LQCD-ext Proposed Selection Strategy for the FY2012 Deployment

Don Holmgren

Fermilab

djholm@fnal.gov

SC LQCD-ext Annual Progress Review

Fermi National Accelerator Laboratory

May 10-11, 2011

Outline

- Overview of LQCD-ext planned acquisitions
- FY12 hardware options
- Deployment scenarios and significant issues
- FY12 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 (not including storage)	$1.60M	$1.69M	$1.875M	$2.46M	$2.26M
Planned/Achieved Capacity of new cluster deployments, Tflop/s	11 / 12.5	9	24*	44*	57*
Planned GPU Deployment Count	–	128	**	**	**

- FY2011 original plan for 12 Tflop/s was changed to 9 Tflop/s plus a GPU-accelerated cluster with 128 NVIDIA “Fermi” GPUs

* FY2012-FY2014 Tflop/s cluster capacities will likely be reduced with some of the budget shifted to GPU-accelerated clusters

** FY2012-FY2014 GPU deployment counts TBD

FY2012 Hardware Options

- BG/Q
- Infiniband clusters based on:
  - Intel “Sandy Bridge”/”Ivy Bridge”
  - AMD “Bulldozer”
- GPU-accelerated clusters based on:
  - NVIDIA “Kepler”
  - AMD (ATI) Radeon HD7000M series

BG/Q Details

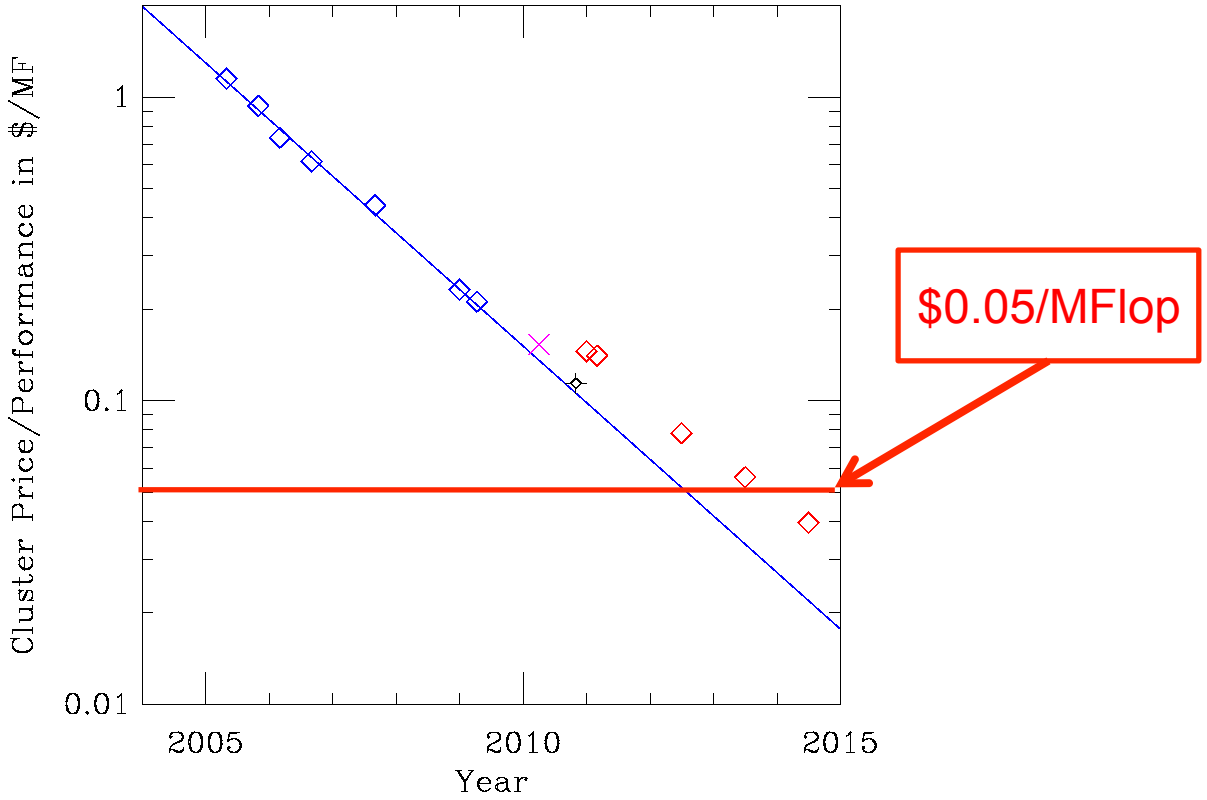
- The following is based on non-NDA material:
  - 16-core, 1.6 GHz CPUs, each core capable of 4 double-precision multiply-adds per cycle (SIMD, 8 flops/cycle)
  - Each rack will have 16384 cores, 209 peak TFlops/rack
  - 5D torus for communications
  - Test system (half rack) scored 65.3 TFlops on HPL Linpack for the November Top500 list; full “Ds” cluster scored 43.1 TFlops
  - ANL “Mira” will be a 20 PFlops peak BG/Q installed in 2012
- Possible pricing and delivery schedule to USQCD are not known
  - Suggestion from last year was ~ $0.05/MFlop on LQCD code
- LQCD performance on all actions of interest has not been measured
  - Some members of SciDAC Software committee will have access to IBM BG/Q hardware this summer for optimization and benchmarking

