Offices of High Energy Physics and Nuclear Physics

Report on the

LQCD-ext/ARRA

2011 Annual Progress Review

May 10-11, 2011

Executive Summary

The Annual Progress Review of the LQCD-ext (LQCD extension) and the LQCD ARRA (American Recovery and Reinvestment Act) projects was held on May 10-11, 2011 at the Fermi National Accelerator Laboratory (FNAL). The purpose of the review was to assess the projects' progress towards their overall scientific and technical goals. Five expert reviewers from the nuclear physics, high energy physics and computer science communities heard presentations on scientific progress, computing hardware acquisitions and operations, allocation of resources, and disseminations of scientific results. In particular, the LQCD-ext/ARRA teams were instructed to address five charges:

1. The continued significance and relevance of the LQCD-ext/ARRA project , with an emphasis on its impact on the experimental programs supported by the Offices of High Energy and Nuclear Physics of the DOE;
2. The progress towards scientific and technical milestones as presented in the LQCD-ext project's Project Execution Plan and the LQCD/ARRA project's Project Execution Plan;
3. The status of the technical design and proposed technical scope for FY 2010-2011 for both projects;
4. The feasibility and completeness of the proposed budget and schedule for each project;
5. The effectiveness with which LQCD-ext/ARRA has addressed the recommendations from last year's review.

The review panel reported that the LQCD-ext/ARRA collaboration had addressed the five charges at the review. The impact of LQCD simulations on experimental programs in precision measurements of the Standard Model, Heavy Ion collisions and spectroscopy has grown dramatically over the last few years. The LQCD-ext procurement plans for the next fiscal year ware still uncertain because of delays in the BlueGene/Q(BG/Q) time table. However, the LQCD-ext team's plan to assess the three competitive hardware choices, 1. BG/Q, 2. Commodity cluster hardware and 3. Graphical Processors Units (GPU's) by July and have a firm acquisition plan by September was endorsed by the review panel. The success of the ARRA project at the Thomas Jefferson National Accelerator Facility (TJNAF) in exceeding its original milestone of 16 TFlops by a factor of ~4.75 by constructing GPU clusters was praised. The review panel urged again this year that the SciDAC program support within the Office of Science be increased because it develops critical software for LQCD. The user survey, which missed its target of a 92% customer satisfaction rating by 11%, was judged to be in need of trimming and focusing in order to increase its user response success. The reviewers concurred with post panels and encouraged the accelerator growth of the number of workshops and conference talks in order to increase the impact of LQCD results within the experimental community.

Introduction and Background

The DOE Offices of Advanced Scientific Computing Research (ASCR), High Energy Physics (HEP) and Nuclear Physics (NP) have been involved with the National Lattice Quantum Chromodynamics Collaboration (USQCD) in hardware acquisition and software development since 2001. The Lattice Quantum Chromodynamics (LQCD) IT hardware acquisition and operations activity, which started in 2006 and ran through 2009, operated a "Quantum Chromodynamics-on-a-chip" (QCDOC) machine at Brookhaven National Laboratory (BNL), and built and operated special purpose commodity clusters at the Fermi National Accelerator Laboratory (FNAL) and the Thomas Jefferson National Accelerator Facility (TJNAF). LQCD met its goal of providing 17.2 Teraflops of sustained computer power for lattice calculations.

The hardware acquisition strategy of LQCD was essential to its success. Each year the collaboration benchmarked the kernels of the QCD code on the newest cluster and supercomputer hardware, and the winner of the price-to-performance competition became that year's provider.

The usage of hardware procured by LQCD has been governed by the USQCD collaboration through its executive board and allocations committee. Members of the USQCD collaboration submitted proposals for computer time, some on general purpose supercomputers run by National Energy Research Scientific Computing Center (NERSC), National Nuclear Security Administration (NNSA), and the National Science Foundation (NSF), and some on the dedicated clusters. The resources were awarded on a merit system. Three classes of computer projects have been considered, ranging from large scale mature projects (allocation class A) to mid-sized projects (allocation class B) to exploratory projects (allocation class C). Suitable computer platforms were assigned to the various projects.

In addition to the hardware project LQCD, USQCD has played a role in software development through the Scientific Discovery through Advanced Computing (SciDAC) program. USQCD was awarded a SciDAC I grant (2001-2006) which developed efficient portable codes for QCD simulations. USQCD now has a SciDAC II grant (2006-2011) which will optimize its codes for multi-core processors and create a physics toolbox. These SciDAC grants provide a user interface to lattice QCD which permits the user to carry out lattice QCD simulations and measurements without the need to understand the underlying technicalities of the lattice formulation of relativistic quantum field theories and its implementation on massively parallel computers.

