Data Production Overview

(mostly processing raw data offsite)

David Lawrence - JLab

Feb. 22, 2012

GlueX Computing Needs

	2017 (low intensity GlueX)	2018 (low intensity GlueX)	2019 (PrimEx)	2019 (high intensity GlueX)
Real Data	1.2PB	6.3PB	1.3PB	3.1PB
MC Data	0.1PB	0.38PB	0.16PB	0.3PB
Total Data	**1.3PB**	**6.6PB**	**1.4PB**	**3.4PB**
Real Data CPU	21.3Mhr	67.2Mhr	6.4Mhr	39.6Mhr
MC CPU	3.0Mhr	11.3MHr	1.2Mhr	8.0Mhr
Total CPU	**24.3PB**	**78.4Mhr**	**7.6Mhr**	**47.5Mhr**

Anticipate 2018 data will be processed by end of summer 2019

Projection for out-years of GlueX High Intensity running at 32 weeks/year

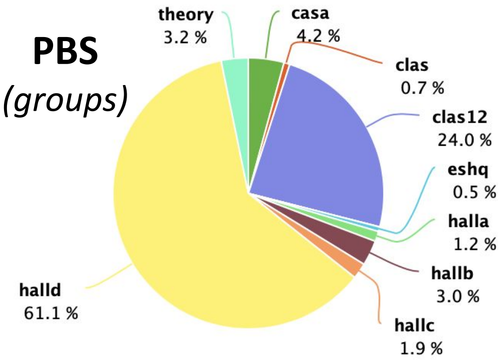
	Out - years (high intensity GlueX)
Real Data	16.2PB
MC Data	1.4PB
Total Data	**17.6PB**
Real Data CPU	125.6Mhr
MC CPU	36.5Mhr
Total CPU	**162.1Mhr**

JLab SciComp Farm

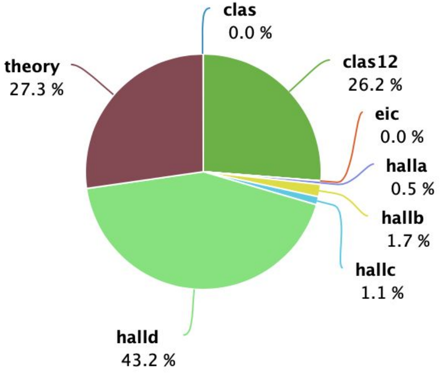
halld total (PBS + slurm):
1.3M jobs
9.7M CPUh

Whole lab (left PBS, right slurm):

PBS
(groups)

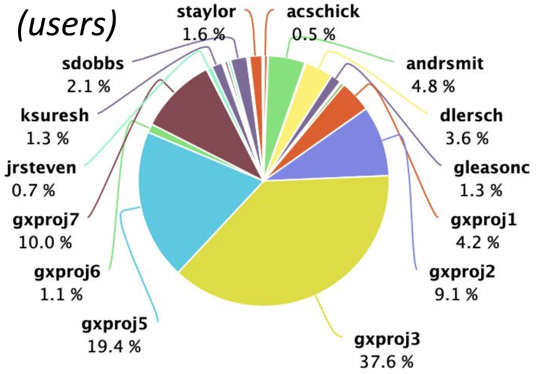


SLURM
(groups)

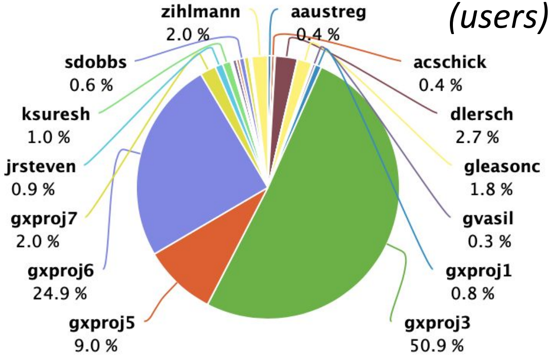


GlueX (left PBS, right slurm):
gxproj3 is calibration
gxproj5 is monitoring and reconstruction
gxproj6 is analysis
gxproj7 is DIRC

PBS
(users)



SLURM
(users)



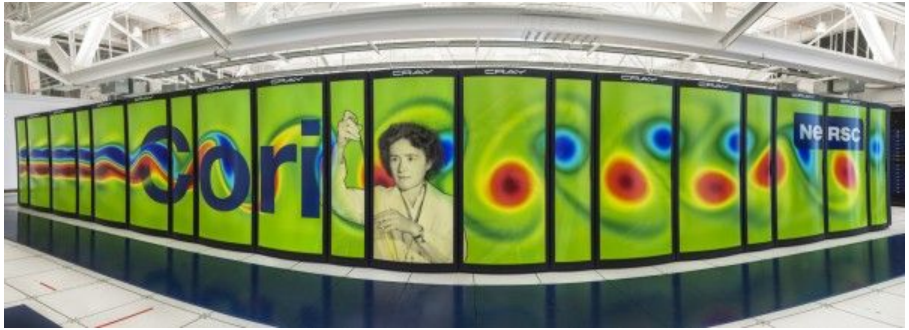
Hi David,

Since March, slurm is in production, but both systems are still running in parallel, so adding the stats is a bit more complicated. Please take the information that you find useful.

Here are the latest numbers between Feb 23 and today:

NERSC = **N**ational **E**nergy **R**esearch **S**cientific **C**omputing center

- Division of Lawrence Berkeley National Lab (LBNL)
- Managed by Univ. of California for DOE
- Multiple large systems, but CORI* is the system we use:
  - Cray XC40 ~**206k full cores** *(Haswell equivalent)*
    - Cori I: Haswell
      - 2,388 Intel Xeon Haswell processors
      - **76.4k full cores**
      - 32 full cores + 32 logical cores/node
    - Cori II: KNL
      - 9,688 Intel Knight's Landing (KNL)
      - 659k full cores (=***129k Haswell equivalent***)
      - 68 full cores + 214 logical cores/node



JLab SciComp Farm **~10.5k full cores**

- 292 nodes
- **~8k full cores** (SciComp only)
- 160 older nodes(**~2.5k full cores**) donated from and shared with HPC

SEMINAR Wed. 5/15 in F224-225 @ 11:00am

(oops, you missed it!)

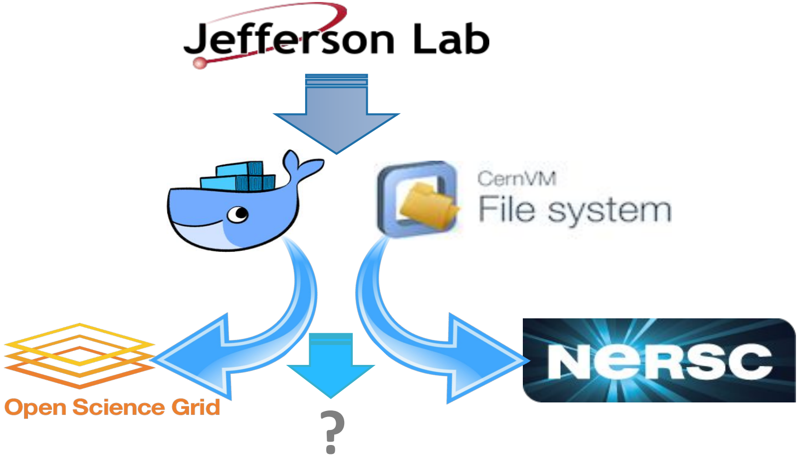
**Cori named in honor of American biochemist Gerty Cori, the first American woman to win a Nobel Prize in science and the first woman ever to win a Nobel Prize for Physiology or Medicine*

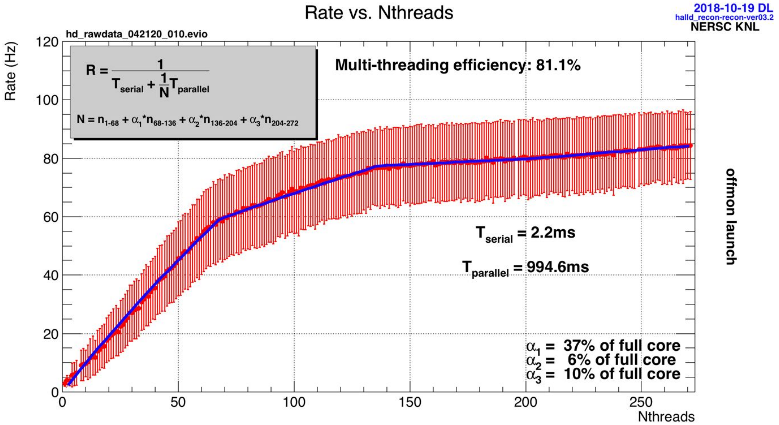
Offsite Computing Resources

Both OSG and NERSC jobs use the same:

- Docker container*
- CVMFS share
  - GlueX Software builds
  - 3rd party software
  - Calibration Constants (CCDB SQLite file)
  - Resource file (field and material maps)

**converted to Singularity and Shifter*



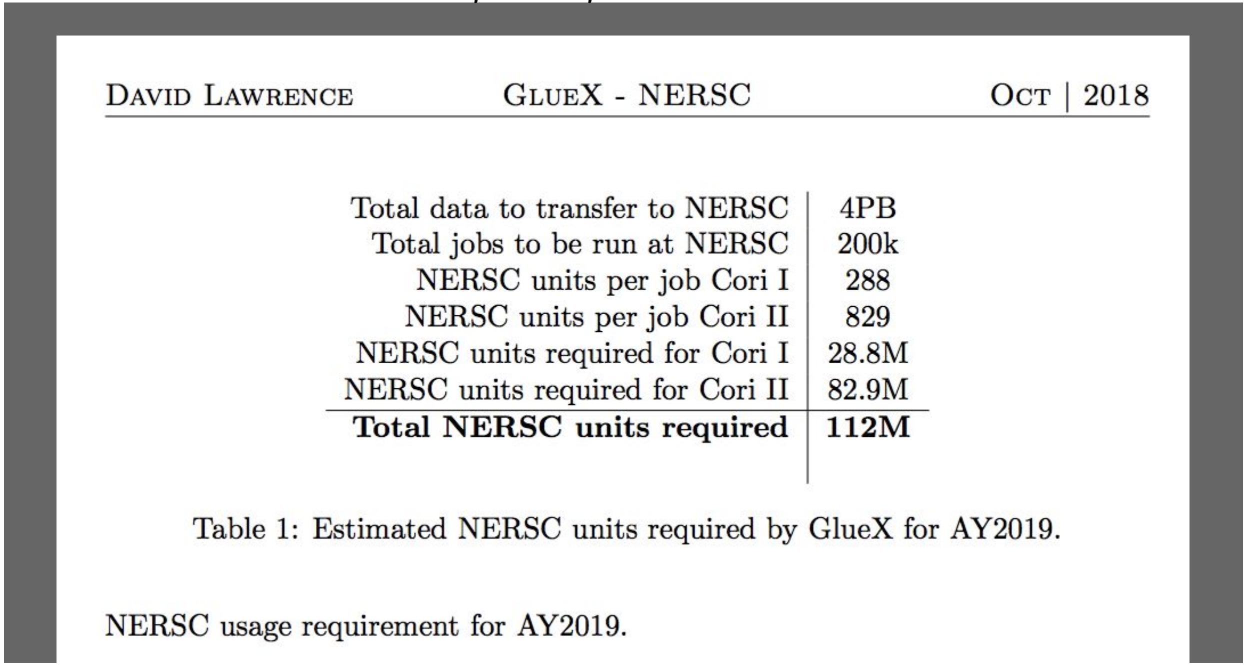


Containerized software runs at NERSC on both **Cori I** *(Haswell)* and **Cori II** *(KNL)*

NERSC AY2019 Request

GlueX DocDB 3793, 3796, 3821

includes estimate of needs for AY2019



DAVID LAWRENCE GLUEX - NERSC OCT | 2018

Total data to transfer to NERSC	4PB
Total jobs to be run at NERSC	200k
NERSC units per job Cori I	288
NERSC units per job Cori II	829
NERSC units required for Cori I	28.8M
NERSC units required for Cori II	82.9M
Total NERSC units required	**112M**

Table 1: Estimated NERSC units required by GlueX for AY2019.

NERSC usage requirement for AY2019.

Requested: 112M units

Awarded: 35M units

Donated: 9M units

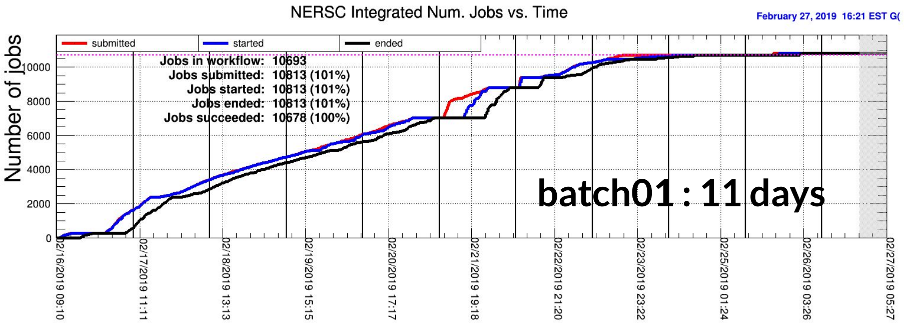
*(**clas12**)* .

Total: 44M units

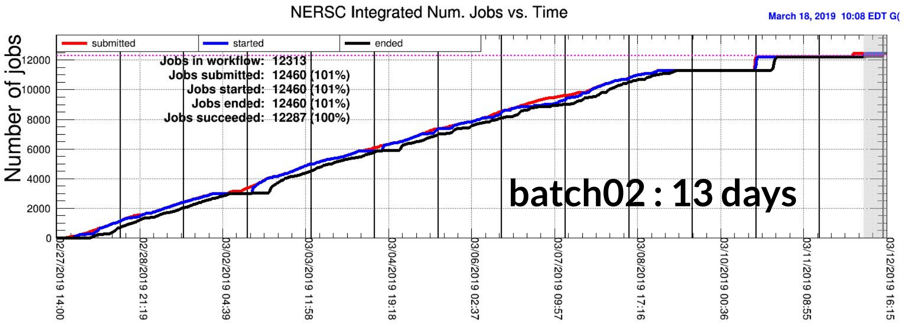
enough for 80% of Spring 2018 data if done completely on KNL

(more on that later)

