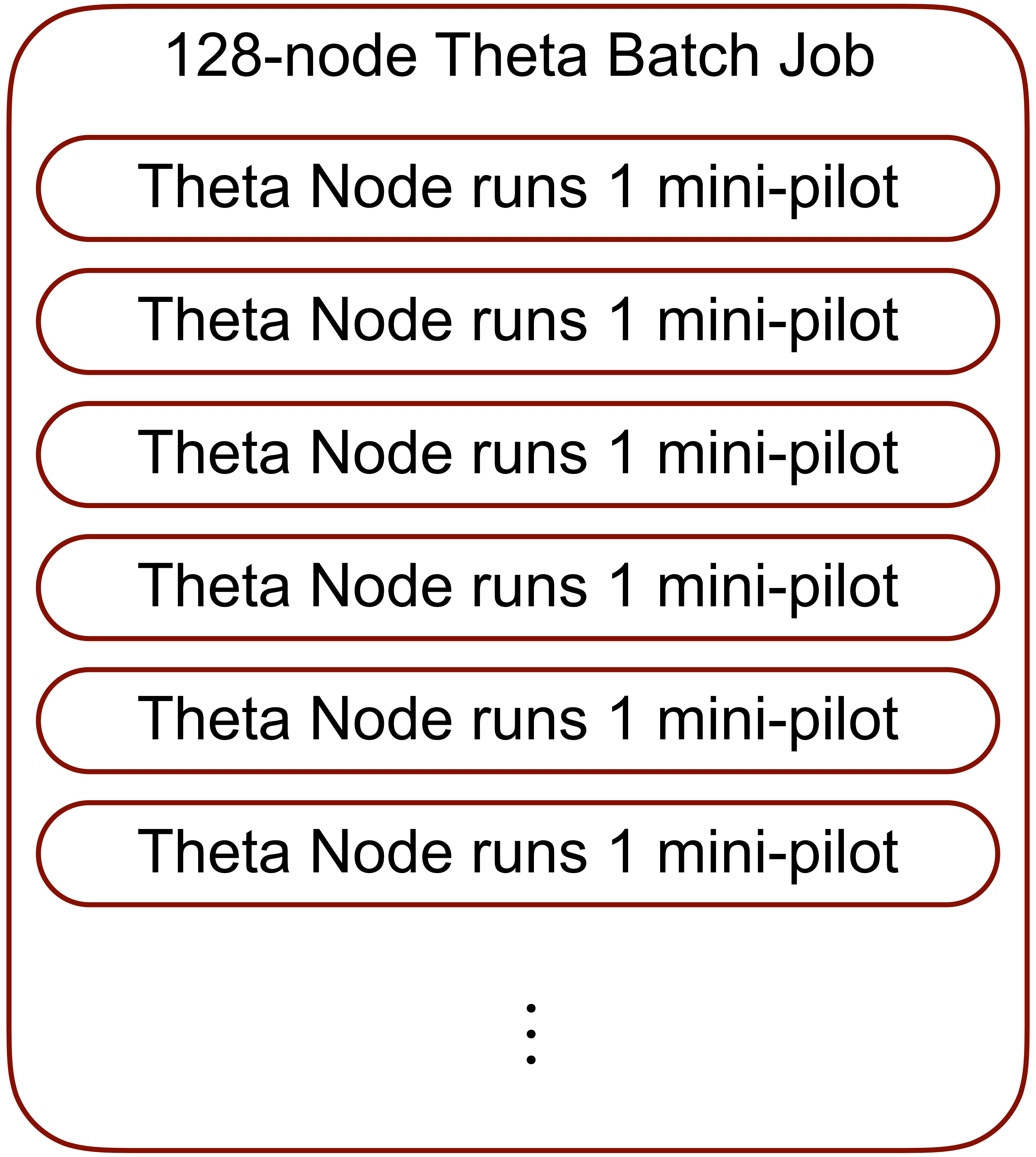
Experiences with Harvester

J. Taylor Childers (ANL), Doug Benjamin (Duke), Tadashi Maeno (BNL)

Harvester setup at ALCF on Theta

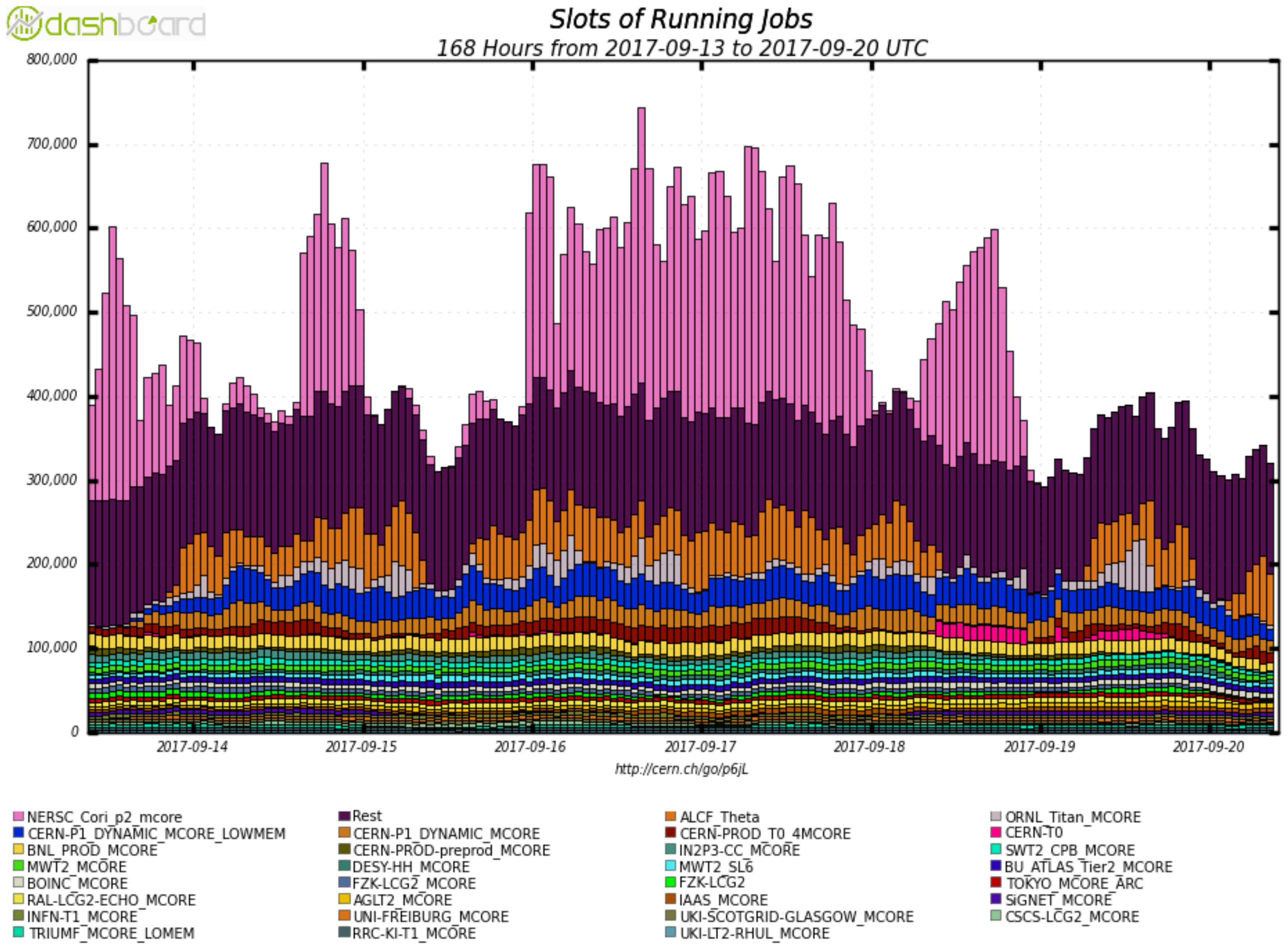
- Began working in August to get Harvester into constant running
- Running in ManyToOne mode:
  - A batch job of, for example, 100-nodes runs one "mini-pilot" per node.
  - The mini-pilot moves data locally, communicates with Harvester via the filesystem, and executes the job transform.
- Harvester handles communication with Panda and remote data staging.
- Harvester relies on users implementing plugins that interact with their local system.
  - Scheduler plugins for submission, monitoring, and killing jobs
  - Stage-in (preparator)
  - Stage-out
  - Batch job submission template script
  - and payload
- In our case we have committed about 1300 lines of code, which is 10% of the Harvester repo (13,000 lines of code in total). (simple count with 'wc -l')


128-node Theta Batch Job
Theta Node runs 1 mini-pilot
Theta Node runs 1 mini-pilot
Theta Node runs 1 mini-pilot
Theta Node runs 1 mini-pilot
Theta Node runs 1 mini-pilot
Theta Node runs 1 mini-pilot
⋮

Theta running validation jobs

- Currently running 128-node jobs with a throughput of 6-8 jobs running on average. At 8 jobs we are at 1024 nodes, or 1/4 of Theta.

The SUNSHINE HARVESTER,
THE BEST PROFIT MAKER FOR THE GRAIN GROWER.
SUNSHINE
Hugh V. McKay, INVENTOR AND MANUFACTURER,
SUNSHINE HARVESTER WORKS, SUNSHINE.
D. W. PATERSON CO., PRINT, MELB.