USQCD proposed to extend the work of LQCD beyond 2009, and submitted a proposal, "LQCD-ext Computational Resources for Lattice QCD: 2010-2014" in the spring of 2008. The scientific content of the proposal reviewed successfully on January 30, 2008 and the scientific vision and specific goals of the project were enthusiastically endorsed in full by the panel of scientific experts. The proposal sought $22.9M over a five year period to achieve its scientific goals.

In the January 30, 2008, review, USQCD argued that the mid-scale computer hardware purchased, constructed and operated by LQCD was a critical portion of its overall strategy to produce the physical predictions of Quantum Chromodynamics. That strategy depends on access to the largest Leadership Class machines for the generation of large lattice gauge configurations. These configurations are then analyzed for accurate predictions of matrix elements and spectroscopy on the mid-scale computers of LQCD and results of interest to the experimental and theoretical communities in high energy physics and nuclear physics are obtained. The mid-scale hardware of LQCD also produces smaller gauge configurations which are critical to studies of Quantum Chromodynamics in extreme environments that are relevant to the heavy ion collision program at the Relativistic Heavy Ion Collider (RHIC) at BNL which is operated by the Office of Nuclear Physics. Many of these calculations are not suited for Leadership Class machines, but run efficiently on mid-scale platforms. Several computer scientists at the January review carefully evaluated and then endorsed the mix of computers advocated by USQCD. The review panel also assessed USQCD's claim that the accuracy of some of its predictions rival the accuracy of the present generation of experiments running at DOE HEP and NP facilities. The review panel also analyzed USQCD's claim that the proposed project, LQCD-ext, was needed to maintain this parity in the future.

The LQCD-ext project then entered the Critical Decision Review process.

The CD-0 Mission Need Statement for LQCD-ext was approved on April 14, 2009.

The CD-1, alternative selection and cost range, review occurred at Germantown on April 20, 2009. The review evaluated the LQCD-ext project's documents on conceptual design, acquisition strategy, project execution plan, integrated project team, preliminary system document, cyber security plan and quality assurance program.

The LQCD-ext team updated its documents following recommendations from the CD-1 review panel and it received formal CD-1 approval on August 27, 2009, through a paper Energy Systems Acquisition Advisory Board (ESAAB) presentation and review.

The CD-2/3, project base-lining and readiness, review occurred at Germantown on August 13-14, 2009. Final approval for the project was granted on October 28, 2009.

The Offices of High Energy Physics and Nuclear Physics produced a planning

Budget for the LQCD-ext CD-2/3 review as follows:

Table 1. Planning Budgets for LQCD-ext (in millions of dollars)

	FY2010	FY2011	FY2012	FY 2013	FY 2014	Total
HEP	2.50	2.50	2.60	3.10	3.20	13.90
NP	0.50	0.75	1.00	1.00	1.00	4.25
Total	**3.00**	**3.25**	**3.60**	**4.10**	**4.20**	**18.15**

The TPC of $18.15 left the LQCD-ext project $4.75M short of the figure of $22.9M which was supported by the scientific review of January 30, 2008, and which USQCD had estimated in their original whitepaper. This shortfall was subsequently addressed, however, by the request of the Office of Nuclear Physics for $4.96M of funding through the ARRA (American Recovery and Reinvestment Act of 2009) to build a 16 Tflop/s commodity cluster at TJNAF and operate it for four years. Although this effort is not a formal part of this LQCD-ext project, the resulting hardware at TJNAF is being governed by USQCD using exactly the same procedures that apply to LQCD-ext and the acquisition, construction and operation of this hardware is being tracked on a monthly basis by the same team that is running LQCD-ext. In this way, the Offices of High Energy Physics and Nuclear Physics are monitoring the full scope of science put forward in the USQCD proposal "LQCD-ext Computational Resources for Lattice QCD: 2010-2014". It was agreed that the two efforts, LQCD-ext and LQCD/ARRA, would share Annual Progress Reviews and this report is the first in a series.

LQCD-ext explained at the CD-2/3 review that the budget of Table 1 would support the new deployments and operations of equipment contained in Table 2:

Table 2: Performance of New System Deployments, and Integrated Performance

	FY 2010	FY 2011	FY 2012	FY 2013	FY 2014
Planned computing capacity of new Deployments, Tflop/s	11	12	24	44	57
Planned delivered Performance (JLab + FNAL + QCDOC), Tflop/s-yr	18	22	34	52	90

