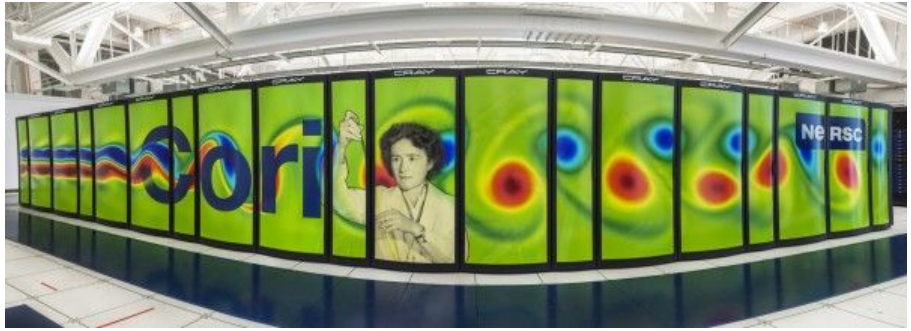
# NERSC + swif2

processing raw data offsite

David Lawrence - JLab
July 24, 2018

# NERSC = National Energy Research Scientific Computing 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 **~8k-10.5k full cores**

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

**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*

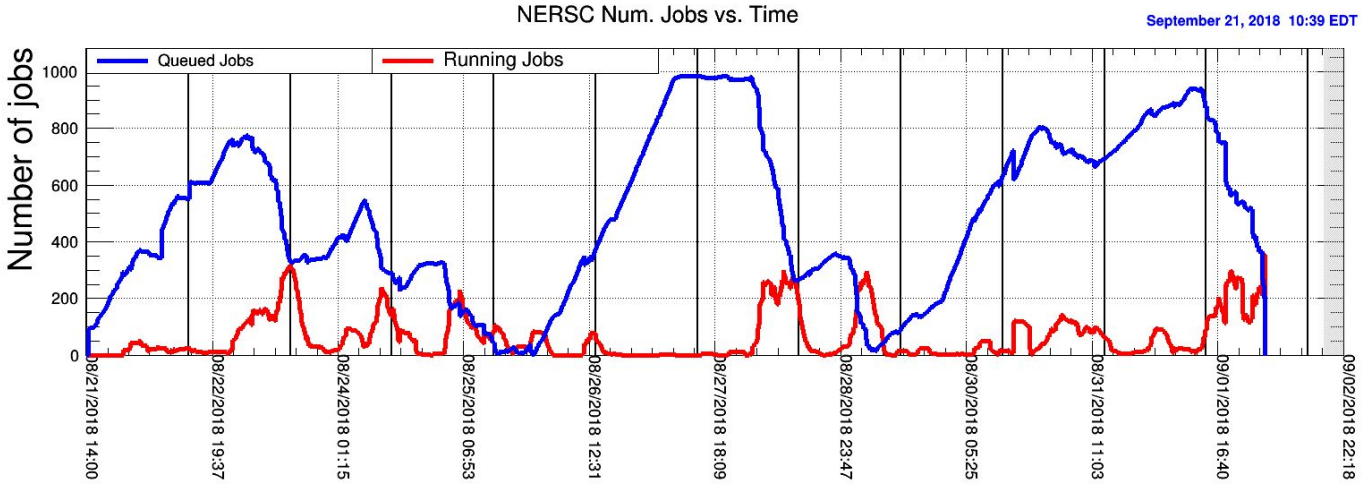
# Anatomy of a swif2 job

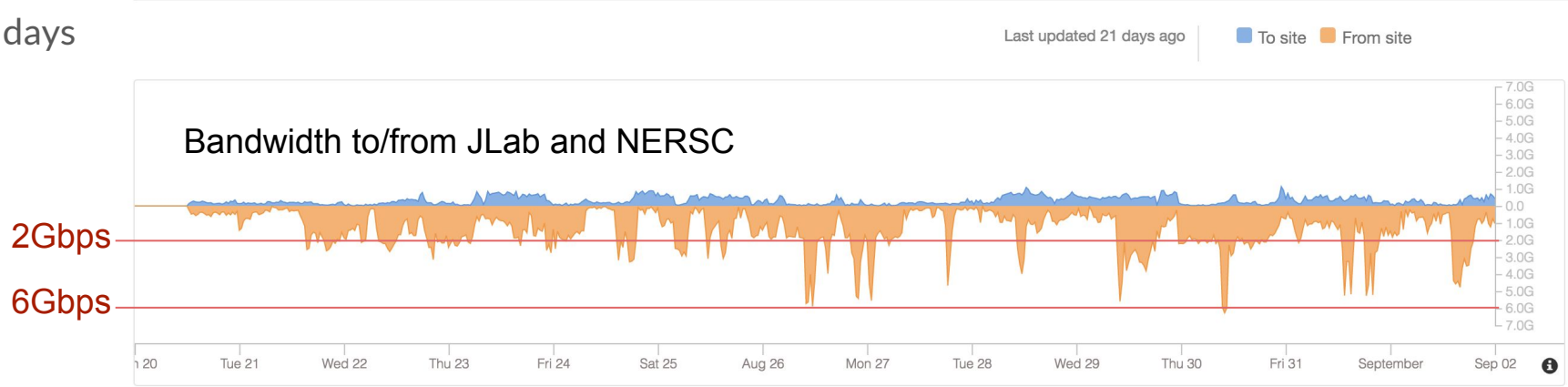
- Jobs are submitted by semi-complex chain of scripts
- Critical information for job passed through combination of arguments, installed script files, and CVMFS
- Very similar to submitting jobs to JLab farm
  - JLab farm jobs can move files directly from job script *(item 5)*
  - NERSC jobs must specify files to move a priori *(item 2)*
- Uses same container and CVMFS mount as OSG jobs

1. launch_nersc.py
2. swif2 add-job ….. ← specify job params and all SLURM options (including container image)
3. run-shifter.sh <args> ← nersc job must be a script
4. shifter --module=cvmfs -- /launch/run_job_nersc.sh /launch/jana_recon.config sim-recon-2.27.0
5. run_job_nersc.sh ← setup environment, copy CCDB, RCDB SQLite files, ...
6. hd_root --config=jana_recon.config hd_rawdata_041137_003.evio

