FY19 Alternatives Analysis
for the
Lattice QCD Computing Project Extension II (LQCD-ext II)

Operated at

Brookhaven National Laboratory
Fermi National Accelerator Laboratory

for the

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

Version 0.1

Revision Date Sep 19, 2018

PREPARED BY:
Amitoj Singh, *Fermi National Accelerator 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 FY19 Procurement

Revision No.	Description	Effective Date
0.1	Document created from FY17 document	Sep 19, 2018

Table of Contents

1. Introduction

This document presents the analysis of FY2019 alternatives for obtaining the computational capacity needed for the US Lattice QCD effort within High Energy Physics (HEP) 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 fit within the current budget guidance of the project for a computing acquisition at FNAL in FY2019:

- $1.2M for computing procurements in FY2019

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

All alternatives assume the continued operation of the existing resources from the FY2015-FY2018 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 FY2019 are: a conventional CPU cluster, a GPU-accelerated cluster, or some combination of these. The interconnect options are either Infiniband or Intel’s Omni-Path 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. GPU-based implementations of multigrid Dirac solvers for clover fermions have been completed and are now in production. KNL (Intel Knights Landing) is currently end of life on Intel’s roadmap but worth mentioning here. 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 as efficiently as optimized codes.

2. FY19 Goals

The project baseline calls for deployment in FY 2019 of 112 Teraflops per second (TF) of sustained performance, based upon extrapolations of price performance of 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. The performance difference between x86 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 and GPU-accelerated nodes, 32-core conventional x86 CPUs from Intel (Skylake) and AMD (EPYC) are now in general availability. In the discussion in this document, the goals are to provide the target of 112 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.

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 hardware solutions for a variety of 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.

The table below shows the portfolio of existing LQCD hardware across all three sites.

Name	CPU/GPU/ KNL	Nodes	Cores	GPU	Network	Online	Offline
Pi0	CPU	314	5,024		QDR 40	Oct 2014 Apr 2015	
Pi0g	GPU K40	32		128	QDR 40	Oct 2014	
Bc	CPU	224	7,168		QDR 40	Jun 2013	Jul 2018
Ds	CPU	420	13,440		QDR 40	Sept 2010	Jul 2016
Dsg	GPU Fermi	76		144	QDR 40	Mar 2012	Jul 2016
16p	KNL	264	16,896		OPA 100	2016	
18p	KNL	180	12,240		OPA 100	Jul 2018	
12k	GPU K20	45		180	FDR 56	2012	Until 19x
12s	CPU	276	4,416		QDR 40	2012	Jun 2018
BNL-IC	GPU K80	40		80	EDR 100	Jan 2017	
BNL-IC	GPU P100	54		108	EDR 100	Sept 2017	
BNL-KNL	KNL	142	9,088		OPA 100	Feb 2018	
BNL-SKL	CPU	64	2,304		EDR 100	Jun 2018	

Table 1. Portfolio of existing LQCD hardware across the three sites.

Per table 1. from FY2016-2018 the project has been decommissioning systems that were purchased from 2010-13, indicating an average 5-year life-cycle for each machine. During this period the project ceased support for the IBM Blue Gene/Q half-rack at BNL in Sept 2017. The 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 and P100 GPUs) that went into production in early January 2017 and an expansion in September 2017 along with Knight's Landing based clusters deployed at JLab in 2016 and 2018. In FY2018 the project also purchased time on the BNL Knight's Landing and Skylake-based Institutional clusters.

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 consider 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 teraflop, or an equivalent cluster sustained performance. Effective GPU TF are based on benchmarks developed 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 upon a budget of $1.2M for computing hardware at FNAL which meets or exceeds the FY19 target goal of 112 TFlops/s. Thus, we are looking for a target price/performance of 10.7k$/TF.

The goal for FY2019 is to install these new resources as soon as possible, using technology that is proven and in general availability such as the Intel Skylake conventional CPU and the NVIDIA Volta GPU. The target date for operations is thus set for Mar 29, 2019.