LQCD-ext/ARRA and was considered in detail by last year's review team. That review had several observations about this development: 1. The success of the hardware project LQCD-ext/ARRA is very sensitive to the continuance of the LQCD/SciDAC software grant because this is where the software that will eventually make GPUs more generally useful to the science community will be developed; 2. A mix of CPU and GPU clusters will be needed in the short term for LQCD-ext/ARRA because many lattice scientific applications are not ready to be ported to GPUs but would be greatly more productive if and when that happens; 3. The initial estimates of TFlops of clusters that can be built for $22.15M will probably be considerably higher than the planning figures shown above, but it is hard to estimate new milestones at this time; 4. The scientific output and impact of LQCD-ext/ARRA may be considerably higher than originally planned for; and 5. The risk associated with the new GPU hardware will exceed that of the more familiar CPUs. All these considerations became part of the discussions of the planning for LQCD-ext/ARRA in FY2010 and 2011, relevant to last year's annual review. Several of these observations have met with fruition: The ARRA GPU cluster is sustaining ~76 Tflops on a fairly diverse set of physics projects, beating the project's original milestone by a factor of 76/16~4.75. The LQCD-ext project is now installing a GPU cluster at FNAL to meet the extra demand coming from proposals submitted to USQCD over the past 12 months.

The Annual Progress Review of LQCD-ext and LQCD/ARRA took place at FNAL on May 10-11, 2011. The review consisted of one day of presentations and a second half-day of questions and answers, report writing, and a closeout session. The appendices to this report provide additional detailed material relating to the review: App.A contains the charge letter to the LQCD-ext/ARRA management team, App.B lists the reviewers and DOE participants, and App.C contains the agenda and links to the talks. The remaining five sections of this report detail the findings, comments, and recommendations of the review committee for each of the charge elements that the LQCD-ext/ARRA collaboration was asked to address.

Continued Significance and Relevance

Findings

The LQCD-ext/ARRA program supports activities in several research areas:

1) Precision calculations relevant to the determination of standard model parameters from heavy quark processes. Calculations of decay constants and form factors which are essential for the extraction of Cabbibo-Kobayashi-Maskawa (CKM) elements from experimental data and for looking for hints of new physics have been made.

2) Exploratory calculations based on "beyond the standard model" (BSM) theories, for which LQCD is at present the only effective technique for extracting quantitative predictions. The emphasis has been on “simple” Technicolor models in which strong dynamics of new generations of quarks and gauge fields generate a composite Higgs which breaks electroweak symmetry. GPU clusters are proving useful in these studies.

3) Hadronic physics quantities such as the spectrum of hadrons, form factors, moments of structure functions, hadron-hadron interactions and scattering. Many of these calculations are aimed at quantities which will be studied at the 12 GeV. upgrade of the Continuous Electron Beam Accelerator Facility (CEBAF) at TJNAF.

4) Calculations of the properties of QCD at finite temperature and baryon density; this regime is explored experimentally in relativistic heavy ion collisions. These simulations are having an impact on the run plans of the Relativistic Heavy Ion Collider (RHIC) at BNL.

USQCD’s scientific goals are focused on carrying out world-leading computations of quantities that are of importance to the experimental HEP and NP programs.

Lattice simulation is the only known way to accurately calculate equilibrium properties of hot QCD matter that is produced in the collisions at RHIC.

LQCD continues to have workshops with the experimental and theory communities to widen its impact and engage in communications with complementary communities of researchers to enhance its influence and impact. The most recent workshops on QCD were mainly focused on Nuclear Physics although High Energy Physics workshops are in the planning stage.

Comments

USQCD now has a significant and growing emphasis on productive interactions with DOE-supported experimental programs. USQCD has organized or participated in a number of workshops with experimental groups over the past two years. In addition, they are producing theoretical results that address areas of current experimental interest.

One prominent example, cited by Richards, is the Nuclear Physics with Lattice QCD (NPLQCD) collaboration’s prediction of net binding for the hypothetical H-dibaryon that appeared in PRL within the last month. The predictions of a Japanese collaboration (HAL QCD Collaboration) also appeared in the same issue of PRL and are supportive of this prediction. Further theoretical work is required to refine these calculations (do the results extrapolate robustly to the physical pion mass?) and give a more precise binding energy. This prediction stands as an excellent challenge to the experimental physics community that will help confirm the predictive power of LQCD when this exotic system is detected.

Richards also cited the USQCD thrust into evaluating Generalize Parton Distributions (GPSs) and Transverse Momentum Distributions (TMDs) as further evidence that the USQCD program is addressing areas of prime interest to DOE-supported experimental programs.

Continuing efforts in the mass spectroscopy of the hadrons and their electroweak properties have refined USQCD's predictions to the stage where they can access states with spin as high as 7/2. Since the excited hadron spectrum is an area of ongoing experimental research, this represents another good example where USQCD is having an impact on DOE-sponsored experimental research.

Richards showed impressive new results for the spin content of the proton with results now at the ~ 2 x physical pion mass. From what USQCD has at present, they infer that ~51% of the spin content is in the glue sector.

