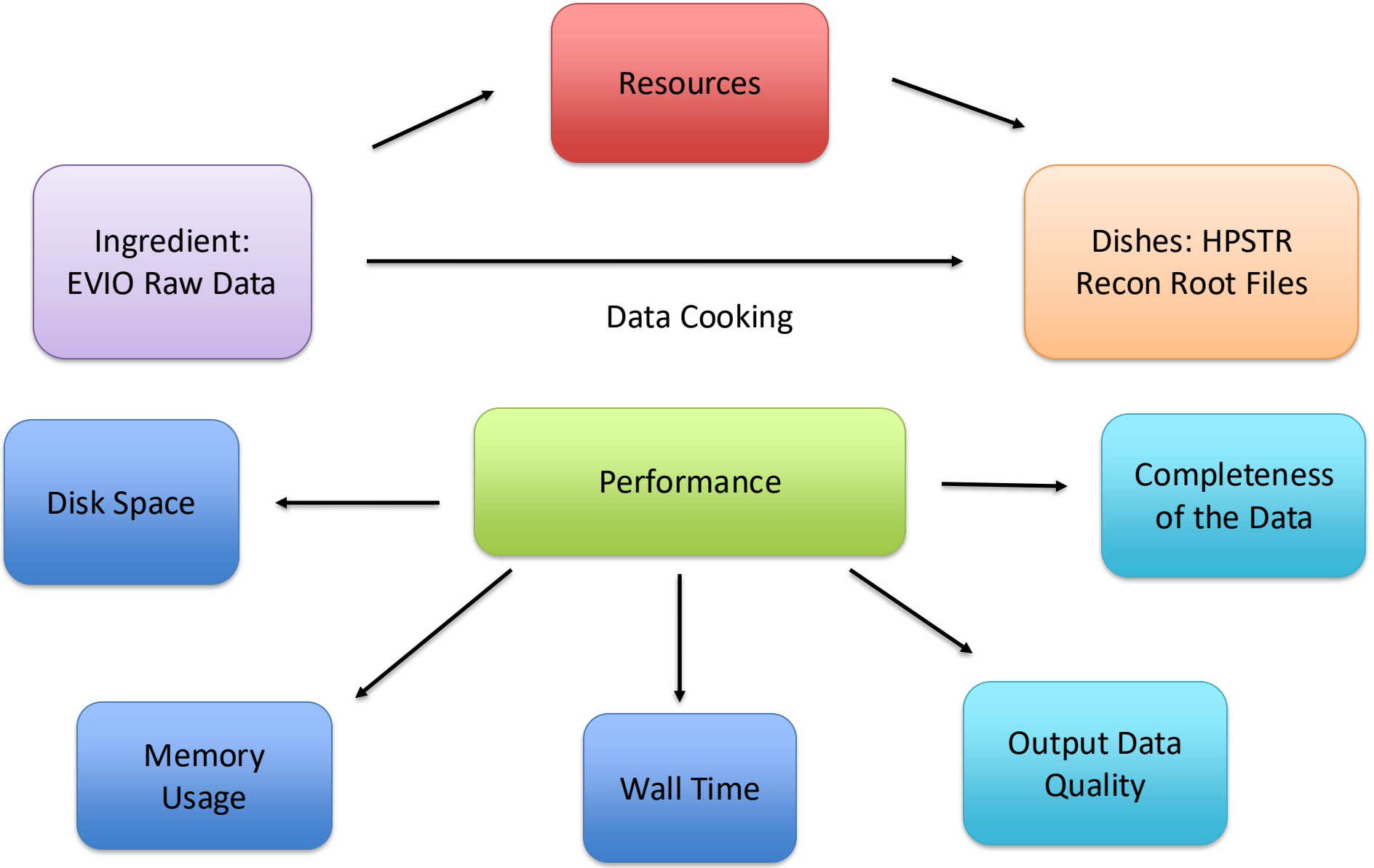
HPS JLAB Collaboration Meeting

Status of the 2021 Data Processing

Matt Gignac and Zhaozhong Shi

06/03/2025

General Philosophy of Data Production Management



Overview of the Raw Data and Final Goal

- Summary of Run 2021 EVIO Raw Dataset

Physics Runs	e^- Beam Energy	Total EVIO Files	Total Events	Computing Time
303	3.74 GeV	405807	49,025,855,852	2451293 hours

