Offsite Data Processing for the GlueX Experiment

David Lawrence - JLab

on behalf of the

GlueX Collaboration

Thursday November 7, 2019
CHEP2019, Adelaide, Australia

Forward Drift Chamber installation in GlueX  Dec. 2013
Electron beam accelerator:
- Continuous wave (250MHz, 4ns bunch structure in halls)
- Polarized electron beam
- Upgraded to 12GeV (from 6GeV)
- 70 μA max @ 12GeV
- 200μA max @ 6GeV

Thomas Jefferson National Accelerator Facility (JLab)
Newport News, Virginia, USA
Hall-D Complex

Electron beam accelerator
  • continuous-wave
    (250MHz, 4ns bunch structure in halls)
  • Polarized electron beam
  • Upgraded to 12GeV
    (from 6GeV)
  • 70 μA max @ 12Gev
    (200μA max @ 6GeV)
The GlueX Detector

- Large Superconducting Solenoid
- Fixed target (30cm LH2)
- Coherent bremsstrahlung polarized photon source
- 38k Detector Channels
- Charged particle tracking, Segmented Calorimetry, PID

Coverage: $1^\circ < \theta < 120^\circ$, all $\phi$
Tracking: $\sigma_p/p \approx 1\% - 5\%$
Calorimetry: $\sigma_E/E \approx 6\%/\sqrt{E} + 2\%$

Liquid Hydrogen Target
## GlueX Computing Numbers

### Data Volume

<table>
<thead>
<tr>
<th>Year</th>
<th>2017 (low intensity GlueX)</th>
<th>2018 (low intensity GlueX)</th>
<th>2019 (PrimEx+ high intensity GlueX)</th>
<th>2020 (high intensity GlueX)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td>0.91PB</td>
<td>3.11PB</td>
<td>0.40PB*</td>
<td></td>
</tr>
<tr>
<td>Model</td>
<td>0.86PB</td>
<td>3.17PB</td>
<td>1.56PB</td>
<td>6.06PB</td>
</tr>
<tr>
<td>Production</td>
<td>1.26PB</td>
<td>1.21PB*</td>
<td>0.62PB*</td>
<td></td>
</tr>
<tr>
<td>Production</td>
<td>0.61PB</td>
<td>3.08PB</td>
<td>1.94PB</td>
<td>4.34PB</td>
</tr>
<tr>
<td><strong>Total Data (actual)</strong></td>
<td><strong>2.17PB</strong></td>
<td><strong>4.32PB</strong></td>
<td><strong>1.02PB</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Total Data (model)</strong></td>
<td><strong>1.47PB</strong></td>
<td><strong>6.25PB</strong></td>
<td><strong>3.5PB</strong></td>
<td><strong>10.4PB</strong></td>
</tr>
</tbody>
</table>

### CPU (Haswell core equivalent from model)

<table>
<thead>
<tr>
<th>Year</th>
<th>2017 (low intensity GlueX)</th>
<th>2018 (low intensity GlueX)</th>
<th>2019 (PrimEx)</th>
<th>2019 (high intensity GlueX)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real Data CPU</td>
<td>21.3Mhr</td>
<td>67.2Mhr</td>
<td>6.4Mhr</td>
<td>39.6Mhr</td>
</tr>
<tr>
<td>MC CPU</td>
<td>3.0Mhr</td>
<td>11.3Mhr</td>
<td>1.2Mhr</td>
<td>8.0Mhr</td>
</tr>
<tr>
<td><strong>Total CPU</strong></td>
<td><strong>24.3Mhr</strong></td>
<td><strong>78.4Mhr</strong></td>
<td><strong>7.6Mhr</strong></td>
<td><strong>47.5Mhr</strong></td>
</tr>
</tbody>
</table>
GlueX Offsite Computing Model

OSG, NERSC, PSC jobs use the same:
• Docker container (converted to Singularity and Shifter)
  • same container used for Cori I (Haswell) and Cori II (KNL)
• CVMFS share
  • GlueX Software builds for CentOS 7
  • 3rd party software (e.g. ROOT)
  • Calibration Constants (CCDB SQLite file)
  • Resource files (field and material maps)

