FY17 Alternatives Analysis
for the
Lattice QCD Computing Project Extension II
(LQCD-ext II)

Operated at
Brookhaven National Laboratory
Fermi National Accelerator Laboratory
Thomas Jefferson National Accelerator Facility

for the
U.S. Department of Energy
Office of Science
Offices of High Energy and Nuclear Physics

Version 0.3

Revision Date
May 12, 2017

PREPARED BY:
Bob Mawhinney, Columbia University
Alex Zaytsev, Brookhaven National Laboratory

CONCURRENCE:

_____ _____
William N. Boroski Date
LQCD-ext Contract Project Manager

Lattice QCD Computing Project Extension II (LQCD-ext II)
Change Log: Alternatives Analysis for FY17 Procurement

Revision No.	Description	Effective Date
0.1	Document created from FY16 document.	April 6, 2017
0.2	Reworked sections 1 to 3	May 5, 2017
0.3	Included comments from May 5 Acquisition Committee meeting. Updated sections 4 and 5	May 12, 2017

Table of Contents

Table of Contents

1 Introduction 1

2 FY17 Goals 1

3 Hardware Options 3

4 Alternatives 9

4.1 Alternative 1: A Xeon Phi / KNL cluster released to production by Sept 30, 2017. 9

4.2 Alternative 2: A conventional x86 cluster released to production by Sept 30, 2017 9

5 Discussion 10

6 Conclusion 10

1 Introduction

This document presents the analysis of FY 2017 alternatives for obtaining the computational capacity needed for the US Lattice QCD effort within High Energy Physics (HEP) and Nuclear Physics (NP) by the SC Lattice QCD Computing Project Extension II (LQCD-ext II). This analysis is updated at least annually to capture decisions taken during the life of the project, and to examine options for the next year. The technical managers of the project are also continuously tracking market developments through interactions with computer and chip vendors, through trade journals and online resources, and through computing conferences. This tracking allows unexpected changes to be incorporated into the project execution in a timely fashion.

Alternatives herein are constrained to approximately fit within the current budget guidance of the project for computing acquisitions at BNL in FY 2017 and FY 2018:

- $0.75M for computing procurements in FY 2017, and
- ~ $0.58M for computing procurements in FY 2018, based on options established on the procurement in FY 2017.

This constraint provides adequate funding to meet the basic requirements of the field for enhanced computational capacity, under the assumption of expanding resources at ANL and ORNL already planned by the Office of Science (SC), and under the assumption that a reasonable fraction of those resources is ultimately allocated to Lattice QCD.

All alternatives assume the continued operation of the existing resources from the FY 2009-FY 2013 LQCD Computing Project until those resources reach end-of-life, i.e., until each resource is no longer cost effective to operate, typically about 5 years.

The hardware options discussed in this document for FY 2017 are: a conventional CPU cluster, a GPU-accelerated cluster, a Xeon Phi Knights Landing (KNL) cluster, or some combination of these. The interconnect options are either Infiniband or Intel's Omnipath network. Conventional clusters can run codes for all actions of interest to USQCD. Optimized multi-GPU codes for solving the Dirac equation are available for HISQ, Wilson, clover, twisted mass, and domain wall fermions, using conventional Krylov space solvers. Recently, GPU-based implementations of multigrid Dirac solvers for clover fermions have been completed. For KNL with Wilson, clover and HISQ fermions, optimized inverter software is available and incorporates JLab's QPhiX code generator. Also for KNL, the Grid software package (Boyle and collaborators from the UK) has highly tuned solvers for domain wall fermions, as well as various types of Wilson and staggered fermions. Unlike GPU clusters, however, KNL clusters can run all codes for all actions of interest to USQCD, though un-optimized codes will not run nearly as efficiently as optimized codes.