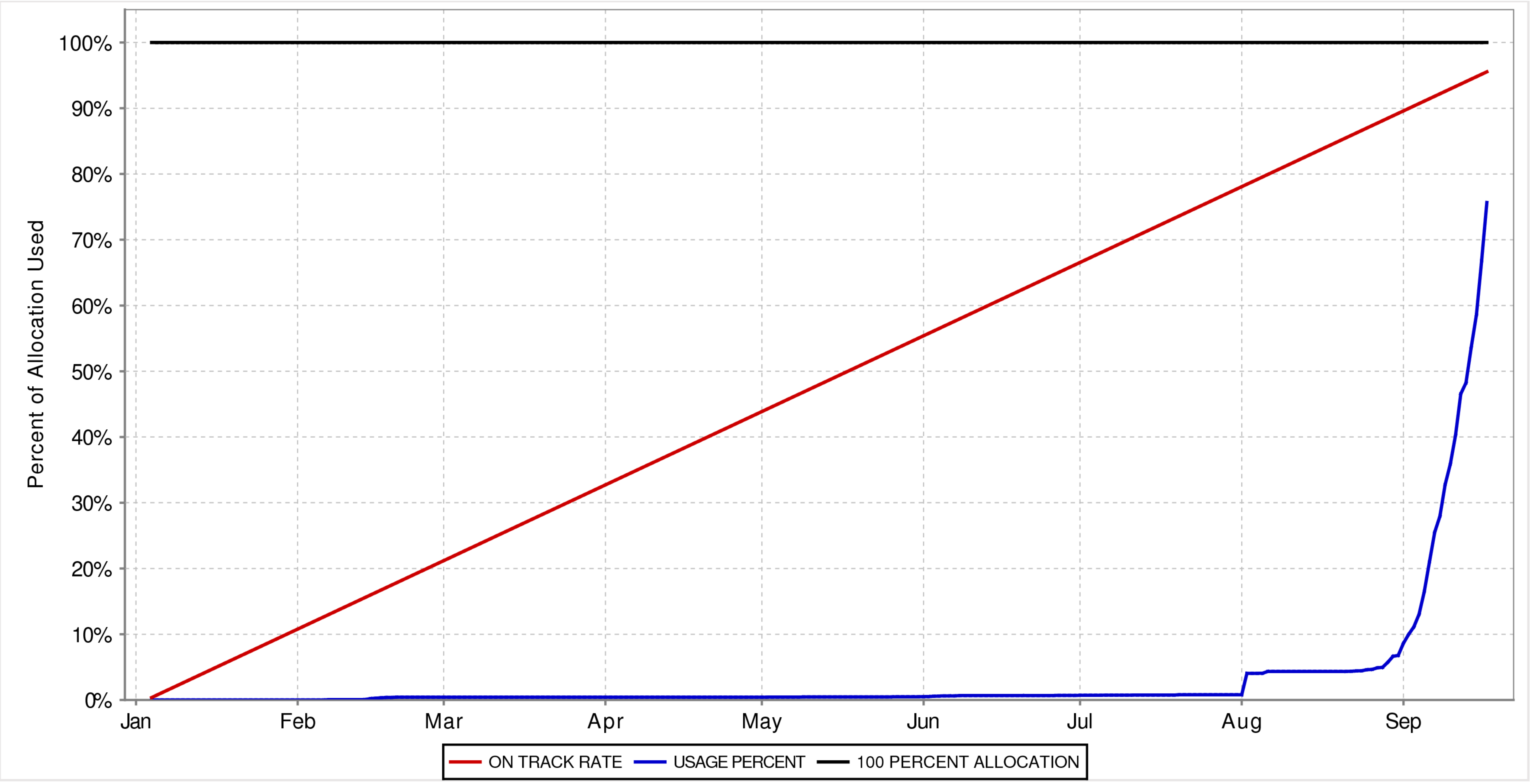


Maximum: 744,072 , Minimum: 293,397 , Average: 470,224 , Current: 320,587

AtlasADSP Machine: THETA

Allocation Core Hours: 20,000,000 | Data Dates: 2017-01-04 through 2017-09-18

Usage Core Hours: 15,716,975 (78.6%) | Allocation Dates: 2017-01-04 through 2017-09-29



Challenges we faced

- Understanding the Harvester design concept such that we could set a reasonable configuration for Theta.
  - Harvester is complex and must be tuned to a system.
- Resource Utilization:
  - Early on, Harvester was typically listed at 100-125% in ‘top’ on the Theta login node.
  - Tadashi has improved this and now sitting in the 25-35% range which I think is acceptable
  - This was likely due to the DB entry sizes and the job log information which Tadashi discussed in his talk earlier today.
- Data Motion:
  - Different simulation jobs sometimes use the same input files and Harvester was not treating duplicate stage-in procedures such that multiple ‘rucio downloads’ would be running against the same file and collide with one another. Tadashi added some checks to avoid this.
  - Using Rucio with Globus URL Copy for this ManyToOne workflow means many rucio transfers happening directly on the login node, adding to the resource utilization. Moved to Globus Online last weekend which removes this and increases the bandwidth for transfers making it more efficient.
- Mini-pilot:
  - For monitoring, it is important to ensure the mini-pilot is providing all the appropriate information to Harvester which reports back to PanDA. We are still filling this in.

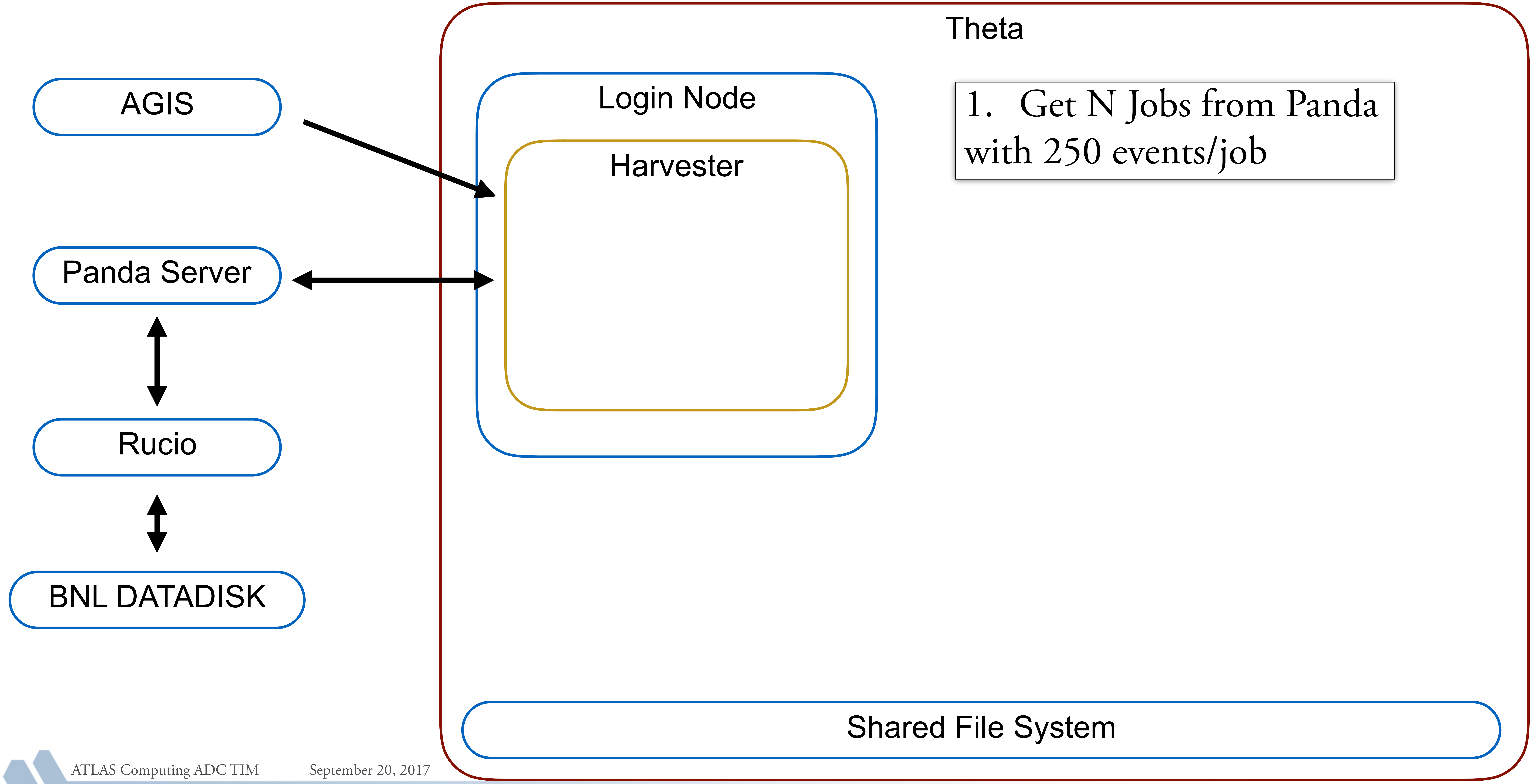
Challenges to Address before distribution ahead of production

