Contents

Chapter 9

Computing

9.1 Overview

GlueX will be the first Jefferson Laboratory experiment to generate petabyte scale data sets on an annual basis (One petabyte, 1 $PB = 10^{15}$ Bytes). In addition, the need to generate physics results in a timely fashion has been identified as a primary goal of the GlueX collaboration since its inception. For these reasons, a well-designed, modern, and efficient computing environment will clearly be crucial to the success of the experiment.

Currently, there are a number of particle physics projects world wide which also will produce very large data sets, and which will function with large dispersed collaborations. It seems quite reasonable, then, to expect that over the coming years, many new tools will be developed which will aid in effectively processing and managing these large volumes of data. As a collaboration, GlueX will undoubtedly make effective use of these tools, which will include such things as grid middle ware, distributed file systems, database management tools, visualization software, and collaborative tools.

Nonetheless, it also is clear that the GlueX collaboration will need to develop a suite of tools which are dedicated to this experiment. This will include data acquisition and trigger software, experiment monitoring and control software, data reduction tools, physics analysis software, and tools dedicated to the partial wave analysis (PWA) effort.

The rest of this chapter outlines in some detail the approach taken by the GlueX collaboration. First, a review the approaches taken by current experiments with similar computing requirements, along with the GlueX specific features and numerical constraints is given. Then an outline of the GlueX strategy to meet these demands, and also the specific tasks that will be divided

up among the collaboration members. Finally, a summary of computing milestone within the GLUEX collaboration will be presented. By keeping abreast of developments and new technologies that may be applicable to the GLUEX software environment, the collaboration will be able to carry the computing effort through from design to implementation and into the steady state running through a steady evolution of the system.

9.2 Background

In developing the GLUEX computing design, one can draw from two experiences, both of which are ongoing activities. These are the experiments using the CLAS detector in Hall B at JLab, and the CERN LHC experiments.

CLAS is of course particularly relevant, as it is also a multi-particle spectrometer arrangement at JLab, and is a good measure of how one may best use the existing infrastructure at the laboratory. An important difference, however, between CLAS and GLUEX is the volume of data acquired and analyzed. Based on the most recent numbers achieved in CLAS, the trigger rates and data volume are still a factor of three less than those projected for GLUEX. (See Sec. 9.2.2). It is clear then that the JLab computing infrastructure will need to be significantly upgraded in support of GLUEX.

As the CERN/LHC experiments, CMS and ATLAS, began to take shape in the 1990's, it was realized that these large international collaborations would be acquiring previously unheard of amounts of data. It was further realized that all members of the worldwide collaborations would need ready access to this data, and that recent advances in computing could in fact make this possible. CERN commissioned the MONARC[1] ("Models of Networked Analysis at Regional Centres for LHC Experiments") project in 1998, to study various configurations of distributed data analysis, based on "regional centers". The results of this study were published in 2000, and it was concluded that a multi-tier system of regional centers was the best solution to the problem.

CMS and ATLAS are now, in fact, following this model in their own computing efforts. Indeed, several large scale collaborations, mainly connecting physicists and computer scientists, have appeared in the U.S. and elsewhere, to realize this computing model for nuclear and particle physics in general. These include the DoE/SciDAC funded Particle Physics Data Grid [2] (PPDG), and the NSF/ITR funded Grid Physics Network [3] (GriPhyN) and International Virtual Data Grid Laboratory [4] (iVDGL). These collaborations are devoted to developing the tools needed to realize the promise of large scale distributed computing and data handling, as it pertains to nuclear and particle physics.

The PPDG, GriPhyN, and iVDGL projects are based on the concept of a "virtual data grid". This concept, which takes its name from the analogy with the public electrical utility network, aims to provide the user with an invisible layer of "middle ware" so that data sharing is carried out straightforwardly and quickly, regardless of the geographic separation of the actual physical data. Grid technology relies on the observation that the rate of increase of deployed network bandwidth is faster than the rate of increase in affordable computing power, and the assumption that these relative trends will continue for a number of years to come. This appears well founded based on historical trends [5], and are presumably driven by economics and the needs of society.