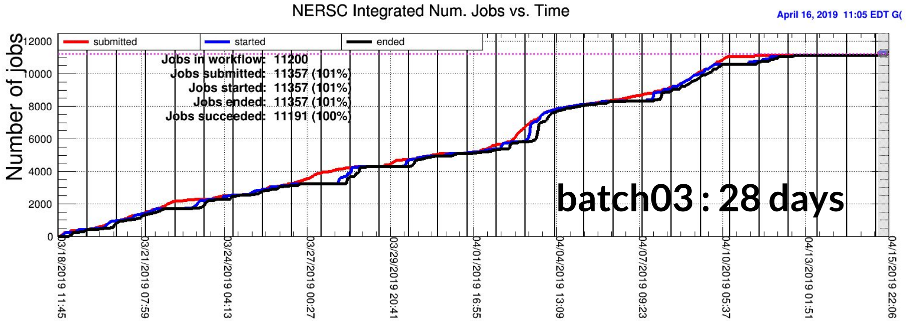
Integrated Jobs vs. Time



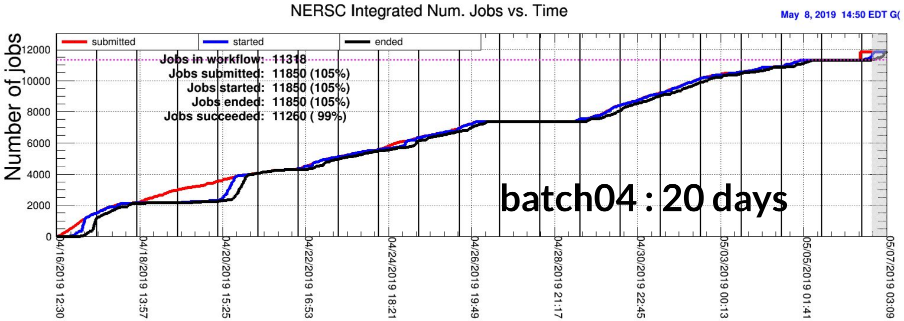
batch01 : 11 days



batch02 : 13 days



batch03 : 28 days



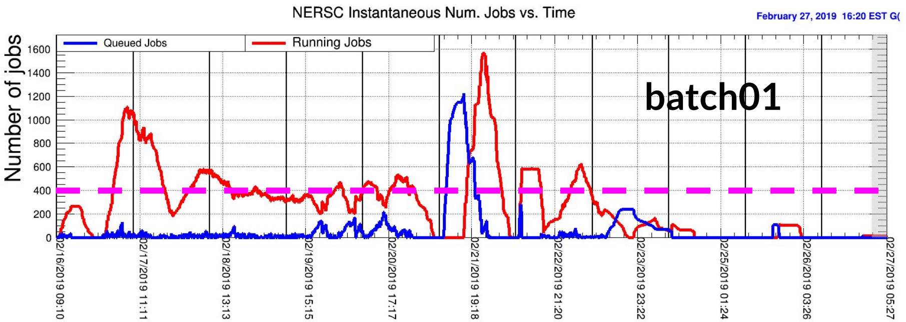
batch04 : 20 days

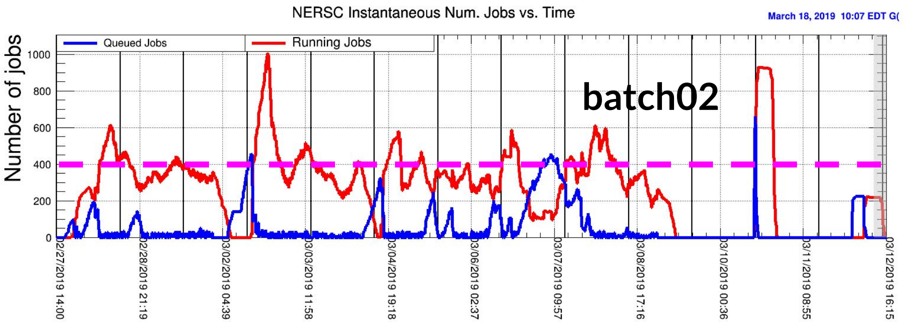
RED: Raw data input file transferred to NERSC and job submitted

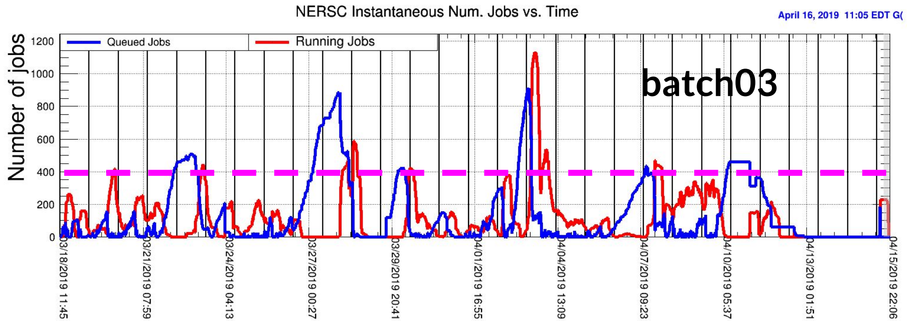
BLUE: Job started at NERSC

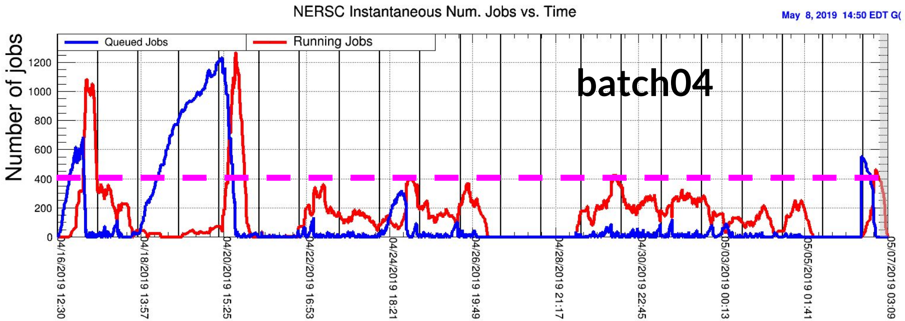
BLACK: Job finished and all output files transferred back to JLab

Instantaneous Running jobs vs. Time



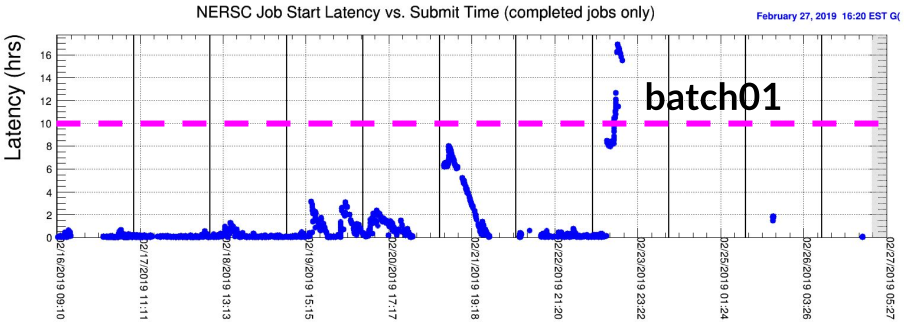


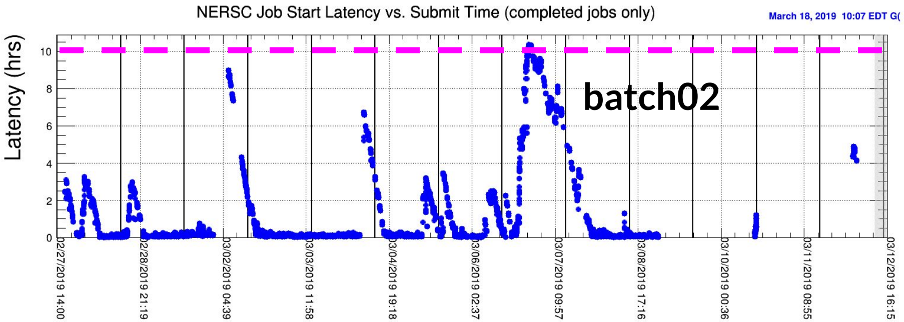


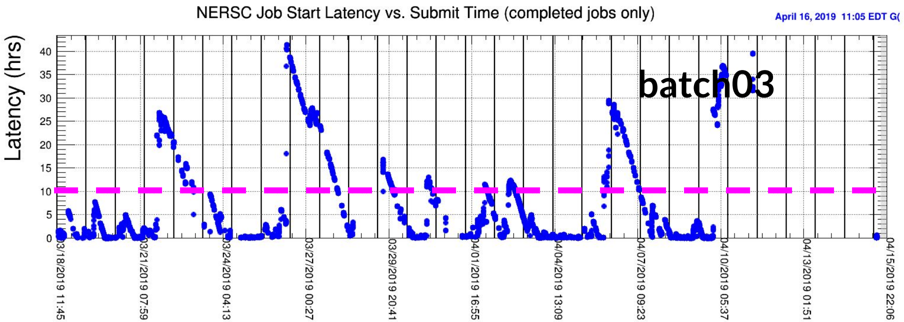


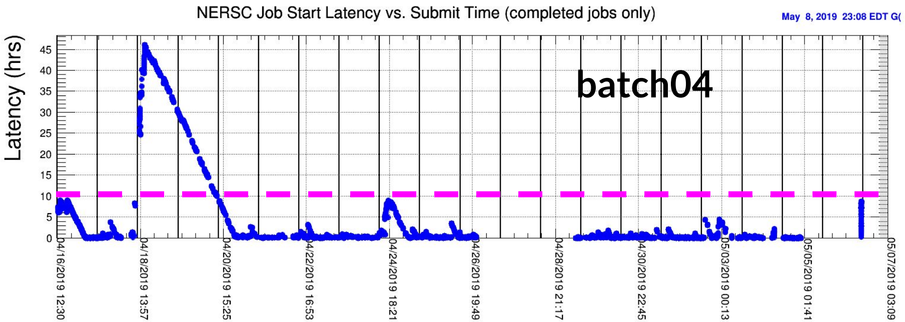
RED: Running jobs at NERSC
BLUE: Queued jobs at NERSC
MAGENTA LINE: 400 jobs

