BNL FY17-18 Procurement

USQCD All-Hands Meeting
JLAB
April 28-29, 2017

Robert Mawhinney
Columbia University
Co Site Architect - BNL

BGQ Computers at BNL

USQCD half-rack (512 nodes) | 2 racks RBRC | 1 rack of DD2 BNL

USQCD 512 Node BGQ at BNL and DD2 Rack

- USQCD SPC allocated time for 4 projects in 2015-2016. Usage as of June 30, 2016. Units are M BGQ core-hours

P.I.	Allocated	Used	% Used
Feng	27.14	31.96	117%
Kuti	14.10	4.74 (DD2) + 11.90 = 16.64	118%
Mackenzie/ Sugar	29.56	2.07 (DD2) + 31.90 = 33.97	115%

- USQCD SPC allocated time for 3 projects in 2016-2017. Usage as of April 26, 2016.

P.I.	Allocated	Used	% Used	Max Usage	Max % Usage
Kelly	50.64	58.98	116%		
Kuti	14.59	7.02	54%	18.02	124%
Mackenzie	5.57	7.77	139%		

- All USQCD jobs run this allocation year have been 512 node jobs.
- Between April 1, 2016 to April 1, 2017, the integrated usage of the half-rack has been 358.6 out of 365 days.
- The LQCD project will run the half-rack through the end of September, 2017.

USQCD Needs: Flops and Interconnect

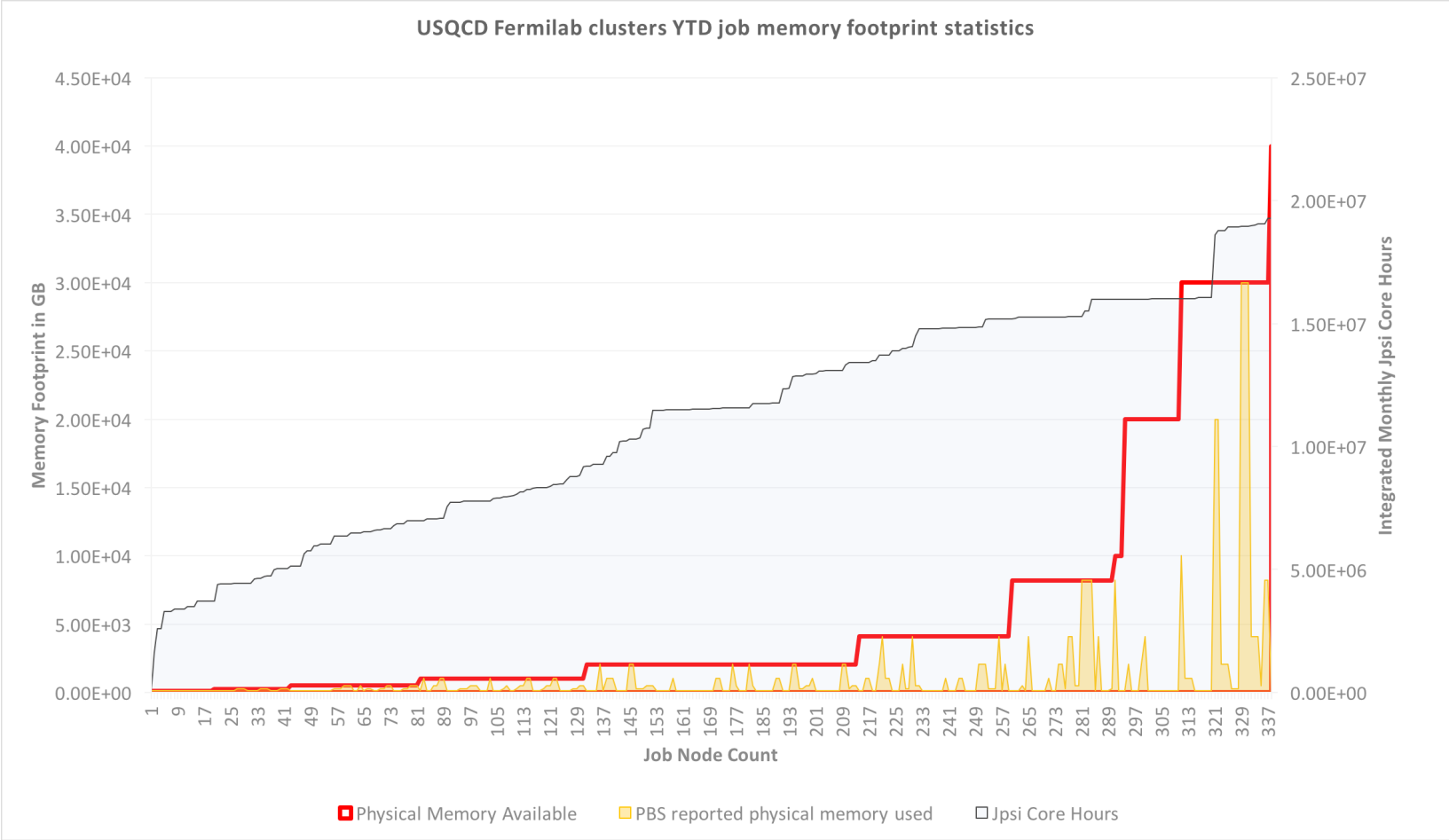
- QCD (internode bytes per second) per (Flop per second) ~ 1
  * BGQ example or DWF with 8^4 per node
  * 20 GBytes/second for 40 GFlops/second on a node
- Now have nodes (KNL, GPU) with ~400 GFlops/sec.
  * With same local volume would need 200 GBytes/second of internode bandwidth
  * Making local volume 16^4 would cut internode bandwidth in half to 100 GBytes/s.
- 100 GBit/second IB or Omnipath gives 12.5 GBytes/sec
- Interconnect speeds limit strong scaling, implying a maximum node count for jobs.
- Size of allocation limits job size. A calculation requiring most or all of the nodes at a site is unlikely to have a large enough allocation to make progress on such a difficult problem.