2 FY17 Goals

The project baseline calls for deployment in FY 2017 of 45 Teraflops per second (TF) of sustained performance, based upon extrapolations of price performance of Intel x86 cores and NVIDIA Tesla GPUs. The project baseline assumes the use of 50% of the compute budget for conventional x86 nodes, and 50% for GPU-accelerated nodes. With the introduction of the KNL, as detailed below, the performance difference between x86 nodes and GPU-accelerated

nodes has changed and, in terms of the price for a given performance, the optimal type of nodes depends on the requirements of the LQCD calculations being done. Further blurring the distinction between conventional x86 nodes and GPU-accelerated nodes, new, 32-core conventional x86 CPUs from Intel (Skylake) and AMD (Zen) are just now appearing. In the discussion in this document, the goals are to provide the target of 45 TF of computing power, using our standard benchmarks and also to optimally meet the overall needs of the user community for the target LQCD jobs for the near future.

Regarding the 45 TF deployment goal for FY 2017, the project has already deployed 16 TF in FY 2017 at TJNAF to expand the KNL-based 16p cluster from 192 nodes to 256 nodes. Since we shifted some funds from staff operations to computing hardware purchase though, we are not including this 16 TF when setting the FY 2017 acquisition "deployed computing" target used in this analysis.

The choice of a 50:50 split between x86 nodes and GPU-accelerated nodes in our baseline forecast was driven by the recognition that not all user jobs can run on GPUs, either due to not (yet) available software or the need for more memory and/or internode bandwidth than is available on GPU-accelerated nodes. Similar restrictions appear for the current analysis, making it important to understand the performance of possible hardware solutions for a variety of likely LQCD jobs of different sizes. The ability of x86 solutions to run all parts of USQCD codes, gives this hardware target an advantage in users ease-of-use.

In FY 2017, the project decommissioned systems purchased in 2010 and 2011. There was also some attrition of systems purchased in 2012. This reduction in capacity was partially offset by a 40-node allocation arranged by the project on the BNL Institutional Cluster (each node includes dual NVIDIA K80 GPUs) that went into production in early January 2017. The project will cease paying for support for the IBM BlueGene/Q at BNL in April 2017, but will continue to support it in-house through the end of FY 2017.

In our baseline model, sustained performance on conventional clusters is defined as the average of single precision DWF and improved staggered ("HISQ") actions on jobs utilizing 128 MPI ranks. In our last cluster procurement at FNAL, the 128 MPI ranks were spread out over 8 nodes, to include the effects of internode communication in the performance. "Linpack" or "peak" performance metrics are not considered, as lattice QCD codes uniquely stress computer systems, and their performance does not uniformly track either Linpack or peak performance metrics across different architectures. GPU clusters or other accelerated architectures are evaluated in such a way as to take into account the Amdahl's Law effect of not accelerating the full application, or of accelerating the non-inverter portion of the code by a smaller factor than the inverter, to yield an "effective" sustained teraflops, or an equivalent cluster sustained performance. Effective GPU TF are based on benchmarks developed in FY 2013 to assess the performance of the NVIDIA GPUs used on the various project clusters on HISQ, clover, and DWF applications, and reflect the clock time acceleration of entire reference applications. As new codes and hardware have become available, we have adjusted our ratings to reflect a balance of LQCD calculations. For project KPI's, effective TF are equivalent to TF when combining CPU and GPU values.

The evaluations below are based up a budget of $750K for computing hardware at BNL in FY 2017 for this acquisition. Thus, we are looking for a target price/performance of 17 k$/TF about

the same target as in FY 2016. We are ignoring the 16p expansion for the sake of this target since we established the KPI's before we found that we could shift $250k in FY 2017 funds from operations staff to the computing hardware purchase, which is about the funding level less carry-over funds used by the 16p expansion. The computing hardware budget has already been adjusted to incorporate unused management reserve, but may be adjusted again based upon the final determination of needed disk capacity relative to the amount to be delivered by BNL according to the LQCD-BNL MOU.

The goal for FY 2017 is to install these new resources as soon as possible, while allowing consideration of the new hardare options, such as the new Intel Skylake conventional CPUor the AMD Zen chip. The target date for operations is thus set to Sept 30, 2017.