Many additional results were presented to compare USQCD predictions with existing experiments and they also demonstrate the impact of these efforts on the broader physics enterprise. Kronfeld cited the example of the precision $\alpha_s(M_Z)$ determination that exemplifies the contributions of USQCD to the determination of the parameters of the Standard Model (SM). He argued that precise SM tests are a key to knowing where the physics beyond the Standard Model (BSM) provides its favorable "windows of opportunity". Kuti followed with a range of examples of BSM models and their simulations. These illustrate a growing enterprise of BSM physics within USQCD that will require major increases in computer resources to fully explore the opportunities.

Karsch showed the relationships between finite temperature (T) Lattice Gauge Theory (LGT) and the research programs at the RHIC/LHC/FAIR experimental facilities. Both theory and experiment seek to simulate/measure features of the QCD equation of state and map out the phase boundaries as a function of temperature and chemical potential. At non-zero chemical potential and large finite temperatures, USQCD requires capability (leadership class) computers while at vanishing chemical potential and finite T, they require moderate lattices that can be run on clusters (capacity computing).

Karsch cited one particular measure of impact when he noted that the most top-cited (HEP-Lat) papers in the last five years are from the hotQCD component of the USQCD program. For a specific example of impact on the experimental physics program, he showed the charge fluctuations along the freeze out line as ratios of moments and compared with results of the RHIC-STAR collaboration. He noted the similarities of the hotQCD results with hadron gas results, at least until the 6^{th}-order moment where significant differences can then be found close to criticality. He argued that theory and experiments need the 6^{th} moment since that is the first moment that goes negative in hotQCD (effect of chiral symmetry) while all moments of the hadron gas remain positive. This motivates both theory (hotQCD) and experiment (RHIC-STAR) to obtain results with higher precision in order to unambiguously establish where the QCD phase boundaries lie.

Karsch also addressed the dilepton production at RHIC (PHENIX data) which currently provides a major mystery in the low-mass region (below 1 GeV) where a prominent excess has been found over the scaled pp collision data. He outlined planned calculations of the lepton spectral function that will provide dilepton and photon yields to compare with these experimental data. This is a forefront area of research and the results from hotQCD are eagerly awaited.

The Beyond the Standard Model effort is growing substantially with eleven US institutions now involved and many projects proposed for the coming year. There have been 15 publications over the past 12 months and 500 citations since 2008. This effort has the potential to attract significant new attention to lattice QCD programs.

In answering one of review panel's questions, Mackenzie portrayed USQCD's "outreach to the community" with many concrete examples of where new members joined the collaboration, including a RHIC experimentalist. This clearly demonstrates that USQCD is open to new members, which is healthy.

Clearly, impact is most easily measured in terms of publications. In answering another of the review committee's questions Kronfeld presented a histogram of publications broken down by project area. He showed a very respectable publication rate with about equal division between regular journal articles and conference proceedings.

Recommendations

USQCD should improve its demographic information so that its impact on postdoc and graduate students training could be addressed quantitatively.

Continued bridges must be built with the community of experimentalists and phenomenologists. Continuation of the "Lattice QCD Meets Experiment" series, invitations of "outside" guest speakers to the annual All-Hands Meeting, and participation by lattice conveners in non-lattice conferences all are helping to foster this aim.

Progress towards Scientific and Technical Milestones

a. LQCD-ext

Findings

Boroski presented a list of milestones numbered 18-21 (slide 16) showing that the LQCD-ext team had met or exceeded all milestone goals and that the milestones were completed on time and within budget.

Two particularly important milestones that were exceeded were:

1. The system scheduled for 12/31/10 deployment was deployed 30 days ahead of schedule with 11% greater performance.

2. The 9/30/10 milestone on cumulative TFlops-yrs was exceeded by 7%.

Comments

LQCD-ext monitors the usage of its clusters effectively.

One future scientific milestone stands out above others and that is the goal to obtain results from simulations with the physical pion mass. In answer to the question of anticipated time scale to reach this future milestone put forward by the reviewers, Mackenzie indicated that the time scale was several years and was dependent on which lattice action is used and on resources available in the future. He also indicated that progress to date involved the implementation of physical mass pions with the Domain Wall Fermion action.

The user survey of customer satisfaction produced a "Customer Satisfaction Rating" of 81% versus the target of a minimum of 92%. USQCD management is aware of this one area of performance deficiency compared to the target and is addressing the perceived underlying issues.

An effort should be made to improve the survey to make it shorter and more focused. Not all questions need to be asked each year. The most recent survey has 42 questions. If that were reduced substantially, the email requesting completion of the web survey could state that it takes (for example) only five minutes to fill out. In an effort to make the survey a more effective tool for improving responsiveness to user concerns, some questions are far more important than other.

Boroski also indicated (slide 33) that LQCD-ext is on target to meet nearly all of the FY11 performance goals. The sequence of Continuing Resolutions has impacted the team's ability to meet some of their performance goals in 2011. Nevertheless, they were able to mitigate the impacts of the delays and make significant progress.