9.2.1 Special features of GlueX

There are important differences between GlueX and the CERN LHC experiments ATLAS and CMS, which can be traced to the primary physics goals. Events in ATLAS or CMS will be very complicated, with very large amounts of data per event, and these will consequently consume a lot of CPU time to reduce. By comparison, GlueX events will be simpler to disentangle. However, the subsequent analysis of GlueX events will be both computationally and data intensive, requiring sophisticated visualization and data handling tools, as large amounts of both "real" and Monte Carlo data are brought together in order to carry through an amplitude decomposition analysis.

The primary goal of GlueX is the systematic identification and categorization of short-lived meson states, unraveled from the raw, multi-particle reaction data using the techniques of "Partial Wave Analysis" (PWA). Achieving this goal requires simultaneous access to two large and independent data sets, namely the actual reduced experimental data and the simulated Monte Carlo data, each sorted for the particular multi-particle reaction(s) under consideration. It is quite probable that these data sets will be distributed physically over multiple locations, and that the access will be from other separated sites, associated with the group who has undertaken that particular analysis.

This not only impacts the structure of the data grid, but also implies that new analysis tools need to be developed. This especially includes visualization tools, as one searches for the appropriate combination of partial waves which best describe the reaction. That is, as one fits the parameters associated with a certain set of partial waves, some visual inspection mechanism is needed to evaluate how well the fit reproduces distributions in angles and invariant mass, for the many possible combinations. A universal set of tools is important in order to come to a more or less standard set of measures that would be applied by the analysis groups.

9.2.2 CPU, Storage, and Bandwidth Requirements

The GLUEX computing requirements are driven primarily by the projected data volume. GLUEX will use a multi-level triggering system, and it is projected that at the peak tagging rate, GLUEX will acquire 15,000 physics events per second which pass the Level 3 trigger requirement, or 1.5×10^{11} events in a live year, (assumed to be 10^7 seconds). The event size will be ≈ 5 kB. Consequently, the data acquisition system must handle 100 MB/sec, which corresponds to storing 1 PB of raw Level 3 data per year.

It is important that the Level 3 raw data be reconstructed somewhat faster than real time, for the purposes of monitoring the detector performance as well as the experiment setup. It takes on the order of 250 msec to process a multi-track event in a detector with complex geometry, on a standard workstation computer available in 2000. Using a conservative interpretation of Moore's Law, i.e. CPU speed doubling every two years, this is reduced to 15 msec by 2008, so 2.25×10^9 CPU·sec to process one year's running. A reasonable goal is to process these data in one-third the time it took to acquire it, i.e. 1.0×10^7 sec. Consequently, 225 circa 2008 CPU's will be required to process the raw data.

An accurate and detailed simulation will be critical for successful partial wave analysis. For any given reaction channel, one needs a greater number of simulated events than actual events, so that the result is not limited by the statistical precision of the generated sample. The goal will be to generate a factor of three times more simulated events than actual actual events for the data sample representing the final states for which one carries out a more detailed analysis. At the same time, one will, at least initially, be interested in analyzing a specific set of reaction channels. Taking both of these factors into account, and assuming a similar event size for reconstructed data, we estimate that the simulations will produce an additional 1 PB/year of simulated data.

