Offsite Data Processing for the GlueX Experiment

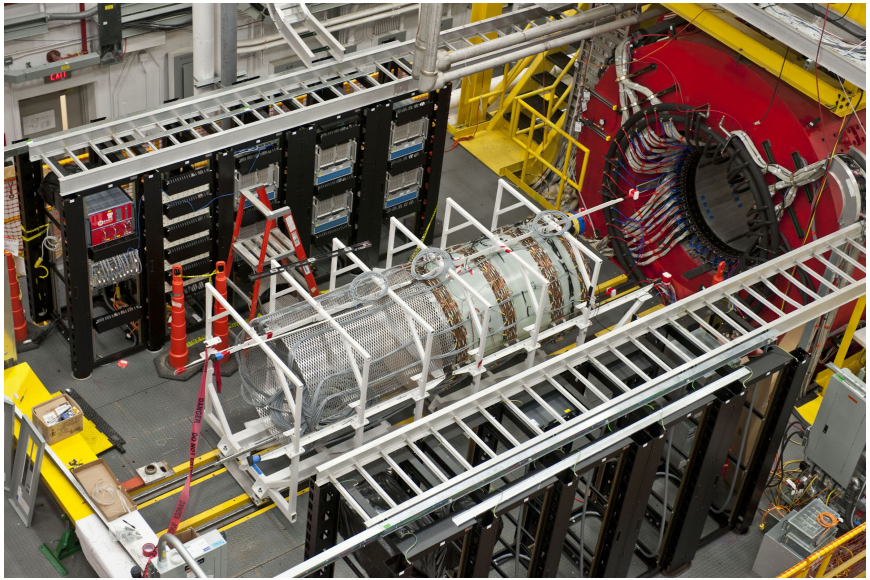
David Lawrence - JLab

on behalf of the

GlueX Collaboration

~~Thursday November 7, 2019~~
~~CHEP2019, Adelaide, Australia~~

Wed. May 6, 2020
NUG Special Interest Group on
Experimental User Facilities



Forward Drift Chamber installation in GlueX Dec. 2013

Aerial p
taken Apr

Hall-D
Comple

Electron beam
- continuous
  (250MHz, 4
  structure in
- Polarized e
- Upgraded t
  (from 6GeV
- 70 μA max
  (200μA max @ 6GeV)

