Evaluating the Networking Characteristics of the Cray XC-40 Intel Knights Landing Based Cori Supercomputer at NERSC

Douglas Doerfler, Brian Austin, Brandon Cook, and Jack Deslippe
Lawrence Berkeley National Laboratory

Krishna Kandalla and Peter Mendygral
Cray Inc

Cray User Group Meeting
Redmond, WA
5/9/2017

Outline

- Introduction and purpose
- Edison and Cori
- Communication microbenchmarks
- Application performance
- Conclusions and observations

Introduction and Purpose

- There are many potential issues with using the Intel Xeon Phi Knights Landing (KNL) manycore processor in a supercomputer
  - The most obvious is the impact on computation
- But what about the impact on communication?
  - We expected there to be some impacts given the Xeon Phi core is a fraction of the performance of a regular Xeon core
- The purpose of this paper is to investigate and quantify this impact
  - Using microbenchmarks and applications we analyze tradeoffs in MPI and OpenMP
  - We also investigate one sided performance using a data intensive UPC application and microbenchmarks

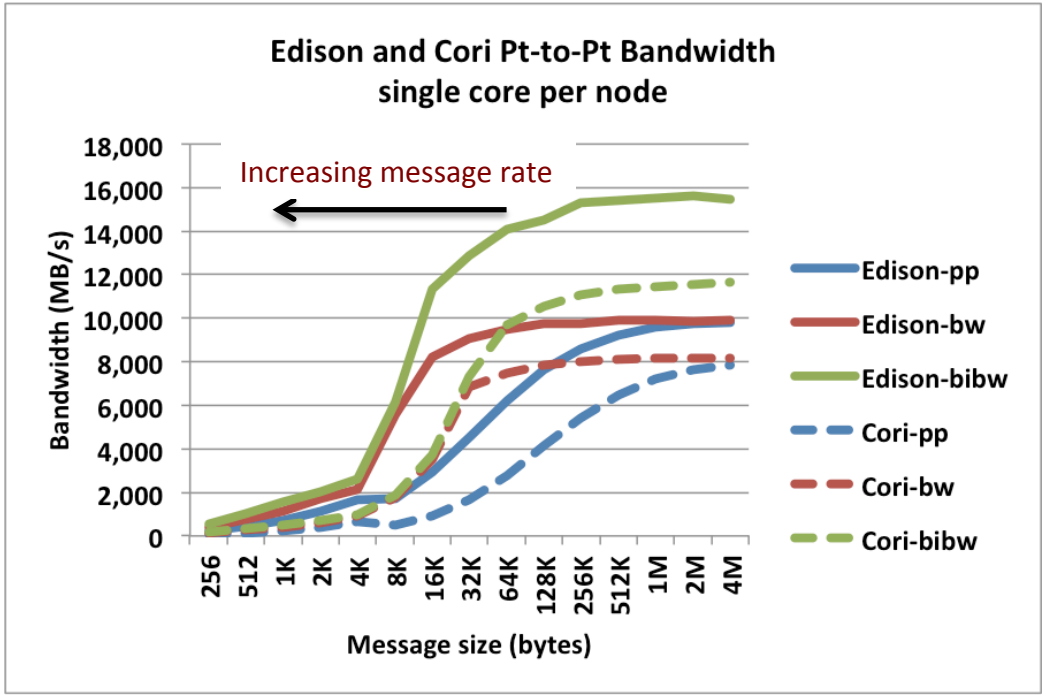
Edison and Cori

- Edison and Cori, located at NERSC, are both based on the Cray XC architecture and both use the Cray Aries high-speed interconnect
- Edison − 5,200 dual socket, 12 core, Intel Ivy Bridge Xeon based nodes
- Cori − 9,688 single socket, 68 core Intel KNL based nodes
  - Cori also has 2,388, dual socket, 16 core, Intel Haswell based nodes
  - But for this talk, when we refer to Cori it's just the KNL partition