Significant CPU resources will be required to generate the Monte Carlo sample. Ideally, one would generate only those events which in fact are accepted by the apparatus, correcting for the fraction of phase space assumed at the beginning. It is very difficult in practice, however, to achieve this optimal "importance sampling". A reasonable assumption is that only 1/2 of the events generated events will actually be accepted by the simulated experimental trigger. Consequently, one must generate a number of events

$$N_{\rm gen} = 2 \times (N_{\rm anal})$$

where $N_{\rm anal} = 1.5 \times 11^{11}$ is the number of (fully) analyzed hadronic events per year from the data stream. Consequently, $N_{\rm gen} = 3 \times 10^{11}$ events. Generating

Table 9.1: CPU, Storage, and Bandwidth Requirements for GlueX

Raw Data Processing	
Level 3 Data Rate	100 MB/sec
Raw data storage	1 PB/year
Reconstruction CPU's	450

Monte Carlo Data Processing	
Simulated data	1 PB/year
Generation CPU's	700

Monte Carlo events requires detailed simulation of various detector components, and then these events must pass through the same analysis program as the raw data. Thus, more CPU time is required per simulated event than for real data. A starting assumption is to use a factor of two, namely 30 msec, or 1.0×10^{10} CPU·sec for a year's worth of simulated data. to generate this data in one-half of a calendar year, ($\approx 1.5 \times 10^7$ sec), translates to approximately 700 circa 2008 CPU's necessary for generating and processing the Monte Carlo data set. Table 9.1 summarizes the CPU and storage requirements for computing in GlueX.

Physics analysis for GlueX will be carried out by a worldwide collaboration, which will require access to both the reconstructed data, as well as the processed Monte Carlo data. It is probable that the reconstructed data, simulated data, and as well the CPU's upon which the physics analysis is carried out, will physically reside at locations separate from one another, and also separate from the typical user. Sufficient bandwidth is necessary to connect the user to these resources in order to make appropriate use of the data grid.

9.3 Computing Strategy

In Fig. 9.1, we show a conceptual plan of the GlueX data processing and computing environment. In the following sections, we will discuss the important features of this plan.

9.3.1 Jefferson Lab Computing Resources

Clearly, the nature of this experiment dictates that a significant computing infrastructure must exist at Jefferson Lab. As shown in Fig. 9.1, the computing facilities at JLab will coordinate the experiment monitoring and control, data acquisition, Level 3 raw data storage, slow controls monitoring, and data reduction.