USQCD Needs: Memory and I/O Bandwidth

- Ensembles are larger and many measurements are made concurrently in a single job.
- Deflation techniques and large number of propagators for contractions are increasing memory footprint.
  * g-2 on pi0 at FNAL: 128 GBytes/node * 192 nodes = 24 TBytes
  * BGQ half-rack: 16 GBytes * 512 nodes = 8 TBytes
  * Jobs of Mackenzie this allocation year just fit on BGQ half-rack.
  * Expect some reduction of footprint via compression and blocking techniques.
- I/O is becoming more important
  * g-2 on pi0 uses all of the available bandwidth to disk when loading eigenvectors.

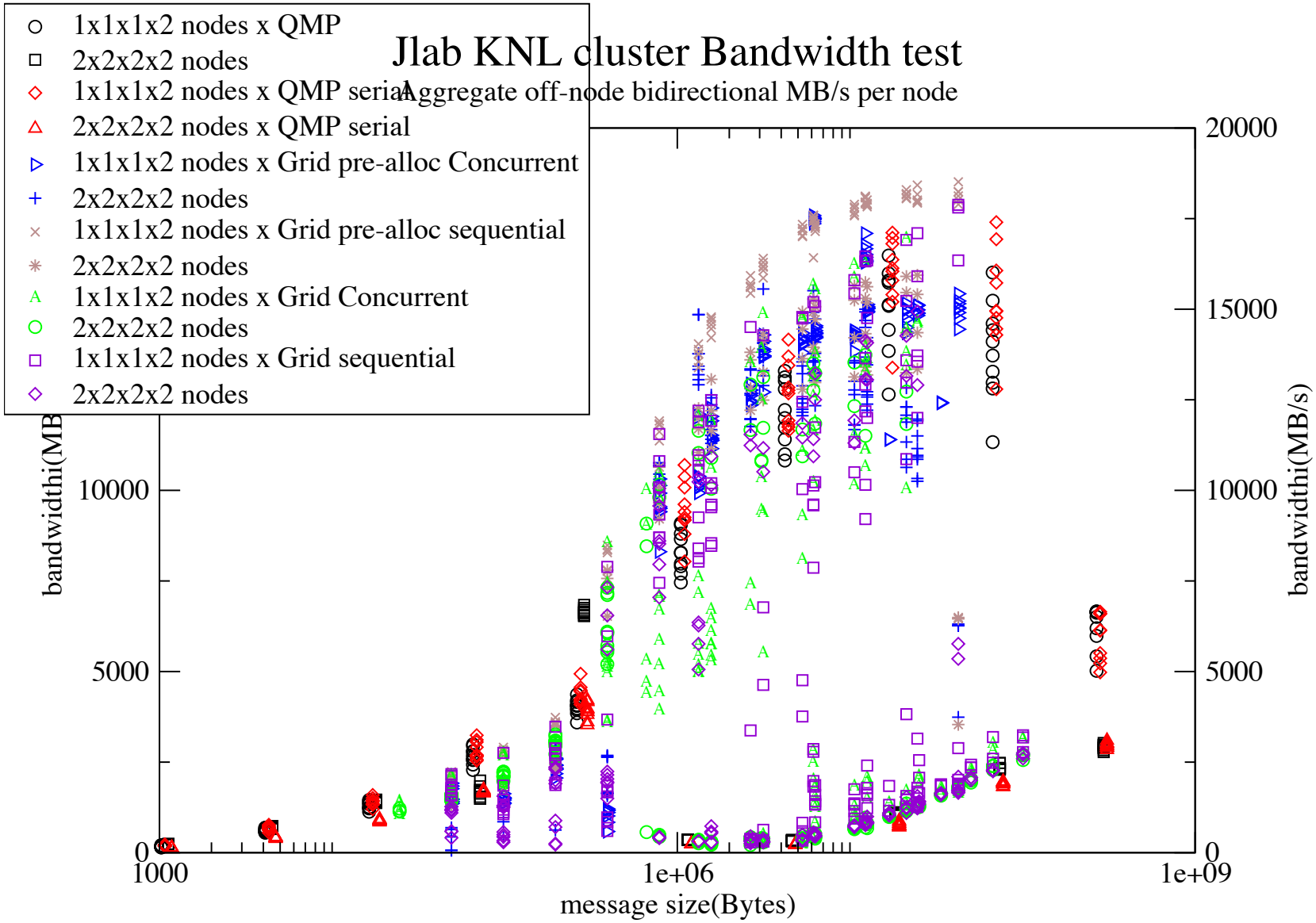
USQCD users need access to 16 to 32 node partitions with ~5 TBytes of memory. Such partitions are intermediate between single node jobs, which run well on GPUs and current KNL, and jobs for Leadership Class Machines.

USQCD History and Requests



- FNAL: 75% of time used for jobs with less than ~5 TBytes of memory.
- KNL at JLAB in 2016-2017 has had 90% of time used to date in single node jobs
- BGQ half-rack at BNL has only ran 512 node jobs this allocation year.
- In this year's requests to the SPC, conventional hardware is oversubscribed by a factor of 2.49, GPUs by 0.98, and KNLs by 2.19. User preference for KNL/Intel clear.