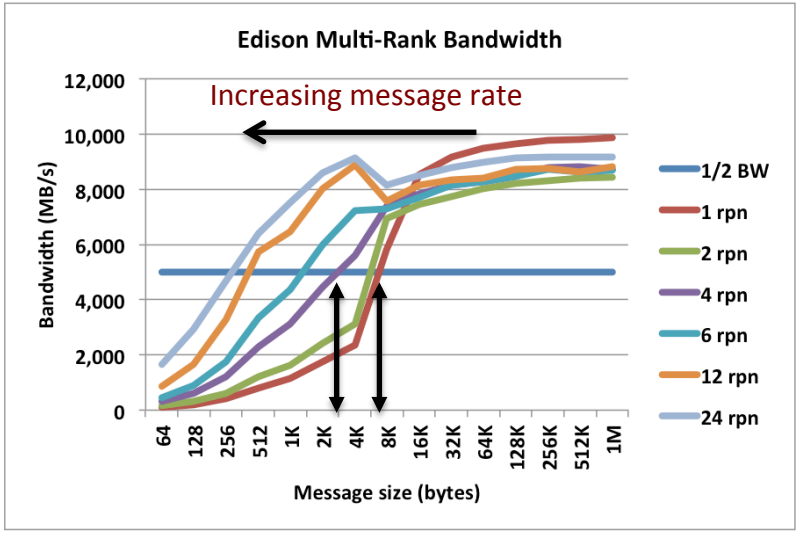
Communication Microbenchmarks

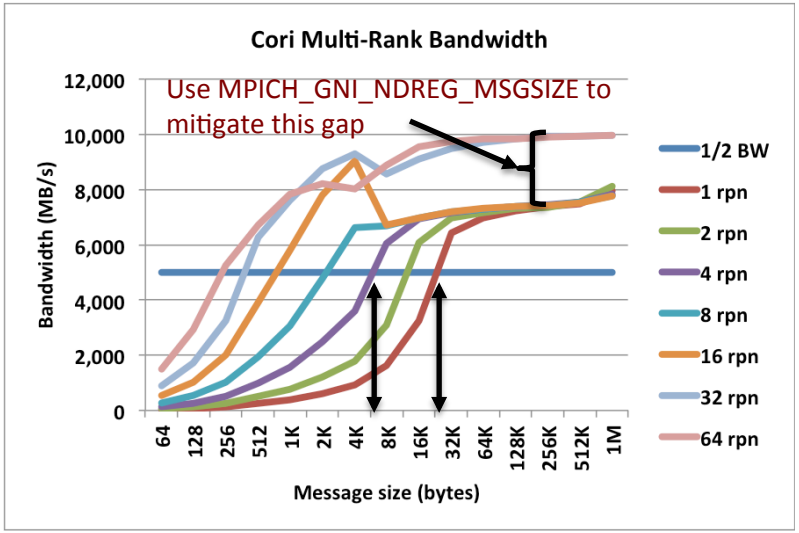
Single rank bandwidth: Pt-2-Pt

- Single core, point-to-point between 2 nodes
- Ping-pong is exactly that, single message ping-ponged between 2 nodes
- Uni-directional is a “streaming” exchange of data with a window of size 64
- Bi-directional is also “streaming”
- Using a single core, Edison achieves a significantly higher bandwidth than Cori
- Cori Latency = 3.1 µS
- Edison Latency = 1.2 µS
- Cori’s latency is ~2.6x greater



