The Hall-D/GlueX Response to the 12-GeV Software Review

GlueX-doc-2097

Curtis A. Meyer
Carnegie Mellon University

October 2012

Abstract

This document reports on the actions taken by the GlueX Collaboration in response to the 12-GeV Software Review conducted in early June of 2012.

1 Introduction

The 12-GeV Software and Computing Review took place on June 7 and 8 of 2012 and the final report of the review committee came out in September of that year. Most of the points brought out in the report were also discussed during the close out part of the review, and because of this, the GlueX collaboration was able to focus effort where it was needed very soon after the review. While many of the actions taken were planned, the review allowed us to more clearly focus on what was needed to move forward. In terms of the collaboration, the feeling was that there were two main areas of focus coming out of the review. The first was ramping up to successively larger data challenges following our first effort in 2011. The second was to lay the groundwork to be able to calibrate the detector elements once they are installed in Hall D.

The response to the first point was addressed by setting up a "data challenge working group" chaired by the Offline Coordinator (Mark Ito) and the Deputy Spokesperson (Matt Shepherd). This working group meets on alternate Mondays and interfaces with both the regular Offline Working Group, the Physics Working Group and the PID Upgrade Working Group.

To respond to the calibration point, each of the Detector Working Groups was charged with developing plans within their own groups. Coordination would be handled at the biweekly Collaboration video conference.

2 Actions of the Collaboration Since the Review

After the software review, the collaboration quickly realized that most of the remaining issues with reconstructing data would be resolved by moving forward with the "next data challenge" as quickly as possible. It also became clear that within the collaboration, there were two competing groups for the output of such a data challenge.

- The physics analysis group.
- The data processing group.

The physics analysis group wants very large background Monte Carlo sample to develop analysis tools and procedures. In addition, they want to study performance and analysis issues associate with an upgrade to the GlueX PID systems.

The data processing group is interested in developing the tools and procedures to allow us to go from the large online volume of Hall-D data into the JLab tape silos. This is then followed by the large processing associated with reading the data in from the tape silos and then producing data summary files for physics analysis.

The collaboration views both of these as vey important activities and plans to proceed on them in parallel, with overlap where it is useful. To first order, the physics analysis group will try to leverage resources on the Open Science Grid to produce the very large data sets they need. This will allow us to develop our ability to utilize the OSG, as well as moving the files around between sites for analysis. The data processing group will focus on using the infrastructure at Jefferson Lab with the goal of trying to push it as hard as they can, and testing new links and computers as they become available.

2.1 Data Formats

One of the things that we learned from our first data challenge in 2011 was that the "HDDM" output format from reconstruction we had in place at the time of the review was large and unwieldy. Its size was severely hampering the analysis efforts both because we could not store enough of it on disk, and second because the time it took to read it back in became prohibitive. At the time of the software review, the DST format we were using doubled the raw data size. However, based on experience from other experiments, we felt that the actual number should be about 10% of the raw data size. However, we note that this 10% was simply the fraction of good physics events in the data stream.

At the time of the review, we also made estimates of raw event size and found that 15 kB per event seemed reasonable. While we had made our best estimate of the raw data size, there is certainly a concern within the collaboration that we need to have a better estimate of this number. To to this, it will be necessary to be write Monte Carlo date in the real raw data format where this data includes all of our current estimates of noise in the experiment. This will give us both a better estimate for the size of the raw data, and allow for exercises such as pushing this data from Hall-D to the tape silos, and then through reconstruction. Thus more precisely defining these data formats rose to the top of our list.

Within GlueX/Hall-D, there are several levels of data formats which we will describe in detail below. For the raw data, it comes out of the data acquisition system in an "Online Data Format". For simulation, the date are stored in a "Simulated Event Storage Format" that can be converted into the online format if desired. Both of these formats go through the event reconstruction and are written in a very compact "Reconstructed Event Storage Format". Finally, users carrying out a physics analysis will use the reconstructed format as input, and most likely write in (possibly) analysis-specific root files.

