Performance and Energy Usage of Workloads on KNL and Haswell Architectures

Tyler Allen[1] Christopher Daley[2] Doug Doerfler[2] Brian Austin[2]
Nicholas Wright[2]

[1]Clemson University

[2]National Energy Research Scientific Computing Center

Exascale Computing

- Exascale is the next supercomputing performance goal (~2021)
- Exascale enables new opportunities in science and technology
- Exascale Computing $\rightarrow$ Cluster with performance 1 ExaFLOP/s
- DoE targets initial power constraint to $\frac{1 \text{ ExaFLOP/s}}{20 \text{ MegaWatts}}$
- Performance under power constraint is the top evaluation metric

Figure: Aurora, to be the first US Exascale computer in 2021 at ANL

Manycore Era

- Current HPC systems are adopting Manycore CPUs
  - KNL is Intel's first manycore self-hosted CPU
  - KNL is in NERSC's Cori, ANL's Theta, LANL's Trinity, etc...
  - Manycore offers performance *and* power benefits
- Research Questions
  - Does the NERSC workload benefit from KNL?
  - Is KNL the right direction for Exascale?

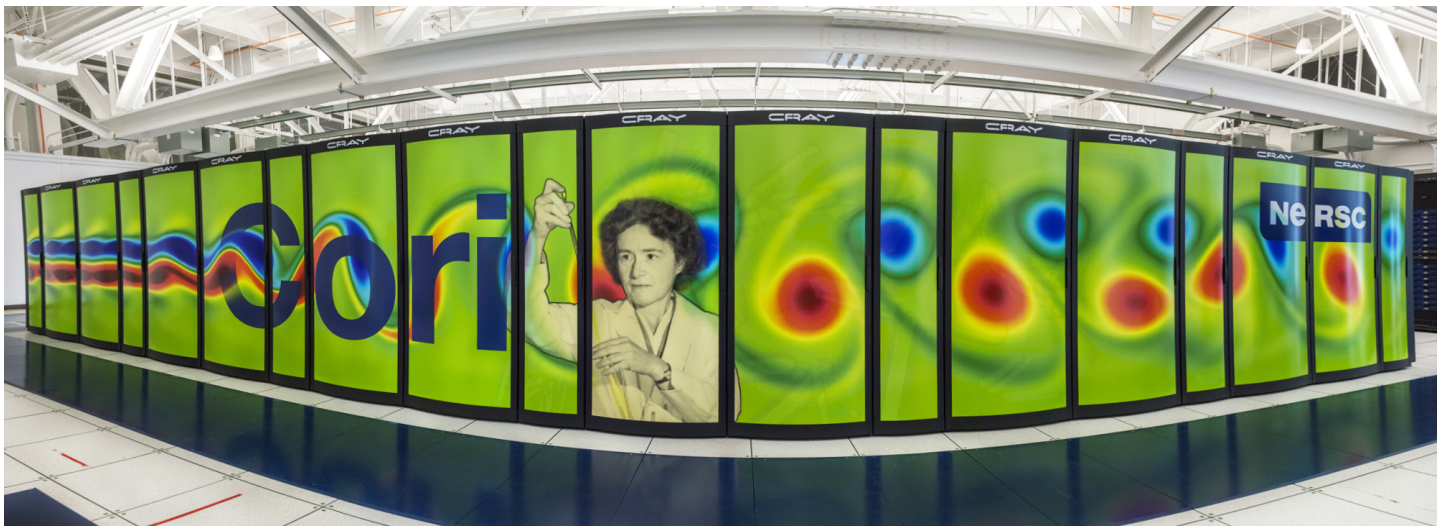


Figure: Cori, NERSC's pre-exascale flagship system, #6 on TOP500

Knight's Landing

- KNL is a milestone on path to exascale
  - KNL is Intel's first manycore self-hosted CPU
  - 64-72 cores, 4 Hyperthreads Per Core
  - Cores "lighter" than traditional server chips
    - Shallower pipelines, less prediction, etc
  - 16GB MCDRAM - Intel On-Chip High Bandwidth Memory (HBM)
    - High Bandwidth, High Latency, High Power-Efficiency