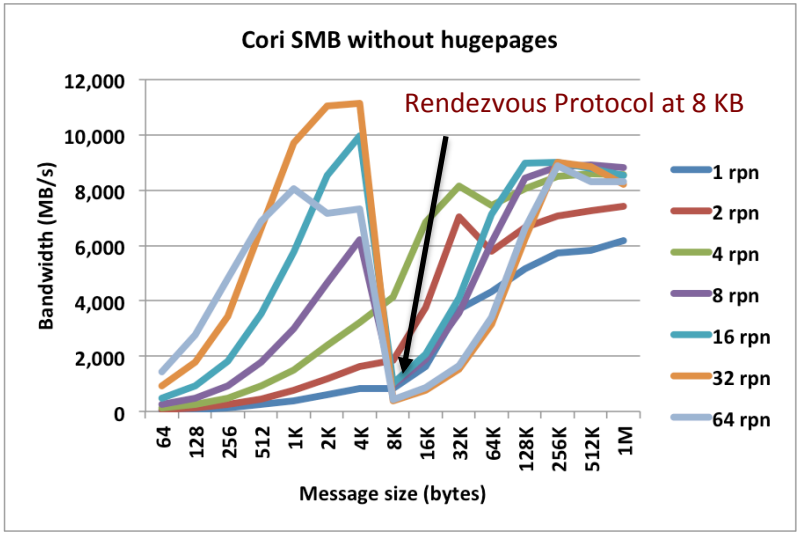
Multi-rank bandwidth: Pt-2-Pt

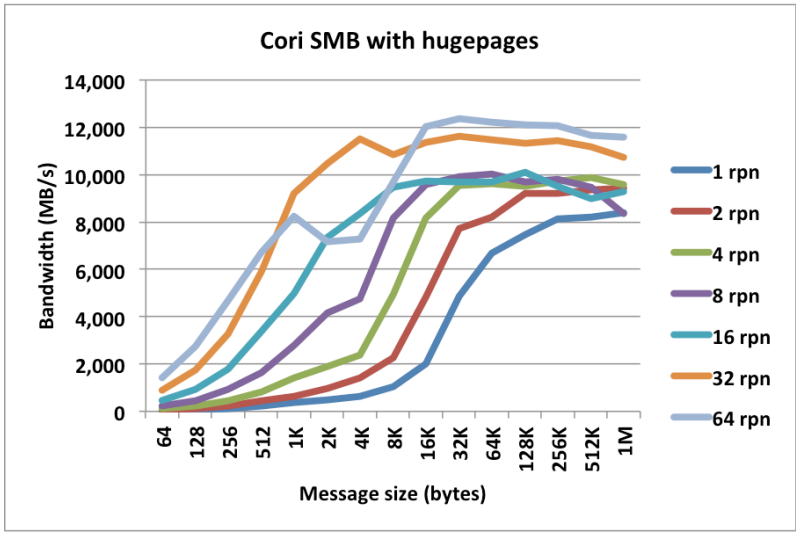




- As ranks per node (RPN) increase in this uni-directional test, higher message rates improve bandwidth
- Cori has a BW "ceiling" below 32 RPN that limits large message performance
  - Cray attributes this to a PCI latency issue between KNL and Aries
  - can be mitigated by moving Aries BTE "put" protocol transition to a smaller message size
  - MPICH_GNI_NDREG_MSGSIZE=65536 (default = 4 MB)
- Cori requires ~2x the message size to achieve bandwidths similar to Edison

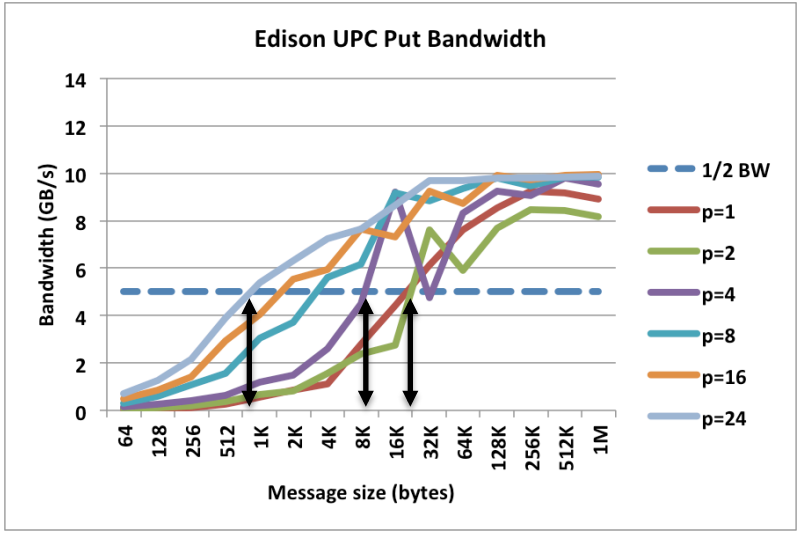
Multi-node, multi-rank bandwidth: 3D stencil