Moller Runs	e^- Beam Energy	Total EVIO Files	Total Events	Total Time
18	1.92 GeV	22700	4,438,137,145	221907 hours

- On JLab Ifarm tape: /mss/hallb/hps/physrun2021/data/
- Performance: 0.18 second/event
- Final goal: 100% pass
- Efficient use of resources that fits into the projected timeline of the collaboration (proposed by Matt Graham in our previous collaboration meeting)

1% Production
Goal: Mid January
✓ completed

10% Production
Goal: Early May
◌ Ongoing, expected by late June

100% Production
Goal: Early July
✗ not started yet, probably in August

- Potentially separate Moller and Physics Runs in different workflows

JLab-SLAC Ifarm/Globus Resources

Disk Storage

- JLab Ifarm cache disk storage available: 150 TB
  - Use for 1%, 10%, and 100% slcio and root file storage
- Volatile disk storage: 10 TB (50/40)
  - Use for intermediate studies and detector testing
- SLAC: about available 100 TB and will go up to 400 TB
  - HPSTR Data storage path: /sdf/group/hps/data

Scientific Computing

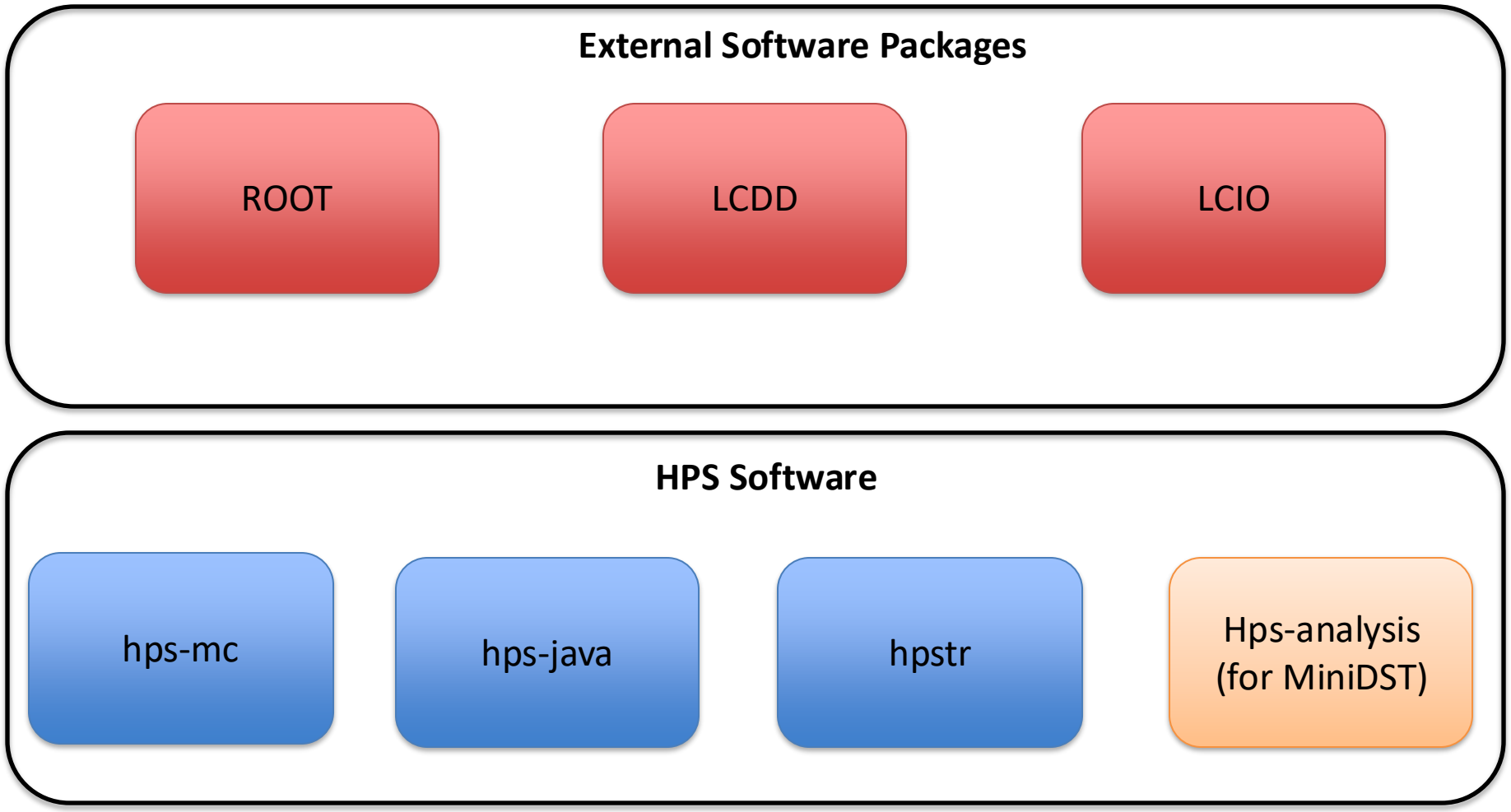
- Swif2 Parallel Processing Jobs: max_concurrent = 5000
- Memory limit: 2 GB
- Wall time limit: 24 h
- 1 EVIO per job in workflow
- Workflow using the hps as user for large scale production
- Use my own user account for V7 detector SVT alignment tuning jobs and Moller run processing previously