- Pooling Globus Online Stage-in/out to avoid GO transfer limits
- Tracking down a problem where Harvester reports job is completed, but output files do not exits.
- Define a way to updates to AGIS and perhaps move Harvester config to AGIS directly
- Agree on a way to handle site maintenance periods, both short (< 1 day) and long (> 1 day)
  - See Doug's Slides
  - Andrej suggested documenting these outages in AGIS somehow.
- Mini-pilot needs to be cleaned up, documented and committed back to GitHub.

Software Distribution using Containers

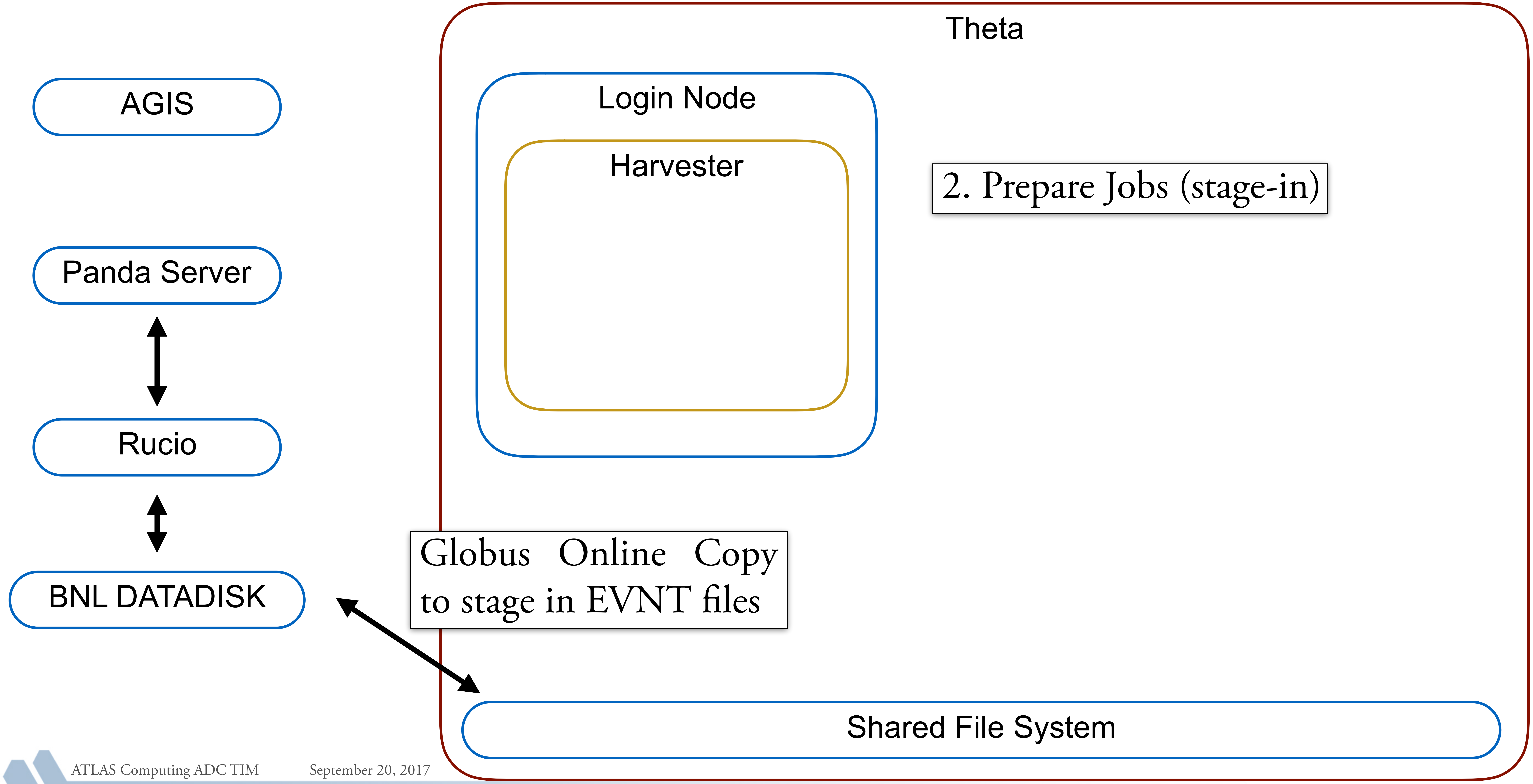
- ALCF has singularity installed on a small test rack of Theta which we will be vetting for them.
- If things go well, we will put pressure on them to deploy on Theta.
- This status is similar to Titan, which means all HPCs in the US will use containers for software distribution