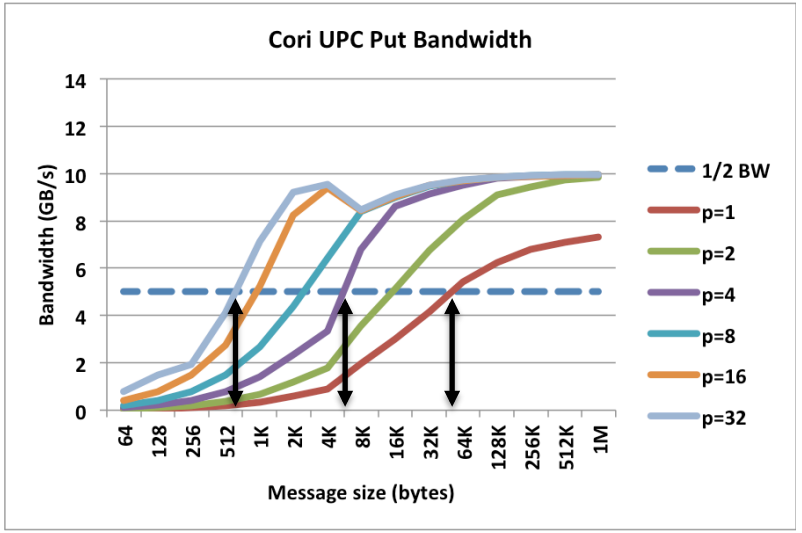




- 16 nodes, 6 neighbors per rank (emulates 3D stencil communication pattern)
  - 6 x RPN (up to 384) total communication pairs per node
- Here we see a severe performance drop when transitioning to the MPI Rendezvous Protocol at 8 KB
  - This is due to TLB thrashing on the Aries network Interface card (the Aries TLB, not the CPU TLB)
- Mitigation is to use huge pages, craype-hugepages2M module (or larger)

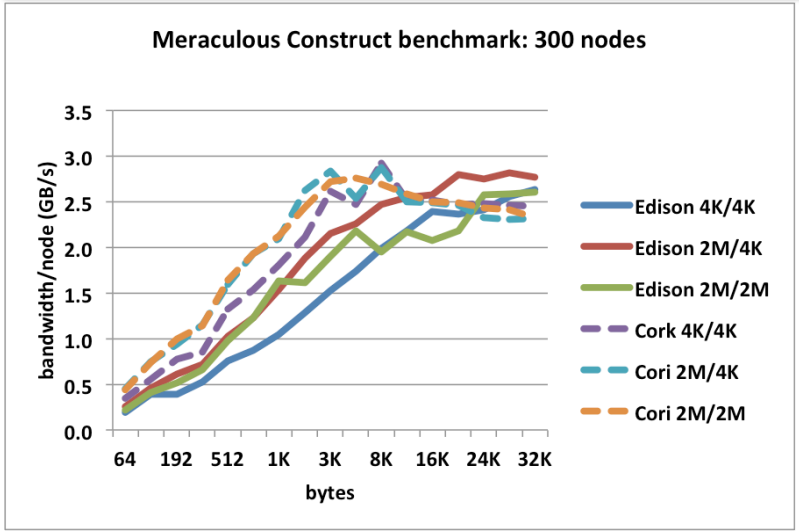
UPC multi-rank bandwidth: Pt-2-Pt

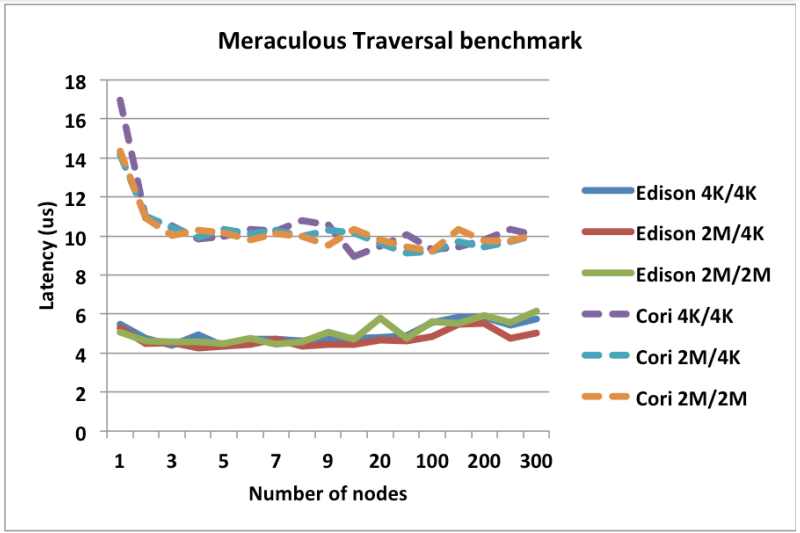




- **Uni-directional bandwidth measured in a way similar to earlier MPI results**
- **As with MPI, Cori is unable to achieve full bandwidth with a single rank-pair**
- **However, Cori's extra cores allow higher bandwidths when all cores are fully utilized**
- **UPC Get performance has similar characteristics, but reduced by ~10%**