3 Hardware Options

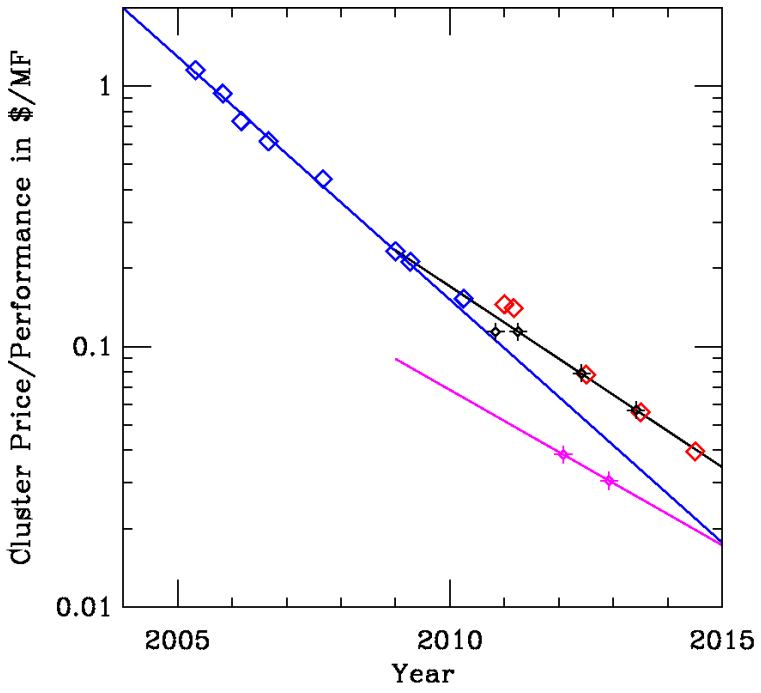
> *Each year the project will optimize the next procurement to yield an ensemble of hardware resources that achieves the highest performance for the portfolio of projects that USQCD intends to execute. This may include procuring two different types of computer systems in a single year.*

The following types of hardware are considered in this analysis:

1. A conventional cluster, based on x86 (Intel or AMD) processors with an Infiniband or Intel's Omni-Path network.
2. A GPU accelerated cluster, based on Intel host processors, an Infiniband network, and NVIDIA GPU accelerators.
3. An Intel Xeon Phi Knights Landing (KNL) cluster with either an Infiniband network or Intel's Omni-Path network

Overview of Hardware Trends

For the LQCD-ext II initial reviews, our baseline performance was developed from our experience with running both conventional clusters (since 2005) and GPU clusters (since 2009). USQCD has tracked price/performance on LQCD Infiniband-based conventional clusters deployed at Fermilab and JLab since 2005. The plot below shows these cost trends, along with exponential fits to two subsets of the data, through 2013. Also included are data and an extrapolation line for GPU-accelerated clusters.

