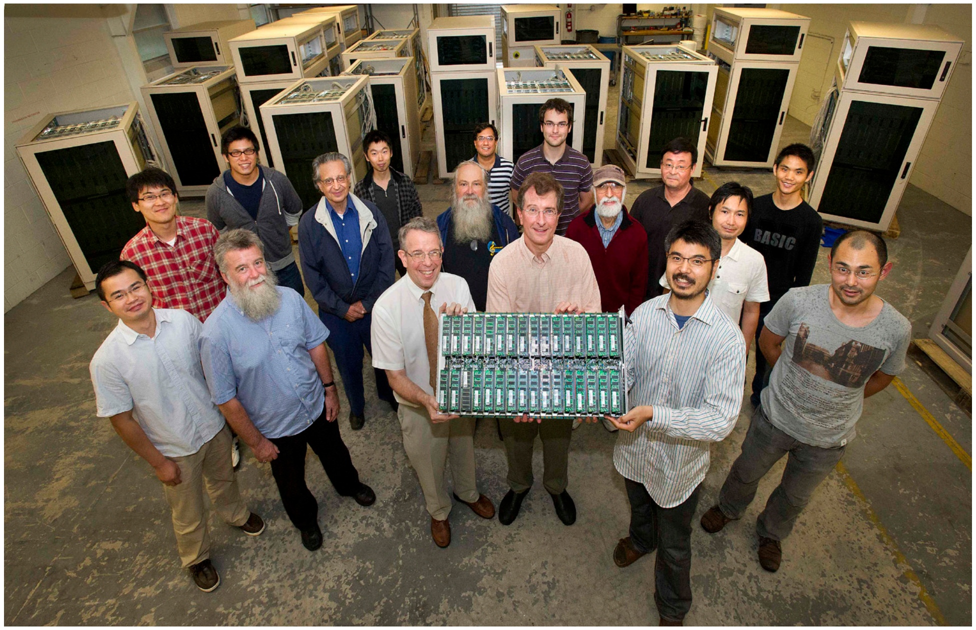
LQCD-Ext: Future Facility at BNL

DOE Annual Progress Review of
LQCD Computing Projects LQCD-Ext and LQCD-ARRA

Brookhaven National Lab
May 16, 2012

Robert Mawhinney
Columbia University

QCDOC at BNL: 2005 to 2012



LQCD tasks

- Generation of ensembles requires many cores applied to evolution
  - * Serial calculation - need one long Markov chain
  - * Need enough cores to produce ensemble in reasonable time
  - * Need to know machine is reliably doing arithmetic
- Measurements generally possible by divide-and-conquer. Some exceptions:
  - * For disconnected quark loop diagrams, gauge noise is very large
  - * Dilution/reduction/all-to-all techniques on large volumes can measure all possible fermion correlators efficiently, but can require many eigenvectors
  - * Calculation of many eigenvectors is CPU intensive and I/O times to write to disk can be prohibitive
  - * One strategy: keep in memory on large machine, calculate eigenvectors, use for a wide range of fermionic contractions and delete.
  - * Example: RBC $K \to \pi\pi$ using EigCG for disconnected diagram calculations. Storing 5d DWF propagators, requires 4 TBytes of machine memory for $32^3 \times 64 \times 32$ volumes.
- Argues for USQCD access to tightly coupled, large machines

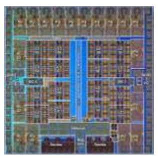
BGQ

- Each rack: 200 TFlops (peak), 1024 nodes (16k cores), 100 kW (max)
- http://www-03.ibm.com/systems/deepcomputing/solutions/bluegene/