Aerial photo

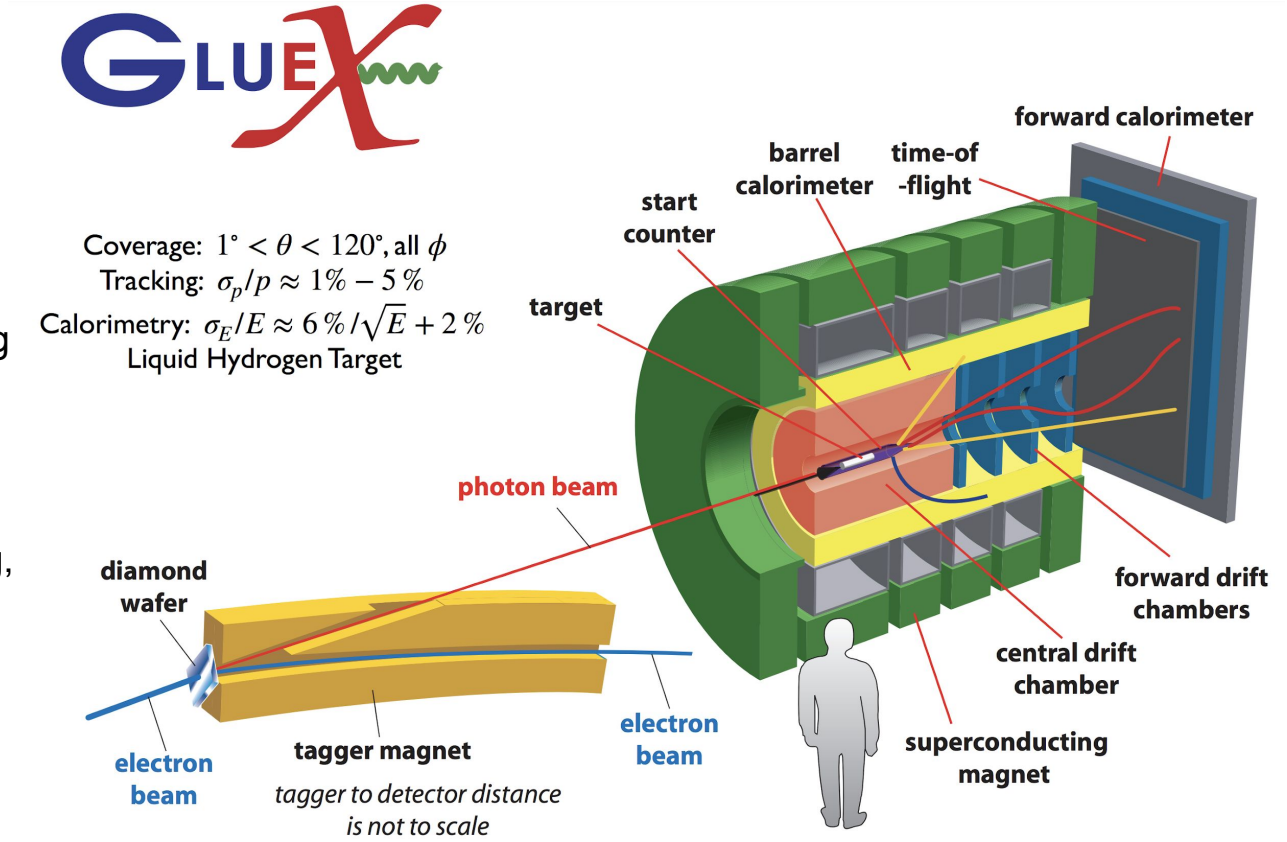
Hall-D Complex

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



GlueX Computing Numbers

Data Volume

	2017 (low intensity GlueX)	2018 (low intensity GlueX)	2019 (PrimEx+ high intensity GlueX)	2020 (high intensity GlueX)
actual (raw data only)	0.91PB	3.11PB	0.40PB*	
model (raw data only)	0.86PB	3.17PB	1.56PB	6.06PB
actual (production)	1.26PB	1.21PB*	0.62PB*	
model(production)	0.61PB	3.08PB	1.94PB	4.34PB
Total Data (actual)	**2.17PB**	**4.32PB***	**1.02PB***	
Total Data (model)	**1.47PB**	**6.25PB**	**3.5PB**	**10.4PB**

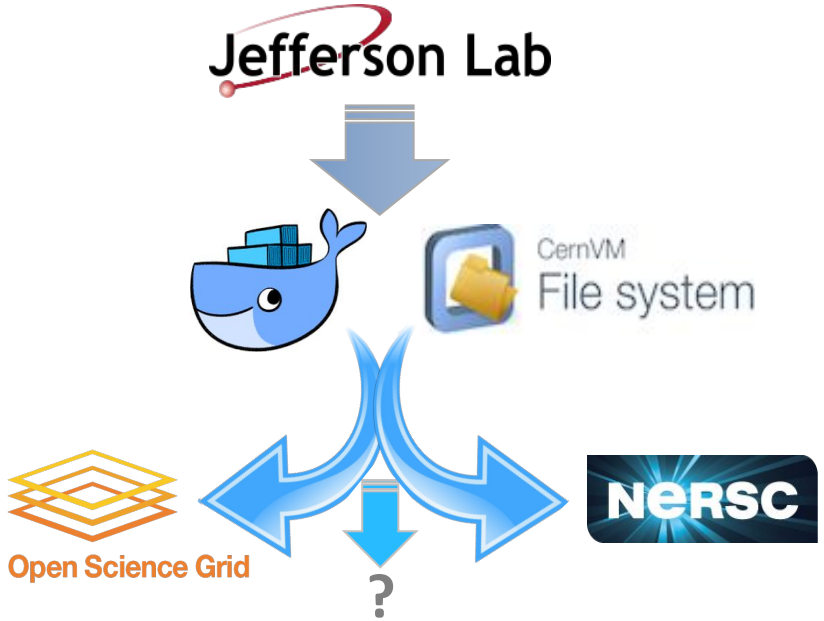
CPU (Haswell core equivalent from model)