3. Hardware Options

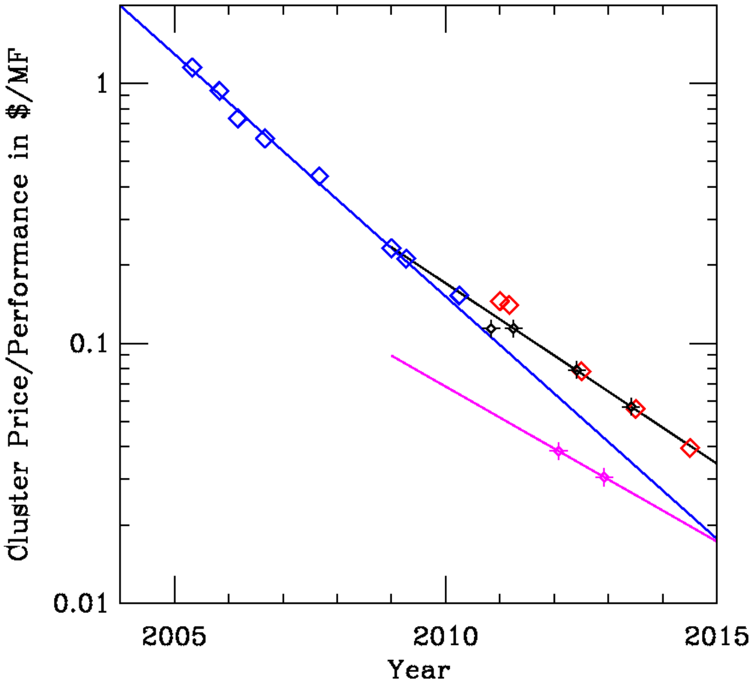
Each year the project will optimize the next procurement to yield an ensemble of hardware resources that achieves the highest performance for the portfolio of projects that USQCD intends to execute. This may include procuring two different types of computer systems in a single year.

The following types of hardware are considered in this analysis:

- A conventional cluster based on x86 (Intel) processors with an Infiniband or Intel's Omni-Path network.
- A GPU accelerated cluster, based on Intel host processors, an Infiniband network, and NVIDIA GPU accelerators.

3.1. Overview of Hardware Trends

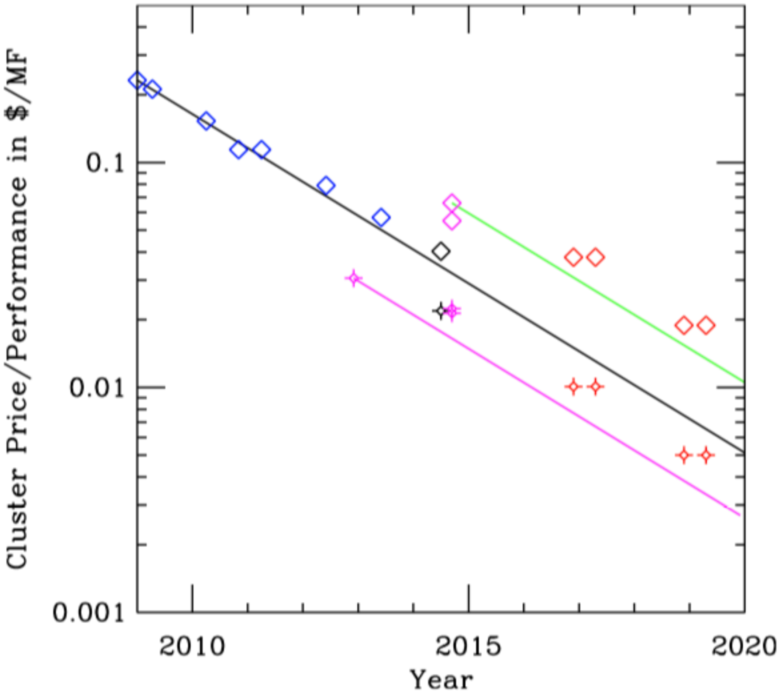
For the LQCD-ext II initial reviews, our baseline performance was developed from our experience with running both conventional clusters (since 2005) and GPU clusters (since 2009). USQCD has tracked price/performance on LQCD Infiniband-based conventional clusters deployed at Fermilab and JLab since 2005. 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 4 times as much performance per dollar but demonstrate 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 and slow return AMD from 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 eigenvectors, or other reusable intermediate solutions, for the operators of interest and these then markedly speed up the calculation of quark propagators and other observables. This change has resulted in the need for larger memory on the computer partition being used, as well as for increased I/O bandwidth to disk and an accompanying increase in disk storage size. In the last year, a number of groups have made substantial progress in reducing the size of these reusable intermediate solutions, decreasing the rate of growth of memory and storage requirements. 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 Project provided computing resources.