Globus

- File transfer speed: ~0.25 GB/s
- Potential parallel processing with multiple jobs
- Any limit for number of jobs or total rate?

- Optimal usage of the resources to achieve efficient data production

Software Organization



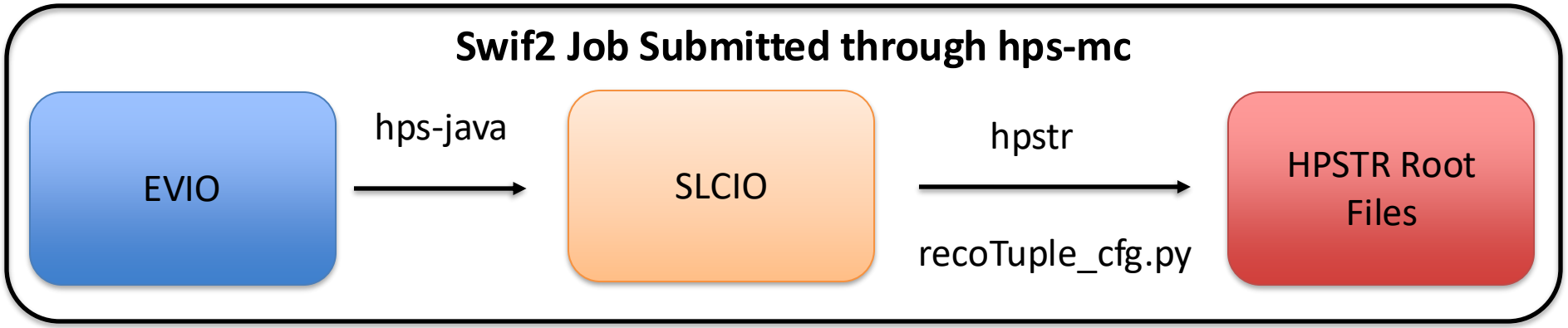
- Currently the are all build together under my own folder located at: /w/hallb-scshelf2102/hps/zshi/sw/
- Need to tag all the software, at least the HPS ones, for future productions for debugging and documentation purposes