Quick overview of the Harvester workflow on Theta



1. Get N Jobs from Panda with 250 events/job

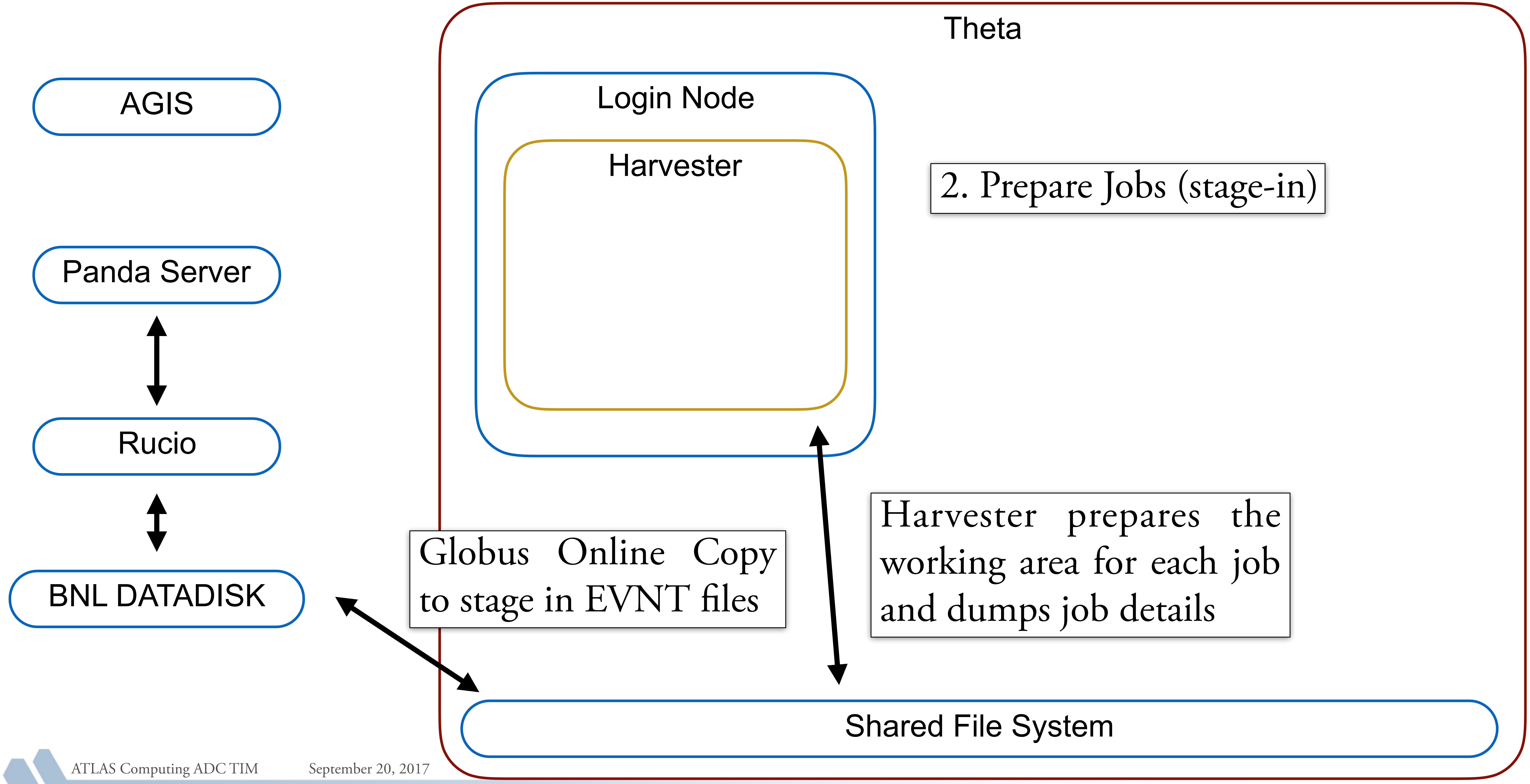
Harvester setup at ALCF on Theta



2. Prepare Jobs (stage-in)