3.2. Overview of Allocation Requests and the LQCD Hardware portfolio

For the upcoming allocation year starting July 1, 2018 through June 30, 2019, USQCD users have submitted proposals that exceed available time on LQCD GPU clusters by a factor of 1.8. For conventional clusters, proposals exceed the available time by a factor of 1.4 and for the KNL clusters, the oversubscription is a factor of 1.2. 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 node requires more carefully crafted code than on a conventional cluster node.

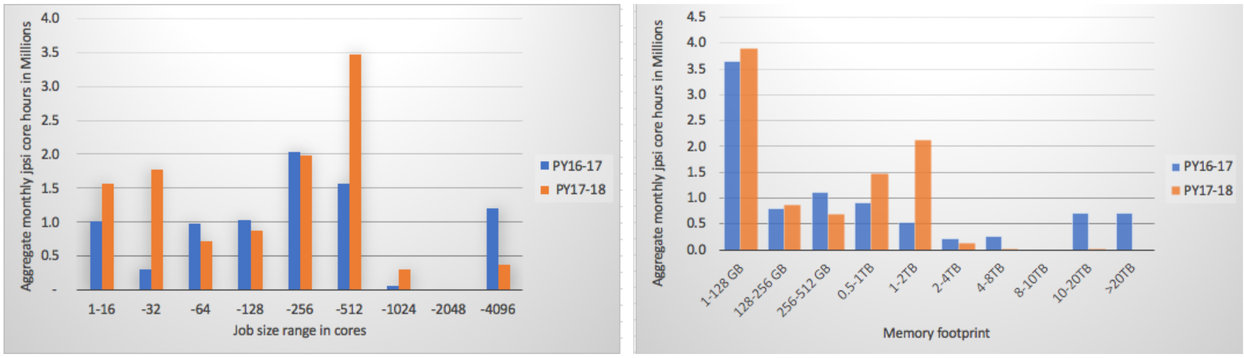


Figure 1. Changes in job size ranges and memory footprint on the FNAL Pi0 (CPU) cluster. PY16-17 is for allocation year from July 1, 2016 through June 30, 2017. PY17-18 is for allocation year from July 1, 2017 through June 30, 2018.

Figure 1 (left) shows demand on the Fermilab Pi0 cluster for 32 node jobs (512 cores) is growing. In addition to user demand for cluster nodes, there is demand for reasonable size memory in the partitions available to users. Figure 1 (right) shows that there is about 20% increase in jobs requiring memory sizes between 1 to 2 TB. During PY16-17 the Muon g-2 campaign put a strong demand on larger node count (256 nodes, 4096 cores), memory hungry jobs (8GB/core). As shown in Figure 1 for the most recent allocation year (PY17-18) this is no longer a requirement.

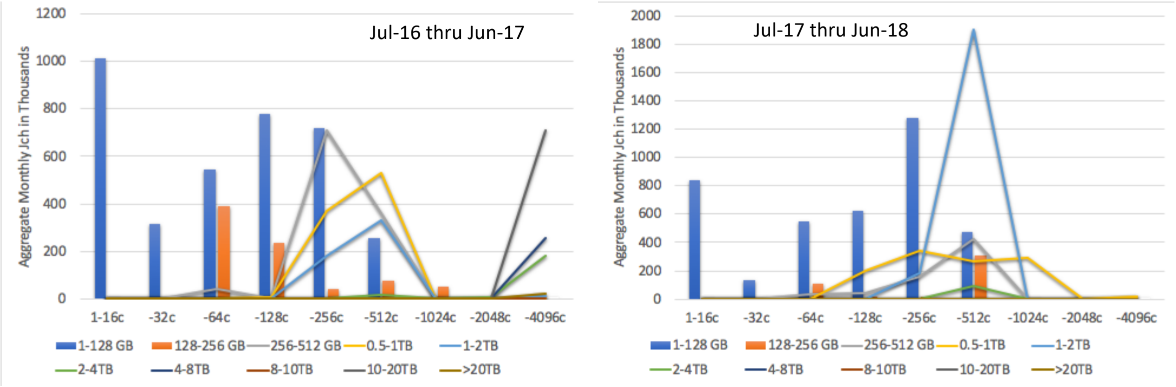


Figure 2. Changes in memory footprint by cores on the FNAL Pi0 (CPU) cluster.