Meraculous (UPC) benchmarks

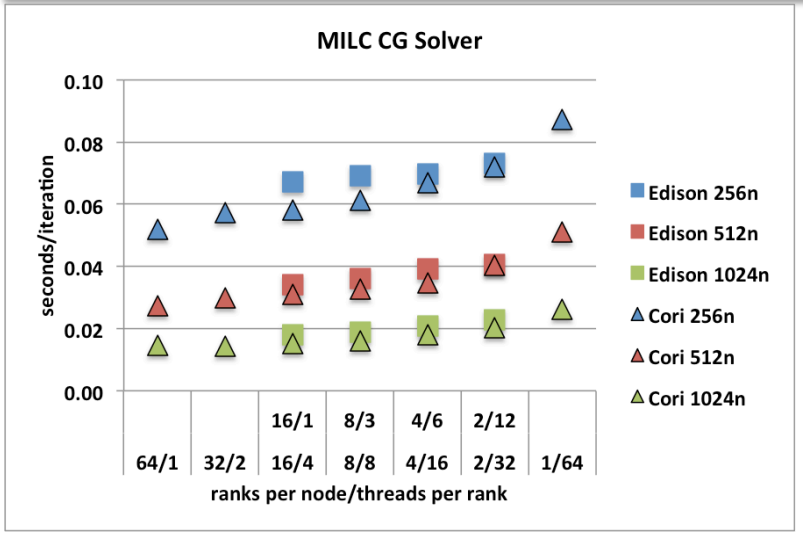


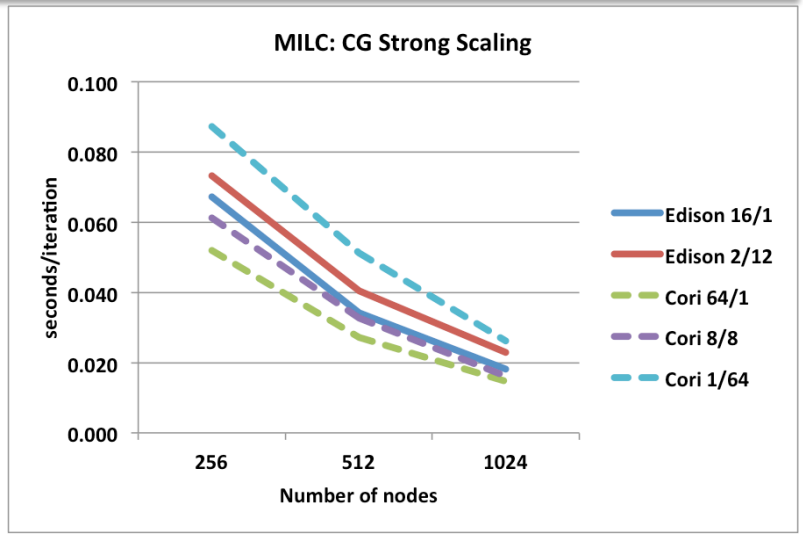


- **Both operate on a 4 GB data array evenly distributed amongst all nodes**
- **Construct: proxy for the construction of the distributed hash table**
  - **Remote *atomic_fetch_and_add* followed by *upc_memput***
  - **Cori shows higher bandwidth per node for small messages, but then tails off at larger sizes**
- **Traversal: proxy for the traversal of the graph**
  - **64 B *upc_memget* followed by *remote_atomic_compare_and_swap***
  - **As with UPC and MPI benchmarks, Edison has a lower latency, ~2x better than Cori**