Data Production Related Tasks Status

Tasks	Expected Completion Date	Current Status
1% data production with pass0 v6 Detector	01/2025	✓ Completed: 01/2025
Steering File Pass 0 → Pass 1 for the 1% Data Production	Discovered in 01/2025	✓ Fixed: 02/2025
Data Reduction	02/2025	✓ Completed 02/2025
SVT V7 Alignment	03/2025	✓ Completed: 04/2025
Java 21 Testing	03/2025	✓ Completed: 03/2025
Local SQLite Database Transition	03/2025	✓ Completed: 03/2025
V0Skim and validation with 0.3% data production	04/2025	✓ Completed: 04/2025
Develop final preselection	04/2025	◌ Ongoing
LCSIM/LCIO event model: new track state	04/2025	◌ Ongoing
SVT V8 Alignment	06/2025	◌ Ongoing

2021 1% Production Pass0 V6 Detector Summary

- Project Timeline: from **12/20/2024 to 01/10/2025**
  - Refer a lot to Cameron's previous work
  - Due to the unfamiliarity with HPS software and Jlab swif jobs and the Christmas break
  - Thanks for the help from Maurik, Matt Gignac, Nathan, and Tongtong to make it possible



- Based on pass0 steering file: PhysicsRun2021_pass0_recon.lcsim

1% evio runs	Number of Files	Output runs	Output slcio files	Output hpstr root files	Total Output Events
323	6013	323	6008	6008	604246912

- Stored at Jlab cache disk: /cache/hallb/hps/physrun2021/production/1percent/
- Resource usages
  - Computing time used: 50k hours (ideally 25k hours)
  - Disk usage on 43 TB

Experience from the 1% Production

Issues in terms of Production

- HPSTR memory usage
- Wrong computing nodes
- EVIOtoSLCIO stuck jobs (5 ultimately failed jobs)
- Cancelled jobs

Issues in terms of Data Quality

- The first batch of the jobs processed two evio files due to the unfamiliarity of the codes by me -> computing time increase by a factor of 2
- Wrong parameter in pass0 steering file: maxResidual
  - Pass1 removes hit collection
  - Pass2 removes unused collections while also filter events
- Wrong B field for Moller Runs (manually modified)

- A lot of lessons learned to improve both the efficiency and quality
- A good experience for me to understand HPS reconstruction and production

V6 and V7 Detectors

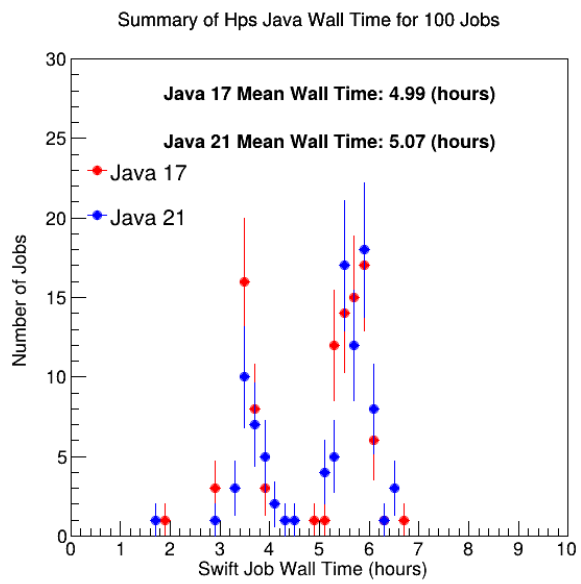
- We have detector version update from V6 to V7
  - Thanks Matt Gignac a lot for tuning the energy scale and Moller mass

- SVT Alignment to the HPS Detector Geometry
  - Detector with alignment organized in run dependent structures tagged individual run ID
  - Coordination of effort for Moller mass optimization in Moller runs through an iterative and feedback process to improve the SVT alignment

- Notable Differences
  - V6 detector we will need to manually modify the lcdd files to change the B field for Moller runs (1.92 GeV) but V7 it is already implemented
  - Runs changed: 11 deleted runs created since 01/27/2025
  - But looks like 1 Moller Run (14646) is marked as deleted but not deleted in V7 detectors → Manually discard in the Moller sample