Globus Online Copy to stage in EVNT files

Harvester setup at ALCF on Theta

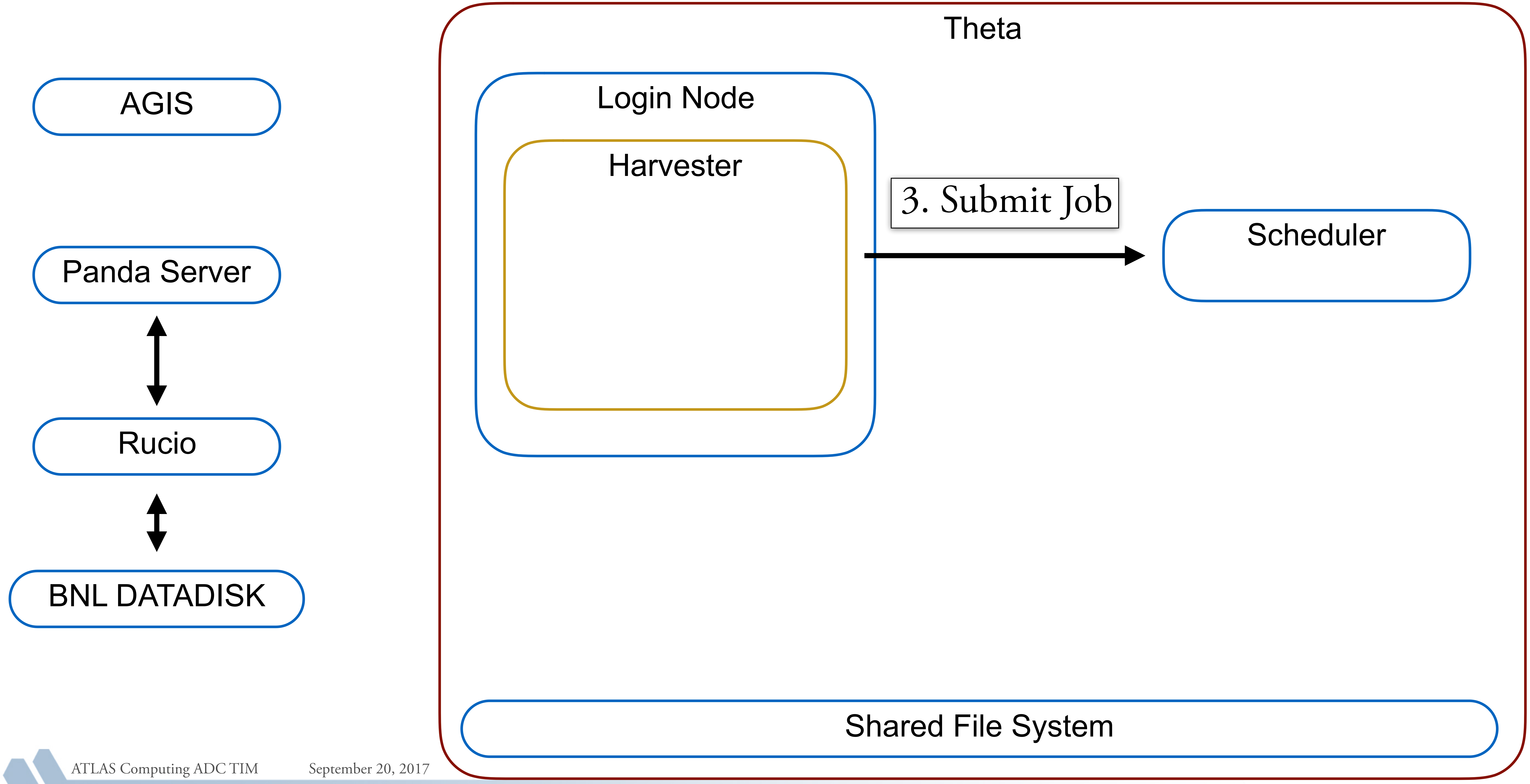


2. Prepare Jobs (stage-in)

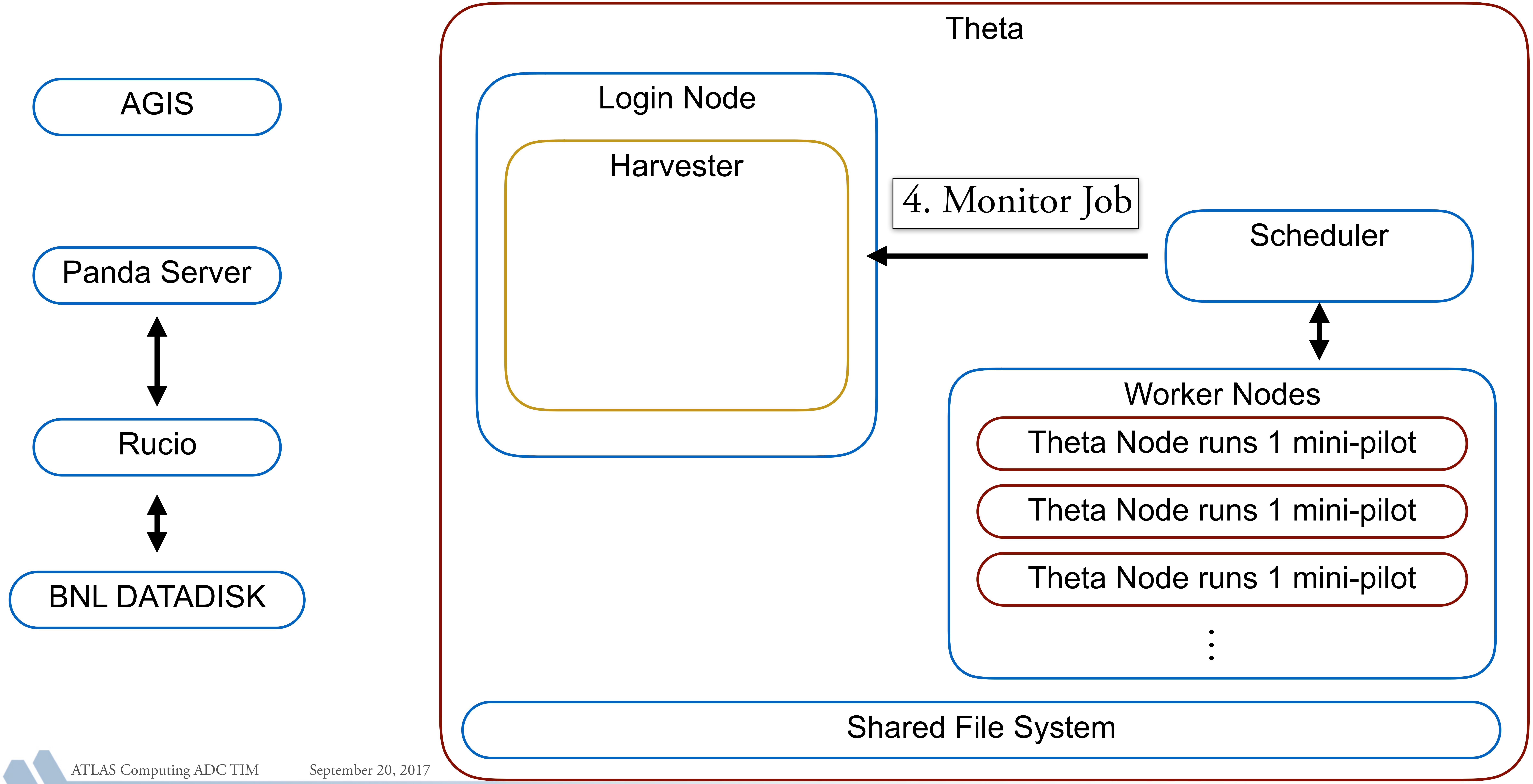
Globus Online Copy to stage in EVNT files