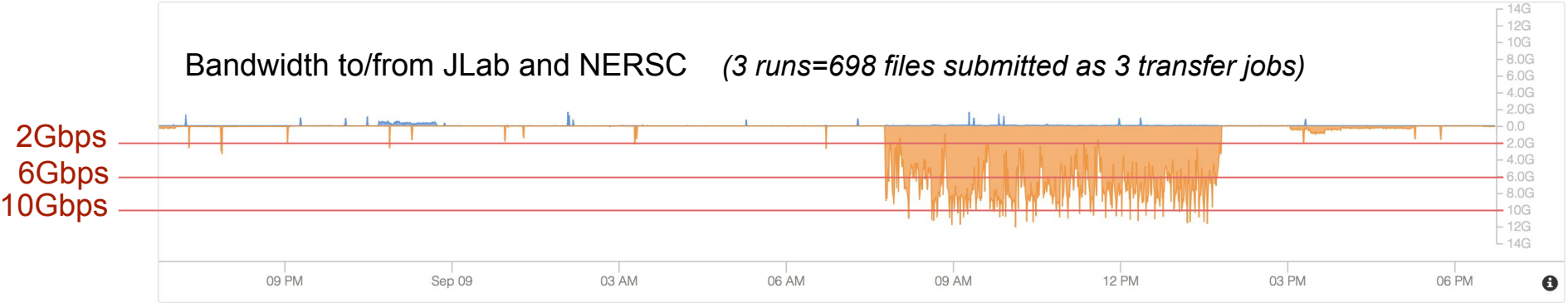
# First Monitoring Launch at NERSC

- RunPeriod-2018-01
- First 10 files of each run
- 5495 jobs
- 12 days

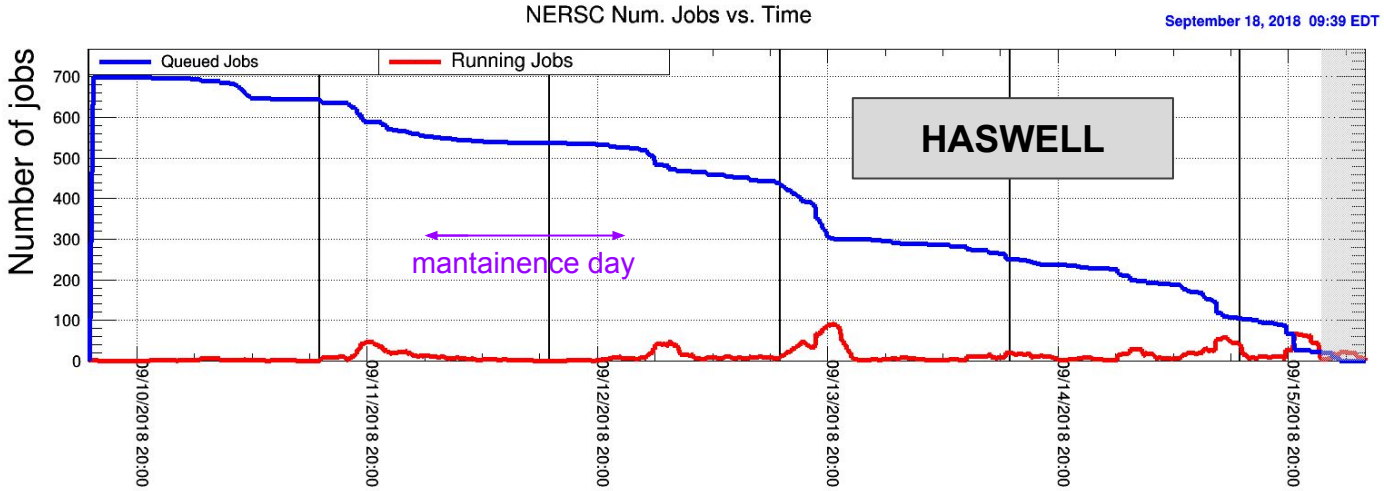


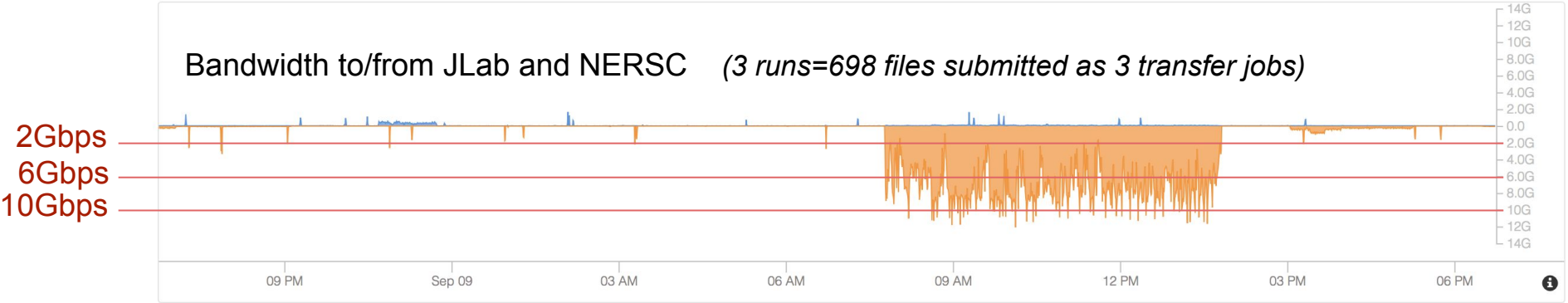
# 14TB test



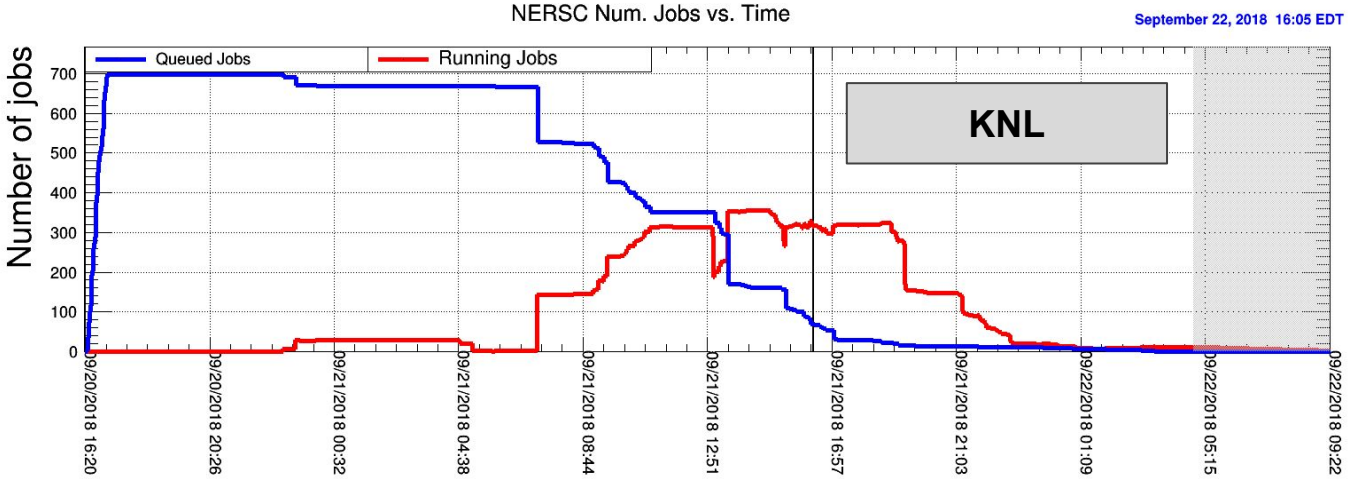

- All jobs submitted at once on Cori I (haswell)
- ~4.5 days for all jobs to complete
- <100 jobs run simultaneously



# 14TB test




- All jobs submitted at once on Cori II (KNL)
- <2 days for all jobs to complete
- ~350 jobs run simultaneously



# 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

# NERSC Accounting

- Cori I (Haswell) nodes: 32 cores + 32 ht
- ~2.9hr wall clock time per job (=20GB file)
- 80 NERSC units per node-hour (haswell)
  - ~232 NERSC units per 20GB EVIO file
- 96 NERSC units per node-hour (KNL)
  - ~557 NERSC units per 20GB EVIO file