Other Improvements for 10% Production

Java21 Testing

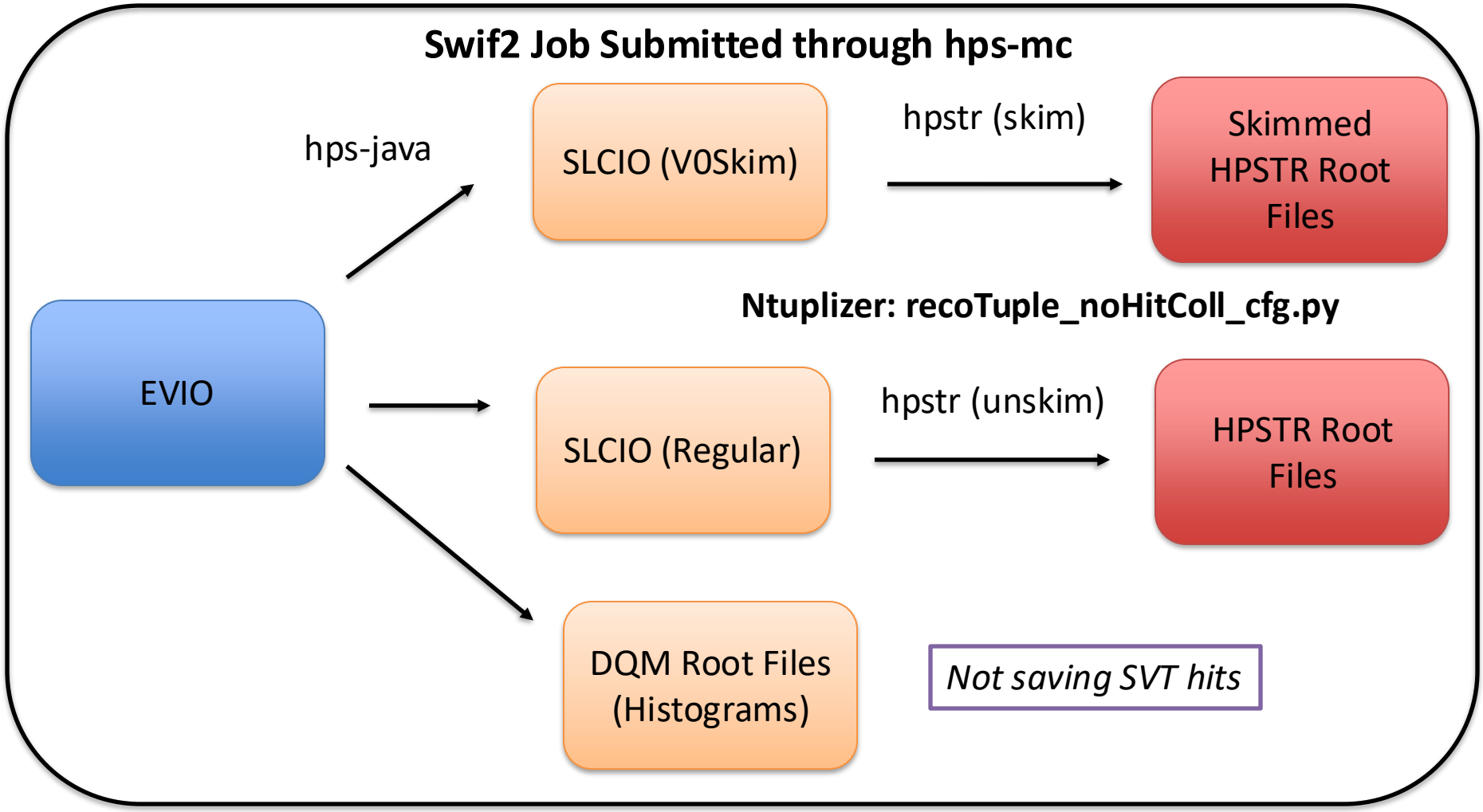


- Wall time performance comparable to java 17
- Can test 1% of the 10% production to see the improvement
- More technical details available in my talk on 02/25 Tuesday Meeting