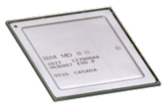
Processor	IBM PowerPC® A2 1.6 GHz, 16 cores per node
Memory	16 GB SDRAM-DDR3 per node (1333 MTps)
Networks	5D Torus—40 GBps; 2.5 μsec latency Collective network—part of the 5D Torus; collective logic operations supported Global Barrier/Interrupt—part of 5D Torus PCIe x8 Gen2 based I/O 1 GB Control Network— System Boot, Debug, Monitoring
I/O Nodes (10 GbE or InfiniBand)	16-way SMP processor; configurable in 8,16 or 32 I/O nodes per rack
Operating systems	Compute nodes—lightweight proprietary kernel
Performance	Peak performance per rack—209.7 TFlops
Power	Typical 80 kW per rack (estimated) 380-415, 480 VAC 3-phase; maximum 100 kW per rack; 4×60 amp service per rack
Cooling	90 percent water cooling (18°C - 25°C, maximum 30 GPM); 10 percent air cooling

Blue Gene/Q packaging hierarchy

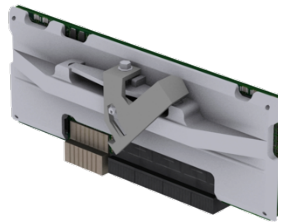
1. Chip
16 cores



2. Module
Single Chip



3. Compute Card
One single chip module,
16 GB DDR3 Memory



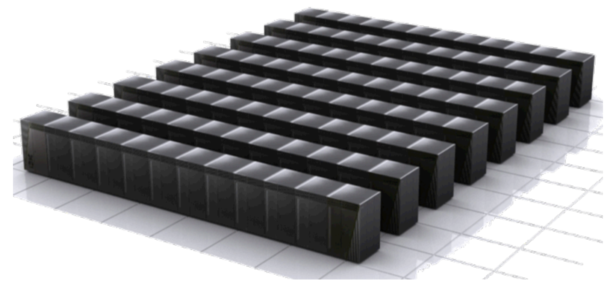
4. Node Card
32 Compute Cards,
Optical Modules, Link Chips,
Torus

5b. I/O Drawer
8 I/O Cards
8 PCIe Gen2 slots

5a. Midplane
16 Node Cards

6. Rack
2 Midplanes
1, 2 or 4 I/O Drawers

7. System
20PF/s



BGQ at Top of Green 500 List (Nov. 2011)

- BGQ built from the beginning to produce many MFlops per watt
- Reliability for very large systems important
  - * BGQ designed with error detection and correction on internode serial links, memory, and all major internal arrays and buses
  - * Extra processor core on each node assists with reporting of any errors