Applications

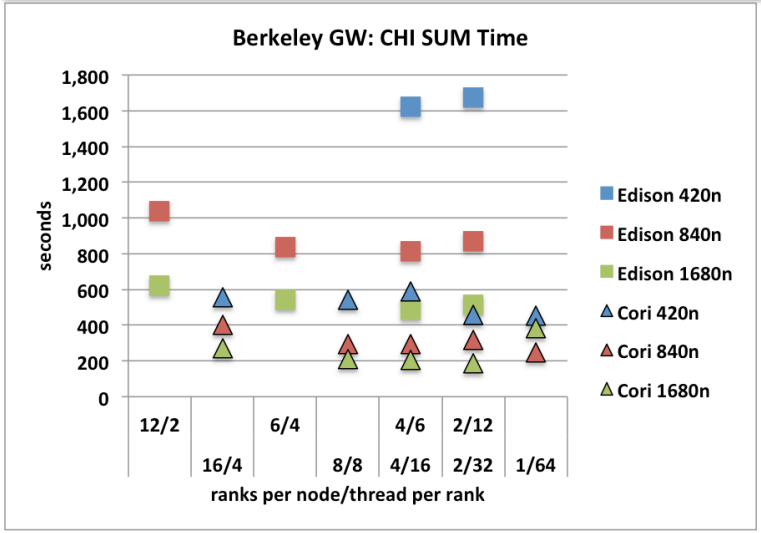
MILC: Quantum Chromodynamics

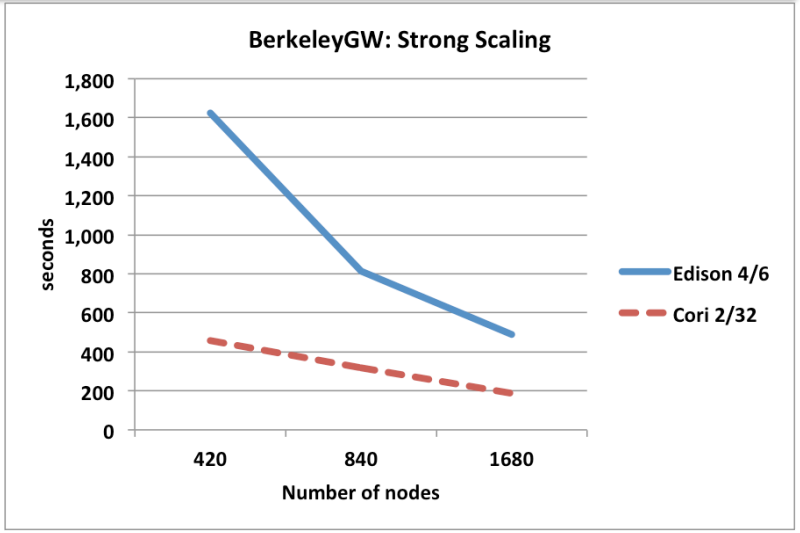




- Lattice size of 128x128x128x128 per node
- Dominated by Conjugate Gradient (CG) solve phase
- Strong scaled from 256 to 512 to 1024 nodes
- Best performance is with MPI-only due to thread scaling limitations of MILC
  - MPI-only performance gains are up to 44% better due to use of craype-hugepages2M
- Cori shows best performance at 256 nodes, but as scale increases performance becomes comparable as communication dominates

