Report from the May 2011 Review of Information Technology for the 12 GeV Era

Summary

A one day internal review of IT for the 12 GeV era of Jefferson Lab was held at the laboratory May 27, 2011 to evaluate the status and progress of IT systems for operations following the 12 GeV energy upgrade.

Presentations were given by each of the four halls and by all other major computing groups at the laboratory. Overall, computing systems are either adapting to small, incremental changes as they occur and thus will be ready for 12 GeV operations, or are under development and have sufficient time to get ready by 2014.

For this initial 12 GeV IT review, presentations were limited to half a day, and so observations were of necessity limited, but were still sufficient to give a good general view of the status today, and should serve as a useful point of reference for future progress reviews.

This report is organized by major software and/or hardware systems, and for each system the most important points of the charge are addressed to the extent possible. Each section has *observations* about what was presented (or in some cases known by the committee), followed by *findings* of importance, followed in some cases by *recommendations* to the relevant group and the laboratory.

Introduction

The Panel was charged to give an assessment of:

- the state of current software and systems developments
- plans for bringing all software to a suitable level of maturity, including software testing for correctness and performance
- plans for an evolution of computing, networking and storage capacity and performance to address the needs of detector simulation and data analysis
- plans for accelerator, light source, theory and lab operations
- the quality and effectiveness of the management of the major preparation efforts
- the resources, budget and staffing, to meet the needs of the program

Presentations were given for the following IT areas:

1. Data Acquisition
2. Hall A Simulation and Data Analysis
3. Hall B Simulation and Data Analysis
4. Hall C Simulation and Data Analysis
5. Hall D Simulation and Data Analysis
6. Scientific Computing Facilities and Lattice QCD Computing
7. Computer Networking and Infrastructure
8. Accelerator Controls
9. Free Electron Laser & future light source needs
10. Management Information Systems

The full charge and the schedule of talks and speakers is given in Appendix A.

General Comments

Findings

It was clear from the presentations that there is an enormous amount of work in progress and that all groups are actively pursuing development of their respective 12 GeV components. Overall, the reviewers saw no critical show-stoppers at this stage. All halls were able to present at least a rough outline of anticipated resource requirements for disk space, computing, and network bandwidth needs. Staffing requirements were less well defined and caught the attention of the reviewers. In the best case, available manpower was deemed "adequate". While a plausible high-level project plan was presented by all halls, it was not clear that the plan and staffing will ensure that all the pieces would come together in time to meet readiness goals.

The following general issues apply to all four experimental halls but were not explicitly addressed in their presentations. Note that this does not necessarily indicate a deficiency in the respective hall preparations, but may simply reflect ambiguity in the charge and time available for this particular review. Nevertheless, we encourage the halls to consider the following ideas and be able to clearly address the associated issues at the time of a formal review.

1. Designate a person to be in charge of the big picture, with the time and resources to maintain a sufficiently high-level perspective.
   - Speaking generally, the software development groups for all halls appeared somewhat understaffed. The larger groups (Halls B & D) appear quite distributed, relying on outside collaborators to develop important components. It was not clear that the organizational structures are sufficient to ensure all components will be delivered on time and integrated into a cohesive whole.
   - Halls A & C have a much lower software burden and are relying on evolving existing packages to support the new hardware. However, neither hall has resources specifically allocated to develop the necessary code. Given the overlap in requirements between the two halls, we recommend that a more formal joint effort be established (perhaps a joint-funded FTE) to ensure the analysis software is ready to meet the experimental requirements.
2. Identify the different pieces of the DAQ and analysis chain, identify those in charge of the individual components, and demonstrate a plan for integration and stress testing of the entire DAQ and analysis chain. This is driven by the following observations.
   - None of the halls have online or offline readiness reviews planned at this time.
   - The most complicated of these systems are in Halls B and D. While there is no question that the respective collaborations have the needed expertise, there is concern among the review committee that it is spread too thin.
   - Low level hardware readout and multi-crate integration tests using the new FPGA-based systems are to begin later this year. Stress-tests of the DAQ chain is planned to occur as each detector is instrumented and integrated into the respective Halls' readout systems. There is concern that this process may bottleneck due to ill-defined schedules and a shared demand on the limited resources available in the DAQ and Fast-Electronics group.