Figure 2 affirms the conclusion that demand for 32-node (512 core on Pi0 cluster) jobs and between 1-2TB of memory, which translates to 4 GB/core is growing. The FY19 acquisition should target a machine with power of two core counts and at least 4GB/core memory

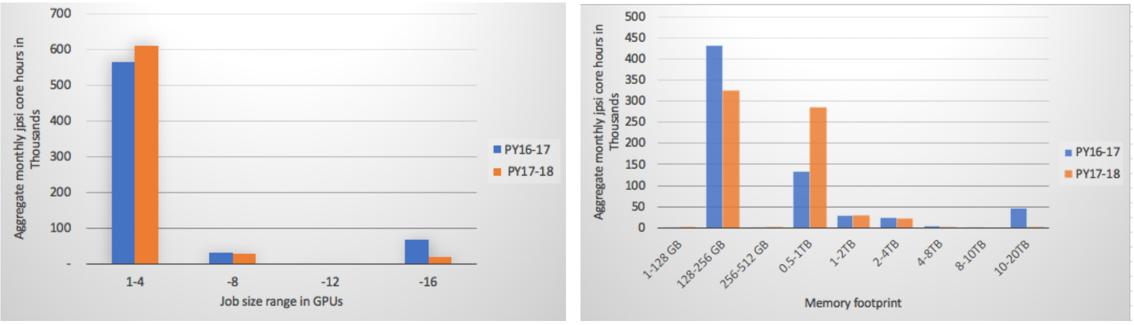


Figure 3. Changes in job size ranges and memory footprint on the FNAL Pi0g (GPU) cluster. PY16-17 is for allocation year from July 1, 2016 through June 30, 2017. PY17-18 is for allocation year from July 1, 2017 through June 30, 2018.

Figure 3 (left hand) shows job size ranges for GPU based jobs. A majority of jobs request between 1 to 4 GPU per job, which on the Fermilab Pi0g cluster translates to a 1-node job. A very small fraction of jobs is multi-node though we are starting to see that change in the current allocation year, starting July 1, 2018.

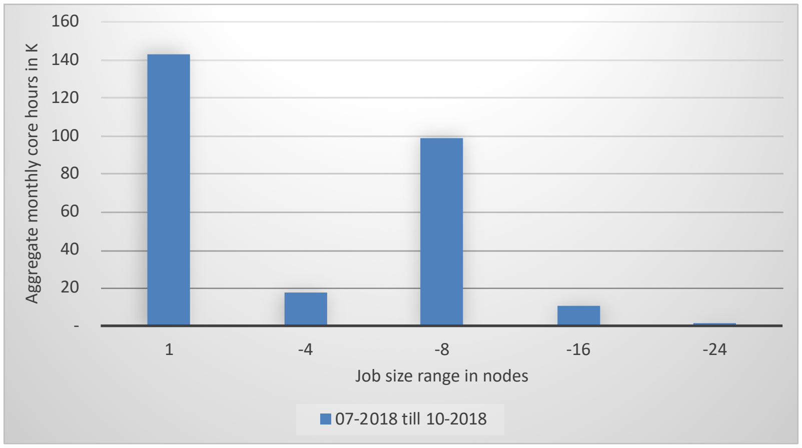


Figure 4. Job size ranges on the BNL Skylake cluster from July to Oct 2018.

Figure 4 shows the job size ranges in nodes on the most recent BNL Skylake cluster from July 2018 till date. The conclusion we draw here is that there is a 50:50 mix of single and multi-node jobs on this cluster, again indicating strong demand for multi-node jobs.

This leads us to target a machine that will run user jobs efficiently in the 16 to 32 node range, with at least 128GB of memory per node.