Recommendations

An effort should be made to improve the user survey in particular by making it shorter and easier to complete in a short time. Users should be urged to complete the survey (or perhaps it should be part of their proposals that all individuals complete their survey).

b. LQCD/ARRA

Findings

Acquisition and implementation is complete. The project completed its hardware component generally on schedule. There was a delay in the 2nd phase of implementation to take advantage of a technology refresh on GPUs.
The LQCD-ARRA project, by adopting GPU systems (a "disruptive technology"), achieved considerable gains over the originally projected goals. In particular, instead of a projected 16 Tflops, they achieved an effective 76 Tflops with a fixed hardware investment.

As a result of their favorable experience with gains in Tflops, the LQCD-ARRA project allocated a larger fraction of their 2010 purchase to GPU clusters and a smaller fraction to Infiniband clusters.

Use of the GPU capability ramped up quickly and the GPUs are well utilized.

Comments

The new GPU technology necessitated larger than expected commitments of manpower resources especially for software development (reprogramming in CUDA). These manpower costs for software development have been borne successfully by other funding sources.

The implementation of GPUs to serve the LQCD community has been very impressive. This work is pioneering in many ways, and to implement it under the time constraints of a Recover Act project is admirable.

The project team presented several examples of where LQCD community access to GPUs enabled exploration of science areas previously constrained by access to adequate computational resources.

Technical design and scope for FY2010-11

a. LQCD-ext

Findings

The FY11 Ds cluster consists of a now "conventional" design for LQCD and was somewhat delayed due to the federal budget continuing resolutions.

The proposed design for FY12 is not complete at this time due to a desire to allow more time for information about and access to the BlueGene/Q.

Storage is a minor cost feature of the overall system and the design and implementation with the Lustre file system is working well.

Comments

Given the experience and success of deploying GPUs at TJNAF in the LQCD/ARRA project, the decision to install 128 GPU nodes at FNAL in FY2011 is reasonable.

The design and scope for FY2010-11 looks very solid. LQCD-ext has a well established procedure for implementing the technical aspects of the proposal. Purchases are designed to minimize overhead, leverage off facilities at national labs and provide a balance between known and new hardware.

For FY2010-11 the plan to reduce the size of the FY2010 hardware in favor of purchasing GPUs as part of the FY2011 acquisition is appropriate as the LQCD community starts taking up the new technology.

Holmgren showed that going down the GPU path has promise for great price/performance results (slide 18). However, Error Correction (ECC) version of GPUs are critical to more general USQCD applications (beyond the inverter application which can be easily cross-checked) on the GPUs, though they are also more expensive by a factor of ~4 at the present time.

The ability to use the GPUs relies on software development from the LQCD SciDAC project and other outside sources not funded within the LQCD-ext project. There is significant community interest in running on the GPUs and some codes have been ported to the GPUs and more codes are being ported at this time.

BG/Q may have the ability to put the project goals of integrated TF/$ ahead of planned projections in future years but, if chosen, could cause some delay in the FY12 deployment milestone.

Given USQCD scientific goals, TFlops are the dominant metric of technical performance for both LQCD-ext and LQCD-ARRA. For this reason, it is useful to measure resources in terms of Tflop-years. In addition, $/Tflop is the appropriate measure of cost effectiveness at purchase. The amount of memory plays a minor role. However, software support for new architectures, especially GPUs, plays a major role. In addition, in the case with GPUs, one measures overall performance of a code, not just the part that is accelerated by the GPU. These factors lead to additional complexity in the assessing overall cost-effectiveness.

In FY11, USQCD will utilize about 10 Tflop-years each in capability computing (INCITE award on DOE's Leadership Class computers – primarily for generating gauge configurations) and 10 Tflop-years in capacity computing (USQCD clusters at BNL, TJNAF and FNAL – primarily for evaluating quark propagation within the gauge configurations). This is a very large amount of computing resources focused on one major area of physics and the USQCD community is well-organized and systematic in its approach to efficient utilization of these resources.

In addition, the USQCD allocates time at "zero priority" on DOE's Leadership Class facilities in order to optimize the science accomplishments utilizing time made available by other users' failure to use their allocations in a timely fashion. This has achieved a tripling of the capability resources for USQCD during the past year. However, as other users come up to speed in their usage of their INCITE resources, these "zero priority" resources are likely to shrink during the coming years.

There was considerable discussion of USQCD's experience with GPUs to date and the impact of that experience, along with the rapidly evolving hardware specifications, on the plans for LQCD-ext procurement plans.

For example, Kuti showed an impressive new strong scaling curve for a particular case (32^3 x 256) with the time dimension segmented onto separate GPUs but full 3-D space on each GPU. These codes and the associated computing strategy are useful for excited states and BSM studies. One noteworthy point is that this strong scaling achievement takes a dedicated programming effort in CUDA and the resulting code is specific to the chosen action and only runs on the GPU produced by the company NVIDIA.