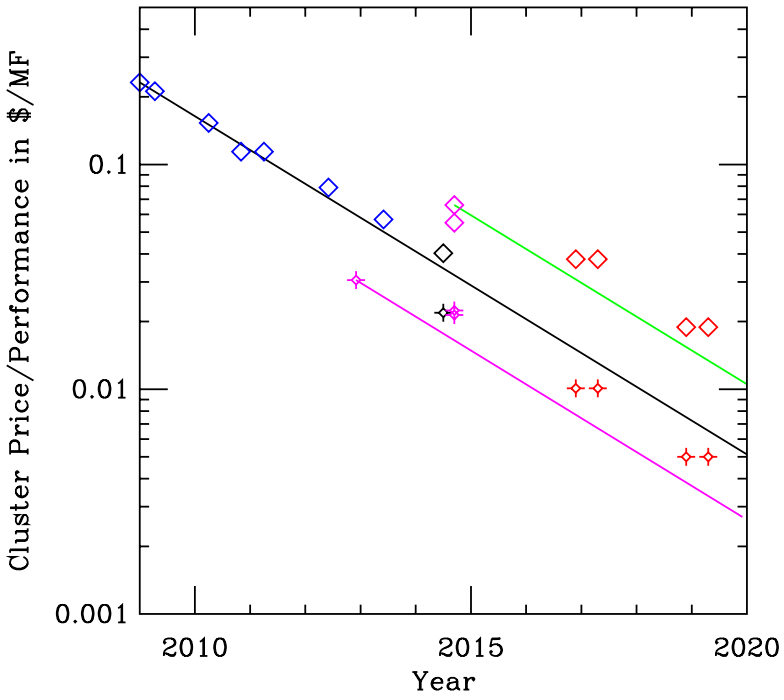


Here, the blue line is the least-squares fit to the clusters purchased between 2005 and 2011, shown as blue diamond symbols. The red diamond symbols are baseline goals used in the LQCD-ext project plan. The black line is the fit to the points from 2009 through the FY13 cluster, Bc. The magenta line connects the points corresponding to the two GPU clusters which were not memory rich, Dsg and 12k.

What is clear from this graph is that the price performance curve has a bend in 2010 such that the performance doubling time per dollar slowed from around 18 months to around 24 months. Tesla class GPUs with ECC provide roughly 4 times as much performance per dollar, but demonstrate roughly the same 24 month doubling time.

For the LQCD-ext II project, we developed our baseline goals using a 24-month doubling time. Dropping data from before 2009, the figure below shows our experience (2014 and earlier) and our forecast (2016 and beyond). (This figure was produced in 2014.) Here the blue diamonds are the LQCD-ext and ARRA clusters. The black diamond is the original estimate for our FY14 purchase. The magenta diamonds, which are noticeably above the trend line, represent the pi0 cluster purchased at FNAL. The lower magenta diamond reflects higher than anticipated costs from manufacturers, due in part to the effective departure of AMD from the the HPC cluster market. The upper of the two diamonds represents the price-performance of pi0 with larger than usual system memory (128 GBytes/node) and a 5-year warranty. The red diamonds are forecasts for "future" clusters (from a 2014 perspective) with purchases split over fiscal year boundaries.

The graph also shows points with magenta stars, representing the two GPU clusters, ARRA 12k and pi0-g, along with a GPU trend line.



It is important to note that the larger memory for the pi0 cluster that was deployed in 2014 was needed for the calculations being started at that time. The trend to larger memory footprint for LQCD jobs has become the norm in much of the USQCD community. The larger memory is used to store eigenvalues of the operators of interest and these eigenvalues then markedly speed up the calculation of quark propagators and other observables. The LQCD community is also generating larger lattices on the Leadership Class Facilities (LCF), and these lattices require larger memory when observables are measured on LQCD hardware.

Overview of Allocation Requests and the LQCD Hardware portfolio

For the upcoming allocation year, 7/1/2017 to 6/20/2018, USQCD users have submitted proposals to use essentially all of the available time on LQCD GPU clusters. For the conventional clusters, proposals exceed the available time by a factor of 2.49 and for the KNL cluster, the oversubscription is a factor of 2.19. Since the KNL is an x86 based machine, codes which run on conventional clusters will run without modification on KNL nodes, although achieving high performance on a KNL nodes generally requires more carefully crafted code than on a conventional cluster node.

In FY 2017, the LQCD-ext II project will be retiring the 512 node BGQ machine at BNL. This machine has run only 512 node jobs for the last year, which means a physical memory size of 8 TBytes was available to the user. A number of users were running jobs for which this memory

size was barely adequate. In addition to user demand for cluster nodes, there is also substantial demand for reasonable size memory in the partitions available to users. About 30% of the cluster time at FNAL from 7/1/16 to 5/1/17 has been used for jobs with memory sizes exceeding 4 TBytes. This, coupled with the large memory size in all the jobs run on the BGQ, indicates that this acquisition should target machines capable of running QCD codes well on partitions with 5 TBytes of memory, or more. This leads us to target a machine that will run user jobs efficiently in the 16 to 32 node range, with between 128 and 256 GBytes of memory per node.

