LQCD-ext Proposed Selection Strategy for the FY2013 Deployment

Don Holmgren

Fermilab

djholm@fnal.gov

SC LQCD-ext Annual Progress Review

Brookhaven National Laboratory

May 16-17, 2012

Outline

- Overview of LQCD-ext planned acquisitions
- FY13 hardware options
- Deployment scenarios and significant issues
- Proposed FY13 hardware selection process

Overview of SC LQCD-ext Acquisitions

Computational capacity goals by year for SC LQCD-ext:

	FY2010	FY2011	FY2012	FY2013	FY2014
Computing hardware budget (excluding storage)	$1.60M	$1.69M	$1.875M	$2.46M	$2.26M
Capacity of new cluster deployments, TFlop/s Planned/Revised/Achieved	11 / 12.5	12 / 9 / 9	24 / 10-15	44 / 15-22	57 / 22-33
Million "Fermi" GPU-Hrs/Yr Planned/Revised/Achieved	0	0 / 1.02 / 1.22	0 / 2.9-4.3	0 / 4.6-6.9	0 / 7.5-11.2

- Baseline computing hardware budgets are shown
- FY2011 original plan for 12 Tflop/s was changed to 9 Tflop/s plus a GPU-accelerated cluster with 128 nVidia "Fermi" GPUs released to production in FY2012 (152 achieved)
- FY2012-FY2014 revised goals reflect 40%-60% ranges in budget allocated to conventional and accelerated clusters
- FY2012-FY2014 GPU-Hrs/Yr figures are based on FY11-model GPUs (NVIDIA "Fermi")
  - New GPU models will deliver more than 1 "Fermi" hour per wall-clock hour

FY2013 Hardware Options

- BG/Q
- Infiniband clusters based on:
  - Intel “Sandy Bridge”/”Ivy Bridge”
  - AMD “Bulldozer”
- Accelerated clusters based on:
  - nVidia “Kepler”
  - Intel “MIC” architecture

BG/Q Details

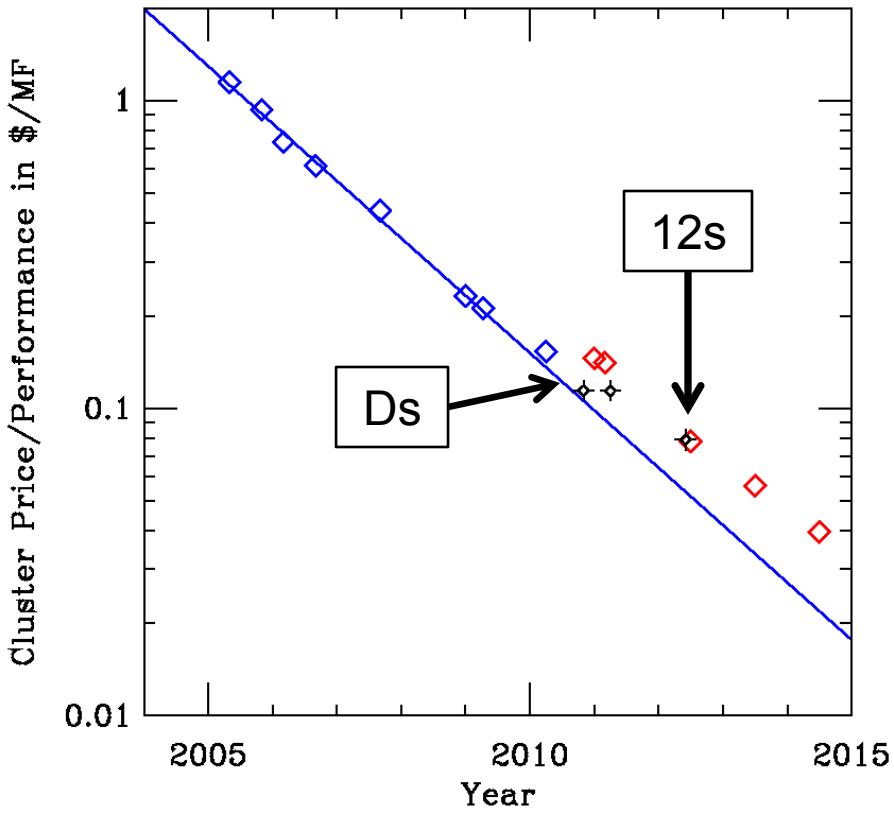
- (See prior talk)
- The following is based on available non-NDA material:
  - 16-core, 1.6 GHz CPUs, each core capable of 4 double-precision multiply-adds per cycle (SIMD, 8 flops/cycle)
  - Each rack will have 16384 cores, 209 peak TFlops/rack
  - 5D torus for communications
  - ANL “Mira” will be a 10 PFlops peak BG/Q installed in 2012
- Possible pricing and delivery schedule to USQCD are not known
  - Estimate from last year was ~ $0.05/MFlop on LQCD code (DWF only)
- LQCD performance on all actions of interest has not been measured
  - Various members of the LQCD SciDAC Software committee have access to IBM BG/Q hardware at Brookhaven, Argonne, and IBM