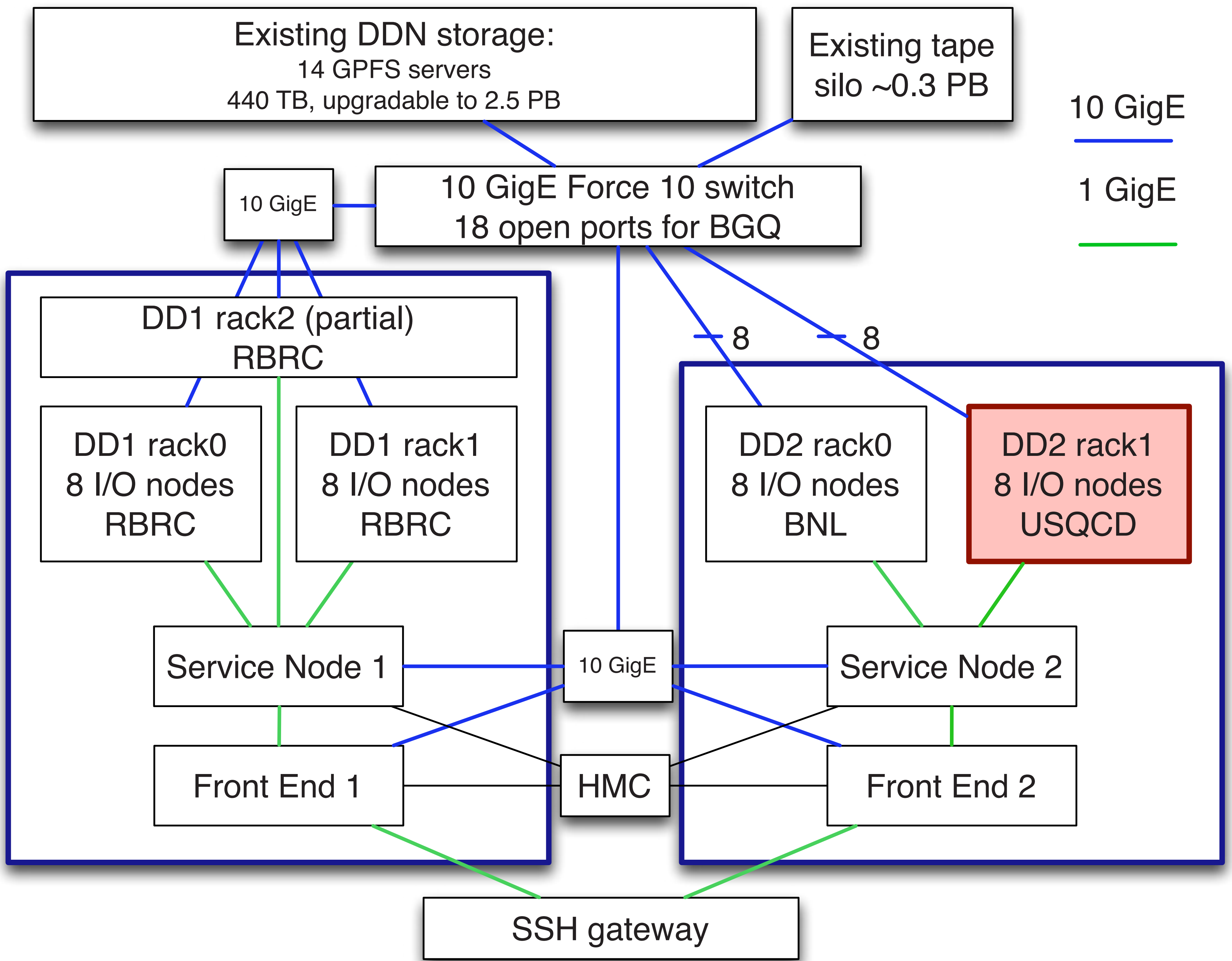
Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	2026.48	IBM - Rochester	BlueGene/Q, Power BQC 16C 1.60 GHz, Custom	85.12
2	2026.48	IBM Thomas J. Watson Research Center	BlueGene/Q, Power BQC 16C 1.60 GHz, Custom	85.12
3	1996.09	IBM - Rochester	BlueGene/Q, Power BQC 16C 1.60 GHz, Custom	170.25
4	1988.56	DOE/NNSA/LLNL	BlueGene/Q, Power BQC 16C 1.60 GHz, Custom	340.50
5	1689.86	IBM Thomas J. Watson Research Center	NNSA/SC Blue Gene/Q Prototype 1	38.67
6	1378.32	Nagasaki University	DEGIMA Cluster, Intel i5, ATI Radeon GPU, Infiniband QDR	47.05
7	1266.26	Barcelona Supercomputing Center	Bullx B505, Xeon E5649 6C 2.53GHz, Infiniband QDR, NVIDIA 2090	81.50
8	1010.11	TGCC / GENCI	Curie Hybrid Nodes - Bullx B505, Nvidia M2090, Xeon E5640 2.67 GHz, Infiniband QDR	108.80
9	963.70	Institute of Process Engineering, Chinese Academy of Sciences	Mole-8.5 Cluster, Xeon X5520 4C 2.27 GHz, Infiniband QDR, NVIDIA 2050	515.20
10	958.35	GSIC Center, Tokyo Institute of Technology	HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows	1243.80

BGQ Systems

- Sequoia at LLNL
  - 96 racks, 20 PFlops peak
- Mira at the ALCF (Argonne)
  - 48 racks, 10 PFlops peak
  - USQCD applying for INCITE time
- Julich (Germany)
  - 6 racks by June 2012
  - "considerable extension in October"
- KEK (Japan)
  - 6 racks, 3 by October 2012
- Cineca (Italy
  - 10 racks, August 2012