2.1.1 Online Data Format

The Online Data Format is how we expect to get events from the GlueX/Hall-D data acquisition system. At present, we have just received the code that will produce the blocked events from the hardware and are in the process of working to put it in the Online Data Format. A goal of this conversion is to make the raw data as compact as possible, hence as little overhead as possible. Data in this format is expected to dominate the storage footprint of GlueX and effort is currently underway to understand more precisely how large we expect this to be. While results are still very preliminary, the 15 kB per event estimate is probably within 35% of the actual size.

2.1.2 Simulated Event Storage Format — SEST

The "smeared" output of the GEANT-based event simulation is stored in our Simulated Event Storage Format, SEST. It contains the simulated hitView structure properly factorized into pure hits and pure truth tags plus a hook for appending the reconstructed data structure (see below for REST) for people who want both hits and reconstructed results for studies. It also contains a small handful of other extended truth tags that have been introduced over the years to facilitate studies. A tool will be available to strip arbitrary SEST files down to bare hits. Such "bare hits" files will also be valid SEST files, but missing the information contained in the extra structures (in particular truth information). The danahddm plugin can read any SEST file and properly handle situations where various types of information are available or not, according to the factory schema. In particular, it will be possible to convert SEST files into the Online Data Format for testing and studies.

2.1.3 Reconstructed Event Storage Format — REST

For the DST formats, we examined the possibility of using the CERN package root, but when we looked in detail at how the structures were handled, particularly with respect to threaded code, it was quickly realized that we would need to write an entirely new package from scratch with little real benefit from the CERN code. It was recognized that at the next step in the analysis, where we moved to MicroDSTs that the root package would be a crucial step, but it was concluded that the DST level was too early in the chain.

Based on this, a format built on the Hall D Data Model (HDDM) was developed known as the REconstructed STorage format (REST). The Online Data and the SEST data becomes

REST data after it has been passed through the reconstruction code, dana, and the hits information and simulation-truth elements are suppressed. This REST format fully contains all information needed for physics analysis, but will yield a DST data footprint about 1% of the size of our raw data files. We gain the factor of 10 noted above for good events, but then we further reduce the information by another factor of 10. This huge reduction over our earlier estimates and opens the possibility of a much more disk-based physics analysis than had been anticipated and will likely substantially reduce the tape storage needs that had been presented at the Software Review. In particular, the tape needs for Monte Carlo and DST are reduces by about a factor of 10.

2.2 Processing of Data

As estimated for the Software Review, full-blown GlueX running will produce on the order of petabytes (PBs) of data per year. These data need to be moved from the Hall-D to the computer center at Jefferson Lab and then written to tape. When calibrations are completed, the data need to be read back in from tape and reconstructed to produce "data summary tapes". We refer to this process as "production" and note that at the time of the review, there was some degree of concern within the GlueX collaboration that the batch-farm systems currently in place at Jefferson Lab would not be able to handle this large scale problem without enhancements. In particular, the reconstruction phase of the production appeared to be lacking the production tools necessary to track jobs and automatically recover from failures.

Experience with other JLab experiments have shown that a great deal of "baby sitting" is necessary to push large scale reconstruction through the JLab farms. Tapes need to be staged, jobs need to be queued, and when a job fails, it needs to be recognized and restarted. Finally, the information produced during the production needs to be stored in some sort of data base.

Our initial approach was to look at the products that were developed to support the LHC experiments at CERN, but our investigations quickly showed that these were very heavily skewed towards grid-based solutions at all levels. These are exactly what are needed for the LHC based on the reliance on the grid. For GlueX this is not true, the grid is not a part of our production environment. Trying to extract the useful features from the existing packages for the non-grid-based use of GlueX was as least as much, and probably more work than evolving existing tools at JLab to try and handle the task. Thus, to first order it was decided that we would try to push at the existing infrastructure using the existing of tools with modest enhancements.

This has developed into our "Mini Data Challenge" where we try to load the JLab farm up with 1000 jobs at once, and then study where the bottle necks and issues are. This work started in late August and is continuing at the moment.

2.3 Large Data Challenges

2.3.1 The Indiana Challenge

The first GlueX Data Challenge was carried out in 2011 at Indiana University in using the Open Science Grid. This challenge simulated on the order of ten hours of GlueX beam and looked for a large cross section channel in the data set. It was found that at the time of this initial work, the reconstruction tools were not fully up to the task of the reconstruction. This led to a great deal of focussed work to improve the analysis.

