Parallel Simulated Event Processing in GlueX: from Local Clusters to the Open Science Grid

I. Senderovich

July 26, 2011

Abstract

An overview of data analysis in particle physics will be presented as well as the processing pipeline specific to the GlueX experiment. Current use of High Throughput Computing resources will be presented as well as the ongoing transition toward broad deployment on the Open Science Grid.

1 Introduction

1.1 Overview of the Generic Particle Physics Data Pipeline

Simulation, data acquisition, reconstruction and analysis of particle collision events are tasks that can naturally be carried out using parallel computing and specifically, using the paradigm of high throughput computing (HTC). Because each event in the detector is statistically independent from the others, and therefore the information in each event is complete, a collection of events can be divided among execution nodes without need for any communication between them. The remaining challenge in this sort of batch processing is the optimal division of events among the parcels sent to the independent execution nodes. The overhead of time to stage, disk space per event processor, scalability of the scheduler (i.e. performance in "shadowing" the many remote jobs) and desired turnaround time for each job have to be considered. (Regarding the last point: while one focuses on the overall throughput in this paradigm, it is useful to analyze the first arriving data for a low-statistics preview of the result.)

The following list represents the processing sequence for simulated and real data.

1. Acquisition of signals from the detector. Their origin and handling depends on whether this is real data or simulated data
   (a) Simulated: Monte Carlo (MC) sampling of the physical probability distribution to determine the energies and momenta of final state particles. Then, MC-sampling is used to simulate their passage through detector material. Models of detector response are included. Event records are built up from this raw data.
   (b) Real: a portion of the live stream of detector signals received on a processing node for raw event recording. [1]

[1] Often, some simple version of the subsequent steps of reconstruction and analysis are carried out as part of the trigger system to reserve more bandwidth to storage for experimentally-relevant events.

2. Reconstruction is performed on the raw data. The collection of signals is interpreted to hypothesize one or several physical outcomes of the reaction. At this stage, particles are usually identified and their most likely momentum and energy determined. Goodness of performed fits are kept for later reference.

3. Analysis is performed. Physical observables are calculated from the energy and momentum data of final state particles (e.g. the mass of their parent.) Some revision to the hypotheses formed in the reconstruction stage may be performed in light of higher-level fits that impose certain constraints from physical laws. Filters are placed on poorly-formed events or those irrelevant to the current study.

1.2 The GlueX experiment

Analysis of a bound state of particles sheds light on the dynamics that keep them together. For example, measuring the difference in binding energies of an electron in a hydrogen atom in different excitation levels allowed a precise check on the quantum mechanical interactions inside this atom.

Similarly, meson spectroscopy: energy level measurements of two-quark bound states, shed light on the physics of the Strong Interaction and offer a test of Quantum Chromodynamics (QCD) that describes it. Specifically, predictions of the theory include excitations of the dense "gluonic field" between the quarks that have not been decisively identified in an experiment.

The GlueX experiment at Jefferson Lab will look for these excitations and other features of the meson spectrum. The experiment will produce mesons in recoil of a high energy photon on a target of protons. The produced states are short-lived and decay through one or several stages into more stable particles that traverse the detector and leave their signature in the form of ionization trails and energy deposition.

The experiment is currently in the construction stage with first calibration runs and preliminary data coming in 2014. At this stage, analysis techniques are being developed using simulated data with as realistic a model of the detector as possible given computational complexity and precision constraints. Reactions due to known physics is simulated in addition to theorized physical processes that generate "exotic states". Exercises are carried out to extract this hypothetical signal from background and perform an analysis on its parameters such as mass, resonance width and phase shifts, as well as angular momentum quantum numbers. A comparison to the parameters between the signal injected into the simulation and that which emerges from reconstruction and analysis informs about the capabilities of the detector (i.e. efficiency, resolution etc.) as well as processing software and quality of the analysis techniques.

2 GlueX Physics Simulations on the Grid

Figure 1 shows the scripted simulation and analysis process chain adapted to the GlueX experiment. The unshaded blocks in the figure represent the generic tasks that are well-suited to parallel batch processing. (Note that there is a generic analysis stage and a more interactive stage, after the job outputs have been condensed, for further work by a physicist.) The generation, simulation, reconstruction and analysis of the signal events are not as costly compared to the background events: the latter is orders of magnitude more plentiful for any process of interest.

