GlueX Computing Resource Model

David Lawrence Mark Ito

November 8, 2018

Abstract

The GlueX Computing model based on an analysis of the Spring 2017 computing resources used to date. The analysis was done in the June 2018 in preparation for the GlueX Phase-II experimental Readiness Review.

1 Introduction

In order to predict future computing requirements for GlueX, an analysis was undertaken of the resource usage in analyzing the Spring 2017 dataset (i.e. RunPeriod-2017-01). This was done in June 2018 while preparing for the GlueX Phase-II Readiness Review as it was the data set whose analysis was at the most mature state. The model tries to capture the use cases for the Scientific Computing farm at JLab and the tape library.

2 Compute Resource Model

The model itself is intended to capture the various uses of compute resources (CPU and tape) for all aspects of data processing for a GlueX run period. This is done for the purposes of projecting future requirements in order to ensure adequate computing resources are available when needed. The model starts with predictions for raw data volumes based on L1 trigger rates and experimental time on the floor (calendar time). The CPU usage is separated into several categories which include various monitoring passes, reconstruction passes, and analysis passes. Some of these are done on just partial data sets which may be based on a fraction of the run or on a fixed number of files from the run. Descriptions of the parameters are given in the following section along with descriptions of how the number was obtained for the RunPeriod-2017 data set. Note that many values cannot be accurately known and were tuned based on the integrated usage statistics reported by the Computing Center. These values are reported in the “Actual Usage” section at the end of the document.

The model itself is implemented as a python script which reads inputs from an XML file. This allows different XML input files to be tuned for different running conditions. The python script itself is maintained on github here:

https://github.com/JeffersonLab/hd_utilities/blob/master/comp_mod/comp_mod.py

The XML input files are kept in the same directory as the python script on github. An example of the XML (content part only) can be seen in appendixB. The full file can be seen on GitHub as “RunPeriod-2017-01.xml”.

3 Parameter Descriptions

The following are desciptions of input parameters to the model and how the values used for describing the RunPeriod-2017-01 data set were obtained. This work was done in June 2018 so values reflect the known state and anticipated resource usage at that time.

3.1 triggerRate

Some fraction of time was spent at 100nA (low intensity) and some at 150nA (high intensity). The trigger rates for these was about 30kHz and 45kHz respectively. The value of 40kHz is a guess based on those two numbers, but leaning higher in order to give a total data volume closer to the 911TB recorded on the tape library for this run period.

3.2 runningTimeOnFloor

Total number of days (40) from Eugene’s collab meeting talk:

https://halldweb.jlab.org/DocDB/0035/003524/001/talk_coll_2018_feb.pdf

Total number of days (35) from Alexandre’s collab meeting talk:

https://halldweb.jlab.org/DocDB/0032/003299/001/Spring17_summary.pdf

We use Eugene’s number since it makes the total data volume closer to the known volume.

3.3 runningEfficiency

Total running efficiency of 48% taken from Alexandre's slide here:

https://halldweb.jlab.org/DocDB/0032/003299/001/Spring17_summary.pdf

3.4 eventsize

Running hdevio_scan on a raw data file gives:

```
> hdevio\_scan /cache/halld/RunPeriod-2017-01/rawdata/Run031034/
    hd\_rawdata\_031034\_000.evio
Processing file 1/1 : /cache/halld/RunPeriod-2017-01/rawdata/Run031034/
    hd\_rawdata\_031034\_000.evio
Mapping EVIO file ...
26500 blocks scanned (18904/19073 MB 99%)
EVIO Statistics for /cache/halld/RunPeriod-2017-01/rawdata/Run031034/
    hd\_rawdata\_031034\_000.evio :
- - - - - - - - - - - -
    Nblocks: 26732
    Nevents: 40372
    Nerrors: 0
Nbad\_blocks: 0
Nbad\_events: 0
EVIO file size: 19073 MB
EVIO block map size: 3772 kB
first event: 1
last event: 1613960
             block levels = 40
         events per block = 1-3,13
                    Nsync = 0
                Nprestart = 1
                      Ngo = 1
                   Npause = 0
                     Nend = 0
                   Nepics = 20
                     Nbor = 1
                 Nphysics = 1613960
                 Nunknown = 0
 blocks with unknown tags = 0
```

which gives for the avg. event size: 19073/1613960 = 0.01182 MB/event or

11.8kB For hd_rawdata_030981_001.evio the number is 13.5kB We use an average of 12.7kB

3.5 eventsPerRun

Number of events (in millions) in a production run. Shift workers took runs based on number of events, thus integrating out any beam trips during the run. For 2017, runs at low intensity were taken at 150M event and for high intensity at 250M. We use an average of 200M events here.