3.3. 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. The presence of both GPU accelerators and the KNL is having some impact on conventional cluster nodes. Both Intel and AMD have 32 core conventional CPU chips in general availability. Intel's 32 core chip, called Skylake has been in production since Q4 2017. In addition to the large core count, Skylake supports the AVX-512 instruction set that was introduced with the Xeon Phi chipset. The Skylake does not 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 EPYC chip is a re-entry of AMD into the x86 server market and offers LQCD users the advantages of a second manufacturer in this market though benchmarking on this platform shows brittle performance.

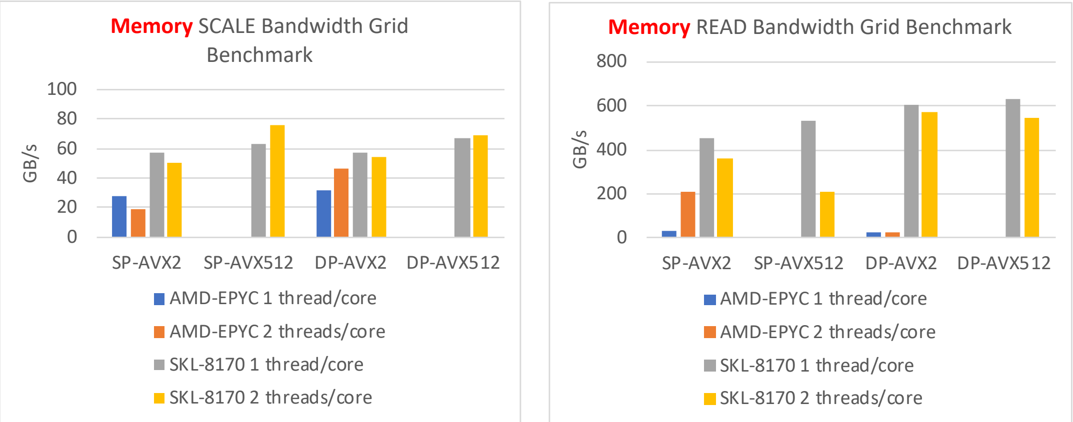


Figure 5. Memory bandwidth LQCD Grid benchmarks. SP = Single Precision, DP = Double Precision.

Results of benchmarks shown in Figure 4 were run on an AMD EPYC 7601 2-socket 32-core 64-thread 2.2GHz, 2666 MHz memory (Figure 4 left plot). The Intel was a Xeon Skylake 8170 2-socket 26-core 52-thread 2.1GHz, 2666 MHz memory, Turbo off (Figure 4 right plot). On both machines no NUMA tricks were used to provide processor or memory affinity to multi-threaded runs. It has been observed that NUMA side effects are worse on AMD than on Intel. AMD's offering of larger core counts per socket compared to Intel translates to additional NUMA domains, thus increasing the complexity of laying compute and data in the associated processor and memory NUMA regions.

For a conventional x86 cluster, single-rail EDR Infiniband (100 GBytes/s) is a well-understood option, with Omni-path as viable option for the Skylake chip. HDR Infiniband (200 GBytes/s) is now in general availability and should be priced as an option during procurement. 128 GBytes/node would be the minimum memory and systems with larger memory (256 Bytes) possible, provided we can meet the baseline delivered TFlops.

3.4. 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 have dual K80 GPUs on each node and 54 nodes of the BNL-IC which have dual P100 GPUs on each node. For USQCD users, and software developers, GPUs 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 latest GPU from NVIDIA, the Volta, offers not only an improvement on the traditional large core count performance associated with GPUs but doubles the memory and also improves 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 NVIDIA Tesla K80 nodes (at the BNL IC) and on P100 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 2.8 TFlops/s of sustained performance (a single node has 2 V100s). This is a very good result and corresponds to 5.3k$/TF.

The difficulty with the powerful GPU nodes comes when one wants to run on a large enough number of GPUs that off-node communications is required. Here our tests show that with 100GBit/s EDR IB, the performance on a 16-node system for the same DWF running as in the previous paragraph drops dramatically to about 0.9 TF per node. This gives a price performance ratio of 16.6k$/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 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 (on a GPU host) and the x86 processor of a non-GPU cluster node. These speeds are not markedly different, between non-GPU and GPU nodes. However, for each node of the cluster that is GPU based (raw cost about $24k/node for a dual V100 based host) we could purchase 3 nodes of a Skylake based system (raw $8k/node), for example, so the parts of the code that do not use the GPUs will run up to 3 times faster on an x86 cluster than on a GPU cluster.