Job Latency in NERSC Queue Before Starting





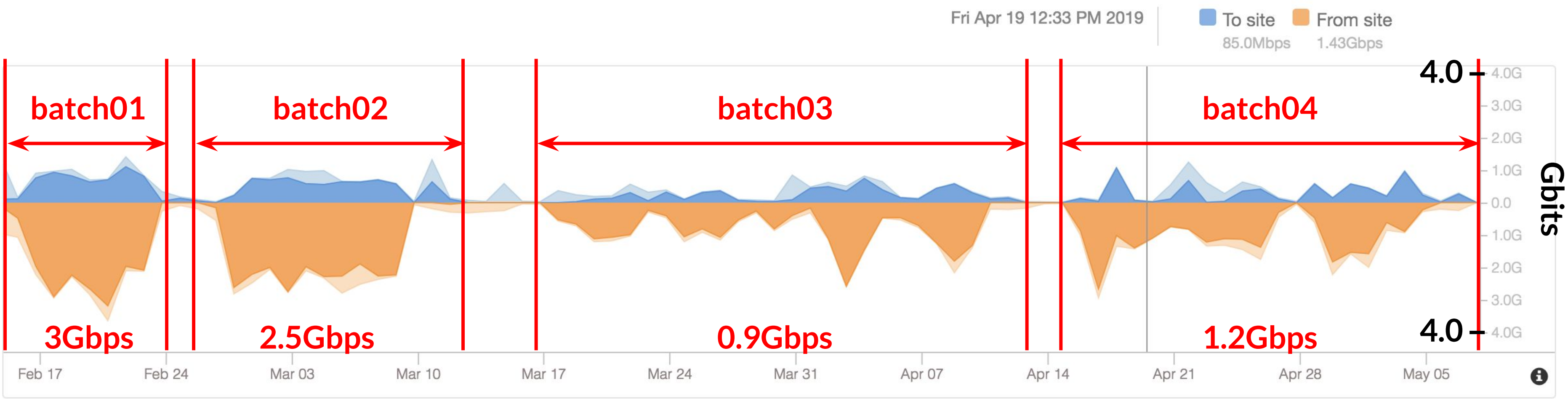




MAGENTA LINE: 10 hours

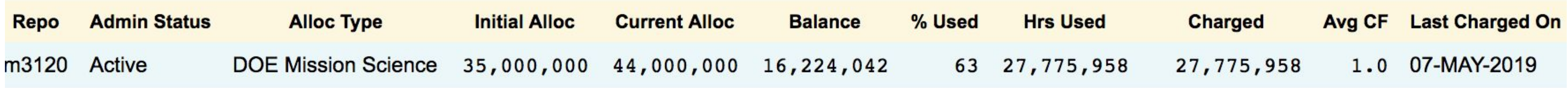
ESNet data transfer rates to/from NERSC

- 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)

Burning through the allocation ...



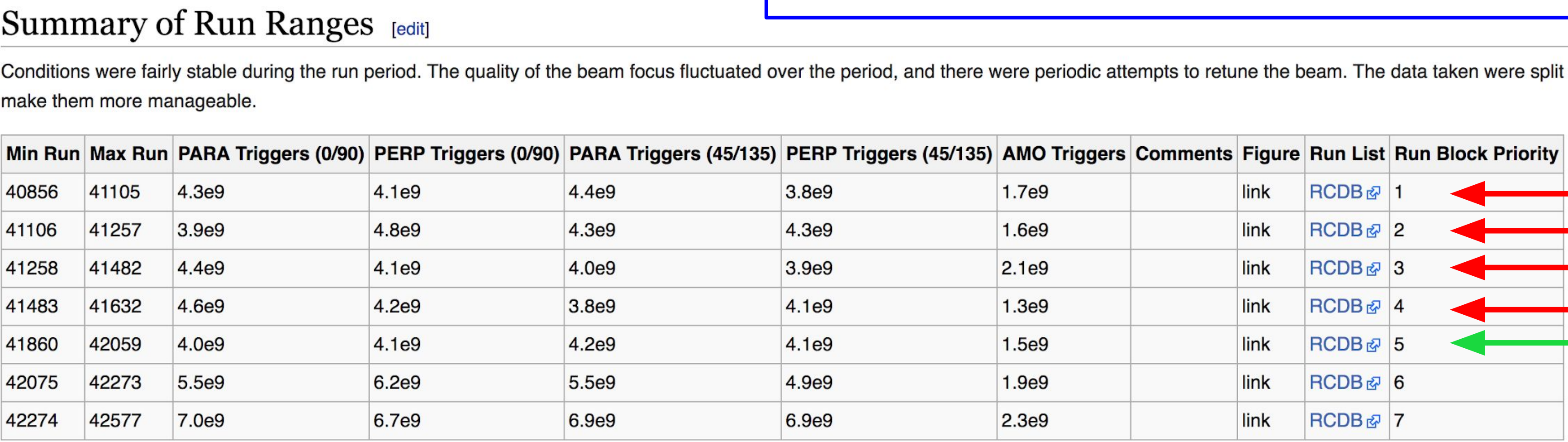
Repo	Admin Status	Alloc Type	Initial Alloc	Current Alloc	Balance	% Used	Hrs Used	Charged	Avg CF	Last Charged On
m3120	Active	DOE Mission Science	35,000,000	44,000,000	16,224,042	63	27,775,958	27,775,958	1.0	07-MAY-2019

used: 63.0% of allocation

batches01-04: 50.7% of data

KNL for all batches will use entire allocation after batch 6 (i.e. only 80% of data processed)

Switch to Haswell for batches 6 & 7 will allow 100% to be done at NERSC



Summary of Run Ranges [edit]

Conditions were fairly stable during the run period. The quality of the beam focus fluctuated over the period, and there were periodic attempts to retune the beam. The data taken were split up into 7 ranges to make them more manageable.

