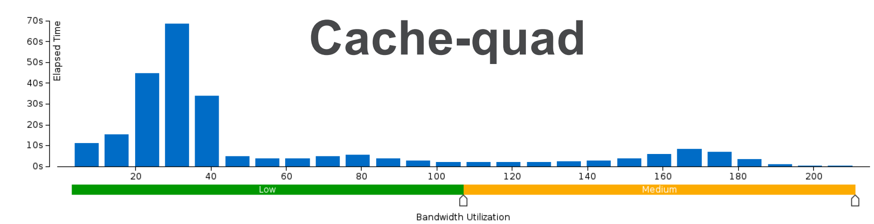
- Molecular dynamics simulation of 32 million particles modeling protein in lipid bilayer up to 3072 nodes of Theta.
- One MPI rank per core in all cases; multiple OpenMP threads used.
- On a per-node basis running identical code, Theta was generally 5.2x faster than Mira.
- Additional 2.2x speedup observed using Intel-optimized code with explicit AVX-512 SIMD instructions.

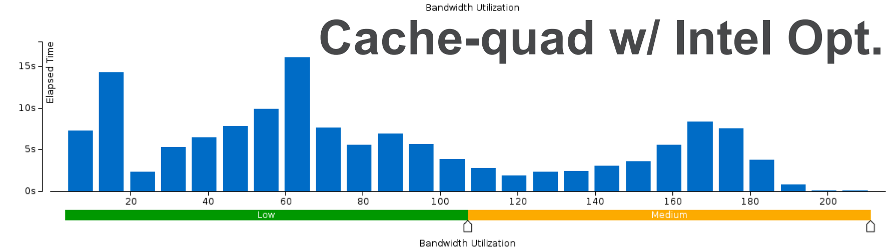


LAMPS – UTILIZING MEMORY HEIRARCHY

- Single-node runs with 256,000 particles and PPPM used for electrostatics.
- DRAM is sufficient to deliver memory bandwidth for pairwise computation and building neighbor lists (~30 GB/s).
- HBM yields up to ~170 GB/s for PPPM stencil and 3D FFT operations.
- Intel-optimized code improves memory bandwidth utilization (30 ➔ 60 GB/s).
- Large-scale runs could default to DRAM with select data structures allocated to HBM.







QUESTIONS?

www.anl.gov

Argonne NATIONAL LABORATORY