**Roughly 1.5PB/20GB = 75k files for RunPeriod2018-01 data set**

**75k*232 = 17.4M NERSC Units for one recon pass *(haswell)***
**75k*557 = 41.8M NERSC Units for one recon pass *(KNL)***

My Repo Usage

| Repo | Type | Repo Balance | My Usage |
|---|---|---|---|
| noexp | EXP | 0 | 0 |
| m2828 | REPO | 89,854 | 10,146 |
| m3120 | REPO | 47,621,762 | 2,378,166 |
| m2828 | STR | 500 | 0 |
| m3120 | STR | 500 | 0 |

Units are in Hrs for REPO accounts and SRU's for STR accounts.

# Numbers Shown at 6/24/2018 Exp. Readiness Review

*Based on High Intensity Running*

| | 1 year | Total |
|---|---|---|
| Real Data Volume | 9.4 PB | |
| MC Data Volume | 6.1 PB | |
| **Total Data Volume** | **15.5 PB** | **34.8 PB** |
| Real Data CPU | 109 Mhr | |
| MC CPU | 160 Mhr | |
| **Total CPU** | **269 Mhr** | **603 Mhr** |

Need: **269** Mhr
OSG: **50** Mhr
NERSC: **70** Mhr
JLab: **149** Mhr

if spread over 9 months of continuous computation, GlueX will need (149 Mhr) / (274 days) =