2.3.2 The Two-billion Event Data Challenge

In moving forward, the collaboration sees a strong need for additional data challenges. The first was limited by the amount of data we could store, and our ability to move it around. All of these size issues have now been removed by the development of the REST format. As we move forward, the next big physics demand by the collaboration has to do with the design of a PID system for GlueX. The studies to date have been limited not by size, but available computer time at outside institutions. In addition, the collaboration would like to move beyond the mini data challenges being carried out at Jefferson lab into a more production oriented challenge. We feel we are now ready for our "two-billion event challenge" which will simulate about 10% of a year of low-intensity running. The goal is to produce 2×10^9 PYTHIA events in the range of 8.4 to 9.0 GeV incident photons. These PYTHIA events represent the full hadronic cross section and minus detailed physics, should simulate all the final states that we will see in GlueX. They are crucial to studies as one then "mixes" physics samples at the appropriate levels into these samples, and then tries to extract them from the data.

At a more basic level, one can study how well a given reaction channel in the PYTHIA sample can be reconstructed. What is the efficiency of extracting that channel from the sample, and how many events that are not really the signal are reconstructed as the signal. Such data will let us understand at what level we can extract exotic signals from the data with our current software and allow the graduate students looking at these to develop fairly sophisticated analyses well in advance of date.

The resources needed to carry out the two-billion event challenge will be a combination of the JLab computer farms as well as the Open Science Grid and university computer clusters. And we hope to generate this sample very soon.

2.3.3 The Online Data Challenge

Moving beyond this challenge, we foresee a challenges based on data in the online data format, potentially trying to move it from the Hall-D area to the tape silo, and then reconstructing it. This challenge would not be for physics, but rather the more fully test the infrastructure for moving a reconstructing data.

2.3.4 Larger Data Challenges

Ultimately, we would like to move up to an even larger physics challenge that would involve data sets corresponding to about a year of running.

2.4 Analysis Tools

An single integrated set of analysis tools which allow a user to select events from a DST file and use them correctly in an analysis are a **CRUCIAL** element of a successful experiment. They insure a needed level of quality control in all analyses and provide a natural wave to steadily evolve analysis procedures through the life of the experiment and make sure that the best vetted tools are always used. The first full version of this software has been developed for GlueX/Hall-D has been developed and deployed since the Software Review. The tools are starting to be used by those doing analysis. This package supplies an interface to get data (uniquely) from the REST format. It also has kinematic fitting and other tools for making event selection, and studying reactions.

2.5 Software Validation and Performance Profiling

The GlueX collaboration has been running nightly builds of its software on multiple platforms for over a year. In addition, we have been running reconstruction jobs on standard data sets several times per week for most of the same time period. These allow us to quickly spot serious problems in the code, but they are not particularly good at monitoring performance issues such as the actual speed of the code, and the memory footprint associated with a reconstruction job.

Other actions that have been taken on a rather irregular basis are profiling the codes performance to try and identify "hot spots" and running code validation tools to look for areas that might impact performance. To move beyond the irregular running of these, we have started to investigate a number of tools that could be run at the time we release a new software version (about once per month). Current work on this suggests that the tools look potentially useful, but in their current form, they report huge numbers of issues that when tracked down, turn out not to be a problem. Because of this very large number of false positives, it is felt that we need to work on methods to suppress these to actually make the validation tools useful. This is work in progress.

2.6 Calibration Issues

Calibration and alignment of the Gluex detector and the Hall-D tagger system is handled in the relevant hardware subgroups. While activity has started in these groups, much of the effort is awaiting completion of detectors. This will become a larger part of our activity as more detector elements are completed and delivered.

3 Report on the Software Review

In this section, we quote the relevant sections of the software review report where we feel some action needs to be taken. We then summarize the action (to date) taken by the collaboration as discussed in the previous section of this report.

3.1 Executive Summary

3.1.1 Summary Comments

Plans delineating specific testing programs and milestones to measure progress towards at-scale production running were mixed. Hall efforts included well developed plans for successive Data Challenges progressively scaling up testing of the software and computing systems; we recommend such plans be made general practice, and make full use of JLabs available computing resources for realistic scaling tests. Such plans will be key not only to software and computing readiness but to a smooth transition from development to operations. When data taking begins, computing operations at realistic data taking levels should not be a new experience.