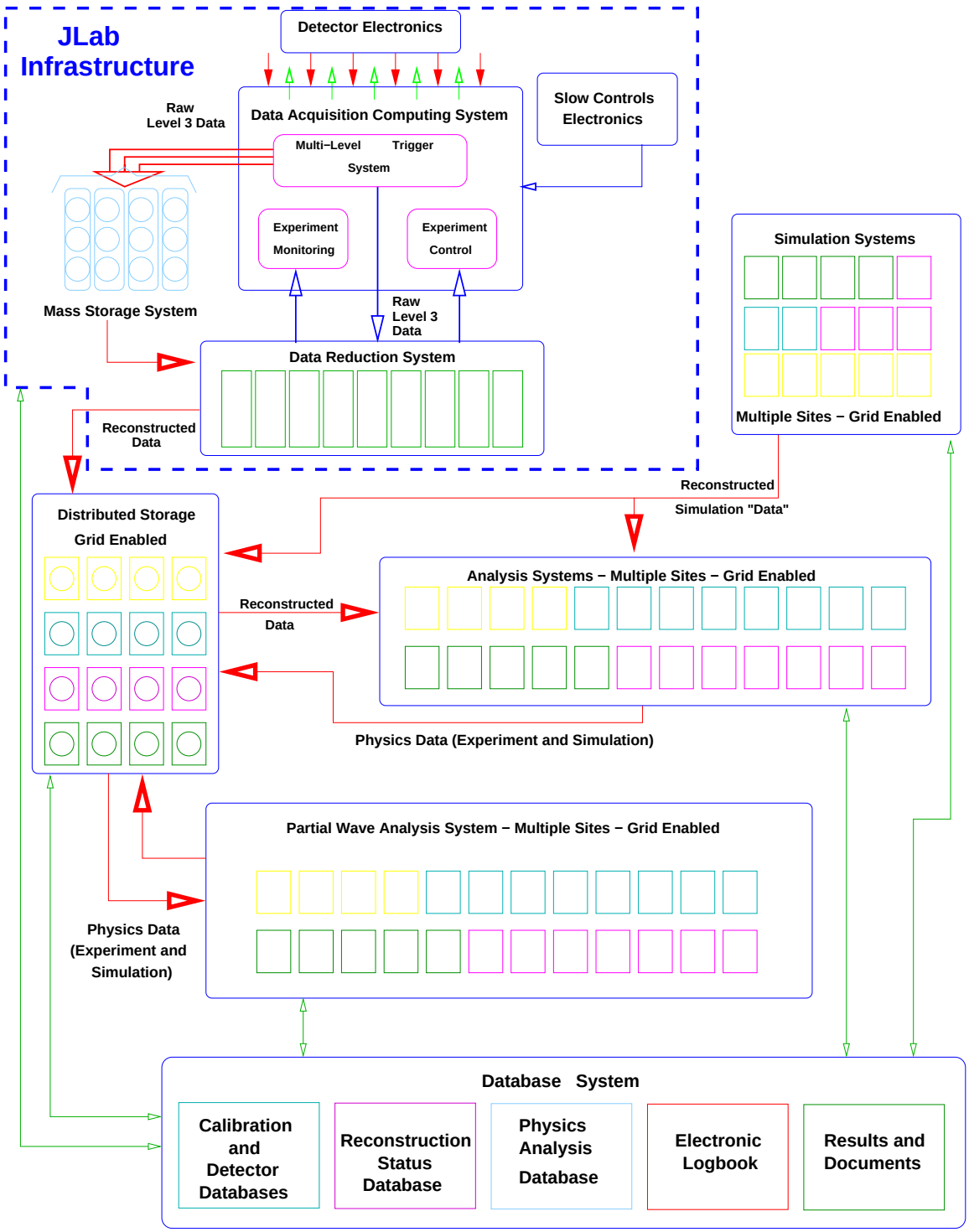

Figure 9.1: The GlueX Computing Environment

Data Storage

Currently, at JLab, raw data from experiments are written to tapes housed in a tape silo in the JLab computer center, and this is one option that we have considered for the GlueX Level 3 raw data. Current tape speeds are 30 MB/sec onto 200GB cassettes, and should exceed 100 MB/s onto 1 TB cassettes when GlueX data taking begins. At a data rate of 100 MB/s, and accounting for tape mount times and redundancy, GlueX would need three to four tape drives dedicated to on-line data recording.

A tape silo typically holds 6000 tapes, or 6000 TB at 1 TB/tape. Thus, JLab would need to purchase one tape silo to store GlueX raw and processed data, and would need adequate tape archive and storage facilities. Tape costs should be less than for CLAS, as much of CLAS data was written to low capacity tapes, and tape costs remain constant independent of capacity.

One should also note that at the present time, the relative prices of tape and disk storage are scaling in such a way that by the time GlueX is in the data taking phase, it may be more practical to store the raw data directly on disk. Even with current RAID technology, high reliability disk storage may be achieved with mirroring or optical archiving techniques.

It is also important to note that while not explicitly shown in Fig. 9.1, the reconstructed data will almost certainly be stored primarily at JLab, and will therefore comprise a significant portion of the grid-enabled mass storage system.

Data Acquisition and Interface to Electronics

The projected raw data rate into the Level 3 trigger system from the detector is 1 GB/sec (5 kB/Event $\times$ 200 kEvent/s). Our goal is a reduction factor of 10 in the Level 3 trigger, resulting in a Level 3 recorded raw data rate of up to 100 MB/sec. There can be no software, or otherwise computing related, impediments to this goal. The computer center staff, working closely with the data-acquisition group, will be responsible for assembling a system that allows direct transfer of the data from the acquisition electronics to the mass storage media, while providing for adequate experiment monitoring and control. It must also provide a natural interface to the data reduction software, which would be used on line for at least a subset of the monitoring activities.