BGQ at BNL

- BNL currently has 3+ racks of preproduction BGQ hardware
  * 1 rack is owned by BNL
  * 2 complete racks are owned by the RIKEN-BNL Research Center (RBRC)
  * A fourth partially populated RBRC rack will be used to hold a few small BGQ partitions for code development and testing.
- 10% of the BNL owned rack is to be used by the USQCD collaboration



Existing DDN storage:
14 GPFS servers
440 TB, upgradable to 2.5 PB
Existing tape silo ~0.3 PB
10 GigE
1 GigE
10 GigE
10 GigE Force 10 switch
18 open ports for BGQ
DD1 rack2 (partial)
RBRC
8
8
DD1 rack0
8 I/O nodes
RBRC
DD1 rack1
8 I/O nodes
RBRC
DD2 rack0
8 I/O nodes
BNL
DD2 rack1
8 I/O nodes
USQCD
Service Node 1
10 GigE
Service Node 2
HMC
Front End 1
Front End 2
SSH gateway

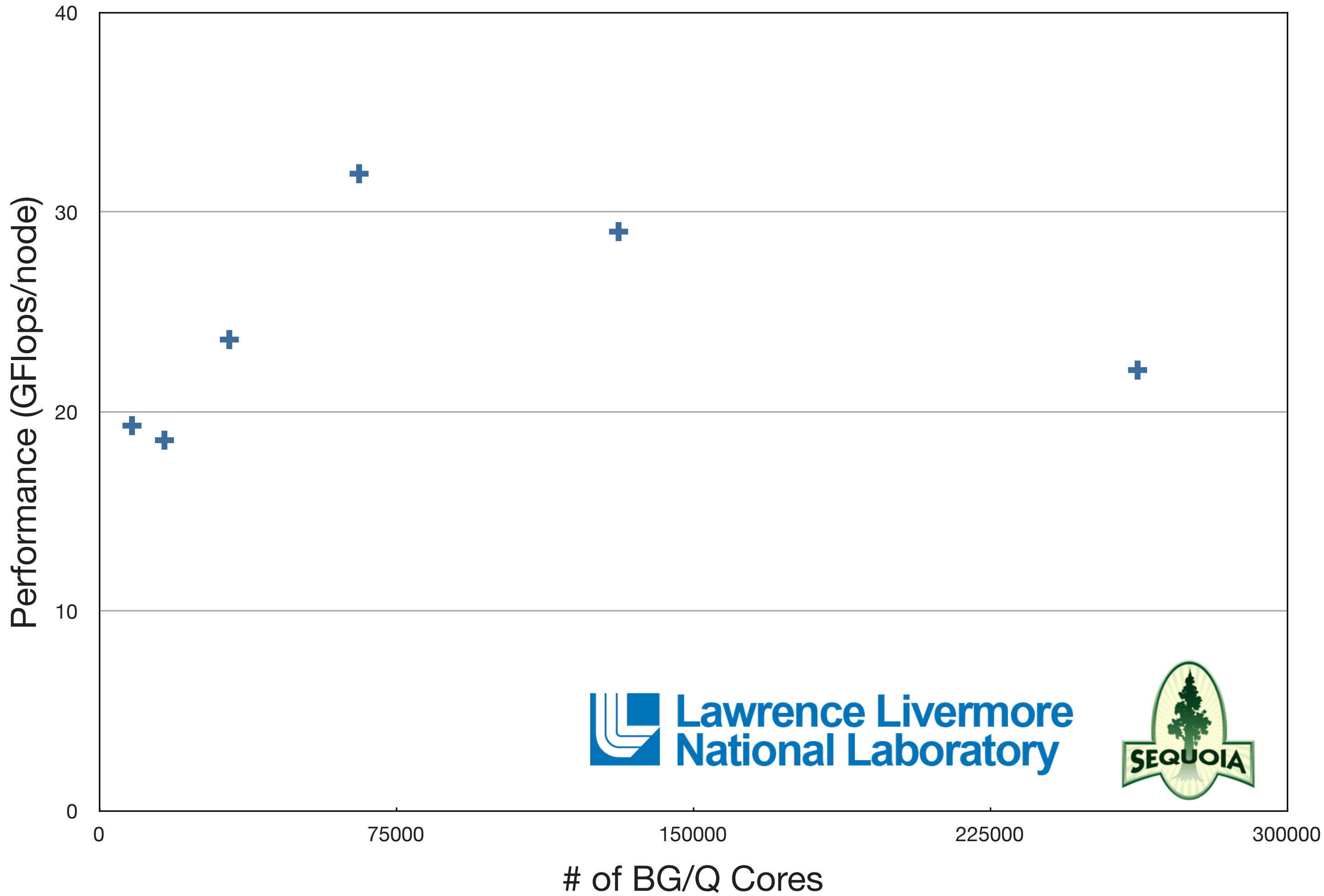
BGQ Infrastructure at BNL

- BNL has cooling and power in place for 5 BGQ racks
  * Transformer required (~ $50k) for power for a fifth rack
- Existing BNL service node, front end node and shared HMC (hardware management console) can drive another BGQ rack (~ $80k infrastructure requirement that already exists).
- Can handle up to eight 10 GigE I/O links from USQCD rack to existing DDN storage with existing infrastructure
  * Have upgraded some tiers of disks in DDN system from 250 GByte to 2 TByte
  * Existing system could go to 2.5 PBytes except some tiers do not recognize 2 TByte disks
  * Not understood why some tiers fail - personnel too busy with BGQ bringup.
- Existing tape silo can be upgraded - on demand licenses. Can increase capacity as needed, if we want to spend the money.
- BGQ project has inherited substantial infrastructure from NYBlue.