Recommendations

Establish a more formal joint effort to ensure that analysis software is ready to meet the experimental requirements.

Establish dates for stress tests of the entire readout through analysis chain (software and hardware) well ahead of detector commissioning (i.e., a data challenge).

The following sections address each area of IT and its corresponding presentation.

1. Data Acquisition and Computing Summary

Graham Heyes gave an overview of the performance requirements for the 12GeV data acquisition system (trigger rates, event sizes and data rates), plus an overview of the DAQ software architecture.

The DAQ group has recently grown by 2 FTE to address the "one deep" problem in a couple of key areas including electronics design. Two FTE on average from the group of seven are currently required for operations support, leaving a team of five for ongoing development. Collectively the group has considerable experience and competence. In addition, each hall has a DAQ expert who is responsible for first line support, and for (helping in) configuring the DAQ toolkit to create the hall's DAQ system.

In 2011 the group plans to test a full DAQ crate with a simulated trigger. In 2012 this will grow to a multi-crate test, perhaps including simulated crates, to test a multi-node event builder. Full scale testing would await the purchase and deployment of the actual DAQ system in the halls. This is expected to begin in 2014.

The software design is for a highly distributed system, with multiple components that can co-exist in a single box or be spread among multiple computers. Software improvements will include a programmable state machine sub-system. Some of the components for the DAQ system have been used to create the CLAS analysis framework CLARA, which will be used by Hall B.

The role of the DAQ group is to advise and support the goals established by the individual halls. It is the responsibility of each hall to maintain clear communication with this group regarding timetables, hardware, and commissioning needs. All four experimental halls plan to take advantage of the cutting-edge hardware being developed by the DAQ group.

This may be a double-edged sword: all halls have direct access to world-class experts on the new hardware; however, it also puts the DAQ/Front End personnel on the critical path for a large-scale roll out of brand new hardware across multiple halls. It is not clear that staffing levels are sufficient to accommodate unforeseen changes in commissioning schedules.

For example, commissioning of the FPGA based hardware (in particular) will require significant support from the DAQ and fast-electronics groups. The hardware is literally brand new and there are limited resources and equipment available for large-scale testing. Failure modes can be subtle, and hands-on experience with these devices is still scarce. Hall staff may not have the expertise or specialized hardware necessary for debugging firmware-related problems that will arise during commissioning. The present staffing plan for the DAQ group over the next several years allocates half the group to Hall B and D projects with the remaining staff supporting A and C along with other more generic projects. This plan requires Hall D support to ramp down in FY13 as Hall B comes online in order for the present staffing to be sufficient. It was noted that that if Hall B requires help earlier or Hall D slips there will be a crunch.

This presentation also included a roll up of the computing requirements for the 4 halls for a single year, FY15.

Findings

The planned system appears to have good scalability and potential performance to address the needs of all four halls. Adequate manpower is in place to bring the system to a robust state in a timely fashion.

Overall, DAQ and offline computing requirements, normalized to Moore's Law, will be less demanding than when 6 GeV began production running in 1995. Twenty years has yielded networks and computers 100 times faster, DAQ rates in 2015 will be about 50x higher, and requirements in 2018 will track Moore's Law another 4x. Similarly, data rates to tape from a single hall will pace what a single tape drive will be able to sustain. System scale will increase, with several times as many front end systems as 6 GeV CLAS currently has. Increased use of digital pipelines in the front end electronics will reduce dependence on hard real time systems (e.g. interrupt response times).

The required computing system size (and hence cost) is overestimated in that the extrapolation from 2010 systems to 2014 systems only assumes that computers will be 3x better, whereas Moore's Law would predict 6x. On the other

hand, the quoted $160/core is reflective of 2011 AMD cores, but most of the halls stated their performances in 2010 Intel cores which cost 2x as much but are higher in performance. These two factors nearly cancel each other out, but it would be good to clarify these matters, and in the future name the reference core by model number.

Disk cost of $500 per terabyte was based on 2010 numbers with no extrapolation, overestimating disk cost by at least a factor of 4 ($450K high). (Actual JLab purchases yielded $673 per formatted TB in 2009 and $350 in 2011.)