Internode Bandwidth on JLAB KNL using Grid

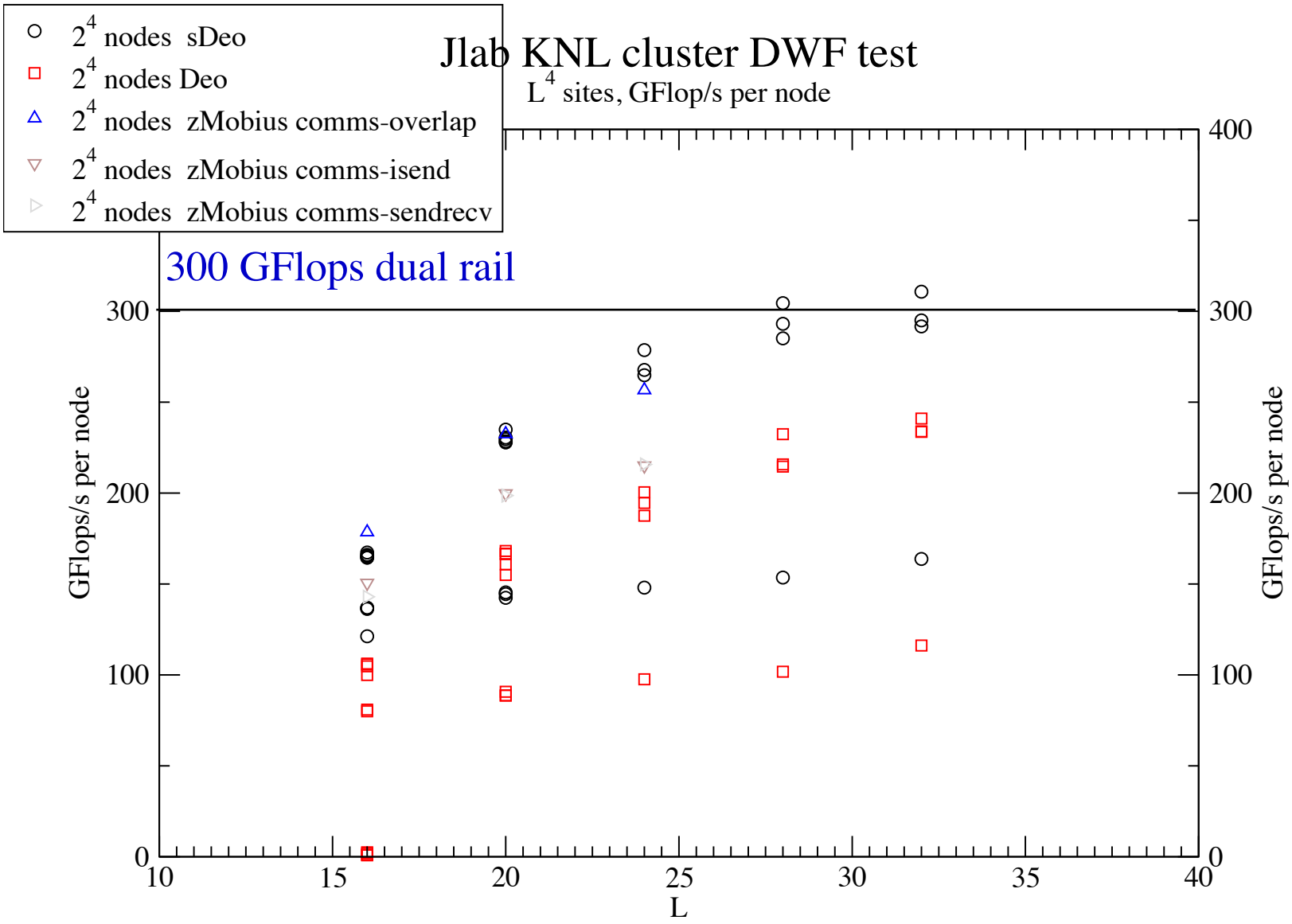


Chulwoo Jung

Scaling on BNL KNL with DWF using Grid

- Grid software, developed by Boyle and collaborators. Benchmarks by Boyle, Jung, Lehner
- DWF reuses gauge fields during D-slash, due to fermions in the fifth dimension.
- Peter has done extensive work to get around MPI bottlenecks. Basically handles communication between MPI ranks on-node by custom shared memory system.
- For a 16 node machine size, with single rail at BNL (Alex unplugged one link on each node), MPI3 and all-to-all cache mode, Lehner finds:
  * 24^4 with overlapping communication and compute: 294 GFlops
  * 24^4 without overlapping communication and compute: 239 GFlops.
  * 16^4 with overlapping: communication and compute: 222 GFlops
  * 16^4 without overlapping communication and compute: 176 GFlops
  * 16^4 and 24^4, dual rail and overlapping communication and compute: 300 Gflops
- On dual-rail KNL at BNL, Lehner reports 243 GFlops/node for a 128 node job. 24^4 local volume, zMobius CG, using MPI3 with 4 ranks per node.