	2017 (low intensity GlueX)	2018 (low intensity GlueX)	2019 (PrimEx)	2019 (high intensity GlueX)
Real Data CPU	21.3Mhr	67.2Mhr	6.4Mhr	39.6Mhr
MC CPU	3.0Mhr	11.3MHr	1.2Mhr	8.0Mhr
Total CPU	**24.3Mhr**	**78.4Mhr**	**7.6Mhr**	**47.5Mhr**

GlueX Offsite Computing Model

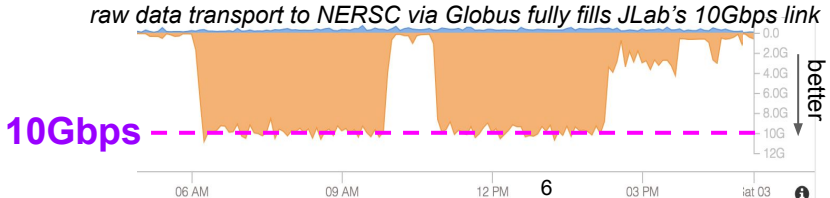
OSG, NERSC, PSC jobs use the same:

- 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

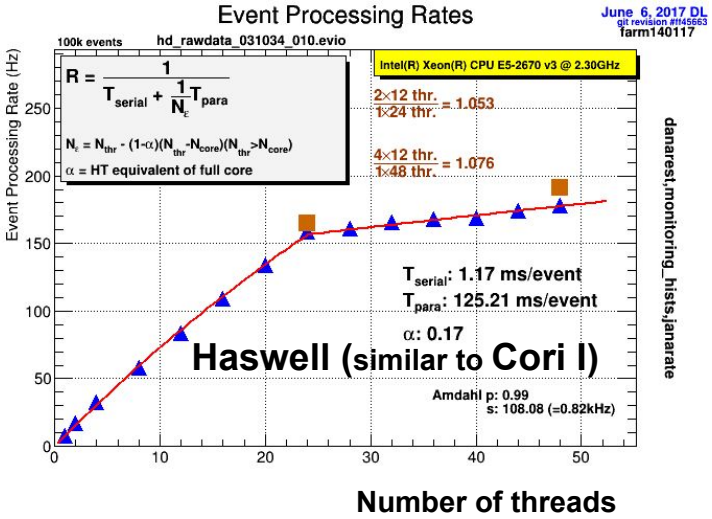
NERSC Cori II

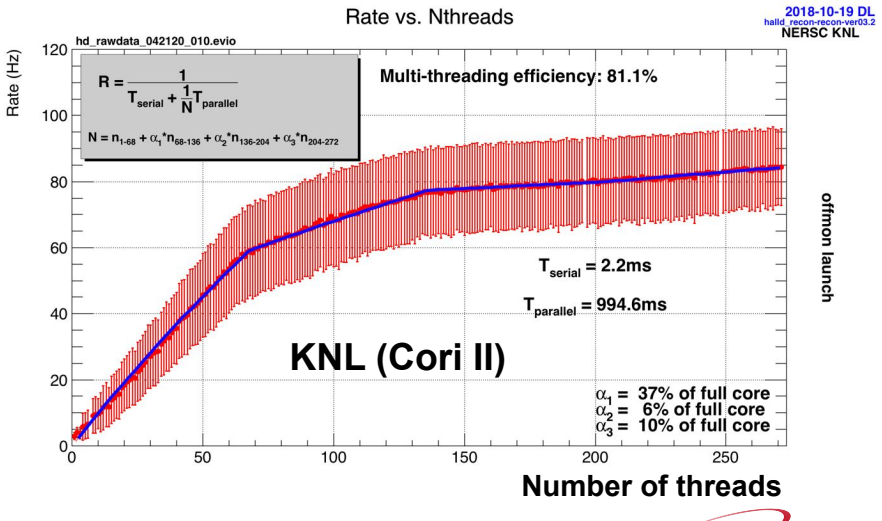
4. After job completion, all output files bundled in Globus task
5. Output files written to cache and backed up to tape. DB updated

- *Currently supports NERSC, PSC, and the JLab farm*
- *Requires automatic login*