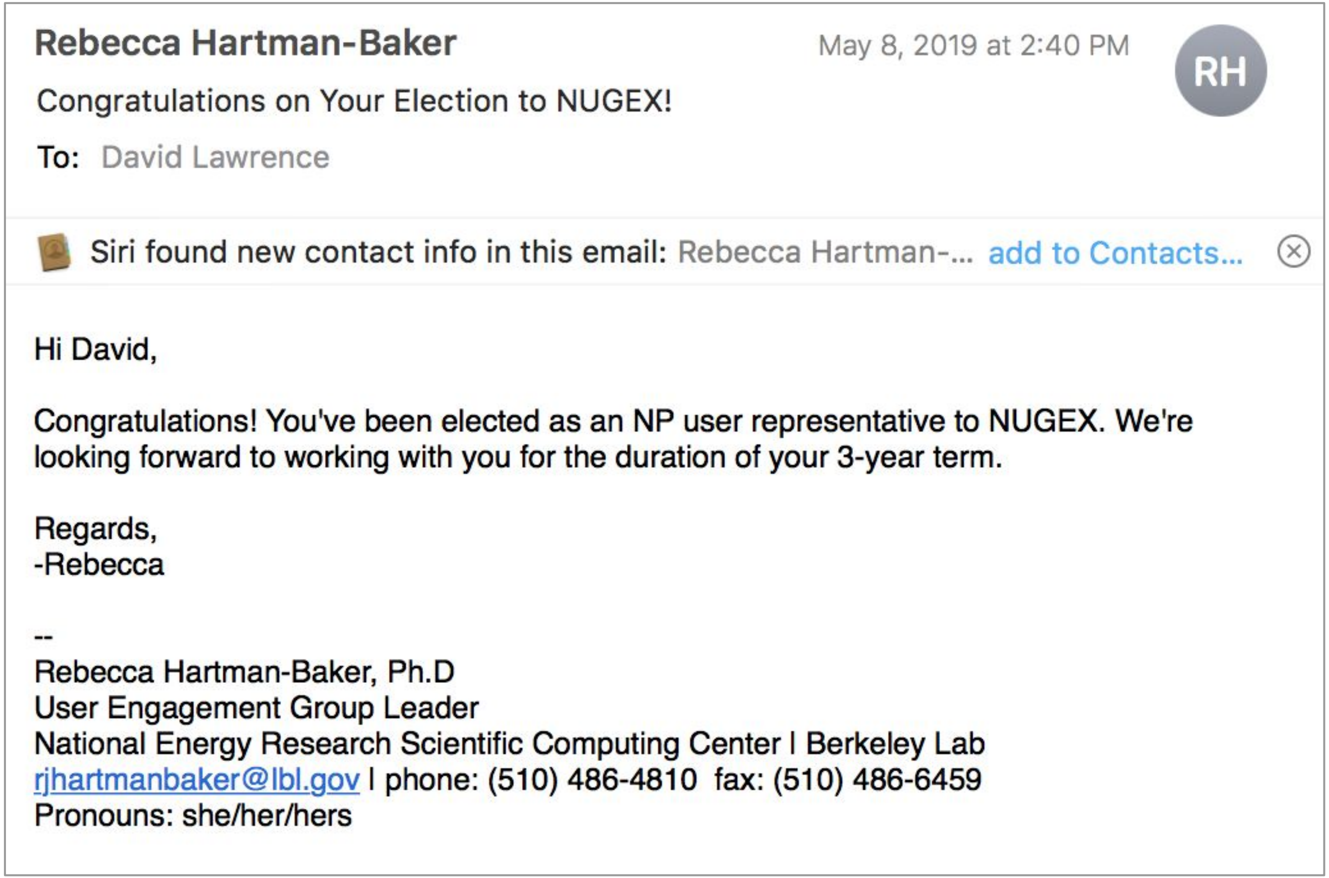
Min Run	Max Run	PARA Triggers (0/90)	PERP Triggers (0/90)	PARA Triggers (45/135)	PERP Triggers (45/135)	AMO Triggers	Comments	Figure	Run List	Run Block Priority
40856	41105	4.3e9	4.1e9	4.4e9	3.8e9	1.7e9		link	RCDB	1
41106	41257	3.9e9	4.8e9	4.3e9	4.3e9	1.6e9		link	RCDB	2
41258	41482	4.4e9	4.1e9	4.0e9	3.9e9	2.1e9		link	RCDB	3
41483	41632	4.6e9	4.2e9	3.8e9	4.1e9	1.3e9		link	RCDB	4
41860	42059	4.0e9	4.1e9	4.2e9	4.1e9	1.5e9		link	RCDB	5
42075	42273	5.5e9	6.2e9	5.5e9	4.9e9	1.9e9		link	RCDB	6
42274	42577	7.0e9	6.7e9	6.9e9	6.9e9	2.3e9		link	RCDB	7

completed

ongoing

https://halldweb.jlab.org/wiki-private/index.php/Spring_2018_Dataset_Summary

NERSC User's Group Executive Committee



Rebecca Hartman-Baker May 8, 2019 at 2:40 PM

Congratulations on Your Election to NUGEX!

To: David Lawrence

Siri found new contact info in this email: Rebecca Hartman-... add to Contacts...

Hi David,

Congratulations! You've been elected as an NP user representative to NUGEX. We're looking forward to working with you for the duration of your 3-year term.

Regards,
-Rebecca

--
Rebecca Hartman-Baker, Ph.D
User Engagement Group Leader
National Energy Research Scientific Computing Center | Berkeley Lab
rjhartmanbaker@lbl.gov | phone: (510) 486-4810 fax: (510) 486-6459
Pronouns: she/her/hers

Other Computing Facilities

Pittsburgh Supercomputing Center (Bridges)

- 752 RSM nodes (128GB, 28 cores (no hyperthreading)), ~21k cores
- ~4k cores on higher end computers with large memory

BRIDGES
A PITTSBURGH SUPERCOMPUTING CENTER RESOURCE

Indiana University:

- Big Red II: ~1k nodes, ~22k cores
- Corbonate: 72 nodes, ~1.7k cores
- Karst: 256 nodes, 4k cores

OSG ?

Proposal to XSEDE for PSC (Pittsburgh Supercomputer Center)

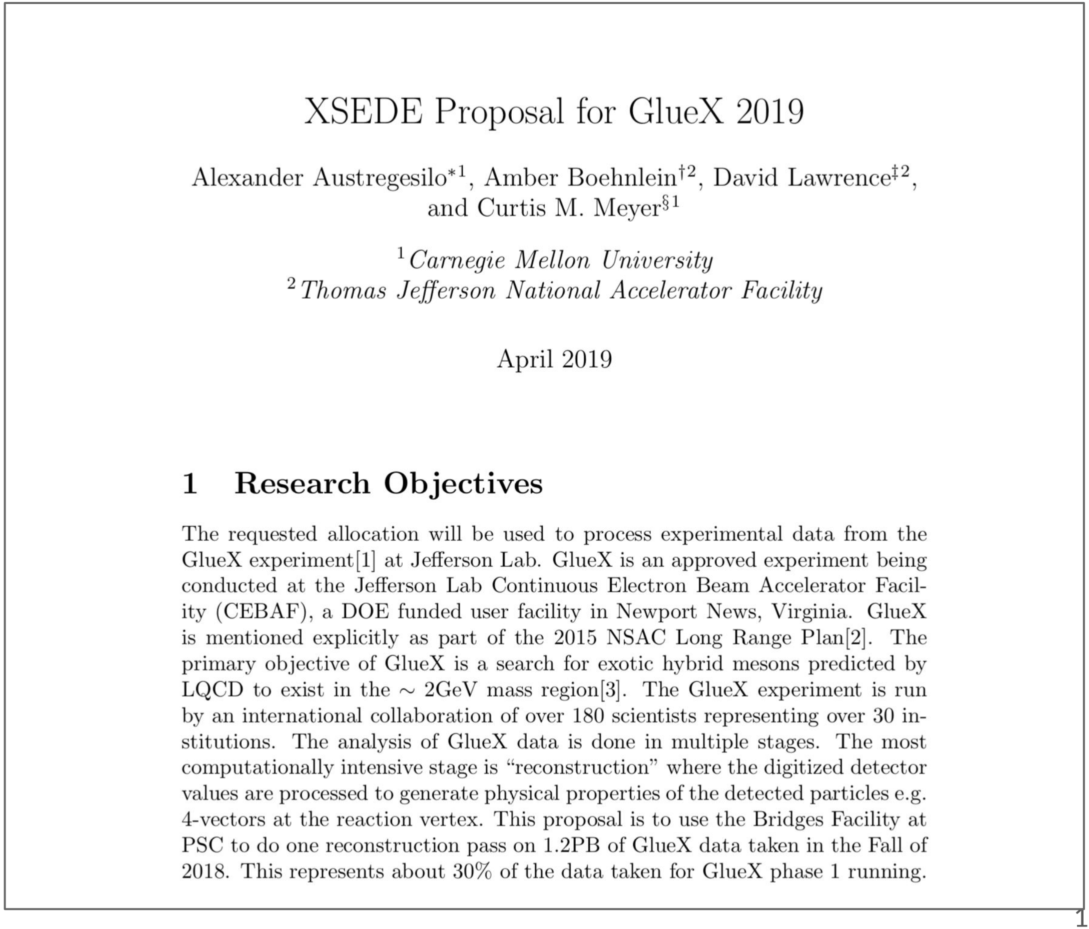
Allocation Request:

- 7.65M SU on PSC RSM
- 34TB storage

where:
SU = Standard Units
RSM = Regular Shared Memory (128GB)

➢ ***Each 20GB file will require 130 SUs to process***

➢ ***7.65M SU required for one reconstruction pass over 2018-08 data***



XSEDE Proposal for GlueX 2019

Alexander Austregesilo*[1], Amber Boehnlein†[2], David Lawrence‡[2], and Curtis M. Meyer§[1]