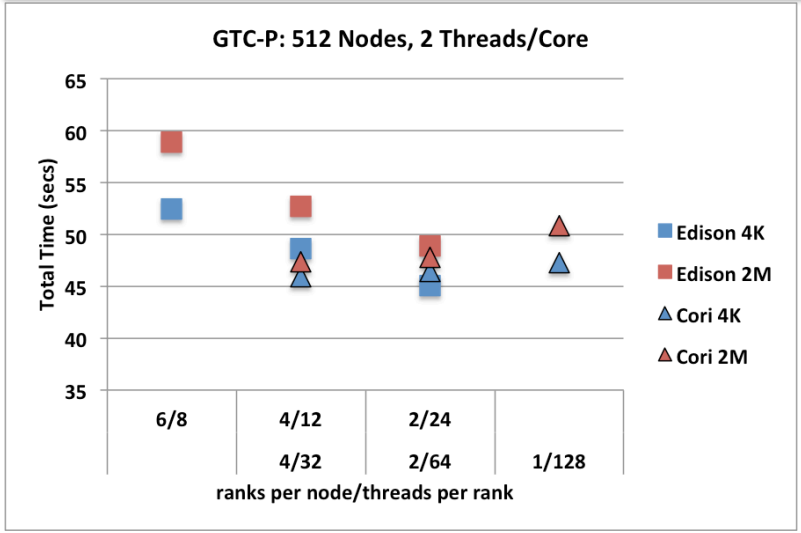
Berkeley GW: Material Science

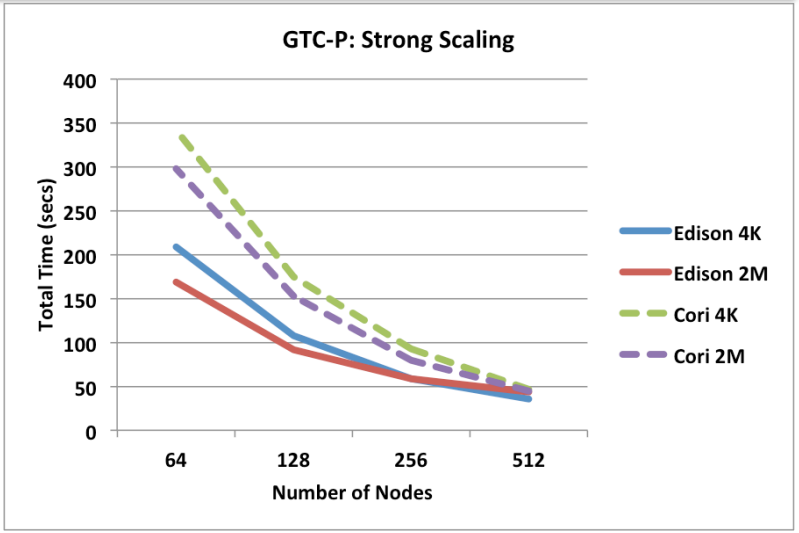




- Dominated by dense linear algebra, in particular fast Fourier transforms (FFT)
- Here we see Cori provides the best overall performance:
  - 3.6x speedup @ 420 nodes, 3.3x @ 840 nodes and 2.6x @ 1680 nodes
- BGW prefers a few ranks per node
  - Larger message sizes per rank maximizes interconnect bandwidth
  - OpenMP scaling is excellent, and OpenMP reduces perform better than an intra-node MPI_Reduce
  - However, 1 rank per node is not sufficient to drive the network as seen in the microbenchmarks