Tape costs were estimated at $60 per tape for a 3 TB tape. Today LTO-5 costs $55/tape, and within 3 years (the current pace for technology changes) the same should be true of LTO-6. However, this ignores the cost of a tape slot, which is $30 today and might not fall. Thus the price of tape is underestimated by 50% ($160K/year low).

There is separately an underestimation of future single server computer performance that results in a high priority being placed on parallel (multi-node) event building, a software complexity that might in the end not be necessary.

Recommendations

Establish a date for full scale multi-crate DAQ tests that is at least 6 months ahead of early detector commissioning.

2. Hall A

Ole Hansen presented event and data rates as a function of time, based on a preliminary running schedule. IT related infrastructure requirements are relatively modest through to the year 2016 (raw data rates on the order of 20—70 MB/sec).

Funding for necessary near-term upgrades to the Hall A online computing cluster is not well defined, but is assumed to come from the Hall A operations budget. Given the modest scope, this is likely sufficient for the near term needs.

Simulation and offline-analysis demands on the farm and related IT infrastructure are expected to be similarly modest for the next five years.

The most demanding experiment, SOLID/PVDIS, will require a very high 500 MB/s data rate, but not until 2021-2022, or 6 years after 12 GeV start. Since this is 4 doubling intervals (16x) beyond 2015, and the rates are only 4x higher, the bandwidth requirements will not be demanding. The trigger rates (500 KHz) will be much more so, possibly requiring new front end electronics. 2022 running includes plans for a level 3 trigger to reduce data rates to tape.

The existing Hall A analyzer (Podd) is a C++/ROOT-based framework that been in production use since 2003. It has a modern, modular design, and has a proven capability to quickly support new detector packages and features through custom software modules written by users. Recognized limitations include a lack of internal parallelization support, and only rudimentary support for JLab FADCs and similar pipe-lined devices with a “blocked” internal data format. It was not clear who would be providing the programming effort to address these issues. Due to similarities in detector hardware and DAQ design between Hall's A & C it seems natural that they “join forces” and develop a shared framework that benefits each. Both halls expressed interest in this approach, but no detailed plans exist at this time.

Software manpower was reported to be a challenge, with expertise thin (dependent on a few key people). They intend to collaborate where possible with Hall C, and will also be dependent upon user contributions and/or new staff.

Recommendations

Progress in software developments towards software maturity by 2014 should be watched. Where possible, firm commitments with MOUs from within the user community should be obtained in the coming year.

Halls A & C should establish a more formal arrangement for shared software development, clarifying the necessary funding and personnel requirements needed to reach their goals.

3. Hall B (CLAS)

Jerry Gilfoyle reported on developments in CLAS simulation and data reconstruction (analysis). While Hall B is comparable to Hall D in terms of DAQ sophistication, IT infrastructure requirements, and analysis chain complexity, the Hall B collaboration has the advantage of being able to rely on the CLAS6 experience and existing personnel to support the planning and execution of their upgrade. Their plans and progress generally appear to be in good shape, however manpower was noted as a potential concern (in particular, there is a fairly small group of core software developers).

The distributed nature of Hall B's 12 GeV analysis software (CLARA, based upon a service oriented architecture) has raised the idea of running services off-site that will require communication between these remote services and the JLab analysis farm and other IT resources. They noted cyber security issues when trying to run this with some components running at Jefferson Lab and other components running off site. It was not clear if the security, bandwidth, latency, uptime, and other associated issues have been fully considered.

There is $100k allocated to upgrading the Hall B Counting House infrastructure for the 12 GeV program, but a detailed plan has been deferred until there is a clearer view of what technologies (electronics, software, etc.) will be needed.

Findings

The committee felt that running CLARA with components distributed across a wide area network is not likely to be a particularly high performance choice, and so the cyber security issues raised are probably not important. No critical service should be located off-site in that it would make the analysis unnecessarily dependent upon WAN performance and robustness.