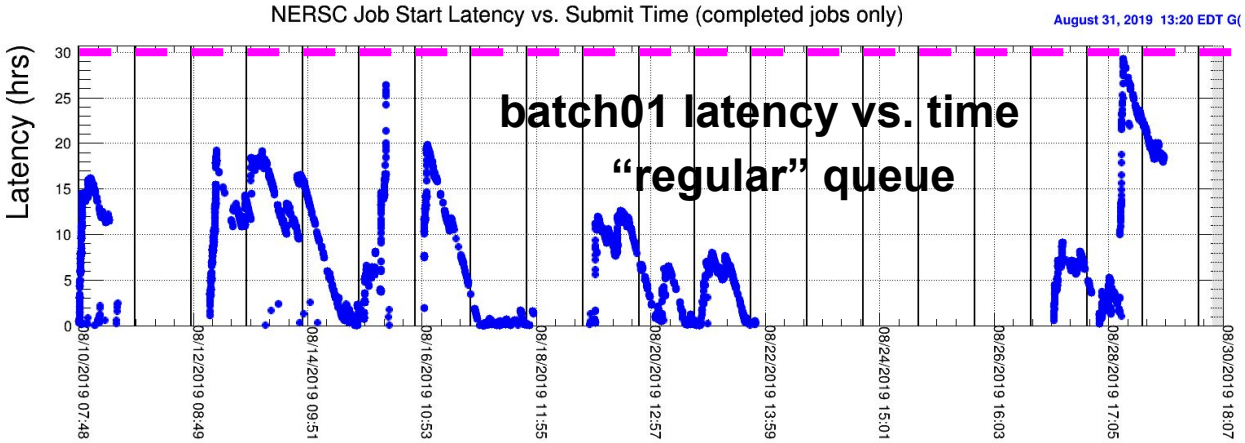
GlueX @ NERSC

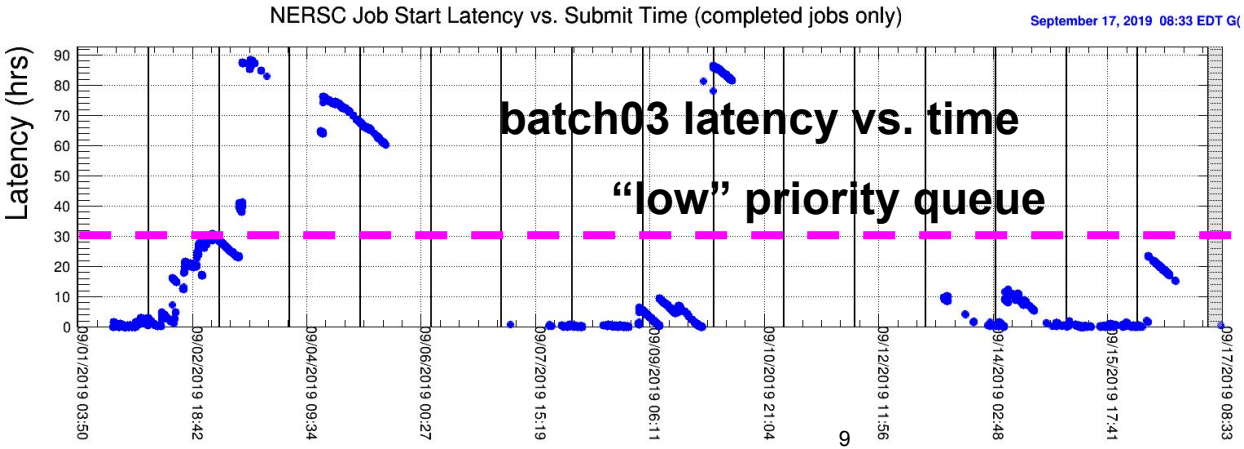
GlueX Allocation AY2019	58.5M NERSC Units
Input file size	20GB (91.9k jobs so far in 2019)
Wall Time/file on Cori I (Haswell)	3 hours
Wall Time/file on Cori II (KNL)	6.9 hours





