Offsite Data Processing for the GlueX Experiment

David Lawrence[1,*] and *GlueX* Collaboration

[1]Thomas Jefferson National Accelerator Facility

Abstract. The Thomas Jefferson National Accelerator Facility (JLab) 12GeV accelerator upgrade completed in 2015 is now producing data at volumes unprecedented for the lab. The resources required to process this data now exceed the capacity of the onsite farm necessitating the use of offsite computing resources for the first time in the history of JLab. GlueX is now utilizing NERSC and PSC for raw data production. Details of the workflow are presented.

1 Software and Calibration Constant Distribution

GlueX[1][2] is a Nuclear Physics experiment being carried out at JLab using the 12GeV electron accelerator. It is a search for exotic hybrid mesons using a $\overrightarrow{\gamma} p \rightarrow Xp$ reaction where X represents the hybrid meson state. GlueX reconstruction software is continually evolving with frequent changes leading to new version tags. This motivates us to use a system that can easily distribute newly compiled versions which can then be run on all production platforms, regardless of the host OS. This is done using a Docker container to provide cross-platform uniformity and the CERN Virtual Machine File System (CVMFS)[3] for binary file distributions. Both NERSC[4] and PSC[5][6] support the use of containers. NERSC supports SHIFTER[7] while PSC supports Singularity[8]. Both container types can be easily generated from Docker containers. The common Docker container used by GlueX for both of these is:

docker:markito3/gluex_docker_devel

Both NERSC and PSC support CVMFS. The CVMFS volume used is /cvmfs/oasis.opensciencegrid.org/gluex. This mirrors sections of the */group/halld* disk on the JLab Central User Computing Environment (CUE). Thus, new builds at JLab are automatically distributed to offsite facilities via CVMFS. The Calibration Constants DataBase (CCDB) is also distributed via CVMFS in the form of a SQLite file. Similarly for "resource" files that contain things like the magnetic field maps.

It is worth noting that while simulation is outside of the scope of this document, most GlueX simulation jobs are run on the OSG using the same Docker container (via Singularity) and CVMFS file system.

2 GlueX Data Volume

In the Spring of 2018 GlueX produced a raw data volume of nearly 2PB. In the Fall of 2018 another 1.2PB was produced. These values are summarized in Table 1. Also shown in the

*e-mail: davidl@jlab.org

table are values from the GlueX computing model used to estimate computing resources based on beam time and running conditions.

For reconstruction jobs, one 20GB raw data file will result in 7GB of output files. For offsite processing, this means an equivalent of about 1/3 of the raw data volume transferred to the remote site for processing will be transferred back to JLab[9].

Table 1. GlueX Data volumes by calendar year. All values are in petabytes(PB). Most years include two run periods. *-marked values indicate partial numbers that are current as of mid-2019, but are expected to increase.

	2016	**2017**	**2018**	**2019**	**2020**
actual (raw data only)	0.624	0.914	3.107	0.400*	
model (raw data only)		0.863	3.172	1.56	6.06
actual (production)	0.55	1.256	1.206*	0.62*	
model (production)		0.607	3.084	1.94	4.34

3 Data Processing at NERSC

Significant portions of the RunPeriod2018-01 and RunPeriod2018-08 reconstruction passes were done using NERSC. The total allocation amount for GlueX at NERSC for calendar year 2019 was 58.5M NERSC hours or 70.5k jobs on Cori II regular queue. Table 2 shows the facilities used for reconstruction of each batch.

Table 2. Locations where reconstruction was performed for 2018 GlueX data. For each run period, the data was broken up into batches for processing that are roughly equal in size.

	RunPeriod2018-01 recon v02	RunPeriod2018-08 recon v02
batch01	NERSC	NERSC
batch02	NERSC	JLab SciComp
batch03	NERSC	NERSC
batch04	NERSC	PSC + JLab SciComp
batch05	NERSC	—
batch06	NERSC	—
batch07	JLab SciComp	—

3.1 Data Transfer

Data transfer from JLab to NERSC is done using Globus[10][11]. JLab has a 10Gbps connection to ESnet[12], a US national research and education network provider, which limits the maximum rate of data transfer. Figure 1 shows the data transfer for the first 4 batches of RunPeriod2018-01 data. The average rate for each of these was limited largely due to contention for the Lustre[13] file system where files are staged from tape before being transferred to NERSC. Prior to starting the first batch of the RunPeriod2018-08 campaign a dedicated data transfer node (DTN) was installed and the software configured to stage directly to it rather than Lustre. Other network parameters were tuned by IT department experts and the throughput was improved substantially. We were then able to reliably fill the 10Gbps pipe with data transfers to NERSC.