Harvester prepares the working area for each job and dumps job details

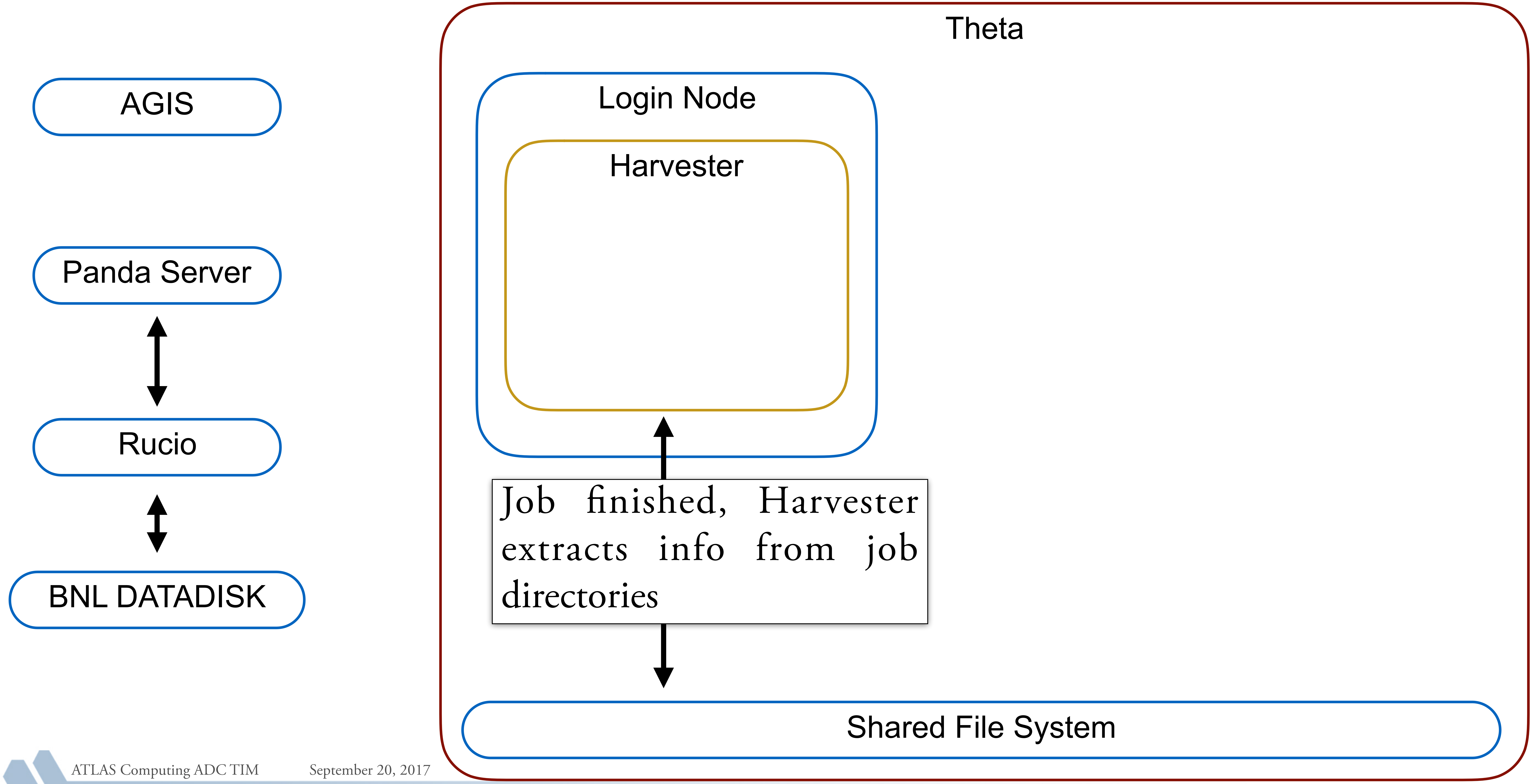
Harvester setup at ALCF on Theta



Harvester setup at ALCF on Theta

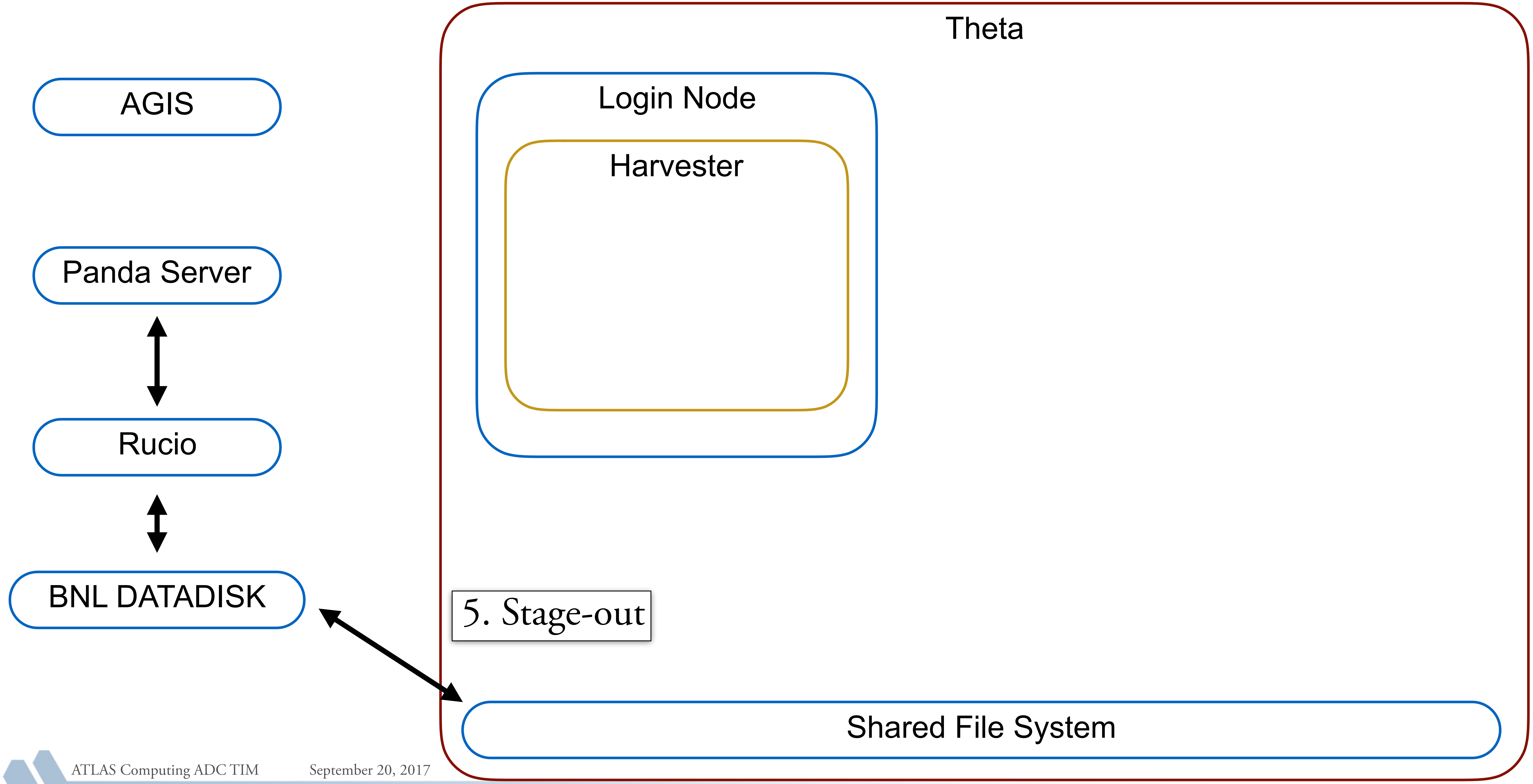


Harvester setup at ALCF on Theta