3.6 RESTfraction

This is based on looking at several REST file sizes on the cache disk e.g. 2867323
/cache/halld/RunPeriod-2017-01/recon/ver02/REST/030739/dana_rest_030739_000.hddm
2950430 /cache/halld/RunPeriod-2017-01/recon/ver02/REST/030749/dana_rest_030749_000.hddm
2857726 /cache/halld/RunPeriod-2017-01/recon/ver02/REST/030769/dana_rest_030769_000.hddm
2915293 /cache/halld/RunPeriod-2017-01/recon/ver02/REST/030787/dana_rest_030787_000.hddm
2857342 /cache/halld/RunPeriod-2017-01/recon/ver02/REST/030788/dana_rest_030788_000.hddm
2796538 /cache/halld/RunPeriod-2017-01/recon/ver02/REST/030823/dana_rest_030823_000.hddm
The raw data files are all very similar in size to: 19570207
thus: 2.86/19.57 = 14.6%

3.7 goodRunFraction

This represents the fraction of the full dataset considered good production runs. We get this from the ratio of the CPU used for the two recon passes from the record (https://halldweb.jlab.org/data_monitoring/launch_analysis/index.html) to that calculated assuming all beamtime was used to collect production data:

(1.5743+1.3669+1.7672+1.5498)/7.4 = 0.85

3.8 reconstructionRate

Directly measured on gluons gives something close to 5.2Hz/core. The 5.0 number is from memory of a calculation I did based on some numbers from one of the launces documented here: https://halldweb.jlab.org/data_monitoring/launch_analysis/index.html I assume the discrepancy is due to inclusion of hyperthreads in the farm number.

3.9 reconPasses

Number of reconstruction passes. We did 2 full recon passes of the 2017 data. https://halldweb.jlab.org/data_monitoring/launch_analysis/index.html

3.10 analysisRate

This is estimated by looking at the total CPU of the first analysis pass of 2017 data compared to the total CPU of the first recon pass and using that to scale the 5Hz recon rate: (5Hz)*(1.5743+1.3669)/(0.1954) = 75Hz Note that this will depend on what channels are included in the pass. Some passes only added channels and therefore took less time. This number represents the rate for the first pass which would have been the slowest rate.

3.11 analysisPasses

For 2017 there were 8 versions, but only 5 had data at https://halldweb.jlab.org/data_monitoring/launch_analys Presumably the other 3 were minor enough as to not warrant bookkeeping. As noted above, not all passes were the same. The first was the most inclusive, but others only added some channels and therefore used much less CPU. The final analysis launch looks to have taken the same amount of time as the first, but was only run on about half of the files. The value here is empirical to rpresent an equivalent number of passes to match the total CPU of the 5 recorded passes. (0.551 Mhr)/(0.1954) = 2.82

3.12 cores

The average number of cores available to us varied from the different launches/batches due to competition for the farm at the time. The number of threads per job was 24. The number of jobs active varied from 100-300 which would correspond to 2400 to 7200 cores. This would include some hyperthreading. The number of 4500 was based partially on the above and partially on an estimate of the time jobs were active in each batch. The following are taken from eyeballing the "active" curve on the plot: "Number of jobs in each stage since launch" 410 + 350 + 350 + 320 = 1430 hr = 8.5 weeks

3.13 incomingData

proportional to number of runs Number of files per run analyzed for the "incoming data" jobs. This is always 5.

3.14 calibRate

proportional to time on floor This value represents the number of Mhr of CPU used per week of running to calibrate the detector. For 2017 data, the gxproj3 account (Sean) used 2Mhr. Additional time was used by individual accounts for calibration that is not as easy to categorize. Tegan B. was the biggest user with 7.4For this value we assume 3Mhr/5.7 weeks = 0.526 It should be noted that during the discussion on this at the Offline meeting on 2018-06-15 there was general thinking that we should be able to calibrate with far less CPU in the future. This number is higher partly because we were still developing technique and partly because the farm resource was not freely available at the time.

3.15 offlineMonitoring

proportional to number of runs A total of about 2.3 Mhr was used for Offline Monitoring jobs of 2017 data. This consisted of a couple of dozen runs with various conditions and amounts of data for each. If we took 289 production runs (based on 0.893PB total data, 24TB/run, and 85offline monitoring used about 0.00800Mhr per run.

3.16 miscUserStudies

proportional to time to process al files of single run This value is used to capture the CPU usage by all of the various users that is attributed to the gluex project. Some of this should probably go under calibRate, but it is very hard to categorize which parts of this should go there. It is assumed here that these are jobs that run over all files from a small number of runs in order to do special studies. The amount of CPU required is therefore proportional to the time it takes to process a single production run. This number is empirical based on 2017 CPU usage. There is about a 9 Mhr descrepency in the total usage (26.3MHr) and the shared account usage (16.4Mhr). We attribute 1Mhr of that to Teagan's calibrations in the calibrateRate value above. 9Mhr/((200M events)/(5Hz)/(3600s/hr)) = 810 Note that this is not to say that there were 810 studies, but rather, this is the proportionality constant for the CPU usage that is proportional to processing a single run. ”””

3.17 simulationRate

This is based on a very rough value Thomas B. gave of 40ms/event for bggen events with real data background mixed in. Note that adding the background this way significantly reduced the compute time required from previous models.