The broader LQCD hardware portfolio includes GPUs and these will continue to be an important part of our hardware strategy. Nvidia Volta GPU, which is a major step in their product line (1.5x in performance over the P100) is part of the DOE's Summit computer. Given the price performance described above, and the clear user preference for x86 platforms, we will be pursuing a GPU solution with some restraint in this year's procurement.

3.5. Xeon Phi / Knights Landing Cluster

Intel has initiated a product discontinuance plan for its Xeon Phi 7200-series processors codenamed Knights Landing (KNL). Given that the existing portfolio of LQCD machines contains a significant fraction of KNL based hardware (444-node KNL cluster at JLAB and 144-node cluster at BNL) the SPC will continue to allocate this resource as long as it is supported by the OEMs and system integrators warranty and service plans. At the same time, Intel will keep offering its codenamed Knights Mill (KNM) solutions for Deep Learning.

For the purposes of this alternative analysis we will not be pursuing a KNL-based.

4. Alternatives

The following sections summarize the alternative technologies considered to achieve some of the stated performance goals of this investment for FY2019 and are listed in order of desirability.

4.1. Alternative 1: A 90% - 10% (by budget) mixture of conventional and GPU-accelerated clusters released to production by March 29, 2019.

Deploy and commission a conventional cluster of ~89 nodes and a GPU-accelerated cluster of ~2 hosts (4 GPUs per host, total 8 GPUs) capable of delivering respectively at least 39 TF and 11 effective TF, 50 TFlops total, with at least a memory capacity of ~12TB for a total M&S cost of $1.2M.

Analysis: The hardware cost for this alternative is within the FY2019 project budget. The 90:10 split between a conventional x86-based cluster which is Skylake based with a small fraction of GPU-accelerated hosts is an option that provides opportunity for in-kind contribution from Fermilab. This option considers the fact that Fermilab has agreed to host the next Institutional Cluster (FNAL-IC) with a design primarily optimal for LQCD but also capable of running CMS and Intensity and Cosmic Frontier jobs. The 90:10 split would allow Fermilab the opportunity to commit additional funds to expand the number of accelerated worker nodes up to a sizeable fraction that is of benefit to both LQCD and Fermilab. Given that both technologies (Intel Skylake and NVIDIA Volta) are in production at various HPC centers and at BNL, we foresee no delay in procuring and deploying this alternative.

The reasoning behind adding two, a tiny fraction, of GPU-accelerated worker nodes is as follows:

1. The two accelerated hosts add ~11 TFlops to the projects FY19 deployment goal.
2. These GPU hosts would be procured with the understanding that we can purchase additional GPU hosts based on available funding, either from the project or Fermilab.
3. The two GPU hosts (8 GPUs total) would provide a "dedicated" pipeline for LQCD jobs with the ability to soak up any idle GPU cycles available on the Fermilab-bought GPU hosts.

This alternative is being discussed based on the fact that during the recent call for proposals for PY19-20, CPUs were over-requested by a factor of 1.4 with GPUs being over-requested by a factor of 1.8 and KNLs by a factor of 1.25.

The conventional piece consists of Skylake based nodes and the accelerated piece consists of the NVIDIA Volta GPU with four GPUs per host. A 100 Gb network will certainly be needed, with lowest price being the determining factor between Intel's Omni-Path or Mellanox's Infiniband solution.

4.2. Alternative 2: A pure conventional x86 cluster released to production by March 29, 2019

Deploy and commission a conventional cluster of ~99 nodes with an initial performance of 43 Tflops with at least a memory capacity of ~13TB for a total M&S cost of $1.2M

Analysis: The hardware costs for this alternative are within the FY 2019 project budget, provided a Skylake solution could be purchased and put into production consistent with our timeline. We anticipate few issues in running a cluster with these chips and expect they would show good performance on the less-optimized parts of our workflow. A Skylake 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.

There is considerable space for detailed optimization of this cluster, including host memory size, network bandwidth and network topology. A 100 Gb network will certainly be needed, with lowest price being the determining factor between Intel's Omni-Path or Mellanox's Infiniband solution. The choice of memory size is based on running at Fermilab and BNL conventional clusters for the most recent allocation year.