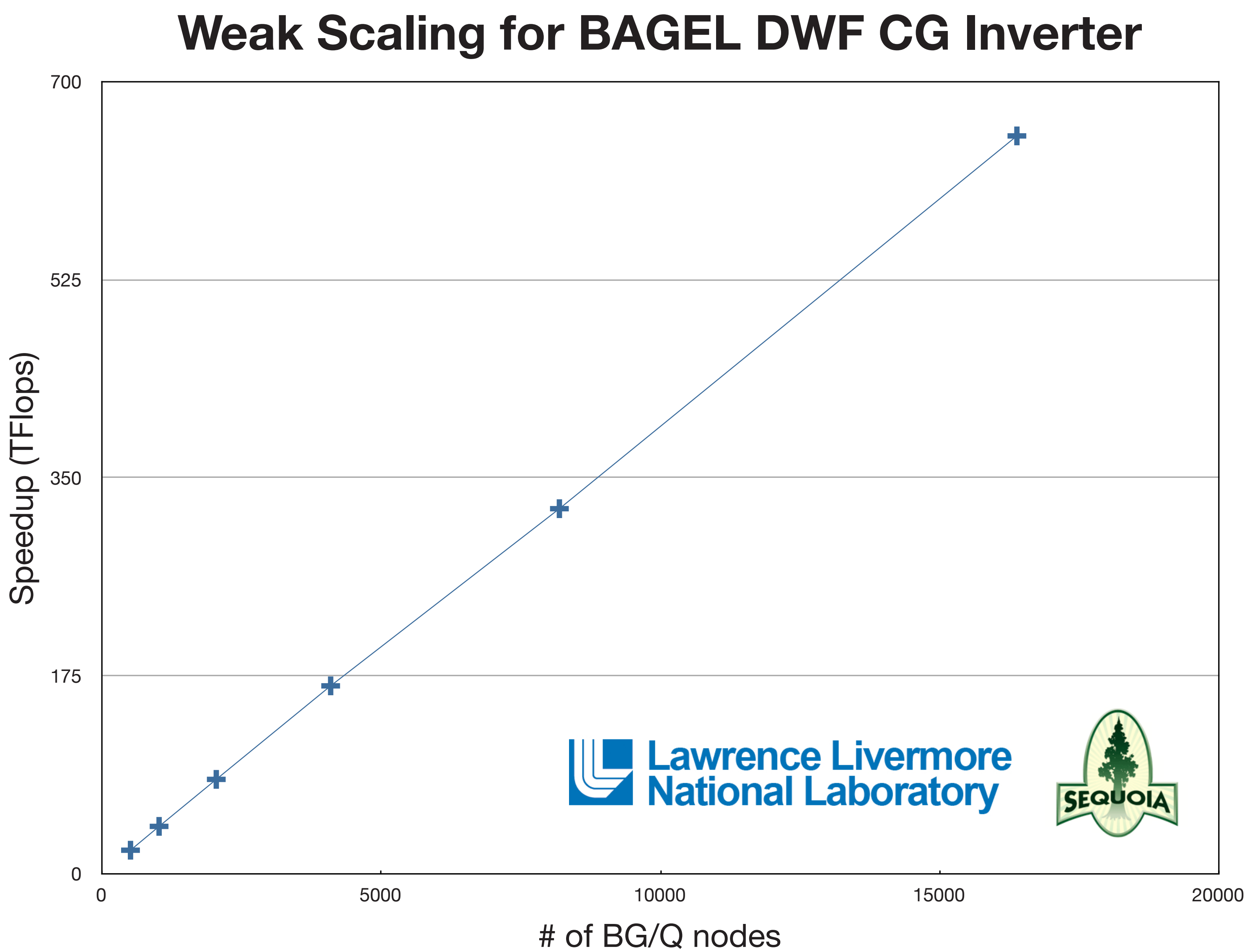
QCD Performance on BGQ

- Peter Boyle (University of Edinburgh) played a major role in designing the BGQ L1 prefetcher and his QCD code was used extensively in debugging BGQ hardware.
- Each node has 16 cores and up to 4 hardware threads per core
  - * Boyle's code, written with his code generation tool BAGEL, uses 64 threads/core
  - * In full double precision conjugate gradient solves for Domain Wall Fermions (DWF) currently sustain 42.5 GFlops per node for $8^4 \times 8$ local volumes
  - * Boyle has a single precision solver and Chulwoo Jung (BNL) is using it in a defect correction solver scheme to achieve a full double precision result. Single precision performance is currently 62.5 GFlops/node.
- Assume BGQ rack costs between $2-3 M and has performance of 62.5 GFlops/node
  - * $2M price gives 0.031 $/MFlop for DWF solver
  - * $3M price gives 0.047 $/MFlop for DWF solver
- Chroma code can readily use Boyle's assembly, with only minor addition of clover term required. Work underway (Balint Joo, JLAB)
- MILC code being optimized via SciDAC libraries. Much work to be done
- USQCD will need highly optimized codes for INCITE requests at ALCF

Strong Scaling of BAGEL DWF CG Inverter on 64^4 volume