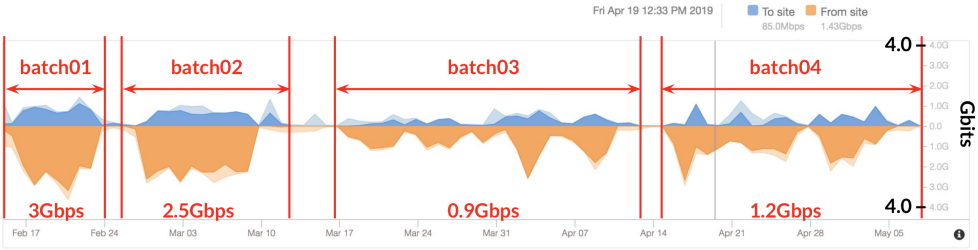

Figure 1. Data transfer rate to NERSC during processing of first 4 of 7 batches of RunPeriod2018-01 data. (~ 55% of data from run period). Transfer rates during this time period were problematic and not stable largely due to contention for the Lustre file system used to stage files read from tape before transferring offsite. The problems have since been corrected and we are now able to transfer data offsite reliably at 10Gbps.

3.2 Disk Space

Disk space required at NERSC is determined by the maximum number of nodes we may feed simultaneously at steady state. Knights Landing (KNL) nodes can consume data at 20GB/6.75hr. It was demonstrated in 2019 that we are capable of filling the 10Gbps pipe from JLab for sustained periods of time. At a rate of 1GB/s we would be able to sustain 1,215 nodes at steady state. This means at least 1215 x 20 GB files on disk for the live jobs and another 1 file for each queued job. This totals 48.6TB of input raw data. Space for the output will be an additional 1/3 of that or 16.1 TB. All of these files will need only temporary storage at NERSC so are best suited for the scratch disk. A total of 64.7TB is therefore required. A special request was made and granted for the required scratch space at NERSC. This must be renewed yearly.

3.3 Job Rates

Figure 2 shows the event processing rate vs. number of threads for a typical GlueX reconstruction job on NERSC Cori II. Note that on Knights Landing (KNL) processors, there are 4 hardware schedulers for each physical core. This results in a progressively slower increase in rate as each additional hardware thread assigned to a core becomes active. The plot shows that some benefit is still observed all the way up to 272 threads.

Figure 3 shows the instantaneous job rate for a production job batch at NERSC. The red curves show the number of running jobs while the blue shows the number of jobs queued. This shows a few periods of time when we were occupying over 1000 nodes on Cori II when using the regular queue.

Multiple options exist at NERSC including whether to use the Cori I (Haswell) or Cori II (KNL) facilities. One can also choose to run in the regular queue or to spend half as much of their allocation by running in the low priority queue. The following sections describe results of some tests using both of these options.

3.4 Haswell queue

We have not used Cori I (Haswell) for production runs due to the greater demand and fewer resources on the system compared to Cori II (KNL). This comes at a cost of 2.4 times as much of our allocation if running on the Cori II regular queue. (Only 1.2 times if running

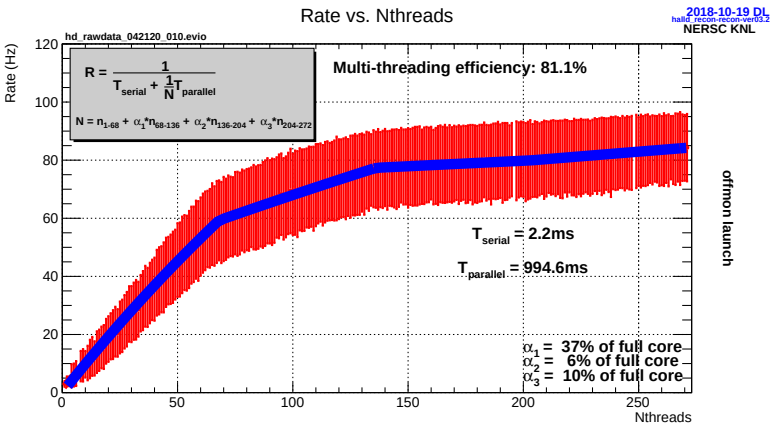

Figure 2. Thread scaling for NERSC job on Cori II (KNL).