4.3. Alternative 3: A 50% - 50% (by budget) mixture of conventional and GPU-accelerated clusters released to production by March 29, 2019.

Deploy and commission a conventional cluster of ~52 nodes and a GPU-accelerated cluster of ~10 hosts (4 GPUs per host, total 40 GPUs) capable of delivering respectively at least 22 TF and 56 effective TF, 78 TFlops total with a memory capacity of at least ~8TB, at an M&S cost of $1.2M.

Analysis: The hardware costs for this alternative are within the FY 2019 project budget. The 50:50 split between conventional and GPU would be adjusted at the time of award based upon science requirements. This mixed resource would augment GPU resources brought online at BNL in 2017-2018, and thus is low risk in that software is already available to run on this platform. Some additional improvements in software would be needed to exploit the AVX512 instruction set and inter-GPU links, but much of this is expected to be contained in the Grid and QUDA packages.

The conventional piece consists of Skylake based nodes and the accelerated piece consist of the NVIDIA Volta GPU with four GPUs per host. A 100 Gb network will certainly be needed, with lowest price being the determining factor between Intel's Omni-Path or Mellanox's Infiniband solution.

4.4. Alternative 4: A pure GPU-accelerated cluster released to production by March 29, 2019

Deploy a GPU-accelerated cluster of up to 22 nodes (4 GPUs per host, total 88 GPUs) sustaining 124 effective TF with a memory capacity of at least ~2.8 TB, with an M&S cost of $1.2M.

Analysis: The hardware costs for this alternative are within the FY 2019 project budget, provided a NVIDIA Volta based solution could be purchased and put into production consistent with our timeline. This alternative is being discussed based on the fact that during the recent call for proposals for PY19-20, GPUs were the most over requested resource amongst the three (GPU, CPU and KNL). The TFlops delivered by this deployment exceeds the FY19 project deployed computing goal of 112 TFlops. We anticipate few issues in running a cluster with these chips and expect they would show good performance for the GPU-optimized parts of our workflow.

There is considerable space for detailed optimization of this cluster, including host memory size, network bandwidth and network topology. A 100 Gb network will certainly be needed, with lowest price being the determining factor between Intel's Omni-Path or Mellanox's Infiniband solution. The choice of memory size is based on running at Fermilab and BNL GPU clusters for the most recent allocation year.

4.5. Alternative 5: Status Quo (no additional deployment in FY19)

Continue to operate the existing project clusters deployed at FNAL and JLab and buy additional runtime on the BNL Institutional Clusters.

Analysis: The cost of this alternative is $1.2M in FY2019 to operate the existing facilities. The incremental cost of this alternative (new investment) is $0. This alternative is included only for completeness and would not be capable of providing the necessary computational capacity to achieve the scientific goals of this project. Specifically, it would not leave USQCD with sufficient capacity to exploit the configuration generation capability of the supercomputers that DOE ASCR Facilities will have released to production around FY18-19.

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.

We propose a conventional x86 solution with a small fraction of accelerated nodes provides the opportunity for in-kind contribution from Fermilab to expand the accelerated piece of the cluster. During the recent call for proposals for PY19-20, CPUs were over-requested by a factor of 1.4 with GPUs being over-requested by a factor of 1.8 and KNLs by a factor of 1.25. Clearly there is a strong demand for both CPU and GPU based clusters. The USQCD user demand for running on the x86 pi0 at FNAL and the latest Skylake Institutional Cluster deployed at BNL shows the importance of this architecture for our community. With new 32-core Skylake chips in general availability, there is the possibility of being able to deploy this machine within a short period of time, with minimum to no delay. With powerful cores that compilers can more easily optimize for, this solution could work very well for LQCD.

The small fraction of accelerated nodes provides a ~3x deployed TFlops compared to conventional nodes thus bringing us closer to meeting a fraction of the FY2019 deployed TFlops goal.

6. Conclusion

The preferred path forward for the LQCD-ext II Project is Alternative 1, in which there is the rapid deployment of a cluster at Fermilab based on chips in general availability. The Project will purchase all of the node-hours on the 91 nodes of this machine from April 1, 2019 through the end of FY19. Fermilab will have the opportunity to provide in-kind contribution to add accelerated nodes to the existing cluster and the project can negotiate purchase of node hours on those nodes as well.