Performance on JLAB KNL with DWF using Grid



Tests by Chulwoo Jung

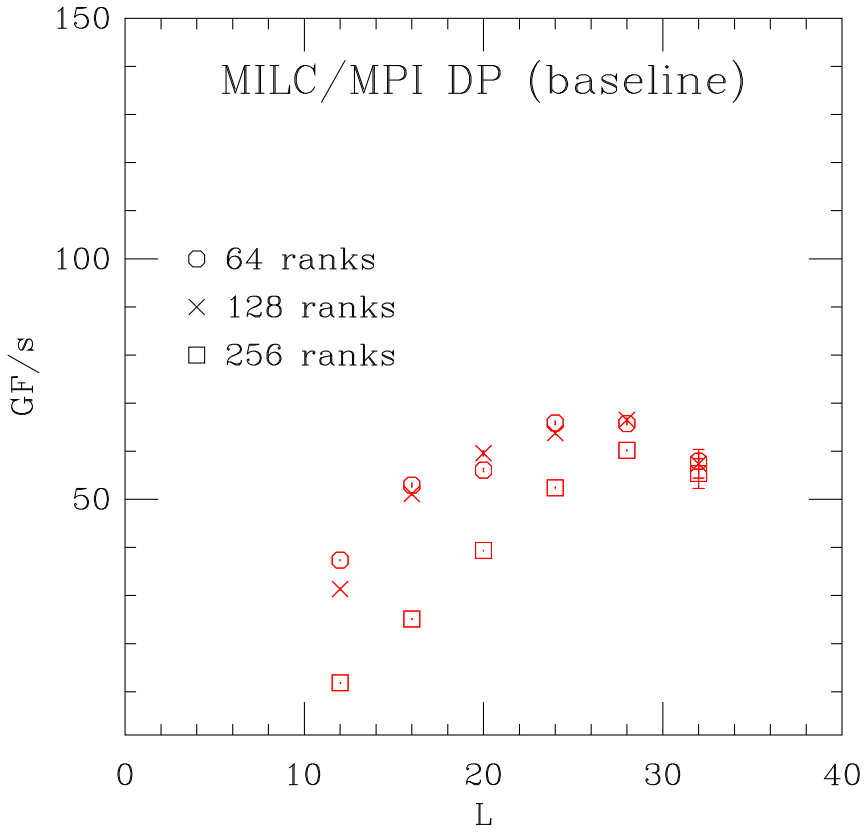
Performance on BNL KNL with MILC and QPhiX

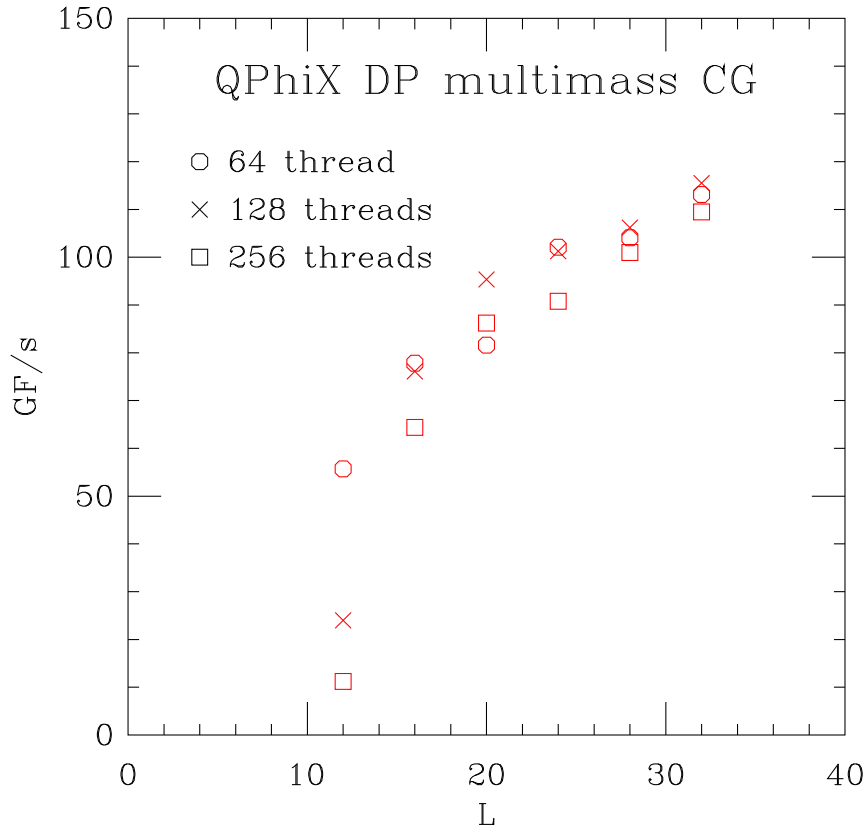
Multi-shift CG performance in Gflops/s/node. Double precision.

MPI ranks	1	2	4	8	16	32	64	128	256
Threads	64	32	16	8	4	2	1	1	
	16 node results with QPhiX, dual rail								
16^4	12.6	12.6	13.1	13.5	14.4	14.0	11.4		
24^4	19.5	20.9	21.4	22.1	21.8				
32^4	24.4	25.2	25.4	26.4	26.4	25.7	22.6		
	16 node results without OMP and QPhiX, dual rail								
24^4							15.2	20.9	
32^4							17.2	29.3	
	1 node results with QPhiX, dual rail								
24^4	35.8	30.4	27.2	25.2					
32^4	38.5	32.1	29.2	28.4					
48^4	34.4	30.8	29.7	29.0					
	1 node results without OMP and QPhiX, dual rail								
24^4							16.6	29.8	36.1
32^4							18.4	34.5	56.0
48^4							22.7	38.3	37.4