Details About Other Hardware

- Intel “Sandy Bridge”
  - Dual-socket motherboards by Q1 FY12, quad-socket systems somewhat later
  - Up to 8 cores/socket, up to 4 channels of DDR3, compared to 6 cores/socket and 3 channels on current CPUs
  - Wider SIMD unit (“AVX”), and a rumor that non-aligned memory access penalties will no longer occur
  - Direct PCIe gen3 access from 2nd generation CPU (“Ivy Bridge”), necessary to fully exploit “Kepler” GPUs and FDR Infiniband)

Details About Other Hardware

- AMD “Bulldozer”
  - Socket-compatible with existing “Magny-Cours” motherboards, such as those on Fermilab “Ds” cluster
  - Up to 16 cores/socket, 4 channels of DDR3 at up to 1600 MHz, compared to 8 cores/socket and 4 channels at 1333 MHz on current CPUs
  - Effectively wider SIMD unit (“AVX”) and fused multiply-add
  - Higher clock speeds (3+ GHz) compared to current processors (“Ds” uses 2 GHz processors)



BG, Intel, and AMD hardware all have improvements on their roadmaps that suggest $0.05/MFlop might be achieved in FY12

- Intel and AMD both improve memory bandwidth and floating point capabilities
- BG/Q will have 15X the peak floating point per CPU as BG/P
- This is an aggressive goal for clusters

FY12 Deployment Scenarios

- BG/Q
  - BNL would be the deployment site, as the lab has prior experience with BG hardware ("NY Blue") and will also house prototype BG/Q hardware late in calendar 2011
  - Purchase could be of a full rack, or a half rack
    - Full rack plus necessary infrastructure (head nodes, parallel file system) might exceed FY12 hardware budget
  - Anticipate delivery no sooner than near the end of FY12
  - Deployment would require reassessment of and changes to the project's operations budgets in FY12 and subsequent years

FY12 Deployment Scenarios

- Clusters
  - JLab would be the deployment site, as the site has significant experience in Infiniband and GPU-accelerated clusters
    - Deployment leverages infrastructure put in place for ARRA clusters (cooling, power, storage)
    - Operating clusters at only FNAL creates the risk of a major disruption shutting down the entirety of LQCD-ext cluster resources, and it would also diminish the effectiveness of ARRA resources
    - Operating clusters at both sites leverages other infrastructure at each lab (*e.g.* robotic tape storage, security)
  - Either a conventional or a GPU-accelerated cluster, or a mixture of both, would be chosen to best match USQCD resource needs

Important Factors for the Decision

- BG/Q hardware would partially satisfy the USQCD need for capability computing, and could also satisfy capacity computing needs
  - However, capability computing needs will be addressed in large part by INCITE time (ANL and ORNL) and NSF time (NCSA Blue Waters)
- GPU-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 project scope), and to how well the characteristics of the hardware match the computations
- Depending on the costs of BG/Q infrastructure, conventional cluster hardware could provide the best price-performance for general purpose computing
  - Some capability computing has always been done on USQCD cluster hardware