Conventional Clusters

The continued demand by USQCD users for conventional cluster time shows the usefulness of this hardware platform for LQCD calculations. Pi0 is entering its fourth year of operation and a successor platform is needed for this workload. (As we will discuss further in the KNL section below, KNL nodes are able to take on this workload without requiring code rewrites.) The presence of both GPU accelerators and the KNL is having some impact on conventional cluster nodes. Both Intel and AMD have new, 32 core conventional CPU chips coming out now. Intel's 32 core chip, called Skylake, is expected to be available in Q4 2017. In addition to the large core count, Skylake will also implement the AVX-512 instruction set that is currently only available on the KNL. The Skylake will not, it is believed, have the 16 GBytes of on-chip MCDRAM that is available on the KNL. On the KNL, the MCDRAM provides a substantial amount of memory with very high bandwidth. The AMD Ryzen chip is a re-entry of AMD into the x86 desktop and server market and offers LQCD users the advantages of a second manufacturer in this market.

In comparing Skylake with KNL, Skylake will have: 1) full-featured Xeon cores which should give better performance for compiled codes, 2) less memory bandwidth which will decrease performance and 3) a larger price per node, possibly by up to a factor of 2. While Skylake seems an attractive possibility for this LQCD procurement, there are two major unknown issues, price and the timeline for availability. To meet our goal of $17k/TFlop, if a Skylake node sustained 0.5 TFlops, which is a reasonable estimate given our experience with KNL, the node would have to cost $8.5k, which is probably slightly optimistic. Our expectations for Ryzen are similar at this point.

An alternative cluster scenario would be a Broadwell based machine. Broadwell only has AVX-2, and not AVX-512, so it would not perform at the level expected by Skylake (or KNL). However, since it is nearing the end of its lifetime, it would become an interesting option if the price were very low.

For a conventional x86 cluster, single-rail EDR Infiniband (100 GBytes/s) is a well-understood option, with Omni-path as an interesting option for the Skylake chip. For either kind of node, 128 GBytes/node would be the minimum memory and systems with larger memory (256 Bytes) possible, provided we can meet the baseline delivered TFlops.

GPU Accelerated Clusters

For those calculations for which optimized software is available, GPU-accelerated clusters offer a substantial improvement in price/performance compared with conventional clusters. The LQCD hardware portfolio includes the 12k cluster at JLAB, the pi0-g cluster at FNAL and 40 dedicated nodes of the BNL Institutional Cluster (BNL IC), which has dual K80 GPUs on each node. For USQCD users, and software developers, GPU nodes will continue to be a major focus, since not only are there substantial GPU resources in LQCD hardware, but the LCFs are deploying large GPU based machines.

The newest GPU from NVIDIA, the P-100, offers not only an improvement on the traditional large core count performance associated with GPUs, but also the NVLINK technology for connecting GPUs and peer-to-peer technology, which allows separate GPUs to access each other's memory, without going through the host.

We have benchmarked optimized GPU codes on K80 nodes (at the BNL IC) and on P-100 nodes, available as test platforms. We have run extensively optimized codes, written by Kate Clark of NVIDIA (Kate did her PhD research in lattice QCD), on these platforms. We see good performance for problem sizes that are small enough to fit on a single node. For example, on a single node of the BNL IC a single precision domain wall fermion solver on a 24^3 x 96 local volume gives about 1.3 TFlops/s of sustained performance (a single node has 2 K80s). This is a very good result and corresponds to 14.3 k$/TF.

The difficulty with the powerful GPU nodes come when one wants to run on a large enough number of GPUs that off-node communications is required. Here our tests show that with 100 GBit/s EDR IB, the performance on a 16-node system for the same DWF running as in the previous paragraph drops dramatically to about 0.7 TF per node. This gives a price performance ratio of 27.0 k$/TF. Given our target of running on 16 to 32 nodes, the scaling of the GPU clusters is not adequate for our purposes.