Speed is a premium, and this software will be dedicated to on-site operation at JLab. Consequently, there are few constraints on the software model used to build it. However, we should also keep in mind that we must have the ability for detector and hardware experts located remotely to monitor detector

performance and provide diagnostic information.

Experiment Calibration and Detector Monitoring

The calibration database will be an important input to both the raw data reduction and to the event simulation. Good indexing will be necessary to track any changes in the detector or its performance over time, and correlate that to analysis and simulation. The database records themselves will be used to monitor detector performance over time, including both long term drifts as well as failure modes.

The calibration procedure will also involve the use of a set of raw data dedicated to detector calibration. It is important that these data have high availability, and thus the calibration data sets would be replicated at multiple sites to achieve this.

Data Reduction: Reconstruction from Raw Data

Event reconstruction will be a CPU intensive task. It will include, for example, accurate particle tracking through the (approximately) solenoidal field to determine the momentum vectors of the individual particles; the event vertex and any secondary vertexes; conversion of time-of-flight and Čerenkov information to particle identification confidences; identification of electrons and photons from the electromagnetic calorimeters; and determination of the corresponding tagging event, with confidences.

The computing hardware requirements for the data reduction facilities at JLab were discussed in detail in the previous section, with the principal motivation being that the Level 3 raw data be reconstructed in approximately real time. To reiterate, it is anticipated that we will require 450 Year 2008 CPU's for this task.

We require this code to be portable, as the same code used for reconstruction of the raw data will be used for reconstruction of the simulated data. These tasks will almost certainly be carried out at different sites with different computers.

Other Tasks

Jefferson Lab needs to extend their high speed network to Hall D, and to establish specific computing resources to acquire and process the raw data from GlueX. This includes storage capacity for the raw data, CPU power to reduce it, and the ability to store the resulting reduced data. A high speed network, capable of sustaining the necessary bandwidth to support the

connections to off-site analysis and simulation centers, must be established at the laboratory.

9.3.2 Off-site Computing Infrastructure

Again referring to Fig. 9.1, the distributed computing facilities associated with GlueX will comprise both distributed mass storage, as well as computational resources devoted to physics analysis and simulation. It is envisioned that the facilities located at these distributed centers will be matched to the specific data-intensive activities, such as detector calibrations, Monte Carlo simulation, and the various stages of physics analysis that are being pursued by the groups located at these institutions. The storage capacity that needs to exist at a center will depend on the specific activity it represents. For example, a typical analysis of 100 GB of reconstructed data may require 300 GB Monte Carlo of simulated data to be loaded and stored at the center simultaneously.

Distributed Data Storage Considerations

The distributed mass storage system (data grid) which we envision is a powerful concept, but it relies on both high speed networks between the centers, as well as networks which are reliable and available. For the purposes of this discussion, we refer to the OC standard for network bandwidth; OC-1 = 51.85 Mbit/second and OC-N = N $\times$ OC-1 rate. Of critical importance will be the connection to JLab, which will be dispensing the reconstructed data to possibly several analysis sites at any one time; and the Monte Carlo center, which would dispense simulated data at about four times the rate of reconstructed data. For example, it takes approximately two days to transfer a 400 GB simulation data set at 20Mbits/sec (13% of an OC-3 connection). With several analysis running at once, it seems clear that we would saturate the currently available OC-3 bandwidth. *It is likely that we would need an OC-24 (1244 Mbits/second) or better connection between the Monte Carlo simulation center, and the physics analysis sites.* Even with high speed networks coming into the universities, it can often be problematic to move the data through the universities' internal networks. However, the few examples that we have within the GlueX collaboration have found that the university computer centers have been very interested in resolving these problems. Nevertheless, this may not always be true, particularly for smaller universities, where the "last mile" problem may still be an issue.