MILC code provided by Steve Gottlieb. Benchmarks run by Zhihua Dong

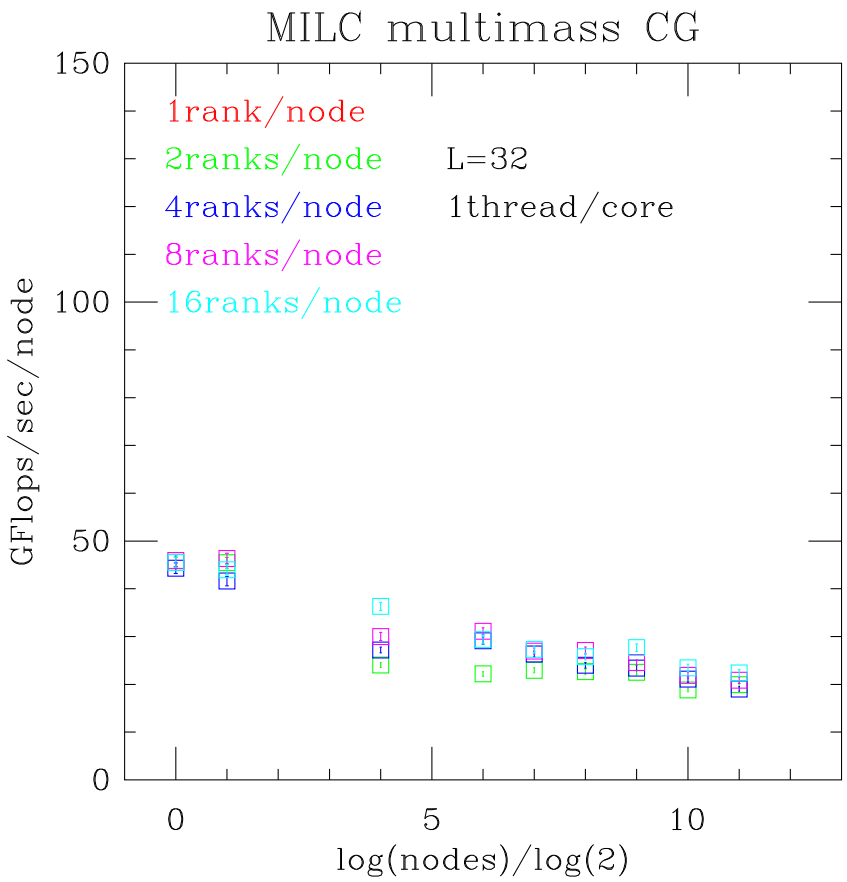
Single Node Performance for MILC and QPhiX

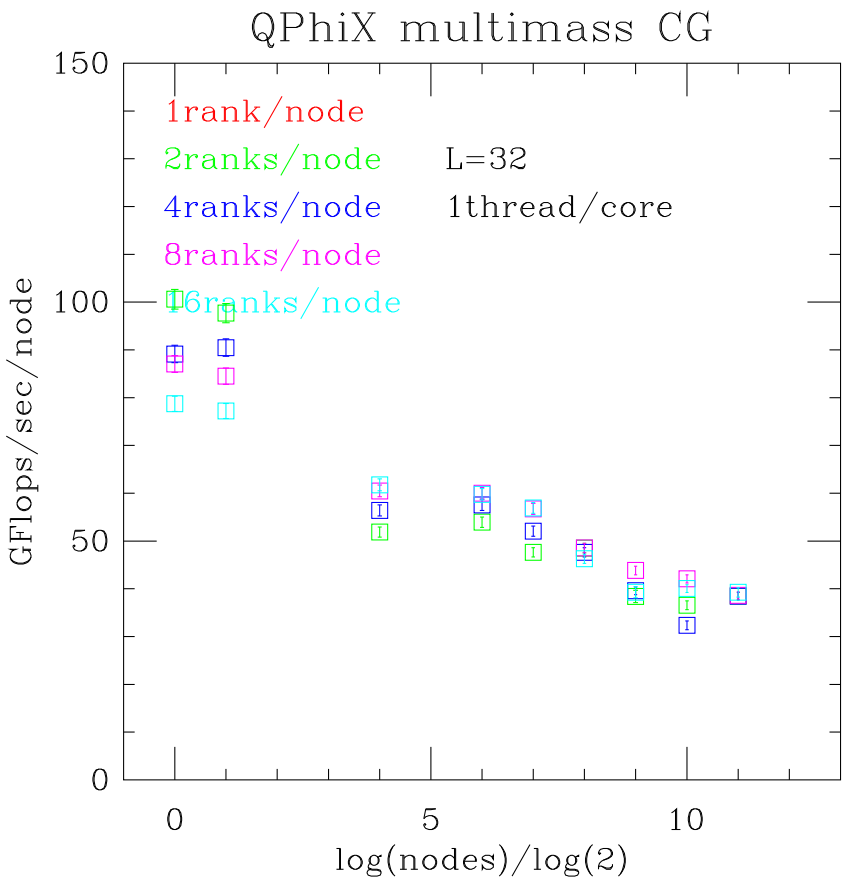




- Single node performance on Xeon Phi 7250 (somewhat old)
- QPhiX is roughly 50-100% faster than MILC
- Using all four hyper threads does not help, but 2nd one can if volume is large enough

Multinode Performance for MILC and QPhiX (Cori)