Possible MCDRAM Configurations

Memory Mode	Clustering Mode
Cache	**Quadrant**
Cache	Hemisphere
Flat	**Quadrant**
Flat	Hemisphere
Flat	SNC4
Flat	SNC2

Table: Cache/Quad and Flat/Quad are easiest to use, most used at NERSC

Experimental Overview

- Goal: Compare Manycore to HPC CPU “norm”
  - HPC Norm: Intel Xeon Server Multicore CPUs
  - Multiple heavy cores vs. Many lighter cores
  - Is KNL the correct path to Exascale?
- Method: Benchmark and Contrast
  - We use microbenchmarks to characterize specific features
  - We use real world apps to evaluate practical benefit
  - Benchmarks run on modern KNL and Xeon systems
  - Metrics: Time to Solution, Power/Energy Consumption
  - Variables: MPI/OMP, threads-per-core, MCDRAM configuration

Test System: Cori

- Our test system was NERSC's Cori Supercomputer
- Cori is a representative modern HPC system

	Haswell Xeon	KNL
CPU Model	Intel Xeon E5 – 2698	Intel Xeon Phi 7250 KNL
Clock Speed	2.3GHz	1.4GHz
Total Cores	32	68
Logical Cores	64	272
Sockets	2	1
Memory	128GB 2133MHz DDR4	96GB 2400MHz DDR4 16GB On-Chip MCDRAM
Total Nodes	2388	9688
Network	Cray Aries Dragonfly	Cray Aries Dragonfly

Profiling Tool - IPM

- IPM: Integrated Performance Monitoring
- IPM is an open-source lightweight profiling tool
  - Source Available at `http://www.github.com/nerscadmin/ipm`
- IPM aggregates low-level profiling interface
  - PAPI performance counters, MPI call data, perf events...
- We added energy/power monitoring to IPM
  - Supported through Cray Power Monitoring and sensors
  - Measures energy over application duration

Experiment Applications

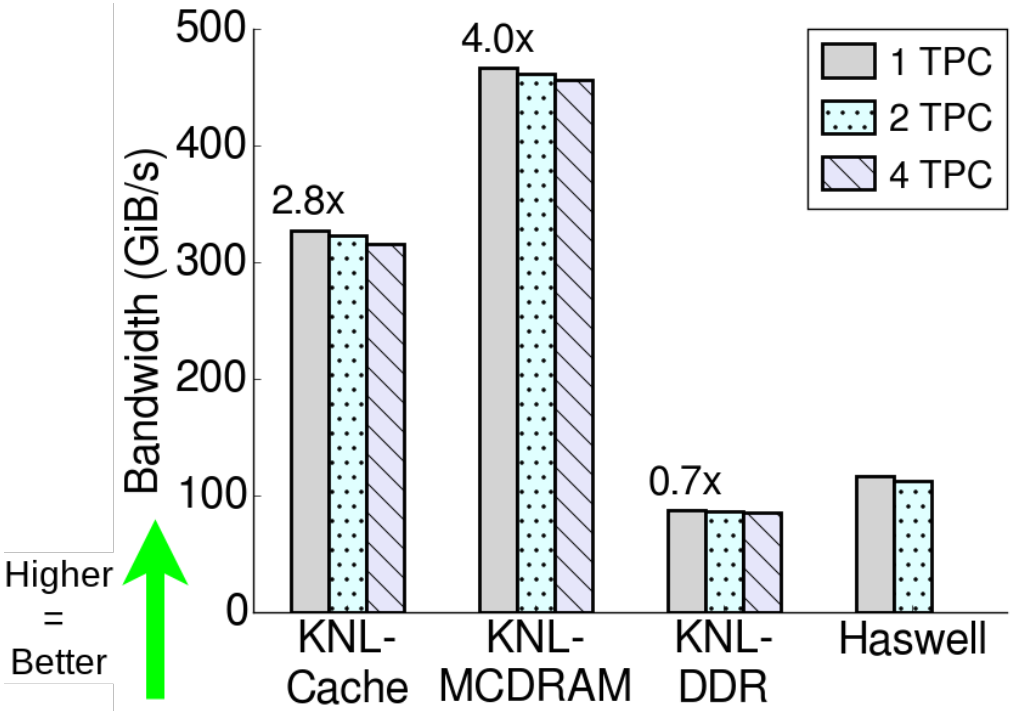
- We use common microbenchmarks to test specific features
- We use applications from NERSC and other DoE labs
  - All applications use hybrid MPI/OpenMP