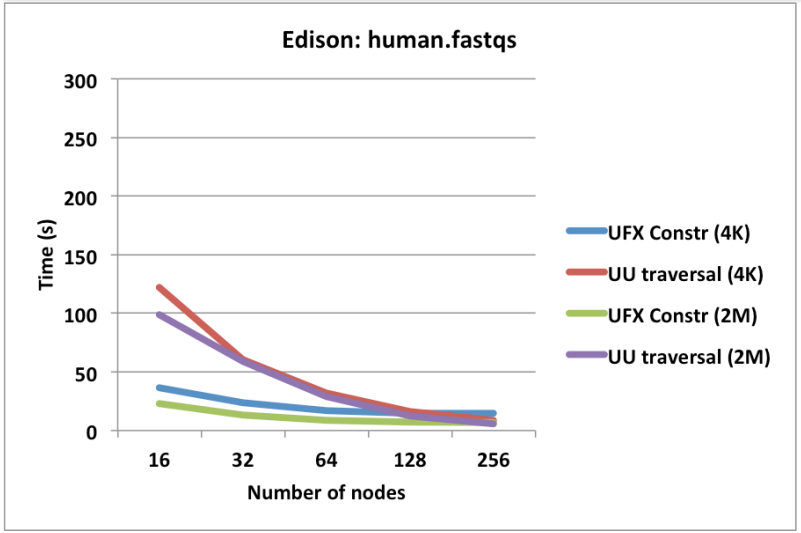
GTC-P: Fusion/Particle in cell

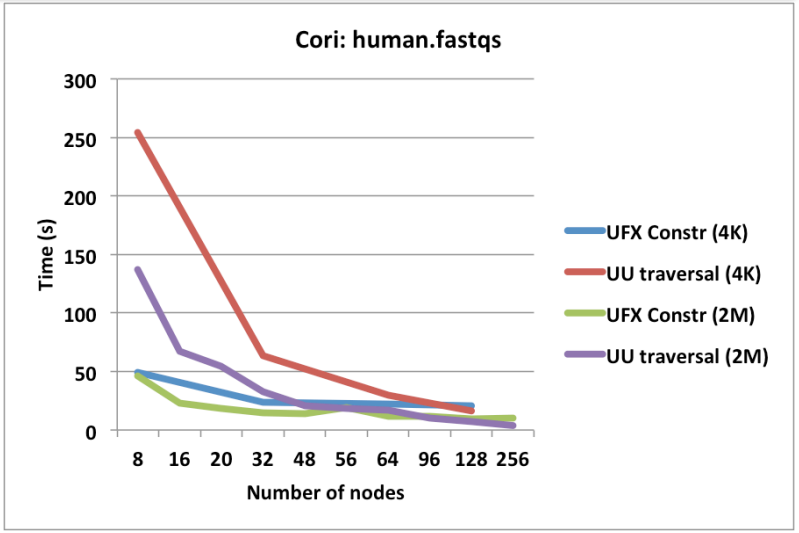




- Dominated by a ”charge” phase that deposits charge from the particles to the grid, and a ”push” phase that interpolates the electric field from the grid to the particles and updates the particle positions based on that field
- In the MPI/OpenMP trade off study, Cori shows good performance at all decompositions, but again we see a performance degradation when using a single rank per node
  - Edison with 2 MPI ranks and 24 threads per core provides the best overall performance, a 1.03x speedup over Cori
- In the strong scaling study, we see that 2M pages give best performance at lower scales, but this advantage diminishes by 512 nodes when 4K pages is actually best, a factor 1.1x for Edison

Meraculous: genome assembly





- Fine-grained random access is a typical feature in this communication heavy pipeline which implements a variety of graph algorithms
  - (4K) uses 4K pages for UPC shared memory segments and application allocations
  - (2M) uses 2M pages for UPC shared memory segments and 4K pages for application allocations
- In this latency sensitive test, Edison provides the best performance at small scales, but that performance advantage diminishes with scale, additionally
  - increased number of UPC threads/node on Cori results in increased contention for reservation of space in destination buffers on Cori relative to Edison
  - as the concurrency is increased the probability of multiple UPC threads accessing the same location in a fixed size table increases

Conclusions & Observations

- It was expected that in order to fully utilize the high-speed network it is necessary to use multiple ranks (or UPC threads)
  - On Cori, we need 2x to 4x more communication contexts in order to achieve the equivalent performance of Edison
  - Cori's latency is roughly 2x to 3x that of Edison
  - At scale, and assuming sufficient number of ranks per node, Cori equals or exceeds Edison, with the exception of Meraculous

Conclusions & Observations Cont'd

- We found that when using a high number of ranks per node, there is an advantage to using huge pages
  - Significant performance gains for nearly all codes, with the exception of GTC-P which favors fewer ranks with 4 KB pages
- We found that we need at least 32 ranks per node in order to obtain maximum bandwidth with Cori
  - Decreasing the RDMA push protocol transition from 4 MB to 64K using MPICH_GNI_NDREG_MSGSIZE improved microbenchmark performance, but we didn't see a significant improvement with a real application
- It is best to test with and without page sizes and protocol changes, these may or may NOT help with your app

Conclusions & Observations Cont'd

- Once you understand how to best utilize the interconnect, performance then depends on the OpenMP (or other thread) scaling characteristics of your application
  - You many find a MPI/OpenMP "sweet spot"
- Advantages to using fewer ranks may include
  - Reduced intra-node communication
  - larger inter-node message sizes (better effective BW), e.g. from decreasing surface to volume rations
  - OpenMP reductions may be faster

Conclusions & Observations Cont'd

- For UPC, we can draw similar conclusions
  - An advantage to using 2 MB pages
  - Cori's latency is 2x to 3x higher than Edison
  - Due to lower latency and reduced problem decomposition and congestion, Edison demonstrated a significant performance advantage on Meraculous
  - But as we strong scaled the problem out, performance differential is reduced

The End

dwdoerf@lbl.gov