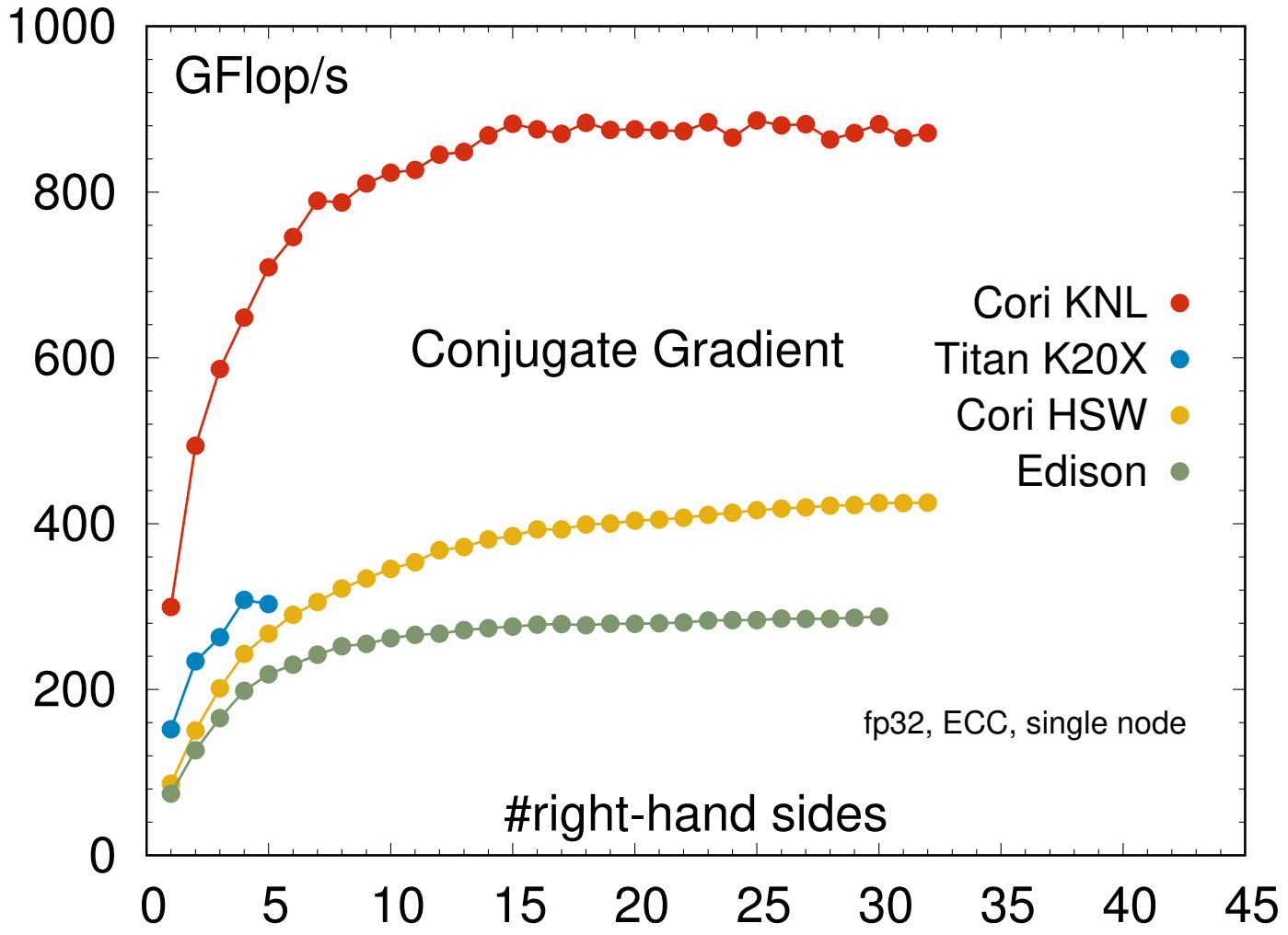
- Performance improves further with increased local volume
- QPhiX is clearly superior to MILC code by about a factor of 2
- Now let's turn to what happens with 2 threads per core

Performance for MILC on KNL

- 32^4 on Cori with QPhiX and 16 nodes sustains ~50 GFlops
- 32^4 at BNLwith QPhiX and 16 nodes sustains ~20 GFlops
- BNL KNL running in all-to-all mode.
  * Nodes not rebooted before these runs.
  * Have seen performance degradation as memory becomes fragmented
  * Does cache mode matter
- Running (now?) on JLAB KNL with 16 nodes freshly rebooted and running in quadrant cache mode
- Need to understand how to get MILC code to run as well on USQCD hardware as Cori II, TACC...
- MILC has benchmarked Grid D-slash and does not find much difference from their results with MILC/QPhiX. Early results, which could change.

Performance for Staggered Thermo on KNL

- Thermo dominately needs fast nodes, with minimal network demands
- Patrick Steinbrecher reports ~900 GFlops for their production running on KNL



Performance for contractions (JLAB)

- Running on JLAB KNL is dominately single node contractions
- Calculation is a complex matrix multiply
- Performance of ~700 GFlops/node their current production jobs. (Robert Edwards)

The followings are zgemm and other benchmarks on KNL, broadwell and K80.

Batched zgemm on KNL, broadwell and K80 for matrix size 384.
KNL: 64 threads, broadwell: 32 threads.

Batched zgemm performance in gflops for matrix size 384

$$\sum_{j=1\ldots N} M^{ij}_{\alpha\beta}\, M^{jk}_{\gamma\delta}$$

batchsize	K80	Broadwell (32 Threads)	KNL (64 Threads)
16	519	597	686
32	522	608	667
64	541	804	675
128	558	938	899
256	558	955	1131
512	559	1027	1394
1024	555	1055	1564
2048		1071	1575