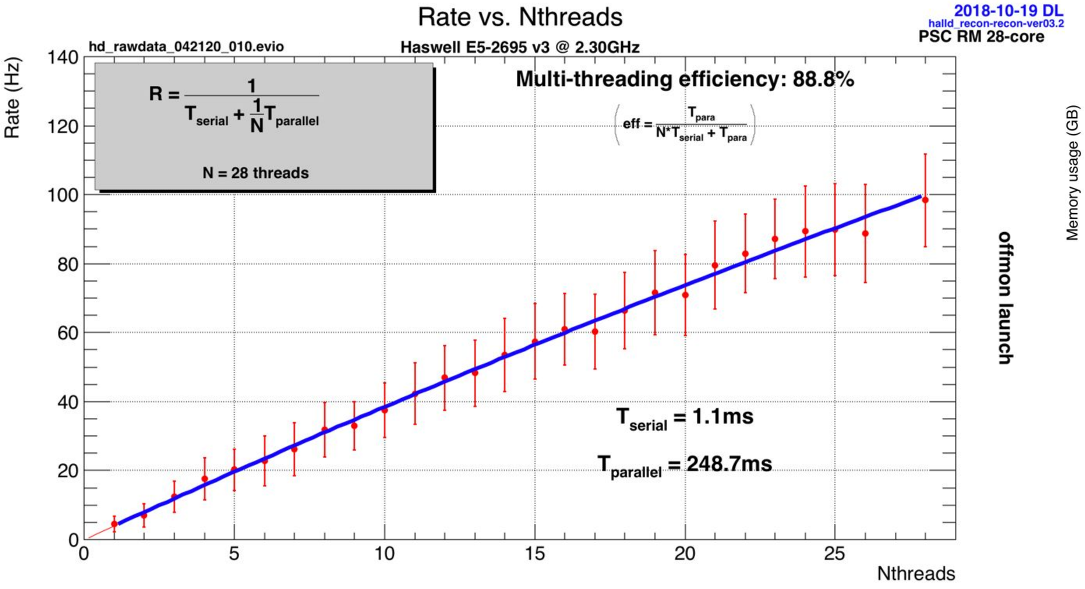
[1] *Carnegie Mellon University*
[2] *Thomas Jefferson National Accelerator Facility*

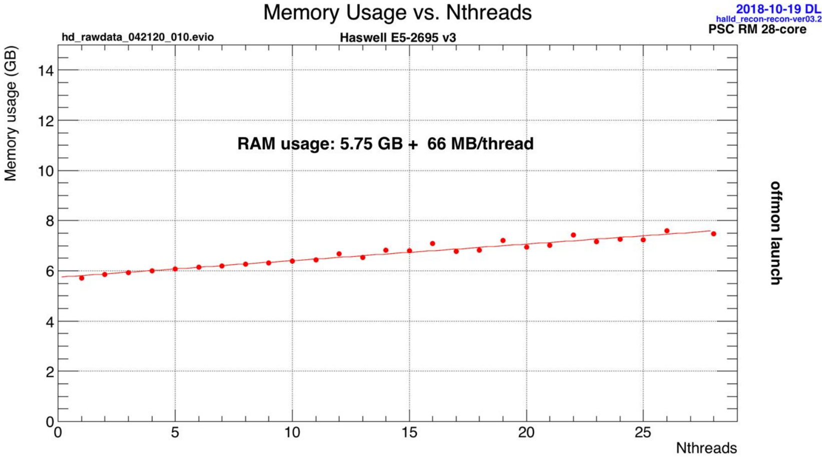
April 2019

1 Research Objectives

The requested allocation will be used to process experimental data from the GlueX experiment[1] at Jefferson Lab. GlueX is an approved experiment being conducted at the Jefferson Lab Continuous Electron Beam Accelerator Facility (CEBAF), a DOE funded user facility in Newport News, Virginia. GlueX is mentioned explicitly as part of the 2015 NSAC Long Range Plan[2]. The primary objective of GlueX is a search for exotic hybrid mesons predicted by LQCD to exist in the ∼ 2GeV mass region[3]. The GlueX experiment is run by an international collaboration of over 180 scientists representing over 30 institutions. The analysis of GlueX data is done in multiple stages. The most computationally intensive stage is "reconstruction" where the digitized detector values are processed to generate physical properties of the detected particles e.g. 4-vectors at the reaction vertex. This proposal is to use the Bridges Facility at PSC to do one reconstruction pass on 1.2PB of GlueX data taken in the Fall of 2018. This represents about 30% of the data taken for GlueX phase 1 running.

Performance Testing on Bridges at PSC





<u>Each node has:</u>

- ➢ **128 GB RAM**
- ➢ **2 Intel Haswell (E5-2695 v3) CPUs; 14 cores/CPU; 2.3 - 3.3 GHz**

Summary

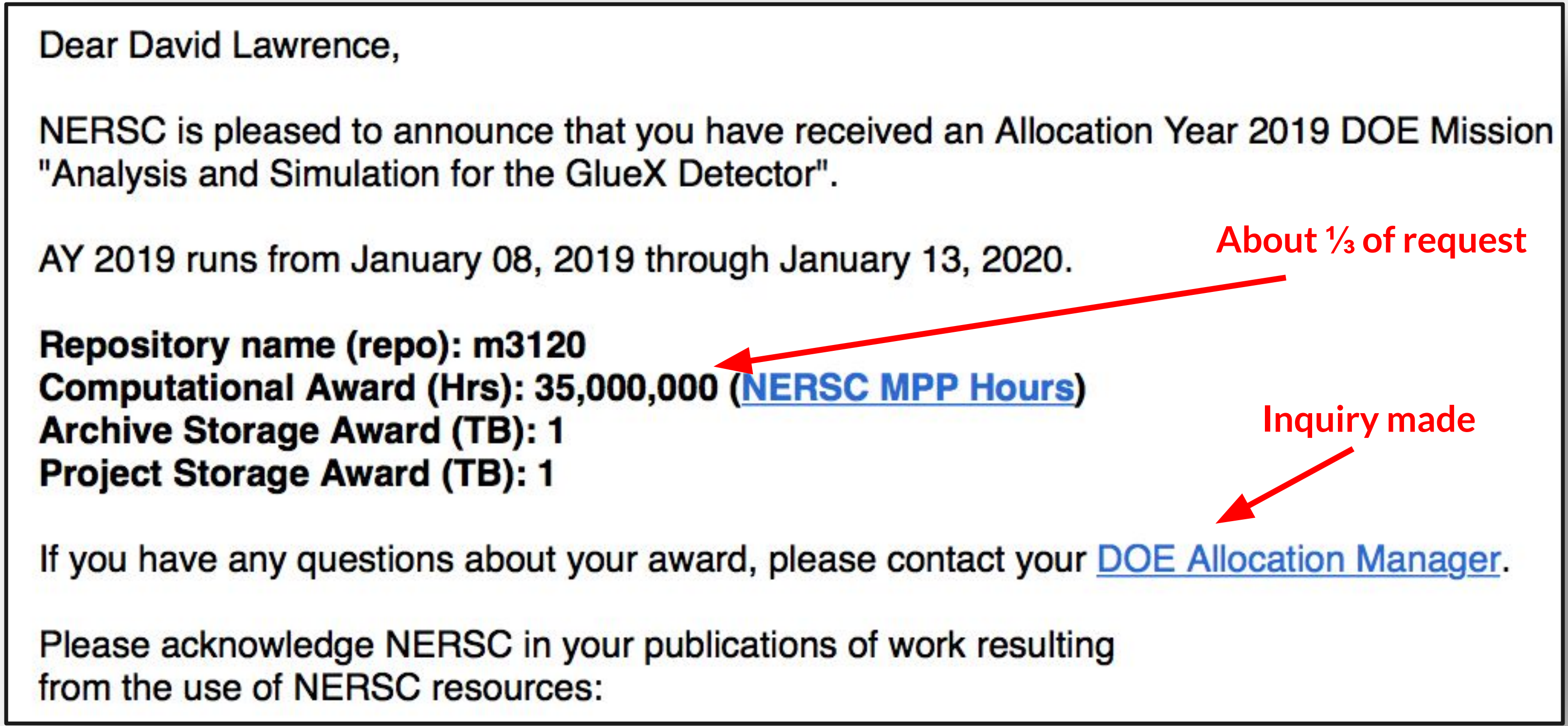
- First recon launch at NERSC for 2018-01 data underway and progressing slower than first anticipated due to transfer rates offsite
  - If rate no improvement made, Spring 2018 data processing will take to the end of June
  - May run one batch at JLab

- NERSC Allocation for 2019 only ⅓ of request and made worse by using only KNL

- XSEDE Proposal submitted for PSC (access should come in July)

Backups

NERSC Allocation Award for GlueX 2019



Dear David Lawrence,

NERSC is pleased to announce that you have received an Allocation Year 2019 DOE Mission "Analysis and Simulation for the GlueX Detector".