Karsch presented an overview of the results of the USQCD allocation process of the last completed allocation cycle. There were 22 "Type A" projects approved (6 are GPU-based) using the largest fraction (90%) of available USQCD computational resources with an average of 11.3 collaborators each. USQCD allocated the remaining fraction (10%) to "Type-B" projects that averaged 4.7 collaborators/project. This dynamic range of resources/project is commensurate with the scientific goals and the needs to engage in multiple and diverse applications to reap the full scientific potential.

The total core-hour resources obtained by the US-LQCD community likely amounts to 50% more than those tracked under the USQCD allocation process (INCITE + USQCD machines) though there appears to be little explicit documentation. For example, USQCD researchers also use NSF (e.g. Kraken) allocations and they anticipate obtaining resources on Blue Waters.

Anonymous user satisfaction survey provided ratings of ~81% overall satisfaction rating compared to the goal of 92%. This could be influenced by the low statistics and by the management focus on the ARRA project that is gaining speed

Recommendations

The impressive successes with GPUs depend upon effective software development which has come through the SciDAC initiative. USQCD and the DOE must find some means of putting this program on a more solid funding base.

b. LQCD/ARRA

Findings

The LQCD/ARRA resources were placed at Jefferson Lab. The acquisition phase of the project is complete. It is in the operations phase.

Comments

The information being collected on GPU performance and reliability is very useful. Time on GPU nodes is included in the USQCD allocation process.

The performance of codes on the GPU enabled systems is very application dependent. The method of assessing the overall system performance based on the equivalent J/Psi cluster time for the application mix actually run on the GPU systems is reasonable.

In aggregate, there are 504 GPUs in production use (May 2011) amounting to ~ 200,000 cores with a practical availability of 67+9 Tflops vs 16 planned Tflops greatly exceeding their original goal (factor of 4.75 increase).

The ARRA team was asked by the review committee if users were ready to use the GPU clusters when they were made available. This question was answered affirmatively by Kuti, who indicated his project, plus one other major project, were ready for production runs once the hardware was on the floor. The USQCD community responded quickly and moved aggressively towards efficient utilization of the GPU facility.

Watson stressed the manpower-intensive nature of testing/installing/maintaining GPUs due to the nature of the fault-tolerant industry that produces GPUs (game applications don't need perfect precision). This will be a continuing problem until vendors improve reliability. He indicated that the more recently installed GPUs have lower failure rates/flop.

Holmgren presented GPU usage statistics in response to reviewers' questions that showed a linear growth of users of GPU time (now at ~30% of USQCD projects) over the first ~16 months of ARRA GPU cluster operations.

Feasibility and Completeness of Budget and Schedule

a. LQCD-ext

Findings

The systems deployed by LQCD so far in FY11 were within the budget and somewhat exceeded the performance targets (Dec. 2010 cluster deployment), and slightly ahead of schedule.

For systems being procured in FY11 there has been some schedule delay due to effects of the continuing resolution, beyond the control of the project. However the LQCD-ext team argued that the performance of the system should exceed the original target while being within budget.

The schedule for the FY12 design & procurement plan is delayed due to uncertainties in the BlueGene/Q timetable. A detailed plan of benchmarking the three alternative platforms (CPU clusters, GPU clusters and BlueGene/Q) is in hand and will be implemented by July, 2011. Hardware acquisition decisions will be in hand by September, 2011

Comments

The plan for completing the FY12 procurement plan by Sept. 2011 is reasonable and should not delay the FY12 milestones unless the choice is made to procure BG/Q.

It is important that the deadline of obtaining benchmark performance data on BG/Q by mid-July 2011 is adhered to, otherwise the project may be delayed.

Reliance on software development external to project is a risk and is increasing with more codes being examined for porting, to GPU's or alternatively to new supercomputer architectures. The LQCD SciDAC project has provided critical software for the project and the continuation of that effort under SciDAC3 is not assured.

Holmgren outlined performance comparisons of BG/Q vs Infiniband clusters vs GPU-accelerated clusters. There are several significant issues complicating the hardware selection process (slide 12) so he did not present a concrete plan at the present time to comment on. Nevertheless, the scope of what is being considered looks reasonable given the mix of science goals.

Holmgren mentioned that the "windfall" of increased productivity from GPUs will work out to increase the science beyond what was originally projected but there are likely to be plenty of needs that will emerge to meet the newly delivered capacity.

In response to questions from the reviewers, Mackenzie focused on some manpower issues which can seriously impact feasibility of long term goals. One of these issues is the rate of replacement of senior leaders as they leave the field.

Budgets are strained and the number of positions in physics departments are experiencing a downward pressure. In order to promote new faculty hires in this field, the reviewers believe that the LQCD community must push to expand its interactions with the experimental NP and HEP communities so that faculties and staffs become excited by the science potential and become willing to advocate for new hires in lattice gauge theory.