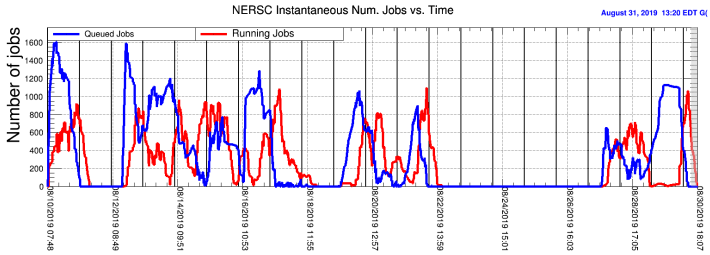

Figure 3. RunPeriod2018-08 data processing at NERSC for batch 01 using the "regular" priority queue. The red curve shows the instantaneous number of running jobs while the blue shows the number of jobs queued, but not yet running.

on the low priority queue). A brief test was done in mid-August 2019 where 10 jobs were submitted to the Cori I queue. Figure 5 shows the results. The first jobs took nearly 20 hours to start. Each job would run for approximately 2.5 hours. Except for a brief period of time where 4 jobs were running, only 2 jobs would simultaneously run. This is believed to be primarily due to the NERSC scheduling policy of starting a maximum of 2 jobs based on priority with additional jobs started via "back filling". The back filling of jobs is done after all jobs have been scheduled based on priority and there are holes in the schedule that lower priority jobs can fit into. GlueX jobs (single node for 8hrs) tend to be small compared to others on Cori II which allows them to be queued via the back filling mechanism easily. This may not be the case on Cori I and this test seems to support that. Based on this, we expect the effective throughput we could get on Cori I would be quite small compared with what has been acheived with Cori II.

3.5 Low Priority Queue

We currently believe most jobs run on Cori II are scheduled via the back filling mechanism (see previous section). If this is the case, then using the *low* priority queue instead of the *regular* queue would not significantly affect our throughput in a campaign. Using the *low*

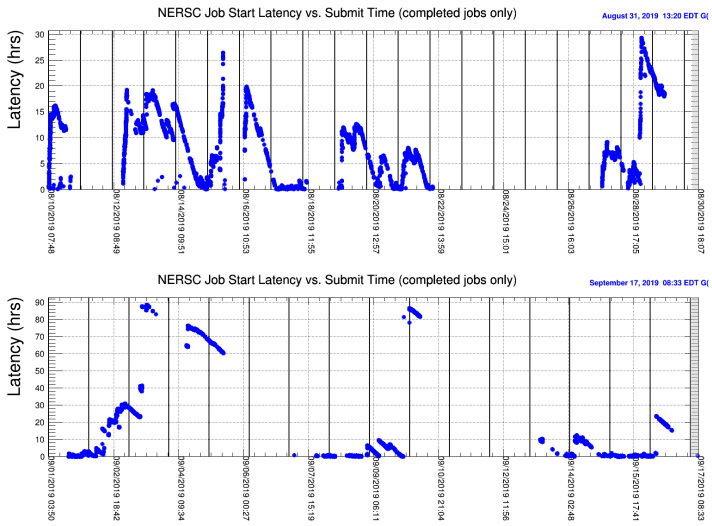

Figure 4. RunPeriod2018-08 job latency at NERSC for batches 01(top) 03(bottom). For the top plot, the "regular" queue was used while in the bottom plot the "low" priority queue was used. It is not clear whether the longer latencies in the bottom plot are due purely to the "low" priority queue or whether other demands on the system at the time played a role.

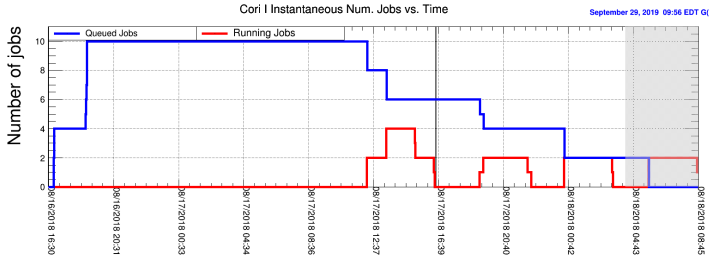

Figure 5. Jobs from short 10 job test using Cori I (haswell). Blue: Jobs queued in the slurm system at NERSC, but no yet running. Red: Number of running jobs.

queue charges half as much of a NERSC allocation as using the *regular* queue. For the RunPeriod2018-08 recon campaign, we used the *regular* queue for batch01 and the *low* queue for batch03 (see Figure 4). The steady-state throughput of jobs using the regular queue (~ 1564 jobs/day) was roughly twice that when using the low priority queue (~ 860 jobs/day). Note that these rates are taken by looking at the throughput during a series of days when the rate was fairly constant. They do not represent total averages. It is not possible to draw any definitive conclusions from these tests since both are strongly affected by what other jobs were queued on Cori II at the times they were run. They do show, however, that it is possible to run on the low priority queue with a reasonable throughput, albeit reduced.