Data Transport:
• NERSC and PSC: Globus
• OSG: Condor

Containerized software runs at NERSC on both Cori I (Haswell) and Cori II (KNL)

OSG = Open Science Grid
NERSC = National Energy Research Scientific Computing Center
PSC = Pittsburgh Supercomputing Center
SWIF2 - Job workflow tool

Manage file transfer and job submission through limited disk resources

1. Request file from tape to DTN
2. Globus transfer task bundles 12 files (=240GB) to NERSC
3. Once file transfer is verified, job is submitted to SLURM

JLAB DTN

NERSC DTN

NERSC Cori II

4. After job completion, all output files bundled in Globus task

5. Output files written to cache and backed up to tape. DB updated

- Currently supports NERSC, PSC, and the JLab farm
- Requires automatic login
GlueX Allocation AY2019

- 58.5M NERSC Units

**Input file size**
- 20GB  (91.9k jobs so far in 2019)

**Wall Time/file on Cori I (Haswell)**
- 3 hours

**Wall Time/file on Cori II (KNL)**
- 6.9 hours
NERSC Job Start Latency

**batch01 latency vs. time**

"regular" queue

**batch03 latency vs. time**

"low" priority queue

Time between submitting job to slurm on NERSC Cori II and job starting

Everything is anecdotal!

- 30 hrs

Offsite Data Processing for the GlueX Experiment - D. Lawrence(JLab) CHEP19, Adelaide, AU
Job scheduler at NERSC is extremely poorly matched to our job shape:

- Schedule at most 2 “jobs” at a time via priority and all others must schedule via backfill
- Scheduler ignores number of nodes and time requested when determining priority
- 64 nodes x 48 hours = 1 node x 3 hours

Suspect most of our jobs run via “backfill” since they are small and fit in cracks. *(Test on Cori I supports this)*
“Regular” vs. “Low” priority queue on Cori II

Aug. 13 - Aug. 18, 2019
~1564 jobs/day

Sep. 14 - Sep. 16, 2019
~860 jobs/day

“normal” queue on Cori II

“low” queue on Cori II
GlueX @ Pittsburgh Supercomputing Center (XSEDE)

PSC Bridges:
- 28 cores/node (no HT)
- 4.2 hours/job
- 6,989 jobs

Smaller than NERSC, but more steady and smaller failure rate
Summary and Outlook

- GlueX is now able to reconstruct large Experimental Nuclear Physics data sets offsite
  - NERSC, PSC
  - Lightweight container used for all offsite HPC computing
  - Software distributed via CVMFS
  - SWIF2 manages workflow

- NERSC
  - Scheduler poorly matched to our natural job size
  - Backfilling saves us (and benefits them!)
  - Job rate fluctuates but averages ~1k/day (=20TB/day)

- PSC
  - Better matched to our natural job size but smaller resource
  - Job rate fairly steady ~0.3k/day (=6TB/day)

- Most simulation is being done on OSG
ESNet data transfer rates to/from NERSC

- Currently have 10Gbit connection
- Will activate second 10Gbit connection this summer
- Proposed 100Gbit upgrade in 2020 or 2021

Anti-correlation observed between transfer rate and Lustre usage
- Test done using OSG16 node, disk speed an issue (longer story, ask Thomas)
- New DTN (Data Transfer Node) being configured with SSD disks for test
- Currently: 10% of files go through OSG node and 90% via cache (=Lustre)
Overview of Jefferson Lab

- Department of Energy National Laboratory with research mission in Nuclear Physics
- In operation since 1995
- Managed for DOE by Jefferson Science Associates, LLC
  - Joint venture of Southeastern Universities Research Association and PAE
- Our primary research tool is CEBAF (Continuous Electron Beam Accelerator Facility) – unique in the world

Jefferson Lab by the numbers:
- 700 employees
- FY2018 Budget: $162.4M
- 169 acre site
- 1,600 Active “User Scientists”
- 27 Joint faculty
- 608 PhDs granted to-date (211 in progress)
- K-12 programs serve more than 13,000 students and 300 teachers annually