NERSC Job Start Latency

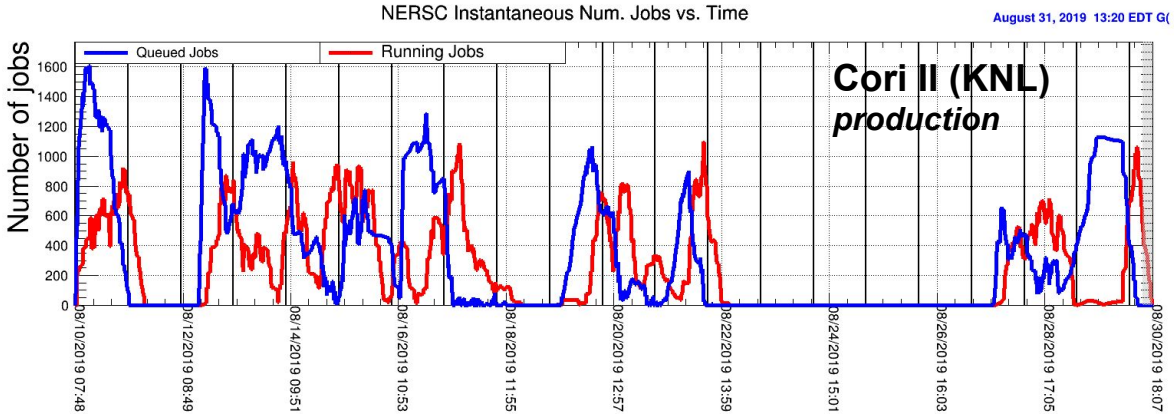


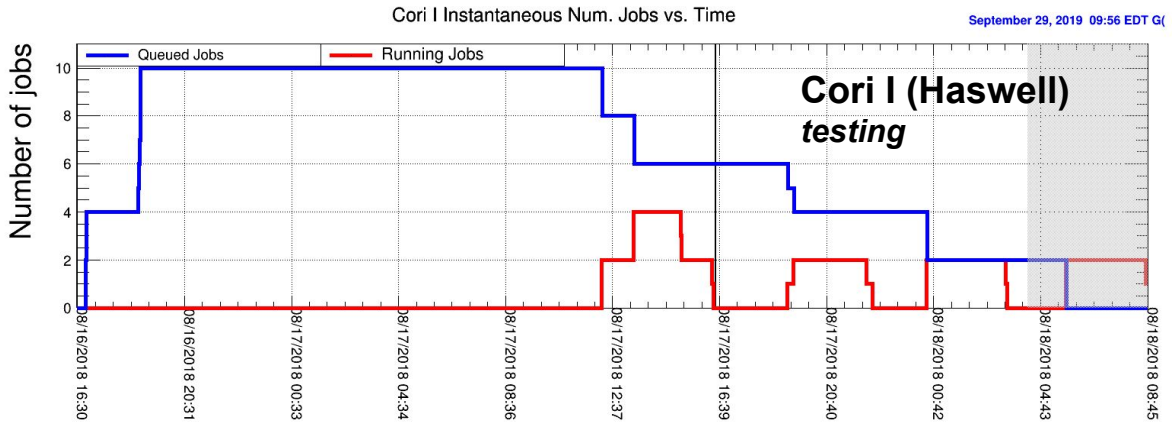


Time between submitting job to slurm on NERSC Cori II and job starting

Everything is anecdotal!

GlueX @ NERSC - Backfilling



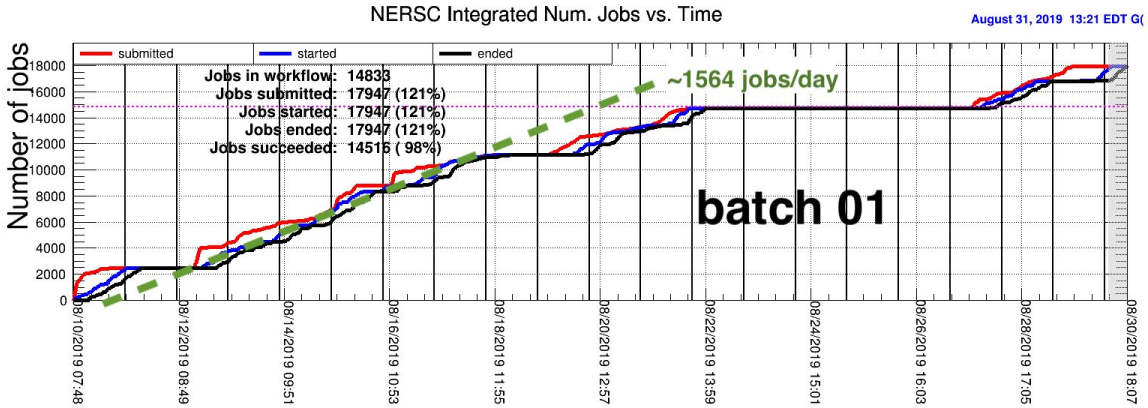


Job scheduler at NERSC is extremely poorly matched to our job shape:

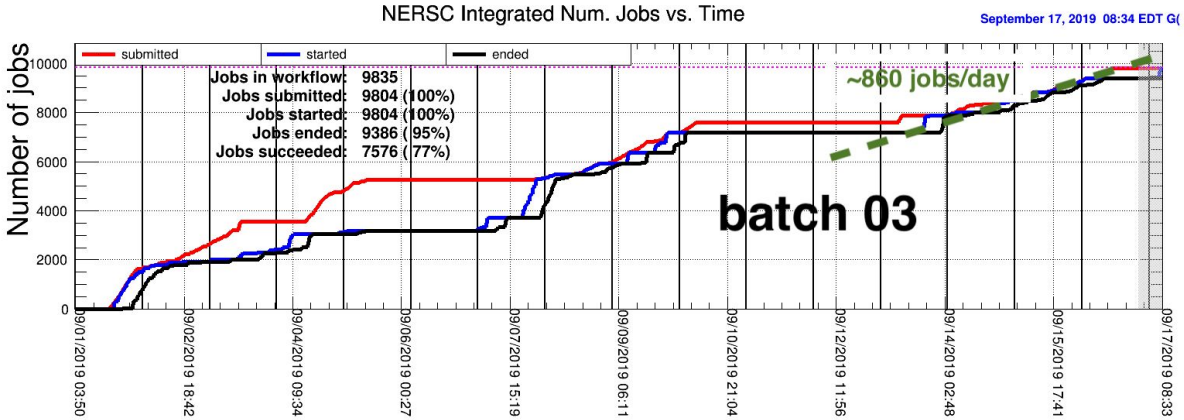
- Schedule at most 2 “jobs” at a time via priority and all others must schedule via backfill
- Scheduler ignores number of nodes and time requested when determining priority
- 64 nodes x 48 hours = 1 node x 3 hours

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

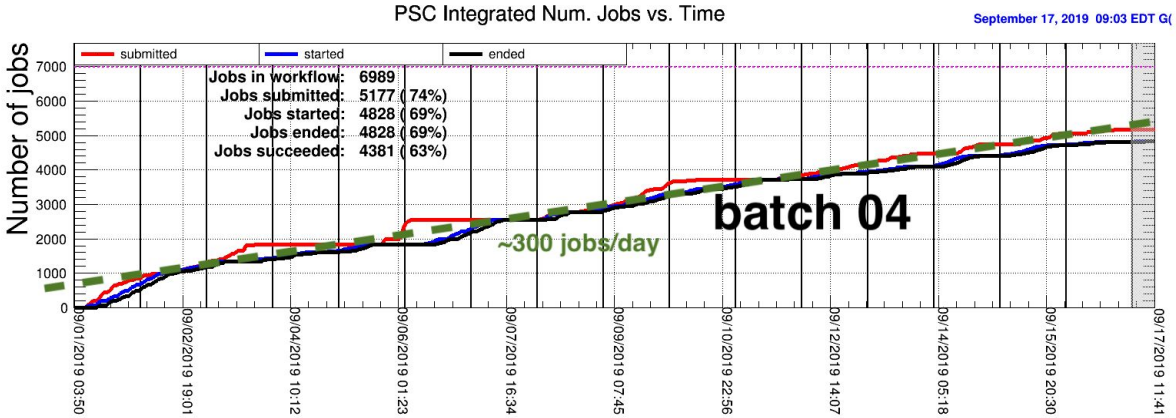


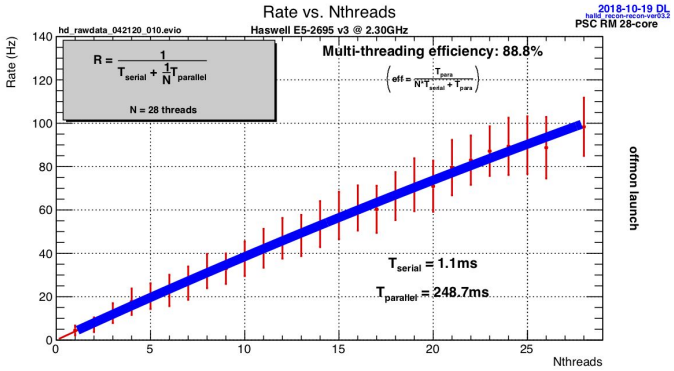
“normal” queue on Cori II
Aug. 13 - Aug. 18, 2019
~1564 jobs/day