**23k cores @ JLab**

# Proposed NERSC Request for FY2019

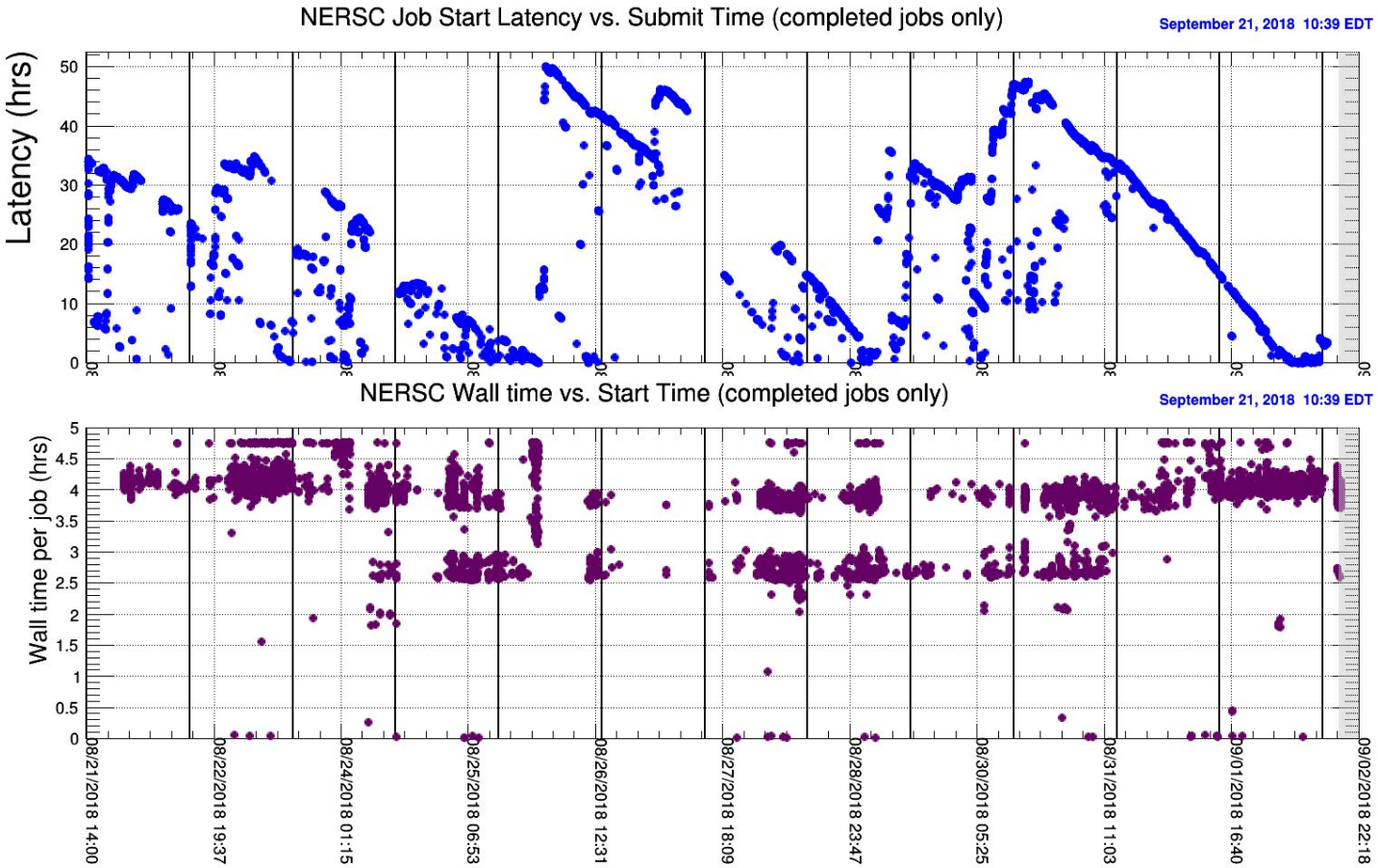
- **4.5PB total transfer to NERSC**
  - 2 passes of 2018-01 data @ 1.5PB each
  - 1 pass of 2018-08 @ 1.5PB
- **90M NERSC units**
  - 4.5PB/20GB = 225k jobs
  - assume ½ jobs on Haswell, ½ on KNL = ½(232 + 557) = 395 NERSC units/job avg.
- **5 weeks/pass (15 weeks total)**
  - assume 5Gbps bandwidth for 4.5PB data
- **270 simultaneous jobs on avg.**
  - 5weeks/(3hr/job) = 280 jobs/node if run continuously over 5 weeks
  - 75k jobs/280 jobs/node = 268 nodes
  - monitoring launch: ~62 simultaneous jobs on avg.
  - 14TB test: ~87 simultaneous jobs on avg.

# Summary

- ~5.5k jobs have been successfully run at NERSC/Cori I using swif2
  - Monitoring launch 2018-01-ver17
  - Limited by combination of data transfer and farm availability
- Transfer rates >6Gbps with present link have been demonstrated
  - Need modification to swif2 to bundle transfer requests
  - Need to test compression of data for transfer
- New request due for NERSC AY2019
  - 90M units
- Must improve avg. number of simultaneous jobs running to ~270
  - Need to test simultaneous running on Cori I and Cori II

# Backups

# First Monitoring Launch at NERSC



# swif2

*commands formed and issued via python script*

```
mkdir -p /volatile/halld/data_challenge/nersc_01/041137/231

chmod 777 /volatile/halld/data_challenge/nersc_01/041137/231

swif2 add-job -workflow nersc_test_01 -name GLUEX_RECON_041137_231 -input
hd_rawdata_041137_231.evio
mss:/mss/halld/RunPeriod-2018-01/rawdata/Run041137/hd_rawdata_041137_231.evio -output
match:* file:/volatile/halld/data_challenge/nersc_01/041137/231 -sbatch -A m3120
--volume="/global/project/projectdirs/m3120/launch:/launch"
--image=docker:markito3/gluex_docker_devel --time=3:30:00 --nodes=1 --tasks-per-node=1
--cpus-per-task=64 --qos=regular -C haswell -L project ::
/global/project/projectdirs/m3120/launch/run_shifter.sh --module=cvmfs --
/launch/run_job_nersc.sh /launch/jana_recon.config sim-recon-2.27.0
```