Computing requirements do not seem excessive or demanding, but the cost of disk and tape will be non-negligible, and the presented plan does not yet include keeping a duplicate of raw data (continuing in the way that CLAS has operated to date). CLAS12 anticipates 1 PB/year of raw data. In FY15 the cost of offline tape (without including the cost of a tape library slot) will be $15-$20 per TB, so this duplicate set will cost no more than $20K, a very low figure for risk mitigation.

It was not clear whether the cost of analyzing simulated events, and the storage of the output of that analysis, was included in the cost of simulation or reconstruction.

Recommendations

Progress in software developments towards software maturity by 2014 should be watched.

Plans should be made to keep a duplicate of all raw data.

4. Hall C

Stephen Wood reported on Hall C requirements and developments. As for Hall A, resource requirements are relatively modest, well under the Moore's Law growth rate. Growing luminosity and event rates will be offset by increasingly tight triggers as experience is gained.

Some large fraction of the existing code is in Fortran, which is problematic moving forward (harder to maintain and/or recruit expertise). Hall C plans to adopt the Hall A ROOT based analysis package as one way to address software staffing requirements.

Recommendations

Progress in software developments towards software maturity by 2014 should be watched.

5. Hall D (GlueX)

Mark Ito reported on Hall D requirements and developments. A high accepted trigger rate will yield the highest data rate to tape of the four halls, 3.2 PB / year compared with about 1 PB / year for CLAS. As in Hall C, growing luminosity (10x) and thus event rates will be offset by increasingly tight triggers as experience is gained. Simulation represented more than 50% of the computing power and 30% of the total storage (much larger than for CLAS). The computing requirements were acknowledged to be tentative, with large uncertainties. Plans do not include keeping a duplicate of raw data.

The GlueX Collaboration intends to use off site resources for simulation (as they already do), but no firm requirements were presented for how much of this data might eventually be stored at Jefferson Lab, and hence what networking requirements there might be. Off site resources are assumed to exist, but no formal agreements are yet in place committing a specific amount of such resources to GlueX.

Software status was presented for Geometry, Simulation, Reconstruction, Partial Wave Analysis, Calibration Database, Event Format, and Utilities. Simulation is currently running and is based upon GEANT3, with a transition to GEANT4 beginning. Reconstruction is based on the (mature) JANA package, which uses a separate thread to process a stream of events (essentially event level parallelism instead of file level parallelism). This package shows good scaling on multi-core boxes.

Partial Wave Analysis software is being developed under an NSF funded grant, with GPU and grid developments included. GlueX's plan is to use off-site resources for PWA.

Hall D software was originally a part of the JLab Baseline Improvement Activities (BIA) Project, and consequently has a detailed break down of needed tasks/components. Documentation was noted to be lagging.

Findings

Approximately 2 FTE of software effort is on staff at Jefferson Lab, compared to 5 FTE for CLAS. This is stated to be 40% of the effort needed, thus 10 FTE for the total, as compared to a total of 15 FTE for the better understood CLAS detector. This might reflect an easier to simulate and analyze event / detector, but could equally well represent insufficient manpower.

Computing requirements should be firmed up, and computing plans should include details on explicit assumptions for offsite computing and associated network requirements.

Hall D anticipates 3 PB/year of raw data. In FY15 the cost of offline tape (without including the cost of a tape library slot) will be $15-$20 per TB, so this duplicate set will cost no more than $60K, a very low figure, especially compared to the cost of running Hall D. It provides a backup for the occasional corrupted data file in addition to mitigating the risk of a complete loss of data.

Recommendations

Progress in software developments towards software maturity by 2014 should be carefully watched. Additional staffing at Jefferson Lab should be added to software development at least until it is clear that all components are identified and good estimates of time to complete are in hand. Where possible, firm commitments with MOUs from within the user community should be obtained in the coming year.

Details for a significant off-site simulation option should be fleshed out (compute, storage, manpower and funding), and where appropriate letters of intent obtained (formal MOUs could be obtained later). Details should include where simulated events will be stored and analyzed.

Plans should be made to keep a duplicate of all raw data.

6. Scientific Computing Facilities and Lattice QCD Computing

Sandy Philpott presented data on the current operations and plans for the scientific computing resources at the laboratory. This includes the experimental physics "farm" and associated disk cache and tape library, as well as the Lattice QCD high performance computing clusters and their disk systems.