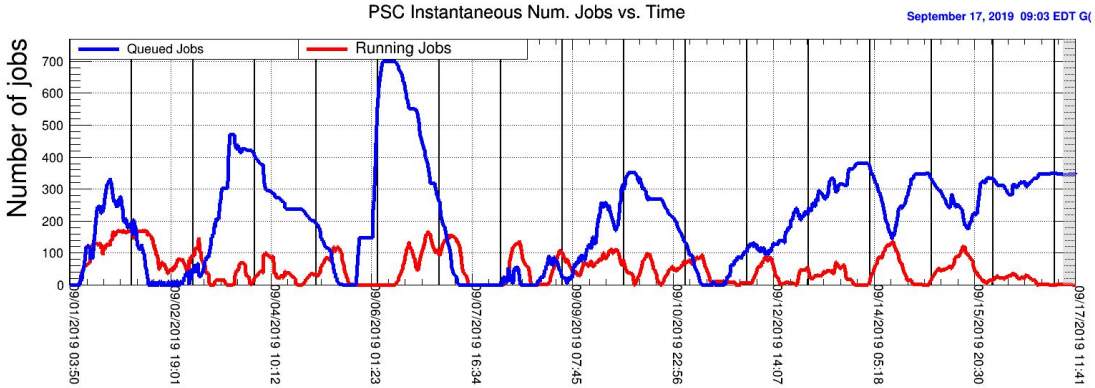


“low” queue on Cori II
Sep. 14 - Sep. 16, 2019
~860 jobs/day

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

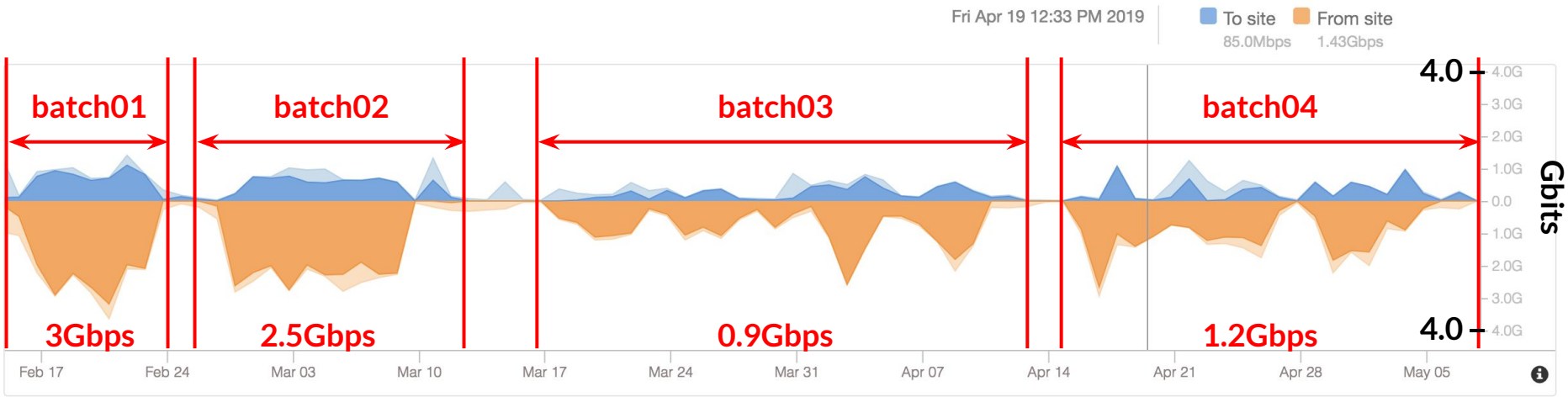
Summary and Outlook

- 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


P
BACKUP

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