Application	Science Area	Nodes	Rnks-Thds/Rnk	
			HSW	KNL
STREAM	Memory bandwidth	1	32t	68t
RandN	Random memory access	1	64t	256t
DGEMM	Dense linear algebra	1	32t	136t
GTC-P	Fusion	8	32r-1t	32r-8t
MILC	Quantum chromodynamics	8	32r-1t	32r-2t
Nyx-AMReX	Cosmology	2	16r-4t	16r-16t
Castro-AMReX	Astrophysics	4	32r-1t	32r-2t
Quantum Espresso	Quantum chemistry	4	4r-8t	4r-16t
BD-CATS	Data analytics for cosmology	16	16r-4t	16r-16t

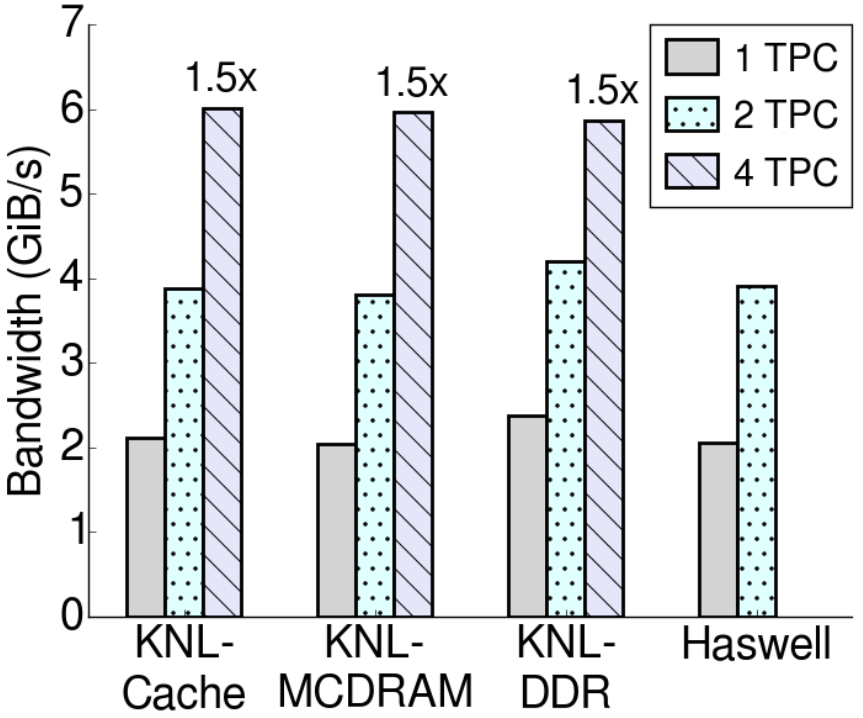
Terminology

- TPC: Threads Per Core
- KNL-Cache: KNL w/ MCDRAM as LLC
- KNL-MCDRAM: KNL w/ MCDRAM as Addressable Memory
- KNL-DDR: KNL w/ MCDRAM not used