Details About Other Hardware

- Intel “Ivy Bridge”
  - 22nm-process update to the “Sandy Bridge” processors purchased on the JLab “12s” cluster
  - No significant architectural differences to “Sandy Bridge”
    - Lower power consumption
    - Perhaps higher clock speeds at similar price points
    - Will put downwards price pressure on “Sandy Bridge”
  - To date USQCD have not released “Sandy Bridge”-specific optimizations, such as
    - Exploitation of wide vector units (AVX)
    - Exploitation of cache-friendly alignment policies
  - By 2013 may see improved LQCD performance through optimizations in addition to any hardware performance gains
    - Quad socket motherboards will lower system prices

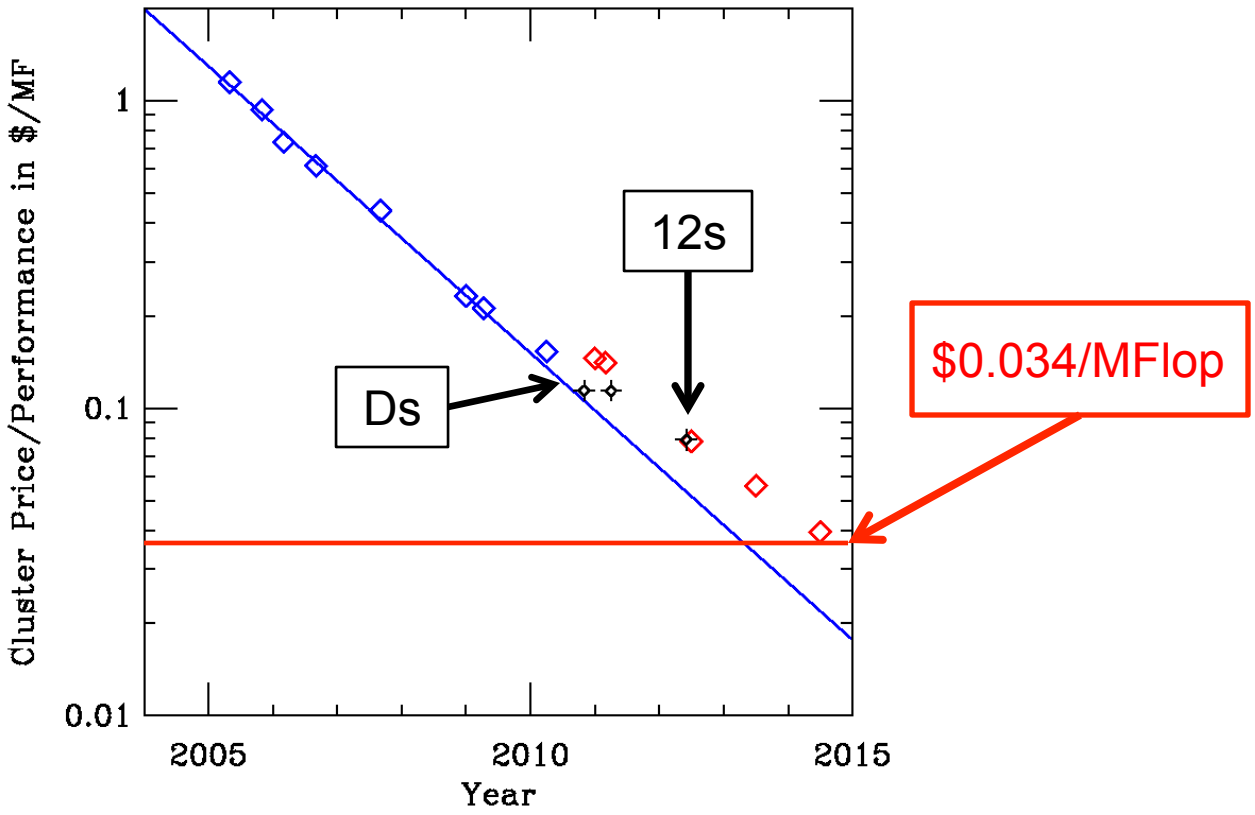
Details About Other Hardware

- AMD “Bulldozer”
  - No significant architectural update to current “Interlagos” processor has been announced
