ATLAS experience using LUMI HPC in Finland

Ievgen Sliusar (UiO)

Fall 2024 CMS Offline Software & Computing Week 31 Oct 2024

LUMI Supercomputer

- LUMI (Large Unified Modern Infrastructure) – also Finnish for “snow”
  - A set of interconnected computing and storage services
- EuroHPC pre-exascale supercomputer
  - Top-500 #5 (Jun 2024), #1 in Europe – Rmax 379.70 PFlop/s
  - HPE Cray EX235a – the same as Top-500 #1 but 4 times smaller
- Hosted at CSC’s data center in Kajaani, Finland
  - CSC is also home for Finnish national HPCs
  - Pilot testing conducted on Puhti HPC
- A GPU-centric machine
  - Fit for CERN workloads?
  - Memory – less than 2GB per CPU core on most nodes
  - No local disks on most nodes

LUMI Pilot Project

- Goal – develop a technical solution to allow the LUMI consortium HEP groups to run LHC computing applications on LUMI
- Resources allocated from Finnish national share
    - Can use other co-hosted services
    - Coordinated by HIP

ATLAS Qualification project – started in Jan 2024

- Make ATLAS Production run on LUMI and other EuroHPC machines
- Make a prototype system for local execution of ATLAS payloads and propose a generic solution
- Study running reconstruction, reprocessing and derivations, not only MC sim

Environment Challenges and Limitations

- Unprivileged remote operation
    - Remote SSH login with access to SLURM and local storage
    - Cannot modify HPC machine configuration: tune kernel, site-wide CVMFS, storage, etc.

- Cannot host any services inside HPC
    - OpenStack cloud provided by CSC close but outside of the HPC

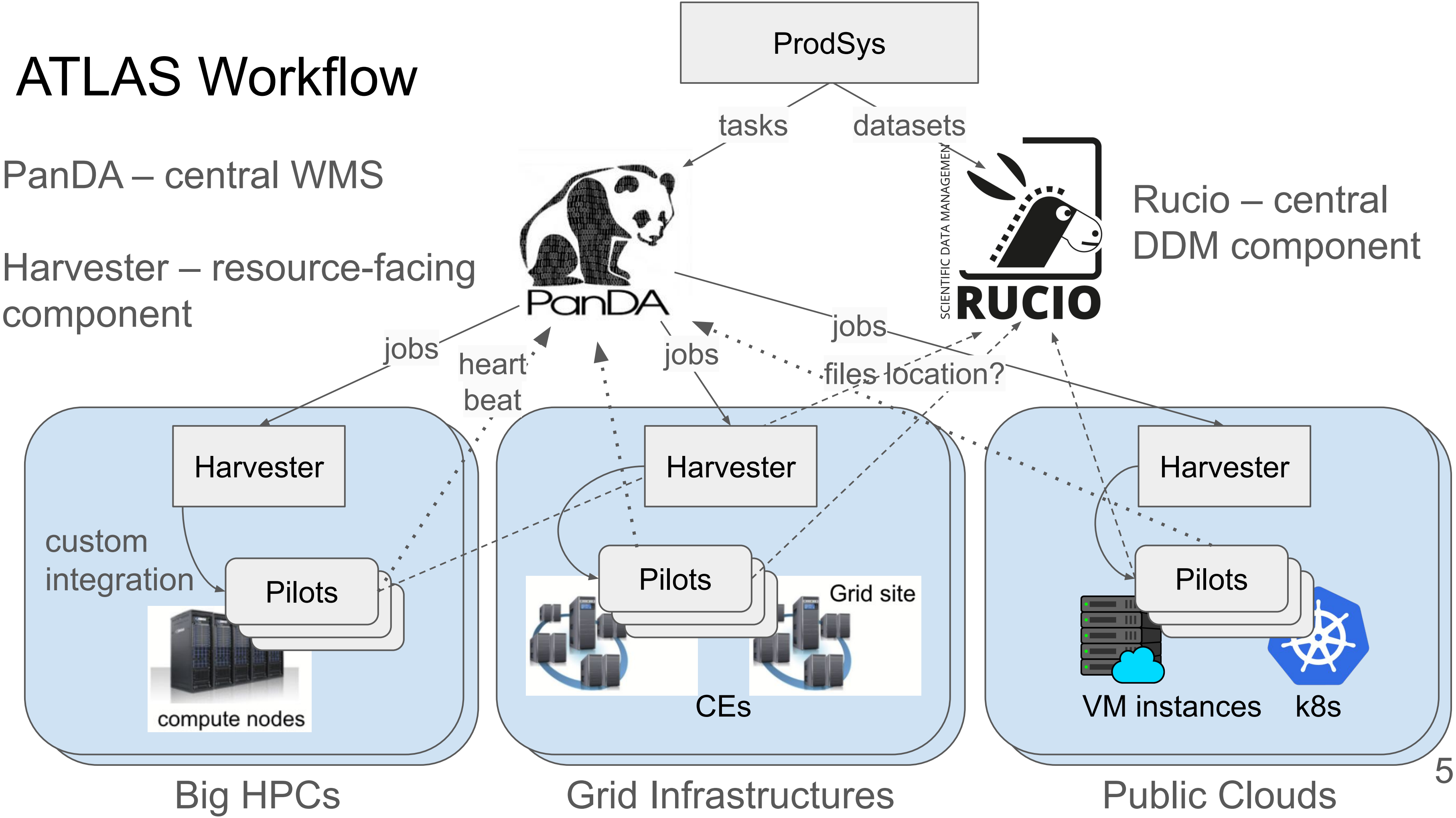
- Limited storage quota
    - 50GB – for software (persistent)
    - 1TB – scratch space (auto-cleaned)
    - Object storage – extendable, S3 protocol, accessible from outside