The SCI (Scientific Computing Infrastructure) group deploys and operates computing resources for both the experimental physics program and for the LQCD program, and develops software in support of operations and in

support of LQCD physics. Computing system capacity and disk and tape capacity for the farm is driven by requirements from the Physics division.

SCI is today operating a computing infrastructure far larger than will be required by the 12 GeV experimental program. That high performance computing system is expected to be refreshed for LQCD twice in the coming 5 years, but will only grow modestly in physical size since older systems will be retired as new ones are added.

Current projections for the experimental physics program show that the computing requirements for all halls will remain much smaller than the theory computing requirements for LQCD (of order 5% today, perhaps 15% by 2015). Adequate space, power and cooling is available, but might be tight in FY13-14 (until additional chilled water becomes available).

The tape library is currently at 6800 slots, 6.8 PB, mostly LTO-4 tape. It can be expanded by an additional 5200 slots (5.2 PB). In the time frame of the 12 GeV upgrade, a transition to LTO-6 media (3.75x denser) can be done yielding a total capacity of 45 PBytes. A new 2nd library with only high density frames would be 60 PBytes in 2015, or 120 PBytes in 2018 with LTO-7 media. With the exception of when the 2nd library is installed, the cost of the capacity growth in incremental through the addition of new frames.

The current library has a bandwidth in or out of 1 GB/second (12 drives). This can be expanded in increments of 140 MB/s for LTO-5 drives, and later 210 MB/s per LTO-6 drive (anticipated before 2014).

The experimental physics program currently uses 75% of the tape library resources. LQCD uses most of the rest.

Plans are underway to migrate the farm to the high performance Lustre file system and to an Infiniband network as used by LQCD.

Findings

JLab IT support overall looks to be in good shape. The 12 GeV computing requirements are not within the formal 12 GeV budget scope, so funding for computing will come from the operations budget (and its growth). It is acknowledged that the recent pattern of 50-50-50 funding for computing requirements (compute+disk+tape library) will be unable to meet 12 GeV requirements at some point in the next few years, but it is not clear when.

Network, file system, and computing infrastructure for 12 GeV will be similar to the large scale clusters already in place. Power, cooling, and floor space plans appear to be in good shape, however this might need to be reviewed once the 12 GeV computing ramp for FY13-FY15 is better defined.

Recommendations

Centralized collection and archiving of experimental results and of supporting information (data provenance) should be formalized and actively supported by the Lab. This includes such things as machine readable data files for cross sections, structure functions, and other extracted observables, as well as associated meta-data such as analysis procedures, calibration data, slow controls data streams, etc. A central repository for such information would greatly simplify secondary analysis of prior datasets both to extract new physics, and to support future experiments.

Experiment-specific databases are becoming increasingly important for meta-data storage (configuration, calibration details, etc). Procedures and policy should be established to streamline Farm and offsite access to these databases.

Estimates of the computing capacity ramp from FY12 to FY16 should be made so that appropriate planning for power and cooling capacity upgrades can be done.

7. Computer Networking and Infrastructure

Andy Kowalski presented the status and plans for the CNI group. Key points of relevance to the experimental program included plans to increase help desk support and increase support for collaborative services, including web and video conferencing. Additional staffing yet to be hired will focus on support for smart personal devices.

On site networking will become more robust as the backbone moves to a ring architecture. An outdoor wifi mesh will support connectivity while moving between buildings, as well as allowing positioning of internet devices away from usual wired connections. WAN bandwidth can grow through an additional lambda in the MAN ring and in the path up to D.C., or through a second path to ESnet, but no firm commitments exist for bandwidth above the current 10 Gbit connection to ESnet.

Findings

Plans to upgrade WAN and increase bandwidth beyond a second 10 Gbit link are not well defined. Increases in distributed computing, video usage, off-site analysis farms, and other bandwidth-intensive applications may become a concern. The CNI group will need to track the WAN requirements of the halls as those become more concrete.

Video conferencing support at both the small working-group and large collaboration-meeting level was a current recognized weakness. There was general consensus that this was becoming an important and cost-effective collaboration tool that should become formally supported by the Lab.