    - Previously (2011 and earlier) referenced “G2012” socket with additional memory channels was canceled for 2012
    - For LQCD, larger core count socket-G34 Interlagos chips have significantly less memory bandwidth per core compared to earlier Opterons
  - To date USQCD have not released “Interlagos”-specific optimizations, such as
    - Exploitation of wide vector units (AVX) and fused-multiply add (FMA)
    - Exploitation of cache-friendly alignment policies
  - By 2013 may see improved LQCD performance through software optimizations
    - Benchmarks using AVX-aware compiler on generic MILC asqtad code show about a 5% performance gain relative to hand-optimized SSE code



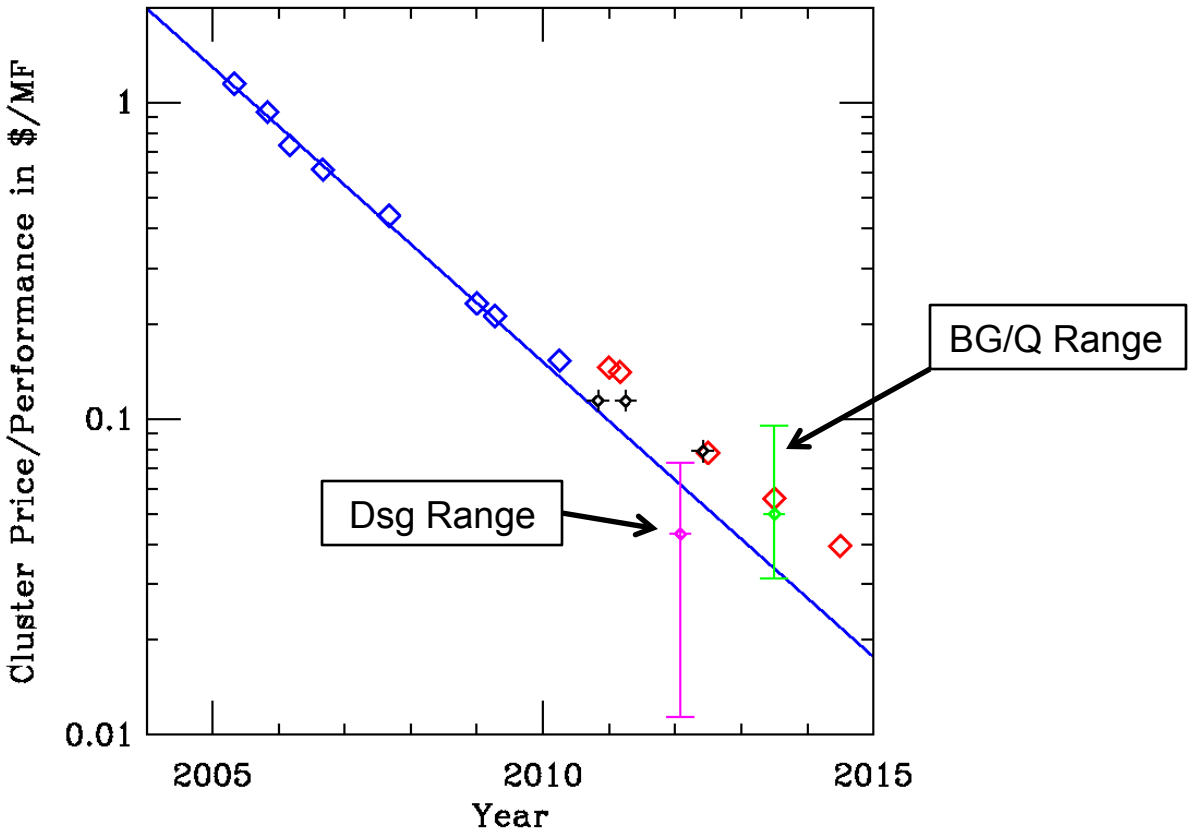
Reasons why LQCD conventional clusters are moving above the trend line:

- Latter half of Ds was later than planned because of budget delays
- Intel was at least one year late in bringing out server-class Sandy Bridge CPUs
- Quad-socket Sandy Bridge systems are still not available
- Roadmap changes at AMD have (indefinitely?) delayed the previously planned “Socket G2012” version of Bulldozer which would have more than doubled memory bandwidth per socket
- AMD changes have resulted in less price pressure on Intel for HPC hardware, and likely enabled Intel to delay Sandy Bridge and Ivy Bridge since Nehalem/Westmere were still profitable



Reaching the FY13 goal of $0.056/MFlop or the trend line of $0.034/MFlop will be challenging:

- BG/Q may achieve $0.031/MFlop for DWF (assuming $2M/rack), but will have higher price/performance for other actions
- Intel/AMD improvements in price/performance would depend on software optimizations and declining prices



With the very important caveat that neither machine type has production software for all actions of interest (but clusters do), we can add the Dsg GPU-accelerated cluster equivalent performance and estimated BG/Q performance (values from high to low price/performance):

- Dsg: MILC HISQ / Isotropic Clover / Wilson (all three are in production)
- BG/Q: 15% Peak @ $3M/Rack (MILC HISQ) / $ 0.05/MF / DWF @ $2M/rack
- Both GPU-accelerated clusters and BG/Q are suitable for calculations requiring capability machines (but GPU-accelerated clusters do not have sufficient memory for the generation of gauge configurations for the largest lattices, nor can they currently be used for DWF gauge configuration generation)

FY13 Deployment Scenarios

- BG/Q
  - BNL would be the deployment site, as the lab has prior experience with BG hardware ("NY Blue"), it operates the prototype BG/Q hardware deployed in 2012, and it has much (and perhaps all) of the specific BG infrastructure required
  - Purchase could be a full rack, or a partial rack
    - Full rack plus any necessary additional infrastructure (storage, power) might exceed FY13 hardware budget
    - Partial rack plus any necessary additional infrastructure could leave sufficient funds to also deploy a cluster at Fermilab

