LQCD-ext II CR16-01 Calculation of BNL IC Allocation

Version 1.0

Robert D. Kennedy, ACPM for the LQCD Computing Project

This document describes how the BNL IC Allocation in CR16-01 is calculated.

Summary

CR16-01 is based on the premise that sufficient BNL IC nodes will be allocated to LQCD to cover the reduced Delivered Computing, a critical project KPI, due to supporting clusters at a third site, taking BNL in-kind contributions into account. The difference in funds available for computing equipment acquisition is estimated using the Cost Forecasts for the 2-Sites model and the 3-Sites FY-Straddle model. The ability of BNL IC nodes to deliver computing with USQCD applications is estimated based on experience with the similar technology in the LQCD Pi0g cluster. In the Performance Forecast workbook, the costs and performance estimate are combined in a time profile for Delivered Computing. Based on these calculations, the number of BNL IC nodes needed to make up the difference in Delivered Computing between the 2-Sites and 3-Sites FY Straddle operating models is determined to be about 40 nodes.

Cost Forecasts

The preparation of the Cost Forecasts is described in the CR16-01 formal document. Note that all in-kind BNL contributions were taken into account to offset the difference in funding available for computing equipment acquisitions between the 2-Sites and 3-Sites FY-Straddle models. The models also include some changes made in consultation with the BNL Site Manager, such as:

- Both Models: Eliminate the IBM maintenance for the BG/Q system in FY17. Dedicate 31k of the funds intended for IBM maintenance to be spent on BG/Q parts as necessary. The BG/Q would be run opportunistically in FY17. If it fails in a way that cannot be repaired within budget, then the affected portion would be turned off. Plan to retire the BG/Q in FY18.
- 3-Sites only: Add funds to the BNL travel budget in project out-years. This funding had been excluded from the budget after the planned retirement of the BG/Q in FY17.

Since there is a deep interplay between equipment costs and subsequent support staff costs which is treated in detail in the "Staffing Model" sheet of the Cost Forecasts, the year-by-year difference in acquisition funding may be misleading. The Performance Forecast evaluates how this funding difference impacts the Delivered Computing KPI for the project. The CR16-01 document package includes the Cost and Performance Forecasts in PDF format.

BNL IC Rating Estimate

The project estimated the rating of the BNL IC nodes using experience on the similar Pi0g cluster. From Don Holmgren, FNAL Site Architect:

> "The effective TF basis used at CD-2/3 was 157 GF per K20 GPU, and 204 GF per K40 GPU. A K80 GPU is roughly a pair of K40 GPUs connected by a PLX PCIe switch. The actual ASIC (GK210) is newer than the GK110 on the K40. There are fewer active cores per ASIC (13 SMs instead of a K40's 15 SMs) and also lower memory clocks (240 GB/s vs 288 GB/s), but there is twice as much thread storage per ASIC (registers and shared

> memory). For some codes, we expect a K80 to have more than twice as much throughput as a K40, and for others, less than twice as much.
>
> “The BNL IC host nodes will have as much host memory as the Fermilab Pi0g quad-K40 nodes, but they will have EDR Infiniband rather than QDR (three times the data bandwidth and less than half the zero-length message latency) and they will have Xeon host processors that are two generations newer ("Broadwell" E5-26xx-v4 vs. "Ivy Bridge" E5-26xx-v2) with higher host memory bandwidth (DDR4-2400 vs. DDR3-1866). Altogether, a dual-K80 BNL IC node should have better throughput than a quad-K40 FNAL Pi0g node, with the exact increase TBD through benchmarking. For the purposes of estimating node hours, I propose a 10% boost to be revised once throughput measurements can be performed. So, in terms of the CD-2/3 deployment milestones, we would initially rate a K80 as 2 * 1.10 * 204 GF = 450 GF (rounding up to the nearest 10 TF), and thus a BNL IC node at 900 GF.”

Therefore, we used 900 GFlop/s-year for the estimated performance rating of the BNL IC nodes and included a condition to revise the allocation should this prove incorrect by more than 10% in either direction.