8. Accelerator Controls

Matt Bickley presented the status and plans for the accelerator control system. Changes to the system primarily involve technology refreshes (from CAMAC to VME for example), with only modest growth in the total number of data channels. The network will evolve from 100 Mbit/s to 1 Gb/s hardware, particularly to support growing use of digitized video and waveforms. Storage will also grow, but at a rate below Moore's Law.

9. FEL and Photon Sciences

Wesley Moore covered the Free Electron Laser and Photon Sciences. While these are not specifically tied to the 12 GeV Nuclear Physics program, they could impact the context for 12 GeV. There are a number of potentially large projects that could emerge from this program, but almost certainly not in a timeframe to impact the initial 12 GeV running.

10. MIS

MIS systems was presented by Kari Heffner. The workload in the 12 GeV era is expected to be heavier than today due to increasing reliance upon automation and a larger workforce, and an increase of 1 FTE is planned to accommodate this.

Appendix A: Charge, Schedule and Panel Membership

Information Technology for the 12 GeV Era – Review

Friday, May 20, 2011

Jefferson Lab will conduct a one day internal review of most aspects of Information Technology that impact preparations for and initial running of the 12 GeV science program. This review is intended to get a good understanding of progress towards 12 GeV, and discover where there are areas that might need increased effort in the coming year. This review is also intended to prepare the laboratory for a full external review a year from now.

Jefferson Lab's 12 GeV upgrade will increase the demands on computing in many ways, from the addition of a fourth hall and thus new staff, an increase in the size of the accelerator complex, to a major increase in the rate at which data will be acquired and analyzed.

By the time 12 GeV beam turns on for physics, it is the desire of the laboratory to have all computing systems and software ready, so that the time from beam on target to physics journal articles is as short as possible. This will require an appropriate allocation of resources (both people and procurements) in the next several years, and an appropriate level of testing and validation prior to operations.

Charge to the Panel

We request that the review panel address the following points for IT in the 12 GeV era:

- An assessment of the state of current software and systems developments
- An assessment of planning for bringing all software to a suitable level of maturity, including software testing for correctness and performance
- An assessment of planning for an evolution of computing, networking and storage capacity and performance to address the needs of detector simulation and data analysis
- An assessment of the IT infrastructure to meet requirements including support for other areas, e.g. accelerator, light source, theory, operations
- An assessment of the quality and effectiveness of the management of the major efforts to prepare
- As assessment of the resources, budget and staffing, to meet the needs of the program

As this will be a one day review, we recognize the difficulty of delving deeply into each relevant computing project. Below is a proposed schedule showing the topics to be covered during the morning's presentations. The committee is asked to address as many of the points above for each of the topics presented as possible given the short time allocated for presentations. The review panel may request additional discussions with any of the presenters immediately after lunch if needed.

IT in the 12 GeV Era – Review Agenda

CEBAF Center – F224-225

8:30	Welcome and Charge – Roy Whitney
8:40	Data Acquisition – Graham Heyes
9:00	Hall A Simulation & Analysis – Ole Hansen
9:10	Hall B Simulation & Analysis – Jerry Gilfoyle
9:35	Hall C Simulation & Analysis – Steve Wood
9:45	Hall D Simulation & Analysis – Mark Ito
10:10	Physics Summary – Graham Heyes
10:20	Break
10:35	Accelerator Controls – Matt Bickley
10:55	Accelerator Physics, FEL, future Light Source needs – Wes Moore
11:15	Scientific Computing and Lattice QCD – Sandy Philpott
11:35	Management Information Systems – Kari Heffner
11:45	Computer Networking & Infrastructure – Andy Kowalski
12:15	-- lunch --
4:00	Closeout

Review Panel:

- Chip Watson, Chair (Jefferson Lab, Deputy CIO)
- Cortney Carpenter (W&M, CIO)
- Graham Drinkwater (CSC, Director, Maximo Services, ATG)
- Brad Sawatzky (Jefferson Lab, Hall C Staff Scientist)
- Richard Jones (UConn, GlueX)
- Karl Slifer (Univ New Hampshire, User Group BOD IT representative)