AY 2019 runs from January 08, 2019 through January 13, 2020.

Repository name (repo): m3120
Computational Award (Hrs): 35,000,000 (NERSC MPP Hours)
Archive Storage Award (TB): 1
Project Storage Award (TB): 1

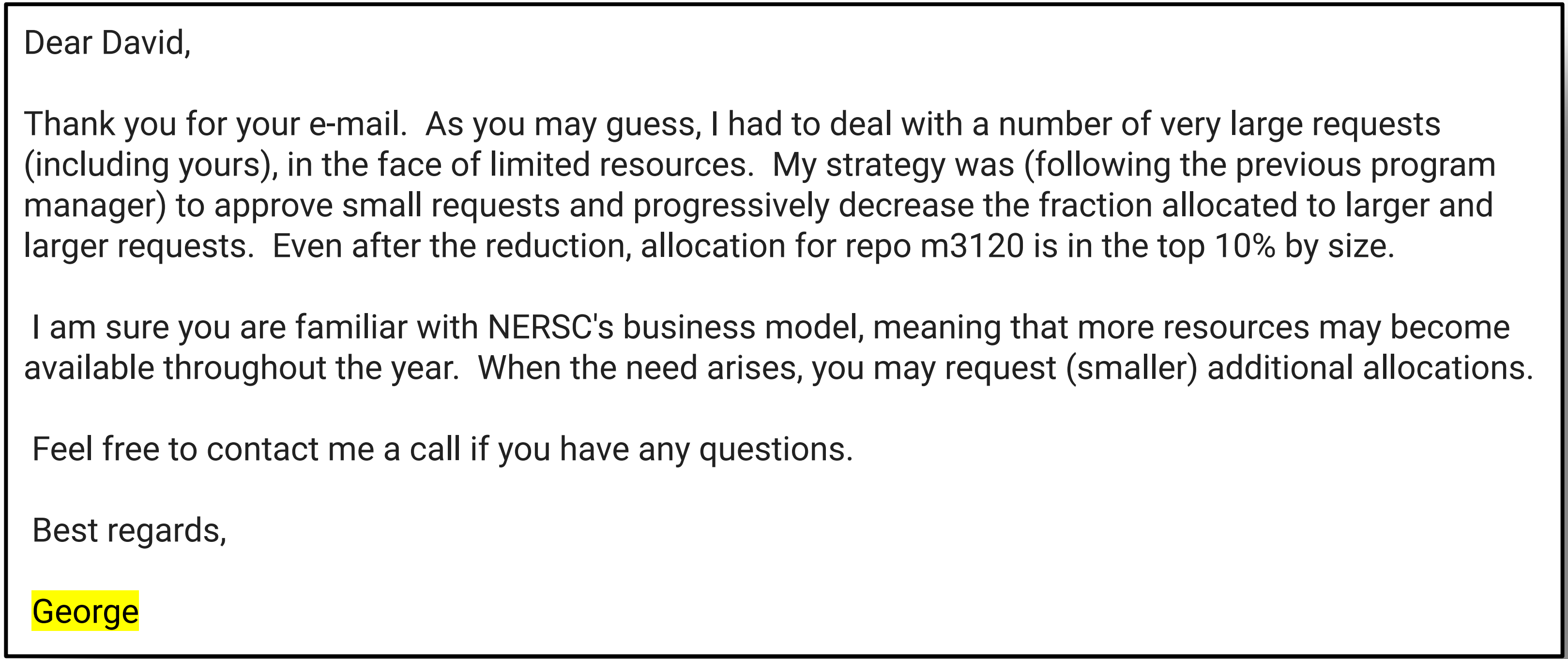
If you have any questions about your award, please contact your DOE Allocation Manager.

Please acknowledge NERSC in your publications of work resulting from the use of NERSC resources:

About ⅓ of request

Inquiry made

NERSC Allocation Award for GlueX 2019



Dear David,

Thank you for your e-mail. As you may guess, I had to deal with a number of very large requests (including yours), in the face of limited resources. My strategy was (following the previous program manager) to approve small requests and progressively decrease the fraction allocated to larger and larger requests. Even after the reduction, allocation for repo m3120 is in the top 10% by size.

I am sure you are familiar with NERSC's business model, meaning that more resources may become available throughout the year. When the need arises, you may request (smaller) additional allocations.

Feel free to contact me a call if you have any questions.

Best regards,

George

Other Issues

NERSC

- Jobs at NERSC tend to be tens of nodes for tens of hours
- Scheduler works (mainly) on job units rather than core-hours which can sometimes block our jobs
- Recommendation was for us to bundle jobs and/or use checkpointing
- Bundling is complicated and would block us if there are “holes” we’d otherwise fit into

OSG

- Single core jobs
- Require splitting existing jobs to smaller pieces and merging outputs

How fast we can process 2018-01 Data

- Transfer of 1.5PB over 10Gbps transfer link would take **~2.5 weeks for one pass**
  - *Factor 2 compression of data may cut this in half*

- One 20GB file job takes ~3 hours = 1.9MB/s

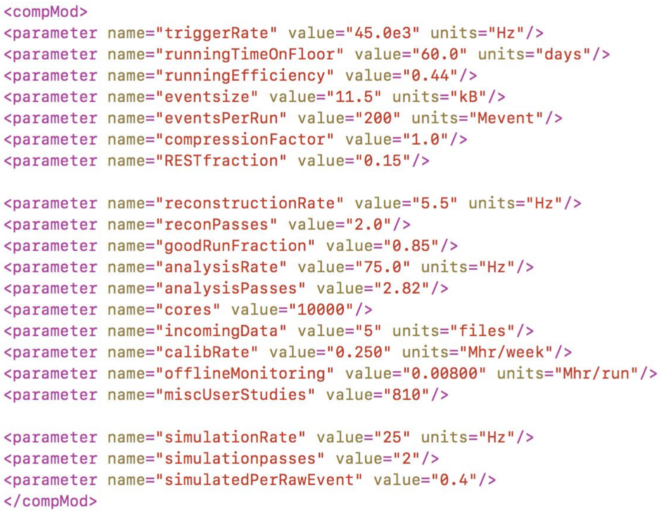
- With 10Gbps offsite bandwidth we can process up to 526 (uncompressed) files continuously

- Realistically, we may only have ~60% of that bandwidth now, but may have x10 as much in 2020

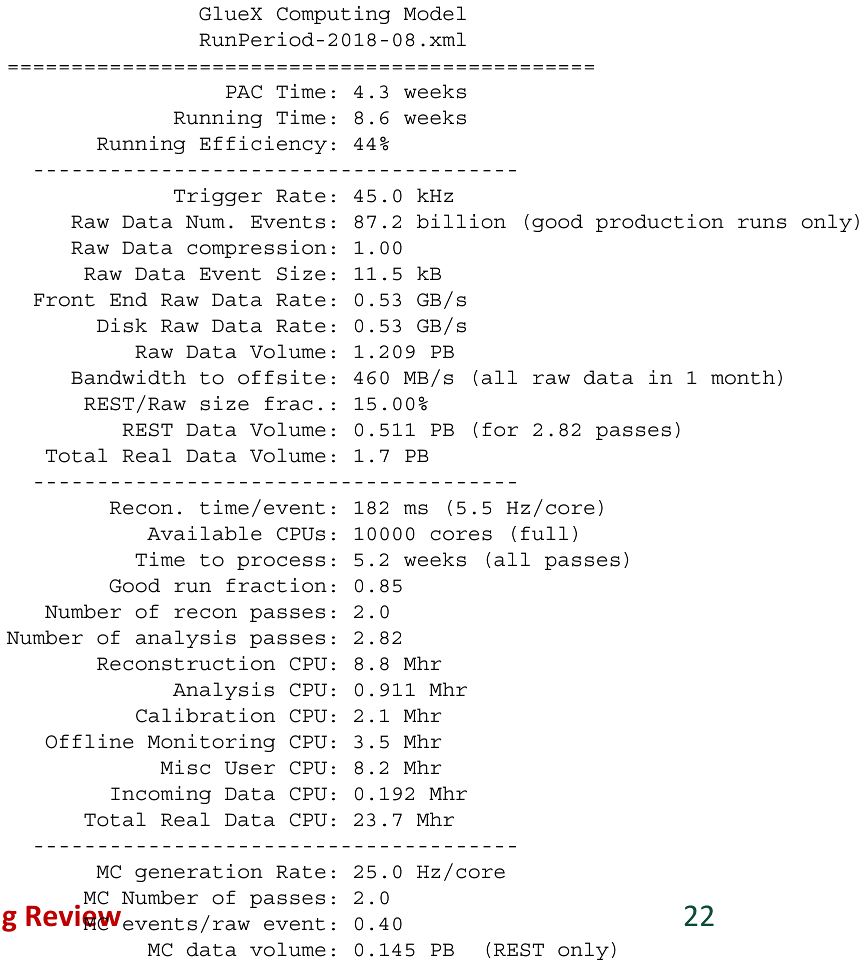
GlueX Computing Resource Model

A model was developed based on experience processing 2017 GlueX data to estimate compute resources required based on several inputs