> As noted in Section 2.3 we are starting a two-billion-event challenge that we estimate to be about 20% of the $10^7\,\gamma/s$ year of running. This challenge requires a combination of the OSG, JLab compute farm and outside users. We expect to follow this up (once the online data format is defined) with a similar scale test involving the tape farms and reconstruction from there.

The Committee heard little on data management plans despite the fact that this will be an important part of the infrastructure. Also plans for workload management were unevenly developed among the Halls. These are particularly important for the more data and processing intensive programs of Halls B and D. Plans in these areas should be carefully developed, and are good candidates for common solutions.

> As noted in Sections 2.1 we have now defined most of our data formats.Given these, we are running through a series of mini data challenges (Section 2.2) to understand the issues associated with the data handling.

3.2 General Comments and Recommendations

3.2.1 Recommendations

Presentations in future reviews should cover end user utilization of and experience with the software in more detail. Talks from end users on usage experience with the software and analysis infrastructure would be beneficial.

With the development of our analysis package as described in Section 2.4, we expect to have an ever increasing number of users carrying out analyses and improving on these common tools. We do not anticipate that this will be a problem in the future.

3.3 Data Acquisition

3.3.1 Recommendations

Once a modest all-way data path is established, plan a mock data challenge with fake data, in particular with nominal data rates from GlueX.

As discussed in Section 2.3, this is planned. It is still waiting for the online data format has nearly been defined with the first tool to carry out the conversion just released. Work now needs to proceed to substantially speed the tool up from its current 4 Hz rate, and to prune the output of redundant and and useless information to more accurately reflect what will come off GlueX. We also need to develop analysis tools to read in this data. FInally, we note that the network infrastructure is not yet in place to allow us to carry out this exercise at full scale, but the last switches are scheduled to be installed soon.

3.4 Experimental Halls - General

3.4.1 Findings

There is good attention to multi-threading/multi-processing support to accommodate new computing architectures. Event-level, but not subevent-level (or at least not before considering event level) parallelism is being pursued, consistent with trends in HENP.

3.4.2 Recommendations

Evaluate standard code evaluation tools, such as valgrind, clangs scan-build, cppcheck, Gooda, ... for inclusion in the software development cycle. We suggest looking at an Insure++ license as well.

As discussed in Section 2.5, we do some of these tests irregularly, but we are now looking at the issues involved in making them part of our release cycle. The biggest issue at the moment is the large number of "false positives" from the tools and filtering them out.

Run a code validation suite such as valgrind as part of the routine software release procedure.

As discussed in Section 2.5, we do some of these tests irregularly, but we are now looking at the issues involved in making them part of our release cycle. The biggest issue at the moment is the large number of "false positives" from the tools and filtering them out.

Give full and early consideration to file management, cataloging and data discovery by physicists doing analysis. Report on this area in future reviews.

This is something that will evolve as we continue analyzing data.

3.5 Hall D Specific

3.5.1 Recommendations

A series of scale tests ramping up using JLabs LQCD farm should be planned and conducted.

As noted in Section 2.3 we are starting a two-billion-event challenge that we estimate to be about 10% of the $10^7\,\gamma/s$ year of running. This challenge requires a combination of the OSG, JLab compute farm and outside users. We expect to follow this up (once the online data format is defined) with a similar scale test involving the tape farms and reconstruction from there.

The data volume and processing scale of GlueX is substantial but plans for data management and workload management systems supporting the operational scale were not made clear. They should be carefully developed.

As noted in Sections 2.1 we have now defined most of our data formats.Given these, we are running through a series of mini data challenges (Section 2.2) to understand the issues associated with the data handling.

Consider ROOT (with its schema evolution capabilities) as a possible alternative for the HDDM DST format.

As noted in Section 2.1 we looked into using root as our DST format. However, we found that the philosophy of using root does not match well onto our threaded environment, and would require substantial development with little if any gain over our HDDM format and the new REST format. We do expect that root will be a crucial part of our analysis after the DSTs are made.

4 Summary