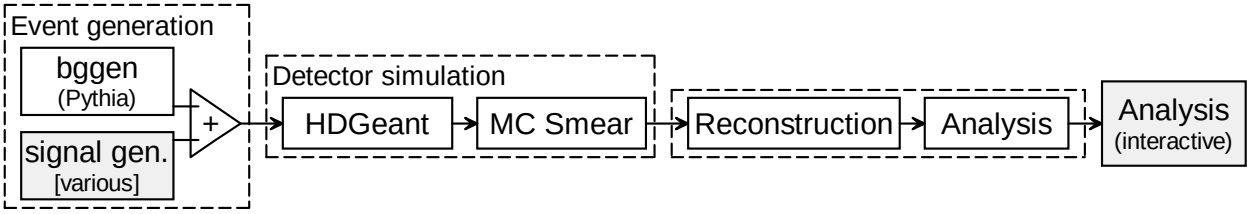

Figure 1: The simulation/reconstruction pipeline used in GlueX. The shaded stages are not part of the batch parallel processing system. Note how the future data source would be a drop-in replacement for the "Event generation" and "Detector simulation" stages. These realistic simulation stages aid in the design of the detector, development of algorithms for reconstruction and analysis to be carried out on real data. Additionally, analysis of simulated data is vital for certain forms of normalization applied in real data analysis.

2.1 Job Load Distribution

Given the variability of computer hardware on which these tasks would be performed, either on clusters within our collaboration or on the Open Science Grid (OSG), it is sufficient to come up with an order-of-magnitude figures for per-job processing time, job segmentation etc. At the time of writing, with the current state of algorithms in the GlueX software, the reconstruction stage takes the most time per event: $\sim 0.5\,\mathrm{s}$. One must consider the time overhead for starting jobs: order of a minute to several minutes for staging and up to a minute for loading all calibration constants, including fine resolution magnetic fields for reconstruction of charged tracks. The other constraint on the lower bound of the number of events to assign per job is the number of *shadows* that would be running on the submission host: significant performance degradation of this host was seen with several thousand jobs. Thus, taking a signal that is 0.1% of background and taking 30,000 signal events for decent statistics leads to about 30 million events. Thus with a limit of 1000 shadows/submission we must process at least 30,000 events per job.

This is a more stringent requirement than that which comes from considerations of the overhead. For example, taking some short processing time that would dwarf the time overhead for staging: $30\,\mathrm{min} * 2\,\mathrm{events/s} = 3,600\,\mathrm{events}$

On the upper bound, one must consider the duration per job (and the inefficiency incurred in having to repeat it several hours into the task) and the size of the intermediate files to be stored locally and final output to be returned. With event size on the order of 20 kB, files before the analysis stage approach 1 GB each when process bundles of several tens of thousands of events.

The hardware of a submission node can (and should) be upgraded and algorithms will be improving, but these figures gives the order of magnitude for segmenting events among jobs. Practical experience with the analysis work cycle and facilities on modern clusters, 50,000 events per job seems to be the best choice.

2.2 Current Implementation

Recent work with these analyses was carried out on single clusters owned by the collaboration universities' groups. The cluster of the Nuclear Physics Group at the University of Connecticut is composed of about 100 machines acquired over the last 10 years and summing to about 400 GFLOPS of processing power. About 330 GFLOPS are in the more modern 64-bit machines (important

consideration for certain optimizations in reconstruction algorithms.) A local dCache pool is in place, distributed among the cluster's nodes for long term, resilient storage of real and generated data. Condor has been used to manage this cluster. This tool alone was sufficient for most procedures:

- seed from random number generator is derived from the process number
- output files are numbered based on the process number
- collation of the results is a trivial command - does not warrant use of a work flow tool such as DAGMan.

2.3 Outlook

Work on simulations in the collaboration is proceeding toward more ambitious quantities of events. It is also recognized that the eventual data sets acquired by this high-statistics experiment (raw data on tape on the order of petabytes, analysis-relevant event sets on the order of 10^9) calling for a transition toward the OSG. Over the last year, the local resources have been added to the OSG under the "Gluex" virtual organization. Test submissions have started, first directly to specific computing elements and more recently through glideinWMS. Through hardware upgrades and tuning, submissions of 10,000 jobs at a time are succeeding. Exchange of terabytes of data from submitted jobs using SRM has been demonstrated, while tuning continues with the use of multiple "doors" in our storage system and distribution of dCache cells for proper load balancing.

Another short-term goal for parallel computing in this collaboration is the ability to recover from failures in the middle of the processing pipeline shown above or to restart from a known position (given intermediate data files) as part of the software design process. Since the pipeline is scripted from several binaries, "checkpointing" is not possible. Use of a manager like DAGman would require transfer of large intermediate files between every stage which are not always needed. (For example, a user may desire to have an intermediate file deleted once its successor has been created in a subsequent stage, and kept only if the process chain halts.) Furthermore, a dependency-conscious system similar to "make" on each execution node would be useful. This is especially important in the OSG environment where software distributions vary widely. To this end, a light software package is under development that allows the user to request a final or intermediate file and ensures that all logical dependencies are met, including building the necessary software. Also, as the GlueX reconstruction and analysis software is undergoing active development, it is useful to have a job layer that tracks the status of the packages present on the remote resource from previous run and dynamically builds the requested branch or release of the software as necessary.