It is also important to note that the previous discussion focused on highly tuned code for the conjugate gradient solvers. For realistic workloads, where part of the code is not highly tuned and the GPUs play little role, the performance is a comparison between the speed of the CPU (a 36 core Broadwell processor on a BNL IC node) and the x86 processor of a non-GPU cluster node. These speeds are likely not markedly different. However, for each node of the BNL IC (cost about $20k/node), we would have 4 nodes of a KNL system ($5k/node), or example, so the parts of the code that do not use the GPUs will run up to 4 times faster on x86 cluster than on the BNL IC.

The broader LQCD hardware portfolio includes GPUs and these will continue to be an important part of our hardware strategy. Nvidia just announced the new Volta GPU, which is a major step in their product line (they claim 1.5x in performance over the P-100) and which will be part of the DOE's Summit computer. Given the price performance described above, and the clear user preference for x86 platforms, we will not be pursuing a GPU solution in this year's procurement.

Xeon Phi / Knights Landing Cluster

We have had substantial experience with KNL clusters in the last 9 months, both through the 264 node KNL cluster at JLAB and also through a 144-node cluster at BNL, used for a number of BNL computing projects, including lattice QCD. (We will refer to the BNL cluster used for its institutional projects as the BNL-IK machine, to keep it separate from a possible BNL KNL cluster procured through LQCD-ext II.) The JLAB KNL cluster is connected by single-rail Omni-path (100 GBits/s) with an oversubscribed network. The BNL-IK cluster has dual-rail Omni-path, a full fat-tree network and 1 TBytes of SSD per node. Both clusters have 192 GBytes of memory per node. As mentioned earlier, the KNL has support for AVX-512 and also has 16 GBytes of high-bandwidth on-chip MCDRAM.

We have benchmarked a number of USQCD application codes on the KNLs at JLAB, BNL and also at TACC and Theta at ANL. Heavily optimized single node codes give sustained performance in the 700-900 GFlops range for staggered, clover and domain wall fermions. The Grid code of Peter Boyle, running for domain wall fermions with a local volume of 24^4 in single precision, gives 300 GFlops/s for single and dual rail systems of 16 nodes. This is a performance of 16.6k$/TF for jobs running on 16 to 32 nodes. On the BNL-IK, the Grid code gives 250 Gflops/s when running on 128 node systems. For the MILC code, with QPhiX optimizations and a 32^4 local volume in double precision, performance is around 50 Gflops. For MILC C code, i.e. without any optimizations beyond what the compiler will do, the multi-node performance is about 30 GFlops.

One conclusion that our benchmarks have led us to is that for current lattice sizes and calculations that run on up to 32 nodes, there is no need for dual-rail networks. If one went farther in the strong scaling limit and tried to run on 128 nodes, dual-rail would be necessary. Jobs of this size would likely not be allocated enough time by the USQCD Scientific Programming Committee to make any real progress on such a large calculation. A single rail system is 10-15% cheaper than a dual rail system.

An important issue with the KNL cluster option is system reliability and usability for 16 to 32 node partitions. As of early May, both the JLAB KNL and the BNL-IK are undergoing upgrades to their BIOS/OS to improve stability and to weed out weak hardware nodes. To date, the JLAB system has been used almost exclusively for single node jobs, with high reliability, and recently (early May) a single user has begun running steadily on a 64-node partition. On the BNL-IK cluster, users are currently running well on 32 nodes and larger partitions. Both KNL clusters show decreasing node performance as more jobs are run, likely do to fragmentation of memory. This is readily fixed by rebooting the node and some codes can run for extended periods with this performance loss, but a more general understanding of the issues is needed. Operational

reliability of the KNL clusters for multi-node jobs has clearly improved, by will need to be monitored during this procurement to fully assess the operational risks involved.