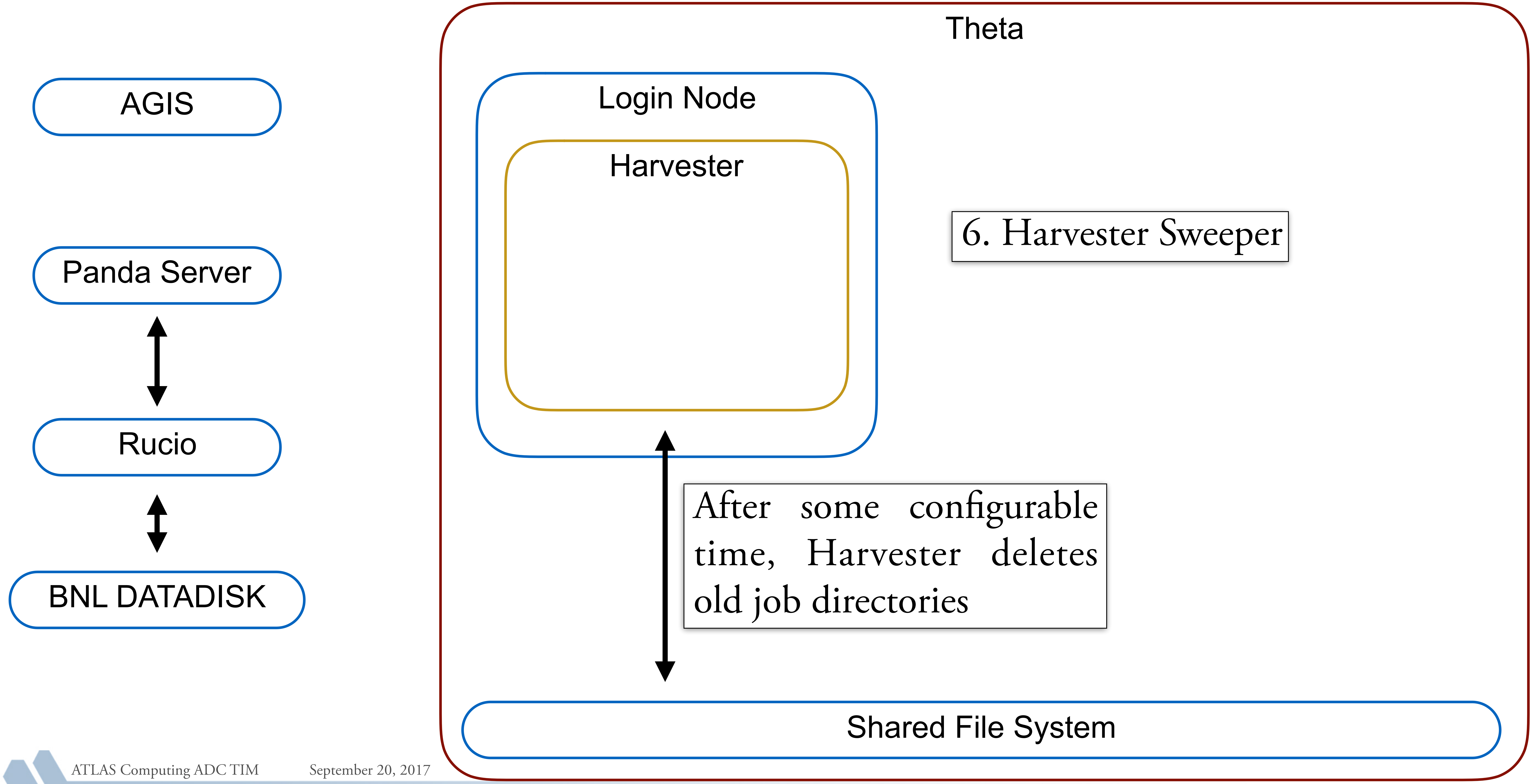


Job finished, Harvester extracts info from job directories

Harvester setup at ALCF on Theta



Harvester setup at ALCF on Theta



6. Harvester Sweeper

After some configurable time, Harvester deletes old job directories