Local SQLite Database Implementation

- Have a fixed location on scratch disk for hps-java to read the database information in each job
- Created a process (SQLiteProcComponent) to copy an existing file to the scratch disk in jobs
- Pull request merged to hps-mc
- Need to set a .hpsmc (or config) in the jobs to specify the customized [SQLiteProc] paths copy of snapshot of .sqlite database file

0.3% Production for Physics Run for V0skim Validation



- Pass2 steering file: [PhysicsRun2021_pass2_recon_skimmed_dataqual.lcsim]
- Collection reduction: 38 → 24
- 2 MB flat DQM root file size

0.3% Production Performance Summary

Summary of the Production (More Details on HPS Tuesday Meeting)

0.3% evio runs	Number of Files	Output Runs	Output Slcio Files	Output hpstr root files
295	1180	295	1178	1170

Average Events Unskimmed	Average Events Skimmed	Reduction Fraction
146647	14633	9.97%

SLCIO

Total Size Unskimmed	Total Size Skimmed	Reduction
1361.96 GB	175.43 GB	12.9%

HPSTR

Total Size Unskimmed	Total Size Skimmed	Reduction
660 GB	95 GB	14.3%

- Job time takes about 10 hours per job.
- Output files storage
  - Jlab Ifarm: /volatile/hallb/hps/zshi/Run2021/prod_0p1_v7_pass2/PhysicsRuns
  - SLAC: /sdf/data/hps/physics2021/data/hpstr/prod_v7_pass2_Prod/PhysicsRuns
- Projection for 10%
  - SLCIO: **5.8 TB** and HPSTR: **3.2 TB** and Job time: **80 hours** (4 days)