FY13 Deployment Scenarios

- Clusters
  - FNAL would be the deployment site, as the site has significant experience in Infiniband and GPU-accelerated clusters
    - Deployment leverages infrastructure put in place in FY08 for JPsi cluster (space, cooling, power, storage)
    - Operating clusters at both FNAL and JLab reduces risk of a major disruption shutting down an entire site's resources
  - Either a conventional or an accelerated cluster, or a mixture of both, would be chosen to best match USQCD resource needs
  - The project would carefully investigate combining the FY13 and FY14 conventional cluster purchases, and/or the FY13 and FY14 accelerated cluster purchases, across the fiscal year boundary to minimize procurement costs

Important Factors for the Decision

- BG/Q hardware could partially satisfy the USQCD need for capability computing, and could also satisfy capacity computing needs ("analysis computing")
  - Such capability computing needs will certainly be addressed in large part by Incite (ANL and ORNL) and NSF (NCSA Blue Waters) allocations
  - Project clusters have also done capability computing (gauge configuration generation) for USQCD, and this use has been demonstrated on GPUs
- Accelerated hardware would provide the best price-performance for a portion of the scientific program
  - The size of this portion is sensitive to the availability of software (software development is not in LQCD-ext project scope) and to how well the characteristics of the hardware match the computations
- Fully optimized software for all of the actions of interest to USQCD may not be available for the BG/Q by the time of the decision, or by the time of deployment
  - DWF single precision inverter sustains 30.5% of peak (~ 62.5 GF/node); current non-threaded MILC HISQ implementation sustains ~ 10% of peak and may increase to 12-15%
  - Since optimized code is strongly desired for Incite & NSF running, the community is well motivated to produce this software (GPU software as well)

Significant Issues

- BG/Q hardware has (currently) unknown cost, and performance that is well understood for only some LQCD codes
- The project submits its request to the DOE for how the next year's funds are to be dispersed in mid-August
  - We must decide by that time what fractions of the FY13 hardware budget will go to BNL and/or to FNAL

Proposed FY13 Selection Process

With the advice of the Executive Committee, the project has adopted the following timeline leading up to a funding dispersal recommendation in August and an FY13 Acquisition Plan in September:

Step	Description	Target Due Date
1	The LQCD-ext Computing Project team (i.e., "the Project") will provide the LQCD Executive Committee (EC) with data summarizing the distributions of job types and sizes during the prior year on the hardware operated by the Project (Infiniband clusters, GPU-accelerated clusters, and the QCDOC). The Project will request that the EC provide the anticipated scientific program requirements for various architectures (i.e., leadership-class machines, BG/Q rack or Infiniband cluster, and GPU-accelerated cluster). Information on USQCD hardware usage will be presented to the collaboration at the 2012 All-Hands Meeting May 5-6.	Apr 15
2	The Project will prepare the F13 Acquisition Strategy document for presentation and review at the FY2012 DOE Annual Progress Review. The Acquisition Strategy will outline the various options under consideration and the proposed process for selecting the mix of computing hardware that will be procured and deployed in FY13 using project funds.	May 16-17
3	The Project will request that the BNL site manager prepare a plan for procuring any additional BG/Q rack or half-rack and operating existing and, possibly future, BG/Q rack(s), detailing estimating hardware, storage, deployment, and operations costs.	Jun 1

4	The EC, with input from the Scientific Program Committee (SPC), will provide the Project with the anticipated scientific program requirements for various architectures (i.e., leadership-class machines, BG/Q rack or Infiniband cluster, and GPU-accelerated cluster). A helpful way of conveying this information would be for the EC to provide an estimate of the relative fractions of "analysis core-hours" and "cost-equivalent GPU-hours" needed to support the scientific program over the next 1 to 2 years. Ideally, the EC will provide the Project with anticipated needs on a per year basis for FY13 and FY14.	Jun 15
5	The BNL site manager will provide the Project with a preliminary plan for procuring and operating a BG/Q, including estimated costs and schedule.	Jul 1
6	The BNL site manager will provide the Project with a final plan for procuring and operating a BG/Q, including costs (hardware, storage, costed manpower for deployment and operations) and schedule.	Jul 22
7	The Project will review the technical landscape, conduct an alternatives analysis of the various options, and propose a cost-effective solution for the FY13 hardware deployment. When considering viable options, the Project will need to factor in the total cost of ownership (TCO) for each solution. In addition to hardware and deployment costs, TCO also includes on-going operations and support costs. Hardware costs will include any necessary storage acquisitions. For solutions involving Infiniband clusters and accelerated clusters, an operations cost model already exists. For a BG/Q option, the Project will need to understand the cost model for operating a BG/Q at BNL. Information on cost and availability of production BG/Q hardware will also be needed. Results of the analysis and an overview of the proposed solution will be summarized in the Alternatives Analysis document. The Project will verify the host laboratory's ability and willingness to provide the necessary space, power, and cooling for each alternative.	Jul 29