Physics and Detector Simulations

An accurate Monte Carlo simulation will be crucial to the success of the detailed partial wave analysis that are the goal of GlueX. This will begin with some physical model for the final states to be studied, followed by "swimming" charged particles through the (nearly) solenoidal magnetic field and then simulating the signal on the various detector components. This will be a CPU intensive task, which will then be followed by the event reconstruction code. The collaboration needs to establish the Monte Carlo farm for generation, reduction, and storage of the simulated data sets. These are critical sites, and the connection bandwidth to JLab and to other users must be realized.

It is likely that event generation will take place at either one physical site, or perhaps a small number of sites, so the portability of the code will not be a large constraint. However, this activity may well benefit from distributed computing, and in that sense, portable code may prove to be a significant asset.

Partial Wave Analysis: Methodology and implementation

The PWA code must be flexible enough to allow for a large number of different final states within the same framework. Further, it is a CPU intensive task, involving the minimization of a complicated, multi-parameter function, as part of the extended maximum likelihood fit. New visualization tools, which need to be interfaced to the raw and simulated data sets through the data grid, should be developed to help assess the degree to which the assumed wave set describes that data.

The code will run on many different computing systems, depending on which collaborator may be using it at any one time. Consequently, the portability of the running code will be important.

Record-keeping and Collaboration Interface

One key to operating an experiment with an active worldwide collaboration is to keep records (including the experiment "logbook") accessible to anyone in the collaboration at any one time. Such a portal can also be used as the basis for virtual meetings over the Internet, and a deposit for presentation materials, publications, internal notes, and other important avenues for information dissemination, both external and internal to the collaboration.

9.3.3 Software Model

An object-oriented framework will be in established for all software that becomes an integral part of the GlueX computing environment. The use of design patterns and other best practices from object-oriented design will encourage maintainable code. Unit testing, static analysis, and similar light-weight additions to the process will encourage a scalable software development and testing cycle.

Grid-based computing environments are in large part described by protocols, interfaces, and schema's. Software components built upon XML interfaces and metadata fit into the notion by providing collaboration access to analysis, simulation, and visualization tools as "web services", a popular theme in current grid computing initiatives. Some work in this direction has already begun at Jefferson Lab [6, 7].

So long as the collaboration adheres to the above framework, it is not critical to decide on any specific programming language. Indeed, a language-agnostic approach will encourage the development of interface compliant, loosely coupled software components. Dependence on legacy code will be limited to the extent that XML interfaces exist (or are written by proponents) which hide the details of the code underneath.

A software distribution and revision control system needs to be set up and maintained. The system should be designed from the outset to not only include code for various purposes, but also documentation, dissemination materials, log books, and other archival information.

9.4 Organization

Clearly the successful development and implementation the of the GlueX computing environment will require extensive coordination between both the GlueX collaboration and the JLab computing center and data-acquisition groups. Crucial to this is both the dynamic definition and the completion of various computing milestones. Figure 9.2 shows the currently identified milestones that need to be achieved to meet the computing requirements for GlueX. Note that Monte Carlo simulations are already in progress and much progress has been made to date in developing the simulation code for detector, beam line, and trigger simulations. In addition, the collaboration is aggressively pursuing the development of the PWA codes and tools which will be crucial in extracting physics results from the data. While it is certainly true that the computing power per dollar invested continues to increase at a dra-