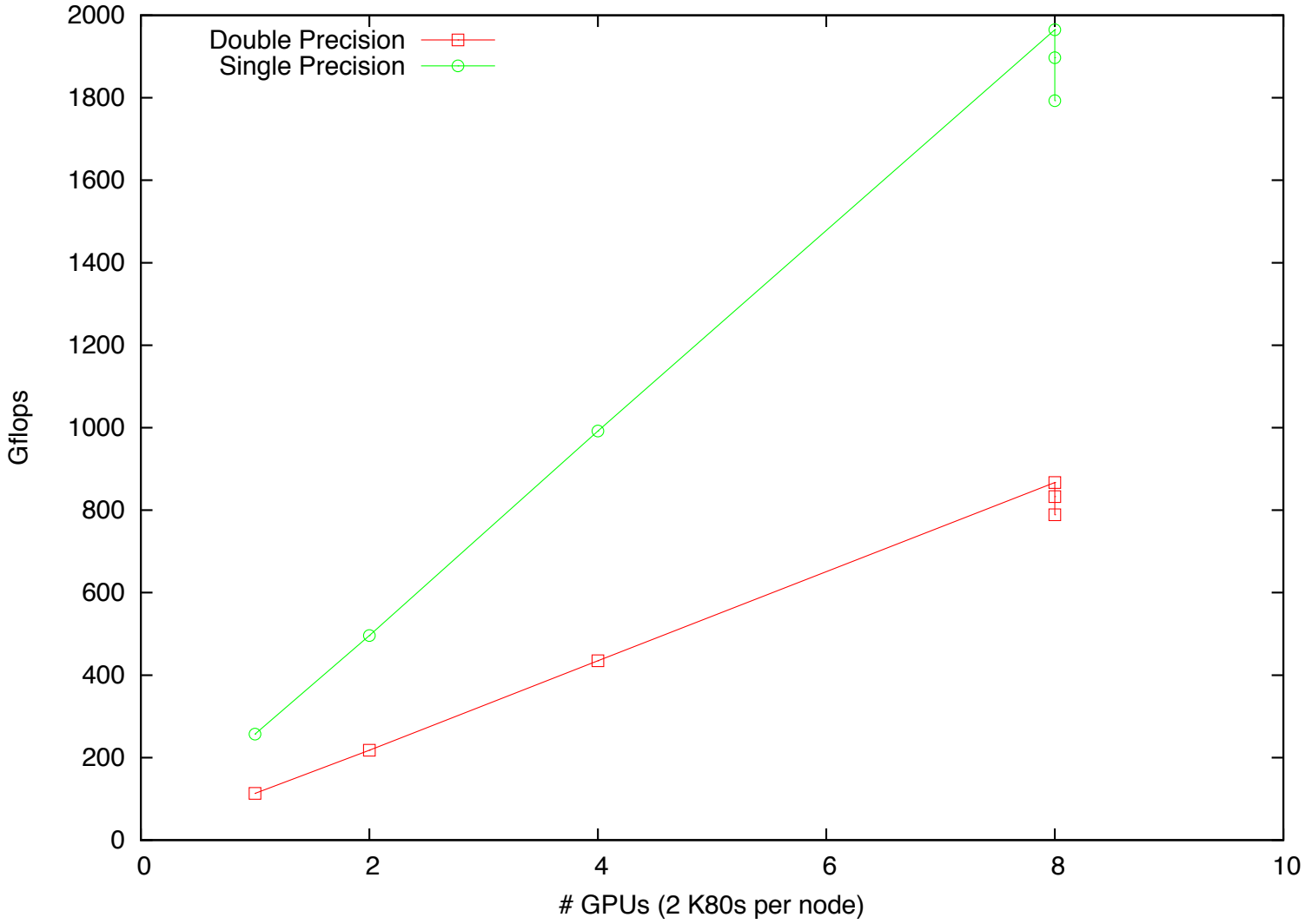
Currently using a batch size of "64" , so 675 GF

Future work:
Can increase performance by working on multiple time-slices in a "batch"

Four time-slices increases batch size to 4 * 64 -> 256, so 1130 GF

DWF Scaling on GPU's: BNL IC (K80's)

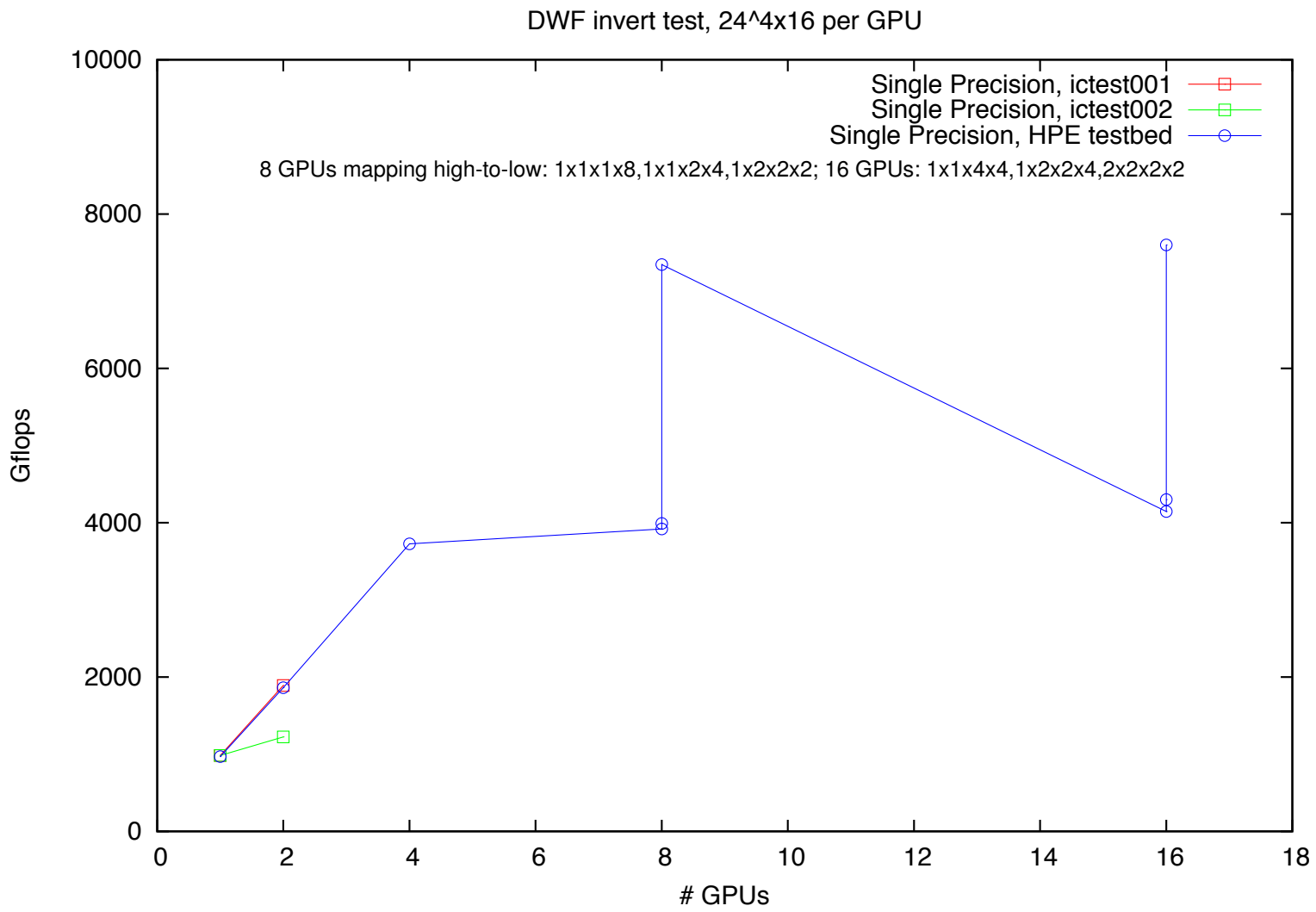
- Tests by Meifeng Lin, running Kate Clark's code, on the BNL IC
- 2 K80s per node running at 732 MHz (base clock freq 562 MHz) (equivalent to 4 GPUs per node)
- dual-socket Intel Broadwell with 36 cores total per node. (Intel(R) Xeon(R) CPU E5-2695 v4 @ 2.10GHz)
- Mellanox single-rail EDR interconnect (peak 25GB/s bidi)