https://github.com/JeffersonLab/hd_utilities/tree/master/comp_mod

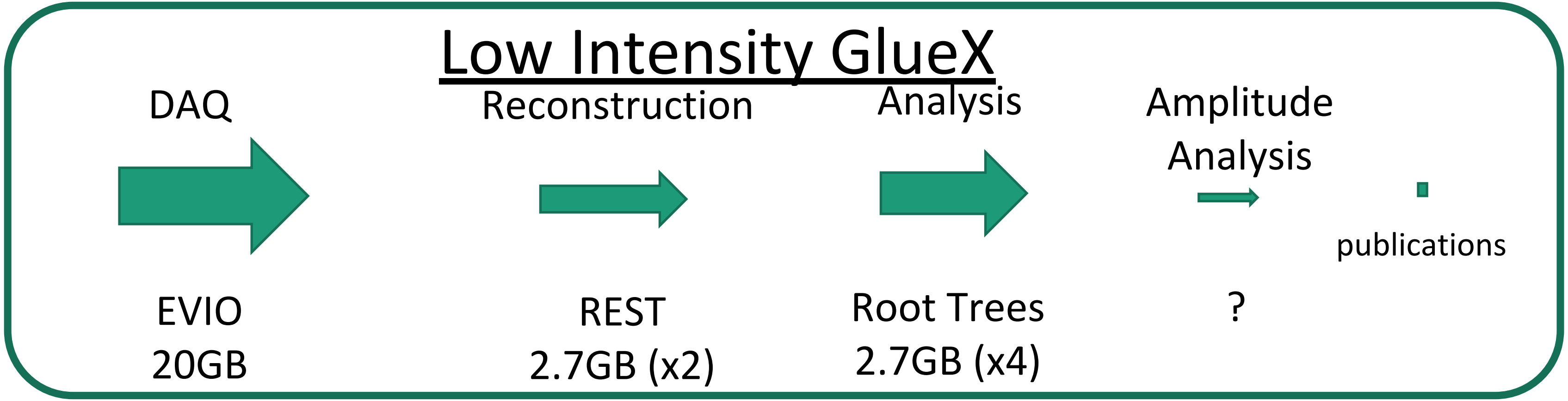


```
<compMod>
<parameter name="triggerRate" value="45.0e3" units="Hz"/>
<parameter name="runningTimeOnFloor" value="60.0" units="days"/>
<parameter name="runningEfficiency" value="0.44"/>
<parameter name="eventsize" value="11.5" units="kB"/>
<parameter name="eventsPerRun" value="200" units="Mevent"/>
<parameter name="compressionFactor" value="1.0"/>
<parameter name="RESTfraction" value="0.15"/>

<parameter name="reconstructionRate" value="5.5" 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="10000"/>
<parameter name="incomingData" value="5" units="files"/>
<parameter name="calibRate" value="0.250" 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="0.4"/>
</compMod>
```



```
               GlueX Computing Model
               RunPeriod-2018-08.xml
==============================================
                 PAC Time: 4.3 weeks
             Running Time: 8.6 weeks
       Running Efficiency: 44%
  --------------------------------------
             Trigger Rate: 45.0 kHz
     Raw Data Num. Events: 87.2 billion (good production runs only)
     Raw Data compression: 1.00
      Raw Data Event Size: 11.5 kB
  Front End Raw Data Rate: 0.53 GB/s
       Disk Raw Data Rate: 0.53 GB/s
          Raw Data Volume: 1.209 PB
     Bandwidth to offsite: 460 MB/s (all raw data in 1 month)
      REST/Raw size frac.: 15.00%
         REST Data Volume: 0.511 PB (for 2.82 passes)
   Total Real Data Volume: 1.7 PB
  --------------------------------------
        Recon. time/event: 182 ms (5.5 Hz/core)
           Available CPUs: 10000 cores (full)
          Time to process: 5.2 weeks (all passes)
        Good run fraction: 0.85
   Number of recon passes: 2.0
Number of analysis passes: 2.82
       Reconstruction CPU: 8.8 Mhr
             Analysis CPU: 0.911 Mhr
          Calibration CPU: 2.1 Mhr
   Offline Monitoring CPU: 3.5 Mhr
            Misc User CPU: 8.2 Mhr
        Incoming Data CPU: 0.192 Mhr
      Total Real Data CPU: 23.7 Mhr
  --------------------------------------
       MC generation Rate: 25.0 Hz/core
      MC Number of passes: 2.0
      MC events/raw event: 0.40
           MC data volume: 0.145 PB  (REST only)
```

Data volumes and high-level data flow



Low Intensity GlueX

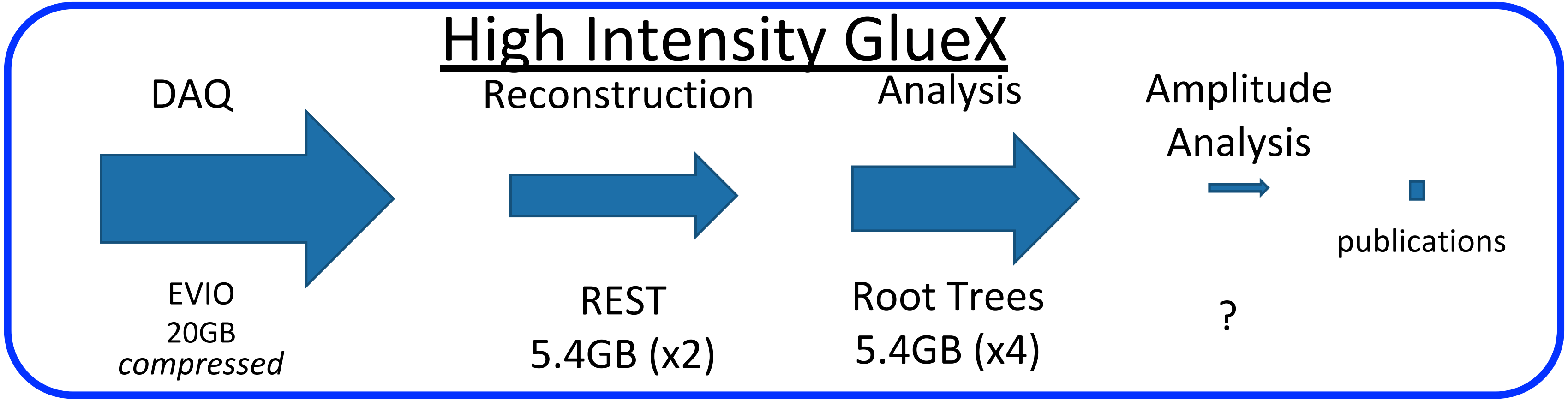
DAQ → Reconstruction → Analysis → Amplitude Analysis → publications

EVIO 20GB

REST 2.7GB (x2)

Root Trees 2.7GB (x4)

?



High Intensity GlueX

DAQ → Reconstruction → Analysis → Amplitude Analysis → publications

EVIO 20GB *compressed*

REST 5.4GB (x2)

Root Trees 5.4GB (x4)

?