On a somewhat related front, the USQCD community does not seem to be well-connected with the Computer Science and Applied Mathematics communities though they have had ASCR funding under SciDAC-2. Given the widespread interest in GPUs a more intense engagement with these communities could be fruitful. If they do not succeed in gaining SciDAC-3 funding, they plan to ask the NP and HEP base programs for funding. In this connection they did not mention asking ASCR for funding. This may be interpreted as indicating that they, as physicists, plan to go-it-alone. This may be a short sighted strategy.

Since the LQCD-ext team presented a "process" rather than a "plan" for hardware acquisitions in FY2012, it is a not possible to evaluate their plan at this time. This delay is due to the stretched timetable for access to the BG/Q. The LQCD-ext team expects to run benchmarks on the BG/Q and other platforms over the summer and produce a detailed plan at that time. If the BG/Q is further delayed, it will be eliminated from the competitive acquisition plan. Overall, it seems that a more complete review of the budget and schedule awaits the presentation of the detailed plan.

Recommendations

The Federal Project managers should review the LQCD-ext acquisition plan as it develops over the July-September window. Further delays in the BG/Q schedule could remove it from the FY2012 hardware plan.

b. LQCD/ARRA

Findings

Operations costs are within budget and on schedule.

Comments

The proposed budget for the remaining years is mostly for personnel to maintain/service the present LQCD-ARRA machines and it appears reasonable.

Effectiveness of Management Structure and Responsiveness to past Recommendations

a. LQCD-ext

Findings

There was a suitable response to the one recommendation from the review last year (turnover in executive committee).

Comments

Overall, the management seems to be performing well under a variety of circumstances – some outside their control. They presented a coherent, consistent and reasonably complete set of summaries for the review.

Although the leadership has made substantial efforts to reach out to the HEP and NP communities via workshops, talks, collaborations, the use of open source software, etc., there may be additional means to attract more attention, interest, and participation in their programs. An outreach effort might include a central effort to communicate via printed materials, a collection of images, and materials for lattice speakers to facilitate their presentations. The three currently scheduled workshops all address nuclear physics issues, so hopefully HEP workshops will be organized soon.

The health and continuity of a field depends on cultivating new leaders with due attention to diversity and professional status. It appears worthwhile to accelerate the rate of building a youthful component on the executive committee.

Recommendations

An outreach effort should be considered. It might include a central effort to communicate via printed materials, a collection of images, and materials for lattice speakers to facilitate their presentations.

b. LQCD/ARRA

Findings

Comments

The management is performing well. They presented a coherent, consistent and reasonably complete set of summaries for the review.

APPENDIX A

Charge Letter to the LQCD-ext/ARRA Team

Dr. W. Boroski

LQCD Contractor Project Manager

Fermi National Laboratory

Mail Station: 127 (WH 7W)

P.O. Box 500
Batavia, IL 60510-0500

Dear Dr. Boroski:

The Department of Energy (DOE) Office of High Energy Physics and the Office of Nuclear Physics plan to conduct an Annual Progress Review of the Lattice Quantum Chromodynamics (LQCD-ext) Computing Project on May 10-11, 2011, at the Fermi National Accelerator Laboratory (FNAL). A review panel of experts in high energy physics, nuclear physics, project management and computer science is being convened for this task.

John Kogut of the Office of High Energy Physics is responsible for this review; he will be assisted by Helmut Marsiske of the Office of Nuclear Physics.

Each panel member will evaluate background material on the LQCD-ext project and attend all the presentations at the May 10-11 review. The focus of the 2011 LQCD-ext Annual Progress Review will be on understanding:

- The continued significance and relevance of the LQCD-ext project, with an emphasis on its impact on the experimental programs' support by the DOE Offices of High Energy Physics and Nuclear Physics;

- The progress toward scientific and technical milestones as presented in the project's IT Exhibit 300;

- The status of the technical design and proposed technical scope for FY 2011;

- The feasibility and completeness of the proposed budget and schedule;

- The effectiveness of the proposed management structure, and responsiveness to any recommendations from last year's review.

In addition, we will also be using this review to assess the plans for, and progress on, the construction and operation of the Thomas Jefferson National Accelerator Facility (TJNAF) LQCD cluster which is funded by the American Recovery and Reinvestment Act (ARRA) of 2009. We are consolidating these reviews because the LQCD ARRA cluster will be operated by the USQCD collaboration like any other hardware platform of the LQCD-ext project. However, since ARRA funding is subject to special scrutiny, it will receive a separate progress report. Chip Watson, the Contractor Project Manager for the LQCD ARRA cluster, should present the relevant information in the LQCD ARRA project documentation in order to allow the panel to evaluate the project according to the above charge elements.