Microbenchmark Performance Results vs Haswell



(a) STREAM performance



(b) RandN performance

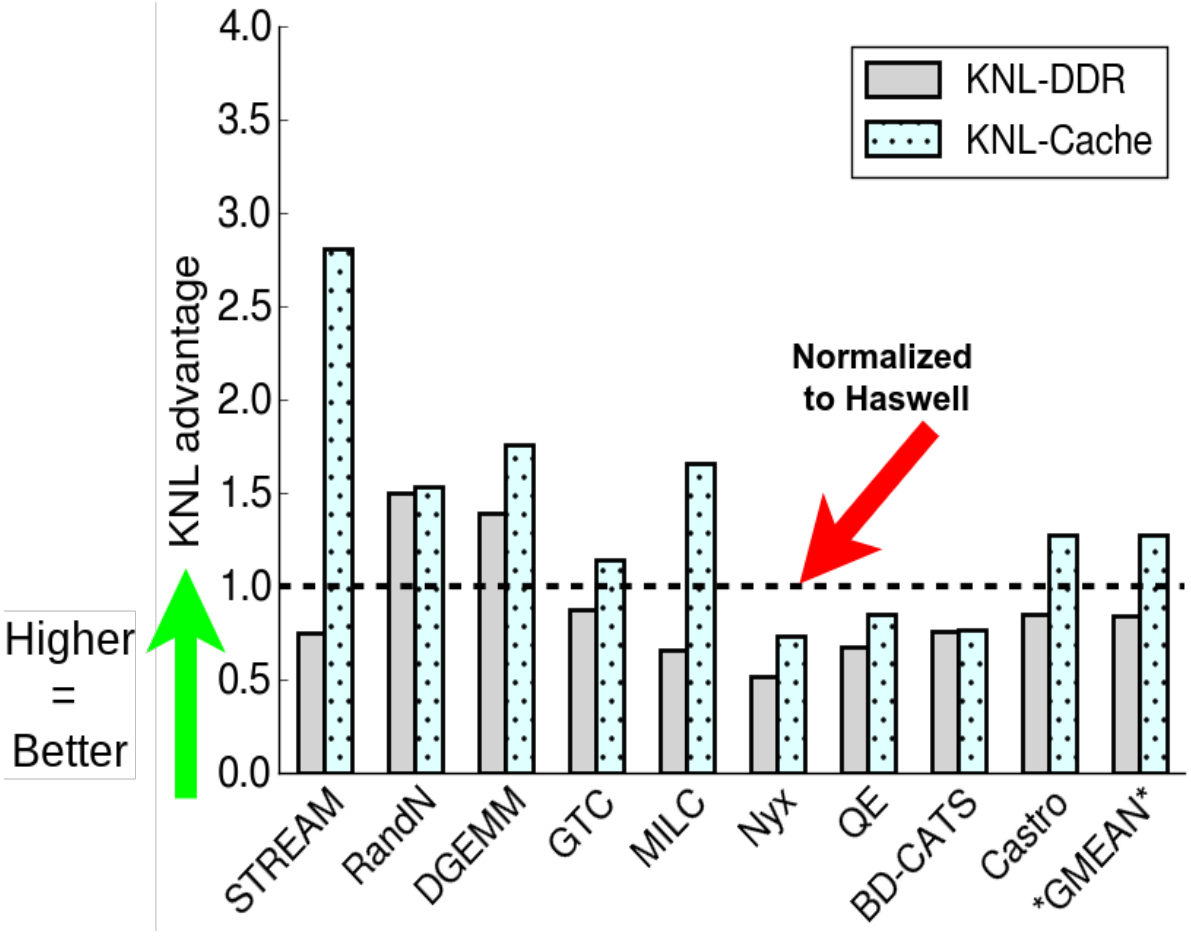
- MCDRAM has significant impact on performance on STREAM
- STREAM is sequential and RandN is random mem access
  - Locality has large impact on MCDRAM/hyperthread value

Microbenchmark Power Results vs. Haswell

Benchmark	Perf Improvement	Energy Improvement
DGEMM	1.9x	2.5x
STREAM	4.0x	4.8x
RandN	1.5x	2.4x

- Note: Table values are with KNL-Cache mode
- Energy efficiency shows greater improvement than performance
- Our results show DGEMM achieves 150pJ/FLOP
  - Exascale Target: 20pJ/FLOP