Good scaling to 2 nodes. Performance around 250 GFlops/GPU single precision

DWF Scaling on GPU's: P-100 test platforms

- ictest002:
  - 2 P100 GPUs per node, w/o P2P
- HPE Pietra:
  - 4 P100 GPUs per node with NVLINK, P2P enabled 2x Intel Broadwell, 14 cores @ 2.6 GHz
  - Single-rail Mellanox infiniband EDR



- If put lattice onto nodes in the "best" way, see good scaling to 8 GPUs.
- With "best" mapping, performance per node close to 2x that for KNL, but nodes cost more. 8 P-100's could be equivalent to 16 KNL's, but would need more memory per node to have total required memory.

Decisions I

- A GPU purchase for this procurement will not be pursued
  * Without NVLink, multinode performance per GPU is comparable to KNL, but the GPUs cost substantially more.
  * With NVLink, performance is perhaps 2x KNL, making up some of the price difference, but one can only scale to around 8 GPUs. Not a large enough partition to compete with 16 to 32 node KNL.
  * Kate Clark has code that scales better to 64 GPUs on the NVIDIA DGX-1. NVLink used between 8 P-100's on a node and then quad EDR IB between nodes. Very powerful device and could be an option in the future.
- A KNL system for the job sizes and node counts we are targeting does not require more than single rail EDR IB or Omnipath.

KNL Issues

- Operational Reliability
  * Node instability: fluctuations in performance and memory degradation
  * Newest kernels look promising - need more integrated time of use
  * Memory degradation - need to reboot nodes
  * Lehner found good stability running multi-node jobs at BNL when he was essentially the only user.
  * Will have more certainty about these issues as BNL operations continue
- MPI between ranks on a node has low performance.
- Bandwidth to disk? Not yet well measured, but will know more soon.
- Performance for more generic C code
  * MILC code without OMP and QPhiX gives 30 to 50 Gflops on a single node (DP).
  * MILC code without OMP and QPhiX gives 15 to 20 GFlops on 16 nodes (DP).
  * This is ~ 1% of peak speed of 3 TFlops (DP)

USQCD Perspective?

- Demand for pi0 at FNAL remains large
- Much demand for single fast nodes (thermo, JLAB contractions)
- BNL IC gives us 40 nodes with dual K-80's for the next few years.
- Substantial need for 16 to 32 node partitions (MILC, RBC, ...)
- Is KNL a reliable successor for pi0 and a replacement for BGQ half-rack?
- Intel Skylake is available for benchmarking and for release in the fall(?)
  * 32 core devices available, including AVX-512, which KNL also has
  * Full capability Xeon cores - more flexible than the cores in KNL
  * Likely runs generic code much better
  * No MCDRAM (high-bandwidth on-chip memory), so there will be memory bandwidth limitations.
  * 1 TFlops benchmarks reported on web for SGEMM.

BNL Procurement

- Intel based system with single rail IB or Omnipath interconnect
- Nodes could be KNL, Skylake or Broadwell
- Benchmarks for evalution of options
  * 1 and 16 node performance for DWF, 24^4 local volume using Grid
  * 1 and 16 node performance of MILC code, 32^4 local volume with optimizations
  * 1 and 16 node performance of MILC code, 32^4 local volume and generic C code.
- How much to weight these in the evalution? 1/3 each for now.
- Working to release procurement in a week or two.
- Many thanks to those who ran/helped with benchmarks: Meifeng Lin, Chulwoo Jung, Christoph Lehner, Zhihua Dong, Steve Gottlieb, Frank Winter
- Also thanks to the Acquisition Advisory Committee: Steve Gottlieb, Carleton DeTar, James Osborn, Chulwoo Jung, Don Holgren, Frank Winter, Gerard Bernabeu, Chip Watson, Amitoj Singh, Robert Kennedy.
- The opinons in this report are those of the author.