Each panel member will be asked to review these aspects of the LQCD-ext and LQCD ARRA projects and write an individual report on his/her findings. These reports will be due at the DOE two weeks after completion of the review. John Kogut, the Federal Project Manager, will accumulate the reports and compose a final summary report based on the information in the letters.

The two days of the review will consist of presentations and executive sessions. The later half of the second day will include an executive session and preliminary report writing; a brief close-out will conclude the review. Preliminary findings, comments, and recommendations will be presented at the close-out. You should work with Chip Watson and John Kogut to generate an agenda which addresses the goals of the review.

Please designate a contact person at FNAL for the review panel members to contact regarding any logistics questions. Word processing, internet connection and secretarial assistance should be made available during the review. You should set up a web site for the review with relevant background information on LQCD-ext, links to the various LQCD-ext sites the collaboration has developed, and distribute relevant background and project materials to the panel at least two weeks prior to the review. Please coordinate these efforts with John Kogut so that the needs of the review panel are met.

We greatly appreciate your willingness to assist us in this review. We look forward to a very informative and stimulating review at FNAL.

Sincerely,

Michael Procario
Acting Associate Director
Office of High Energy Physics

Timothy Hallman
Associate Director
Office of Nuclear Physics

APPENDIX B

<u>Reviewers for LQCD-ext/ARRA Review, May 10-11, FNAL</u>

<u>Computer Scientists</u>

Doug Olson (LBNL, Computer Research Division, NERSC) dlolson@lbl.gov

Kevin Regimbal (High Performance Computing at PNNL) kevin.regimbal@pnl.gov

<u>HEP</u>

Michael Barnett (LBNL, PDG) rmbarnett@lbl.gov

John Rosner (UoC) rosner@bquark.uchicago.edu

<u>NP</u>

James Vary (Iowa State) jvary@iastate.edu

<u>List of DOE program managers</u>

J. Kogut (HEP, LQCD-ext Federal Project Director)

H. Marsiske (NP, LQCD/ARRA Project Director)

T. Barnes (NP)

APPENDIX C

https://indico.fnal.gov/conferenceTimeTable.py?confId=4148#201
10510

Agenda for May 10-11 Review at Fermilab

Tuesday 10 May 2011

Executive session - Wilson Hall, One North (08:30-09:15)
Welcome - Vicky White - Wilson Hall, One North (09:15-09:25)
- Conveners: White, Victoria
Logistics and Introductions - Bill Boroski - Wilson Hall, One North (09:25-09:30)
- Conveners: Mr. Boroski, William
LQCD-Ext Project Overview - Paul Mackenzie - Wilson Hall, One North (09:30-10:15)
- Conveners: Mackenzie, Paul
Science Talk 1: Lattice QCD and the Search for New Physics - Andreas Kronfeld - Wilson
Hall, One North (10:30-11:00)
- Conveners: Dr. Kronfeld, Andreas
Science Talk 2: Beyond the Standard Model Physics – Julius Kuti - Wilson Hall, One North
(11:00-11:40)
- Conveners: Kuti, Julius Kuti
Science Talk 3: Hadron Spectroscopy, Structure and Interactions – David Richards -
Wilson Hall, One North (11:40-12:20)
- Conveners: Dr. Richards, David
Science Talk 4: High Temperature/Density QCD – Frithjof Karsch - Wilson Hall, One North
(13:10-13:40)
- Conveners: Prof. Karsch, Frithjof
LQCD-Ext Project: Management and Performance - Bill Boroski - Wilson Hall, One North
(13:40-14:10)
- Conveners: Mr. Boroski, William

LQCD-ARRA Project: Management and Performance - Chip Watson - Wilson Hall, One North (14:10-14:30)
- Conveners: Watson, Chip
LQCD-ARRA Technical Performance - Chip Watson - Wilson Hall, One North (14:30-15:00)
- Conveners: Watson, Chip
LQCD-Ext Technical Performance & Remaining Plans for FY2010/2011 Deployments - Don Holmgren - Wilson Hall, One North (15:45-16:15)
- Conveners: Dr. Holmgren, Don
LQCD-EXT Proposed Selection Strategy for FY2012 Deployment - Don Holmgren - Wilson Hall, One North (16:15-17:00)
- Conveners: Dr. Holmgren, Don
Executive Session - Wilson Hall, One North (17:00-18:00)
Committee request for additional information - Committee/Project Leadership - Wilson Hall, One North (18:00-18:30)

Wednesday 11 May 2011

Executive Session - Wilson Hall, One North (08:30-09:00)
Committee questions and discussion - Wilson Hall, One North (09:00-10:00)
Executive Session / Preliminary Report Writing - Wilson Hall, One North (10:10-12:00)
Executive Session / Closeout Preparation - Wilson Hall, One North (13:00-14:00)
Closeout - Wilson Hall, One North (14:00-15:00)