KNL vs. Haswell Performance

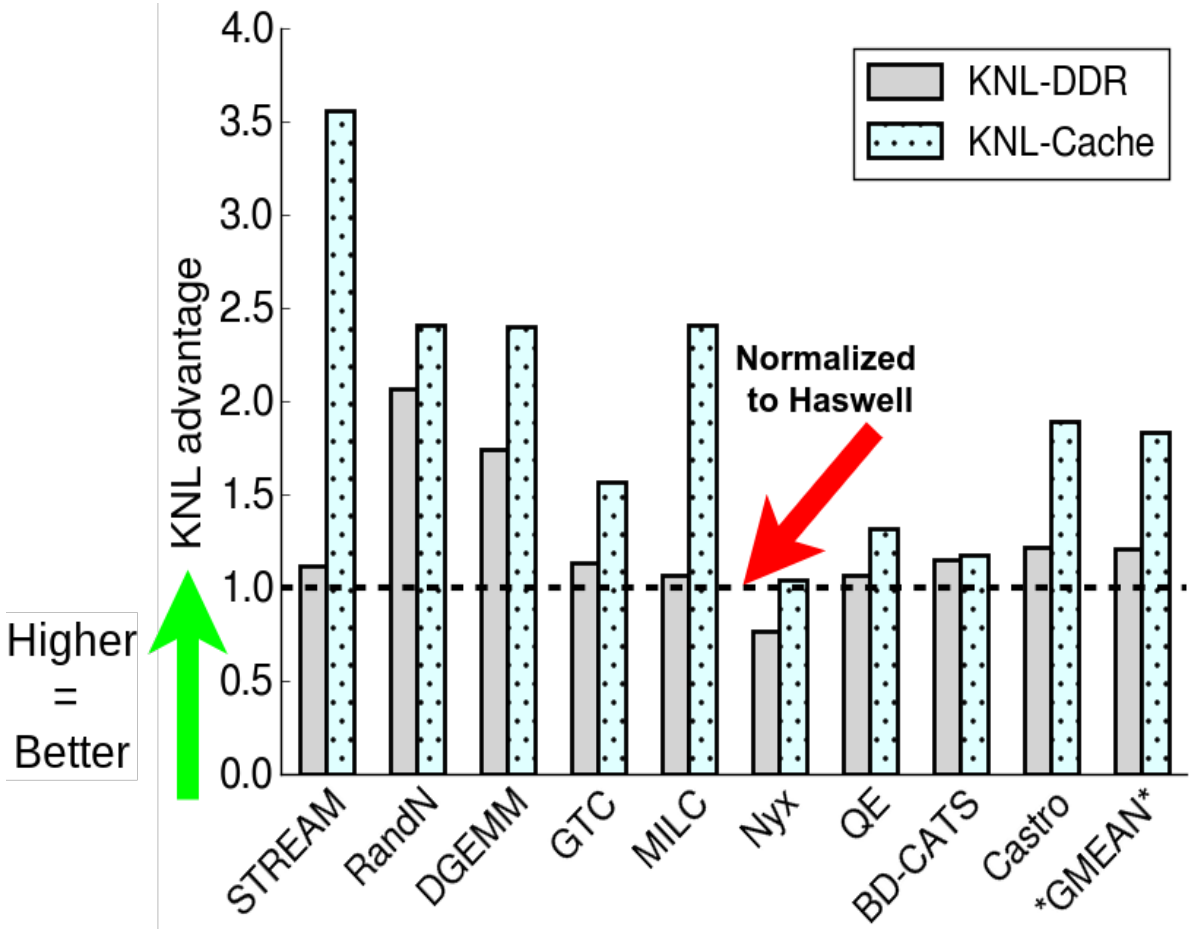


(c) Geometric mean: 0.84 and 1.27

Figure: Best KNL configuration against best Haswell

- Value above bar indicates speedup relative to Haswell
- Real Apps on KNL-DDR *always* worse than Haswell
- The best KNL configuration is *always* KNL-Cache