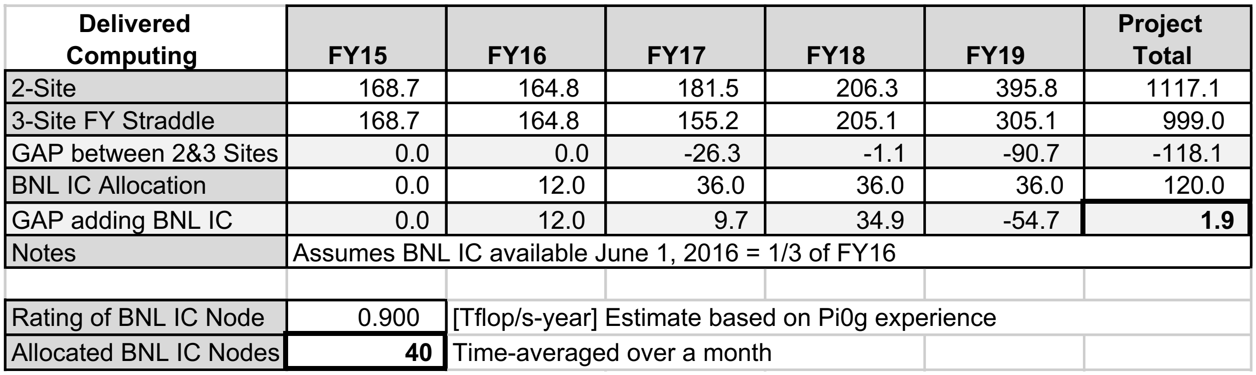
Performance Forecast

The translation of acquisition funding and operations timeline into Deployed Computing and Delivered Computing is performed in the Performance Forecast. The current workbook includes tables for each of the models for which we produced Cost Forecasts, as well as a table for the 3-Sites FY-Straddle model which includes the BNL IC allocation. We assumed the BNL IC allocation would become available in June 2016 (though we now know this is unlikely to be the case) and continue through the lifetime of the project. Since the project is not operating the cluster, but rather receiving an allocation of computing cycles, we used 100% for the allocation’s “uptime.” We have chosen not to fine-tune the timing of the availability of the BNL IC allocation for availability delays through summer 2016 since such changes are proportionately small compared to the uncertainty in the BNL IC node rating. However, the project does expect the full BNL IC allocation to be available by September 1, 2016, as the BNL IC capacity has been factored into this year’s scientific computing allocation process.

The acquisition funding profile and the BNL IC node rating are entered into the Performance Forecast in the table “3-Site FY Straddle with BNL IC” (bottom of worksheet). The number of nodes (cell E195) was varied to find the smallest integral number which produces a zero or negative difference (neutral or good for the project) in Delivered Computing between 2-Sites and 3-Sites FY-Straddle (cell U193). This value is 40. Given the uncertainty in the BNL IC node rating, we have chosen to state this as “about 40” BNL IC nodes until we have some operational experience on those nodes. We have agreed to make this allocation threshold time averaged over a month to allow large programs to be runnable by LQCD and other BNL IC users. This process is summarized in Table 1 below.

The CR16-01 document package includes the Performance Forecast in PDF format.

Table 1: Addressing the Gap in Delivered Computing with the BNL IC Allocation



Delivered Computing	FY15	FY16	FY17	FY18	FY19	Project Total
2-Site	168.7	164.8	181.5	206.3	395.8	1117.1
3-Site FY Straddle	168.7	164.8	155.2	205.1	305.1	999.0
GAP between 2&3 Sites	0.0	0.0	-26.3	-1.1	-90.7	-118.1
BNL IC Allocation	0.0	12.0	36.0	36.0	36.0	120.0
GAP adding BNL IC	0.0	12.0	9.7	34.9	-54.7	**1.9**
Notes	Assumes BNL IC available June 1, 2016 = 1/3 of FY16					
Rating of BNL IC Node	0.900	[Tflop/s-year] Estimate based on Pi0g experience				
Allocated BNL IC Nodes	**40**	Time-averaged over a month				

Conclusion

A BNL IC allocation of about 40 nodes will offset the additional costs of adding a third cluster hosting site to the LQCD Computing Project, taking BNL in-kind contributions into account as described by CR16-01. If the performance rating of the BNL IC nodes varies by more than 10% in either direction, we have agreed to revise this allocation level.