Tests were performed with the STFC funded DiRAC facility at Edinburgh



Tests were performed with the STFC funded DiRAC facility at Edinburgh

Summary

- BNL has successfully managed QCDSP, QCDOC, BG/L, BG/P and now BG/Q
- Unique opportunity for USQCD to get up to 1 rack of BGQ for dedicated usage - about 50 TFlops of performance for optimized QCD codes
  * The RBC collaboration is thermalizing a 2+1 flavor, $48^3 \times 96 \times 32$, $(5.5\,\mathrm{fm})^3$ domain wall fermion ensemble, with physical pions, on 1 rack of BG/Q.
- This computing power comes with a powerful network for tightly-coupled parallelism
- BNL infrastructure in place to support a USQCD rack - small costs beyond rack itself
- Code optimizations well underway via USQCD SciDAC support
- Will support USQCD evolution and measurement jobs that may be too small or specialized for inclusion in an INCITE proposal.
- In 2012, USQCD has an INCITE allocation on ALCF BG/P of 50 M core-hours = 17 BG/P rack/months = 3.5% of ALCF BG/P resources
- If USQCD gets 3.5% of ALCF, 48 rack BG/Q via INCITE this is 1.7 BG/Q racks
- A 1 rack BGQ at BNL is a substantial addition to the BGQ INCITE resources, at historical levels of support for USQCD.