KNL vs. Haswell Total Energy



(a) Energy Consumption
Geometric mean: 1.21 and 1.84

Figure: Best KNL configuration against best Haswell

- Except BD-CATS, *all* show significant efficiency improvement w/ MCDRAM
- Cache mode *always* improves energy efficiency over Haswell

Summary

- We evaluated HPC applications on KNL vs. Haswell
- We explored the parameter spaces for KNL and Haswell
- Main Findings
  - KNL improves performance for 6 of 9 apps vs Haswell
  - KNL reduces energy consumption for all listed applications
  - Apps with locality show significant improvement from MCDRAM
  - Geometric Mean Perf Improvement vs. Haswell: 1.27x
  - Geometric Mean Energy Improvement vs. Haswell: 1.84x
  - DGEMM achieves 150pJ/FLOP vs. Exascale Target of 20pJ/FLOP
- KNL is a step in the direction of Exascale
- We still need much greater efficiency gains

Acknowledgement

- This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.
- Christopher Daley, Doug Doerfler, Brian Austin, Nicholas Wright NERSC and Lawrence Berkeley National Laboratory

Future Work

- Compare KNL to more recent Intel architecture (Broadwell, etc)
- More thorough, specific characterization using perf counters
- Derive specific indicators of application sensitivity to parameters

Stream Variability in KNL-Cache

- Cache mode has issues with consistency
- Performance reduction correlated with DDR Traffic
  - Indicating a Last-Level Cache Miss
- However, STREAM should fit in MCDRAM
- Direct-map KNL cache can cause **significant variation**

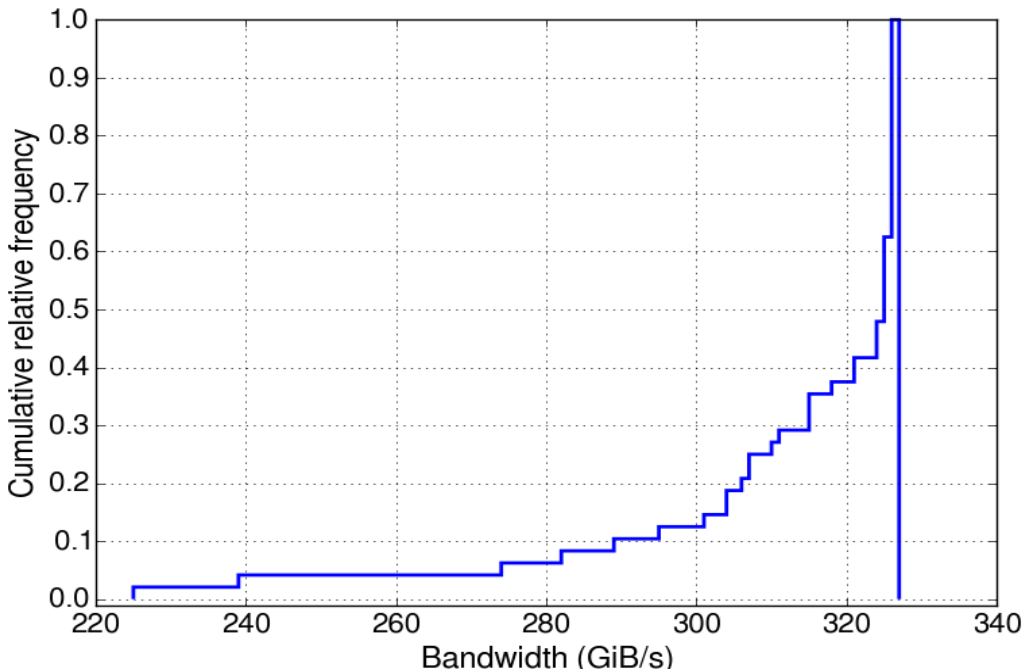


Figure: STREAM bandwidth in KNL-Cache mode over 48 trials

Cray Power Monitoring vs RAPL

- Cray power monitoring using physical sensor
  - RAPL uses perf event avg estimation
  - Physical measurement far more accurate
- Cray monitors on the rail at input source before voltage drop
- Supported by default on Cray systems (Cori)