As seen in the purchase of the JLAB KNL in FY 2016/2017, such a system will meet our target performance goals. Since it is an x86 architecture, all of USQCD code will run on this platform, with optimized code getting very good performance. At least 192 GBytes of memory per node will be needed, and the option of adding 0.5 TBytes of SSD to each node for local storage will be investigated. As mentioned, we will continue to monitor the operational reliability of the JLAB and BNL-IK KNL nodes, as we proceed with this procurement.

4 Alternatives

The following sections summarize the alternative technologies considered to achieve some or all of the stated performance goals of this investment for FY 2017, and are listed in order of desirability.

4.1 Alternative 1: A Xeon Phi / KNL cluster released to production by Sept 30, 2017.

Deploy and commission a KNL cluster of ~140 nodes with an initial performance of 50 Tflops and a memory capacity of ~27 TB for a total M&S cost of $0.75M.

Analysis: The hardware cost for this alternative is within the FY 2017 project budget. The benchmarks show this gives the best price performance ratio for currently available hardware and there are optimized codes in existence for staggered, clover and domain wall fermions. It is also a platform that runs all of our codes base, although with reduced performance for codes without optimization. This solution also supports the part of the USQCD portfolio that only needs fast, single nodes to run on.

We have extensive, and growing, experience with this platform at both BNL and JLAB. A major issue that is ongoing and will become clearer during the procurement, is the operational reliability and stability of this platform for jobs in the 16-32 node range.

4.2 Alternative 2: A conventional x86 cluster released to production by Sept 30, 2017

Deploy and commission a conventional cluster capable of delivering at least 45 TF with at least 20TBytes of total memory at an M&S cost of $0.75M.

Analysis: The hardware costs for this alternative are within the FY 2017 project budget, provided a Skylake or Ryzen solution could be purchased and put into production consistent with our timeline, or a very inexpensive Broadwell based solution is proposed by a vendor. We anticipate few issues in running a cluster with these next-generation chips and expect they would show good performance on the less-optimized parts of our workflow. A

Skylake/Ryzen solution is expected to show better performance per node, leading to fewer nodes and the possibility of larger memory per node, to keep the total memory available to our jobs large enough.

5 Discussion

The goal of this alternatives analysis is to select the purchase scenario which best optimizes the portfolio of USQCD dedicated resources. The estimates of procurement costs are only approximate and the project plan provides estimates of operational costs.

The focus on KNL or conventional x86 nodes has been driven by the job sizes that need to be run on this platform to maintain support for the USQCD user requests and the optimal price performance ratio for the KNL for the targeted job sizes. Given our experience with the BNL-IK and JLAB KNL, we are confident that the KNL solution will meet our performance goals. There is still operational uncertainty surrounding the KNL for larger node-count jobs, which will be monitored closely during this procurement. We would expect to get a better price on a KNL procurement now than either BNL or JLAB received in Q3 2016.

A conventional x86 solution remains an interesting possibility. As mentioned, the USQCD user demand for running on the x86 pi0 at FNAL shows the importance of this architecture for our community. With new 32-core Skylake and Ryzen chips appearing, there is the possibility of competition between vendors, which can help us achieve the best price. With core counts only 2x below KNL and with more powerful cores that compilers can more easily optimize for, this solution could work very well for LQCD, provided the hardware is available on time. Since the schedule for new chips generally has uncertainty, we would likely penalize the price performance for a Skylake/Ryzen solution by 5% per month for each month that it is late.

Finally, a Broadwell solution is also a possibility, if these chips are viewed as sufficiently old, with the introduction of Skylake, that they are substantially discounted.

6 Conclusion

We plan to pursue a procurement strategy that keeps both a KNL and an x86 solution active. We have done extensive benchmarks on KNL and will do some of our own benchmarks on x86 solutions based on new chips, along with asking for vendor benchmarks. We will stay in close contact with the operations staff for the current BNL and JLAB KNL clusters, to stay abreast of improvements and any operational risks associated with that solution.