Significant Issues

- BG/Q hardware will be:
  - Available only very late, if at all, in FY12,
  - Currently of unknown cost and performance
  - Available only in half-rack or full-rack size, so fitting to a single year's budget may be difficult
  - Optimized software for all actions might not be available at the time of deployment
- NVIDIA "Kepler" GPU may become available only late in FY12
- In a typical year the project would present the acquisition plan for FY12 at this review
  - The potential significant impact of a BNL BG/Q has lead us to defer a decision until August
  - We are presenting our strategy that leads to a detailed acquisition plan in September

FY12 Hardware Selection Process

- With the advice of the Executive Committee, the project has adopted a process to determine by August the breakdown in the FY12 deployment between BG/Q hardware, and cluster (conventional and accelerated) hardware
- The project will request FY12 funding dispersal between BNL, FNAL, and JLab according to this breakdown
- If we were to learn prior to August that the BG/Q will not be available for delivery in time, the project will short-circuit this process and proceed with planning for cluster deployment at JLab

FY 2012 Hardware Selection Process

Step	Description	Due Date
1.	Gather information on future computing needs from experience of previous 12 months (e.g., distribution of job types and sizes)	May 5-6 (done)
2.	Prepare Acquisition Strategy for annual progress review outlining the options for the (mix of) computing hardware to be procured in FY12	May 10-11 (done)
3.	Obtain estimate of the relative fractions of "analysis core-hours" and "GPU-hours" needed for the science program over the next 1-3 years	June 15
4.	BNL Site Manager will provide a with a final plan for procuring and operating a BG/Q, including costs (hardware, storage, costed manpower for deployment and operations) and schedule.	July 22
5.	Create draft Alternatives Analysis document on various options: Infiniband cluster, GPU cluster, BlueGene/Q, or some combination	July 31
6.	Executive Committee reviews Alternatives Analysis	Aug 10
7.	Final update of AA and Procurement Plan (type of machine, site)	Aug 20
8.	Update FY12 budget plan and allocations to host institutions	Aug 20
9.	Prepare detailed Acquisition Plan (IB cluster / GPU cluster split can be determined later)	Sept 30

FY12 Schedule After Acquisition Plan

- The BG/Q schedule would be determined as part of the hardware selection process
- Preliminary cluster schedule:
  - Benchmarking in summer/fall to measure performance of AMD “Bulldozer” and Intel “Sandy Bridge” (and/or “Ivy Bridge”)
  - Release Request for Information to prospective vendors by Oct 31, covering clusters and GPUs, in time to generate discussions at SC’11 conference
  - Assess USQCD requirements and determine split between GPU-accelerated cluster and conventional cluster by Dec 31
  - If budget allows, issue RFP by Jan 31

FY12 Schedule After Acquisition Plan

- The budget lesson from FY11:
  - We should plan on similar budget chaos in FY12
  - We may not be able to obligate all funds until late fiscal Q3
- Release new hardware to production by Sept 30, rather than by our typical June 30 date (although for cluster and/or GPU hardware, work to release by June 30 if budget allows)

Summary

- BG/Q, Infiniband clusters, and GPU clusters could each fulfill needs in FY12
- We have initiated a process to determine by August the breakdown by hardware type, and the corresponding dispersal of hardware funds among the labs, with the detailed FY12 acquisition plan to follow
- FY12 and later acquisitions will be guided by balancing the portfolio of hardware and external resources 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

Questions?

Backups

Proposed FY12 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 FY12 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 2011 All-Hands Meeting May 5-6.	Apr 15
2	The Project will prepare the F12 Acquisition Strategy document for presentation and review at the FY2011 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 FY12 using project funds.	May 10-11
3	The Project will request that the BNL site manager prepare a plan for procuring and operating a BG/Q rack, 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 FY12 and FY13.	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 FY12 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 GPU-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

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 FY12 hardware deployment. The Project Manager will advise the EC, the host laboratories, the Federal Project Director, and Project Monitor of the planned FY12 hardware acquisition.	Aug 15
10	The Project Manager will revise the project budget as necessary to accommodate the FY12 hardware solution. Depending on the alternative selected, changes may be required in the planned allocation of funds across the three host laboratories.	Aug 20
11	The Project Manager will provide the Federal Project Director with the FY12 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 FY12 architecture solution.	Sep 30, 2011
13	The Project will execute the FY12 acquisition plan in a manner that meets approved performance goals and milestones.	Sep 30, 2012