8	The EC will review he Alternatives Analysis document and proposed FY13 hardware solution, and will provide advice on how to proceed to the Project Manager.	Aug 10
9	The Project will analyze the advice of the Executive Committee as well as any new data that might have been obtained, and will produce the final plan for the FY13 hardware deployment. The Project Manager will advise the EC, the host laboratories, the Federal Project Director, and Project Monitor of the planned FY13 hardware acquisition.	Aug 15
10	The Project Manager will revise the project budget as necessary to accommodate the FY13 hardware solution. Depending on the alternative selected, changes may be required in the planned allocation of funds across the three host laboratories.	Aug 20 (est.)
11	The Project Manager will provide the Federal Project Director with the FY13 Financial Plan, containing the requested distribution of project funds to the three host laboratories.	Aug 20 (est.)
12	The Project will develop a detailed acquisition plan, with timeline, based on the approved FY13 architecture solution.	Sep 30, 2012
13	The Project will execute the FY13 acquisition plan in a manner that meets approved performance goals and milestones.	Sep 30, 2013

FY13 Schedule After Acquisition Plan

- The budget lesson from FY11 (and to a lesser extent, from FY12):
  - We must plan on budget chaos in FY13
  - We may not be able to commit all funds until late fiscal Q3
- Consequence:
  - Release new hardware to production by Sept 30, rather than by our typical planned June 30 date (but work to release by June 30 if budget allows)
    - Note that use of combined FY13/FY14 procurements would also push release to production of the FY13 hardware to late in the fiscal year

FY13 BG/Q Schedule

- Preliminary schedule:
  - Benchmarking and code development now through July to determine performance of the major actions
  - Interact with IBM through July to obtain pricing ranges and possible delivery timeline
  - Once FY13 funds are available, initiate RFQ process
  - System release to production as soon as delivery, integration, and acceptance tests are completed
  - If DOE funding and IBM production timeframes cooperate, BG/Q hardware could be released to production well ahead of the project target date of June 30

FY13 Cluster Schedule After Acquisition Plan

- Preliminary cluster schedule:
  - Benchmarking in summer/fall to measure performance of AMD "Bulldozer", Intel "Sandy Bridge" and "Ivy Bridge"
  - For the accelerated cluster, benchmarking in fall of NVIDIA Kepler GPUs and, if available, Intel "MIC" accelerators
  - Release Request for Information to prospective vendors by Oct 31, covering clusters and accelerators, in time to generate discussions at SC'12 conference
  - Assess USQCD requirements and determine split between accelerated cluster and conventional cluster by Dec 31
  - If budget allows, issue RFP(s) by Jan 31

Summary

- BG/Q, Infiniband clusters, and accelerated clusters could each fulfill USQCD needs in FY13
- We have initiated a process to determine by August the dispersal of hardware funds among the labs, with the detailed FY13 acquisition plan to follow
- FY13 and FY14 acquisitions will be guided by balancing the portfolio of hardware and external allocations against USQCD scientific needs
  - The project will work with the community through the Executive Committee to determine the fractions of work that can or must be done by each of the three classes of hardware