Readiness for 10% Data Production

What we have now

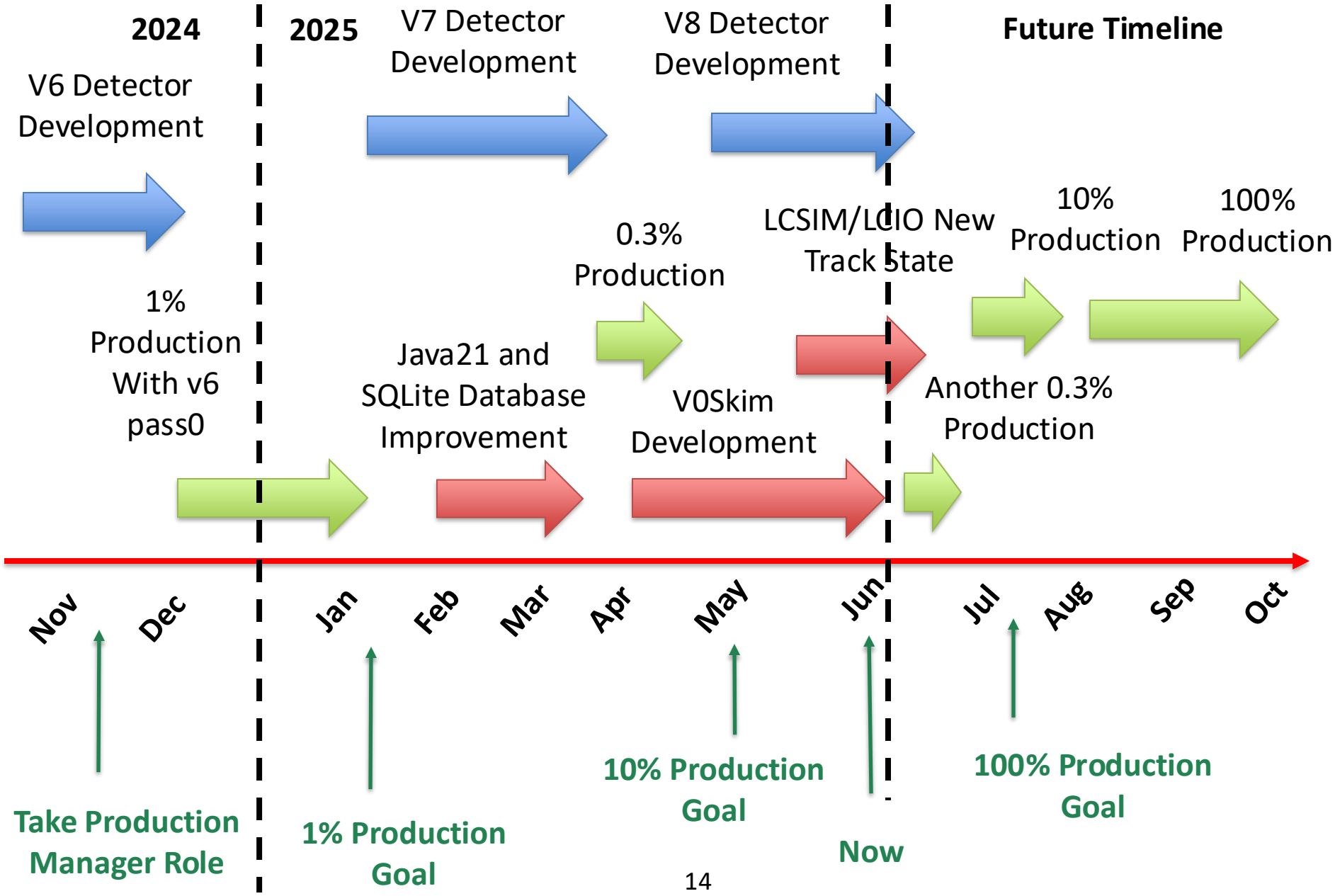
- Well developed v7 detector
- Local SQLite Database
- Java21
- Functioning V0skim
- Pass2 Steering File

What we need

- A new validated v8 detector with better alignment
- Finalized preselection
- New track state data
- Validation of the final setup
- Software tagging
- Optimization usage of resources

- An extra iteration with 0.3% data set with v8 detector for SVT hit killing studies
  - Coordinated with Matt Gignac, Matt Graham, and Elizabeth
- We should be able to push the start button for 10% production in **Late June**!

2021 Data Production Timeline Recap



Summary and To Do List

- Overall, we have a satisfying data production status
  - Resources available to efficiently handle data production
  - Output data quality keeps improving
  - Developing right procedures and good documentation

- 1% Pass0 production **complete** in 01/2025
  - Took 20 days for completion
  - Not usable due to multiple issues
  - Lots of lesson learned for future improvement

- 10% production **work-in-progress**
  - Still waiting for some improvement and another 0.3% iteration
  - Proposed to start in **late June**
  - Should take about 4 – 5 days to run

- 100% production **not yet started**
  - Depends a lot on 10% production
  - Optimistically happen **in August**
  - More optimization for the production needed

Back Up

Experience: Errors and Solutions

SLURM_OUT_OF_MEMORY

- Frequency: quite often with a large number of failed jobs in eviotoslcio + hpstr workflow
- Solutions: (1) increase the job memory limit to a larger value (2) break down the workflow to allow the job to finish

SLURM_REAP_FAIL

- Frequency: quite often with a small number of jobs
- Solution: keep retrying job, usually succeed after 1 retry

SLURM_CANCELLED

- Frequency: occasionally happens due to over time (24h) or job killed manually
- Solution: simply resubmit the jobs

SLURM_TIMEOUT

- Frequency: always happens due to over time (24h) for a small number of jobs
- Solution: can retrying job but eventually converge to several problematic input files

SLURM_NODE_FAIL and SWIF_SYSTEM_ERROR

- Frequency: quite often with a large number of failed jobs
- Solution: keep retrying job until all succeed