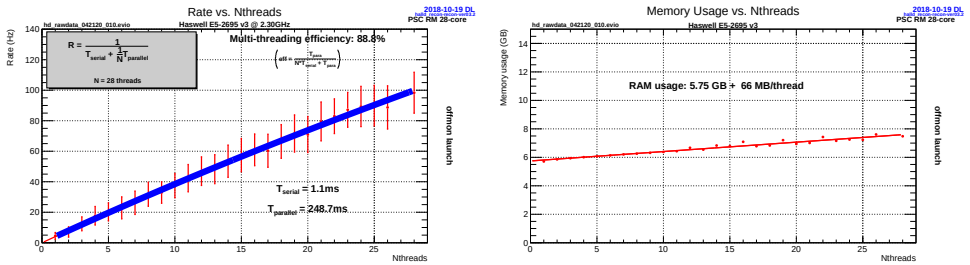

Figure 6. (Top)Thread scaling for PSC Bridges RM node. (Bottom)Memory usage vs. number of threads for GlueX reconstruction job on PSC Bridges RM node.

4 Data Processing at PSC Bridges

This year marked the first ever production processing of JLab experimental data at the Pittsburgh SuperComputing Center (PSC) Bridges facility . Figure 6 shows the event processing rate and memory usage as a function of the number of threads. These were run on the PSC "Regular Memory" nodes which contain 28 core Intel x86 CPUs. The nodes are configured to *not* use hyper-threading so the scaling curve in figure 6 has only one slope.

We were awarded an XSEDE allocation on PSC for 5.9M SUs (Standard allocation Units) for the term starting October 1, 2019. Prior to receiving the award an advance was requested for 0.85M SUs, 10% of the total amount requested in the full proposal. The advance was granted and it was used to process 70% of the RunPeriod2018-08 batch04 data in September 2019. This consisted of 6989 jobs which used 805k of the 850k advance.

4.1 Job Rates

The job throughout at PSC was very steady during the course of the campaign. A rate of roughly 300 jobs/day was maintained during a more than 2 week period. The job processing times were very steady at about 4.25 hours. There was a brief period of time on September 5, 2019 when jobs started being timed out. This was due to an issue at PSC with CVMFS. It was corrected and a portion of our allocation refunded so that those jobs could be re-run without penalty. Aside from that, the overall failure rate for jobs at PSC was only about 0.2%, much lower than what has been observed at NERSC (~2%).

5 Summary

The GlueX experiment at JLab has now begun large scale event reconstruction of experimental Nuclear Physics data at offsite HPC facilties. These include both NERSC and PSC Bridges. A workflow using containers and CVMFS has been successfully deployed and used at both facilities. Over 1.4PB of experimental data was transferred to NERSC and processed in a series of campaigns spanning approximately 8 months in 2019. Another 140TB of data was transferred to PSC Bridges and processed in Sept. 2019.

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Nuclear Physics under contract DE-AC05-06OR23177. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National

Science Foundation grant number ACI-1548562. Specifically, it used the Bridges system, which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC).

References

[1] H. Al Ghoul, E.G. Anassontzis, F. Barbosa, A. Barnes, T.D. Beattie, D.W. Bennett, V.V. Berdnikov, T. Black, W. Boeglin, W.K. Brooks et al., AIP Conference Proceedings **1735**, 020001 (2016), https://aip.scitation.org/doi/pdf/10.1063/1.4949369

[2] S. Adhikari et al. (2020), 2005.14272

[3] J. Blomer, G. Ganis, N. Hardi, R. Popescu, *Delivering LHC Software to HPC Compute Elements with CernVM-FS*, in *High Performance Computing*, edited by J.M. Kunkel, R. Yokota, M. Taufer, J. Shalf (Springer International Publishing, Cham, 2017), pp. 724–730, ISBN 978-3-319-67630-2

[4] NERSC, *National Energy Research Scientific Computing Center* (2020), https://ror.org/05v3mvq14

[5] PSC, *Pittsburgh Supercomputing Center* (2020), https://ror.org/04tac1482

[6] N.A. Nystrom, M.J. Levine, R.Z. Roskies, J.R. Scott, *Bridges: A Uniquely Flexible HPC Resource for New Communities and Data Analytics*, in *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure* (ACM, New York, NY, USA, 2015), XSEDE '15, pp. 30:1–30:8, ISBN 978-1-4503-3720-5, http://doi.acm.org/10.1145/2792745.2792775

[7] NERSC, *Shifter* (2020), https://docs.nersc.gov/development/shifter/how-to-use

[8] SYLABS.io, *Singularity* (2020), https://sylabs.io/docs/

[9] JLab, *Thomas Jefferson National Accelerator Facility* (2020), https://ror.org/02vwzrd76

[10] I. Foster, Internet Computing, IEEE **15**, 70 (2011)

[11] B. Allen, J. Bresnahan, L. Childers, I. Foster, G. Kandaswamy, R. Kettimuthu, J. Kordas, M. Link, S. Martin, K. Pickett et al., Communications of the ACM **55**, 81 (2012)

[12] JLab, *Energy Sciences Network* (2020), https://ror.org/0382bxa43

[13] , *Lustre Filesystem* (2020), https://www.lustre.org