3.18 simulationpasses

Number of times we will need to repeat simulation. This value of 2 is an old estimate.

3.19 simulatedPerRawEvent

Number of simulated events needed for each raw data event (production runs only) This value of 2 is an old estimate.

4 Actual Usage

4.1 Actual Farm CPU usage

By way of comparison of the calculation to the actual farm usage for the recon launches, recon numbers are obtained from:

https://halldweb.jlab.org/data_monitoring/launch_analysis/index.html

```
Full Recon.
ver01:  3.19Mcore-hr
   batch1:  mean CPU/job=68.55hr  Njobs=23337
   batch2:  mean CPU/job=71.16hr  Njobs=22411
ver02:  3.37Mcore-hr
   batch1:  mean CPU/job=76.95hr  Njobs=23262
   batch2:  mean CPU/job=80.68hr  Njobs=19569
Total: 6.56 Mcore-hr  n.b. this will include hyperthreads and failed jobs
```

4.2 Actual Tape usage

The total amount of raw data was 911TB (from memory since the scicomp page is down). This number includes special runs, including some tests by Sasha after the beam was gone. Other non-production running was also mixed in that would cause this number to be higher than the estimate calculated by the model.

A RunPeriod-2017-01 Computing Resources

```
               GlueX Computing Model
               RunPeriod-2017-01.xml
==============================================
                 PAC Time: 2.9 weeks
             Running Time: 5.7 weeks
       Running Efficiency: 48%
  --------------------------------------
             Trigger Rate: 40.0 kHz
     Raw Data Num. Events: 56.4 billion (good production runs only)
     Raw Data compression: 1.00
      Raw Data Event Size: 12.7 kB
  Front End Raw Data Rate: 0.52 GB/s
       Disk Raw Data Rate: 0.52 GB/s
          Raw Data Volume: 0.863 PB
     Bandwidth to offsite: 328 MB/s (all raw data in 1 month)
      REST/Raw size frac.: 14.60%
         REST Data Volume: 0.355 PB (for 2.82 passes)
   Total Real Data Volume: 1.2 PB
  --------------------------------------
        Recon. time/event: 200 ms (5.0 Hz/core)
           Available CPUs: 4500 cores (full)
          Time to process: 8.3 weeks (all passes)
        Good run fraction: 0.85
   Number of recon passes: 2.0
Number of analysis passes: 2.82
       Reconstruction CPU: 6.3 Mhr
             Analysis CPU: 0.589 Mhr
          Calibration CPU: 3.0 Mhr
   Offline Monitoring CPU: 2.3 Mhr
            Misc User CPU: 9.0 Mhr
        Incoming Data CPU: 0.123 Mhr
      Total Real Data CPU: 21.3 Mhr
  --------------------------------------
       MC generation Rate: 25.0 Hz/core
      MC Number of passes: 2.0
      MC events/raw event: 2.00
           MC data volume: 0.504 PB  (REST only)
        MC Generation CPU: 2.5 Mhr
    MC Reconstruction CPU: 12.5 Mhr
             Total MC CPU: 15.0 Mhr
  ---------------------------------------
         TOTALS:      CPU: 36.3 Mhr
                     TAPE: 1.7 PB
```

B Summary of values in XML format

```
<compMod>
<parameter name="triggerRate" value="40e3" units="Hz"/>
<parameter name="runningTimeOnFloor" value="40.0" units="days"/>
<parameter name="runningEfficiency" value="0.48"/>
<parameter name="eventsize" value="12.7" units="kB"/>
<parameter name="eventsPerRun" value="200" units="Mevent"/>
<parameter name="compressionFactor" value="1.0"/>
<parameter name="RESTfraction" value="0.146"/>

<parameter name="reconstructionRate" value="5.0" units="Hz"/>
<parameter name="reconPasses" value="2.0"/>
<parameter name="goodRunFraction" value="0.85"/>
<parameter name="analysisRate" value="75.0" units="Hz"/>
<parameter name="analysisPasses" value="2.82"/>
<parameter name="cores" value="4500"/>
<parameter name="incomingData" value="5" units="files"/>
<parameter name="calibRate" value="0.530" units="Mhr/week"/>
<parameter name="offlineMonitoring" value="0.00800" units="Mhr/run"/>
<parameter name="miscUserStudies" value="810"/>

<parameter name="simulationRate" value="25" units="Hz"/>
<parameter name="simulationpasses" value="2"/>
<parameter name="simulatedPerRawEvent" value="2.0"/>
</compMod>
```