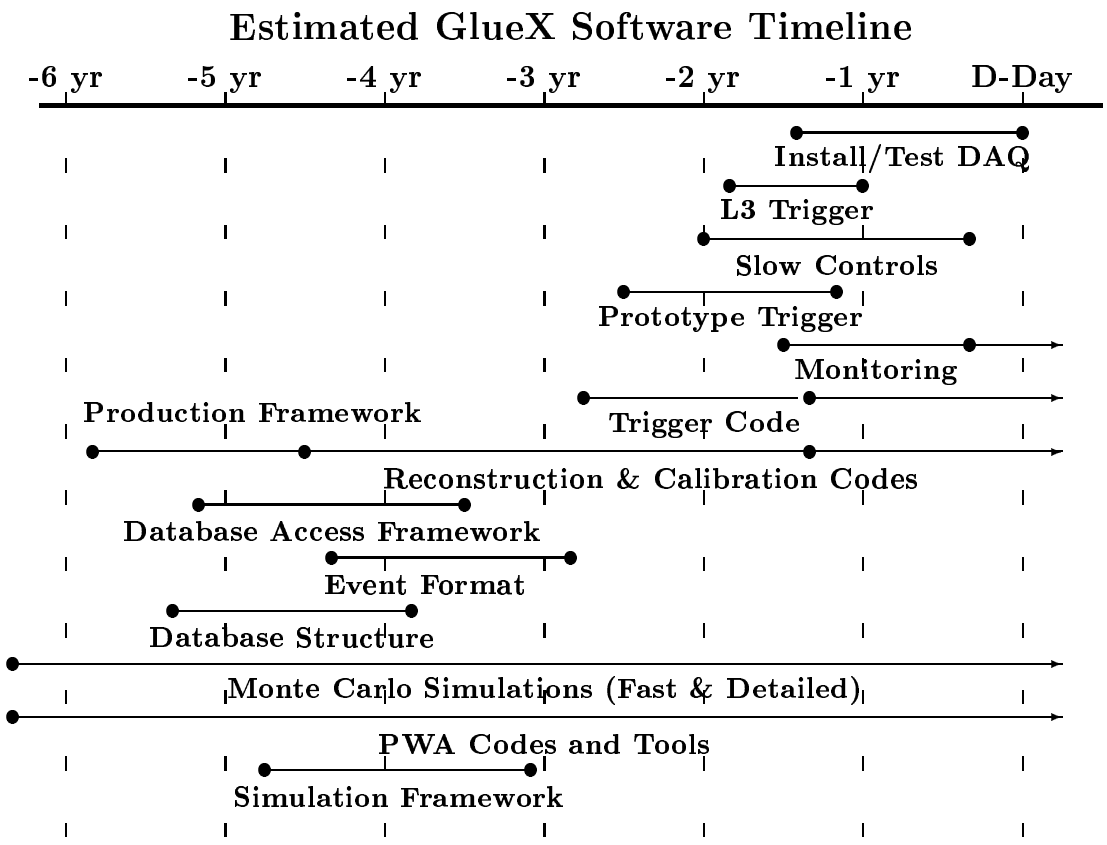

Figure 9.2: Milestones for GlueX computing projects and tasks.

matic rate, it is not a viable option for the collaboration to wait until the last possible moment to purchase the necessary computing hardware infrastructure. The reason is that a large fraction of the software that will be needed to carry out the project must be developed by the collaboration. One cannot simply use a set of "canned" packages. In order to develop this software, as well as the associated physics analysis techniques, the computing infrastructure, both at JLab and at the university centers, must be at least partially in place well ahead of time. Thus, this infrastructure must be ramped up in the upcoming years. Indeed, a segment of the collaboration is in the process of securing funds to develop a dedicated center for PWA studies (Indiana University). As well, integration of several of the already existing and future computing clusters for initial grid computing studies (Carnegie Mellon, Connecticut, Indiana, JLab Regina) will be tested in the coming months.

List of Figures

List of Tables

Bibliography

[1] `http://monarc.web.cern.ch/MONARC/`.

[2] `http://www.ppdg.net/`.

[3] `http://www.griphyn.org/`.

[4] `http://www.ivdgl.org/`.

[5] Jim Gray and Prashant Shenoy. Rules of Thumb in Data Engineering. In *Proceedings of the 16th International Conference on Data Engineering*, pages 3–12, 2000. Microsoft Research Technical Report MS-TR-99-100.

[6] Chip Watson. Web Services Data Grid Architecture, March 2002. PPDG documentation.

[7] Ian Bird, *et al.* Common Storage Resource Manager Operations version 1.0, October 2001. PPDG documentation. http://www.ppdg.net/docs/documents_and_information.htm#Reports.