ATLAS Workflow

PanDA – central WMS

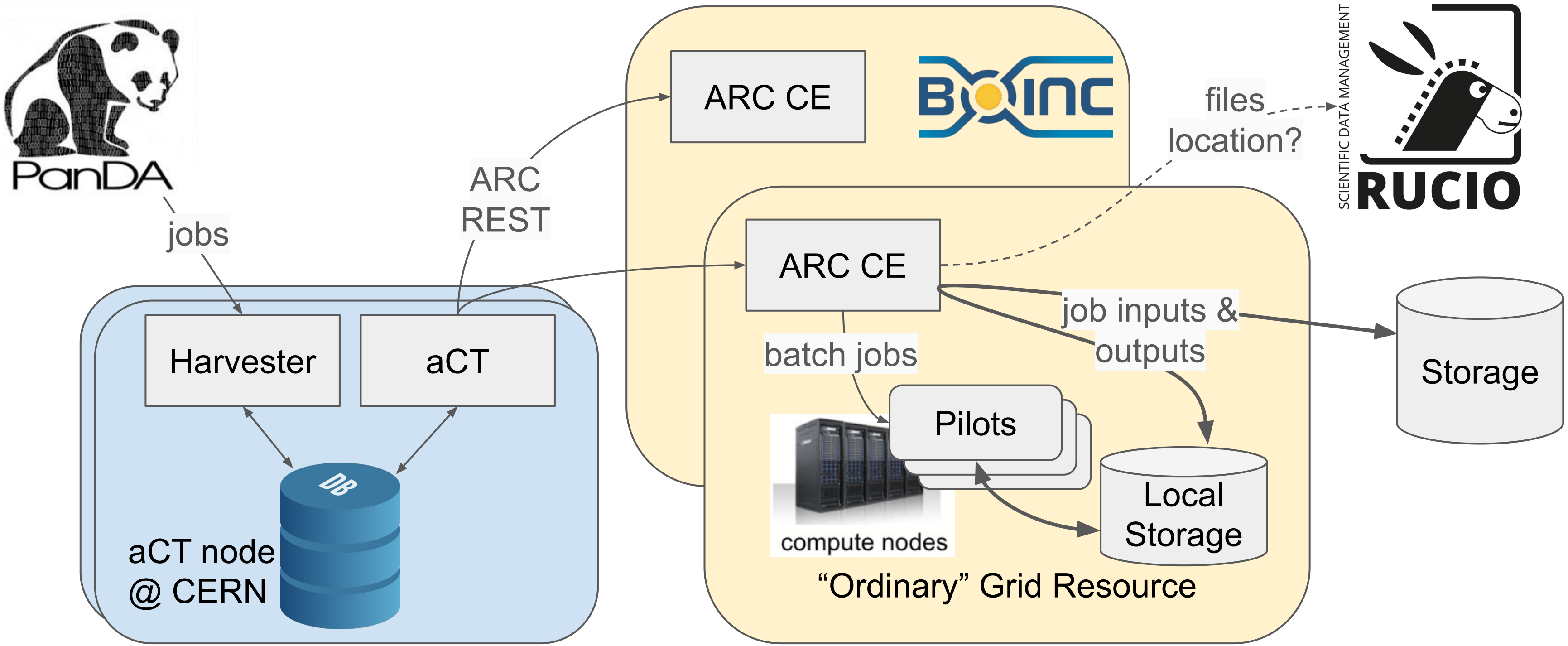
Harvester – resource-facing component

Rucio – central DDM component



aCT – ARC Control Tower

- ARC-based Grid Infrastructure integration for Job factories (e.g. PanDA for ATLAS)



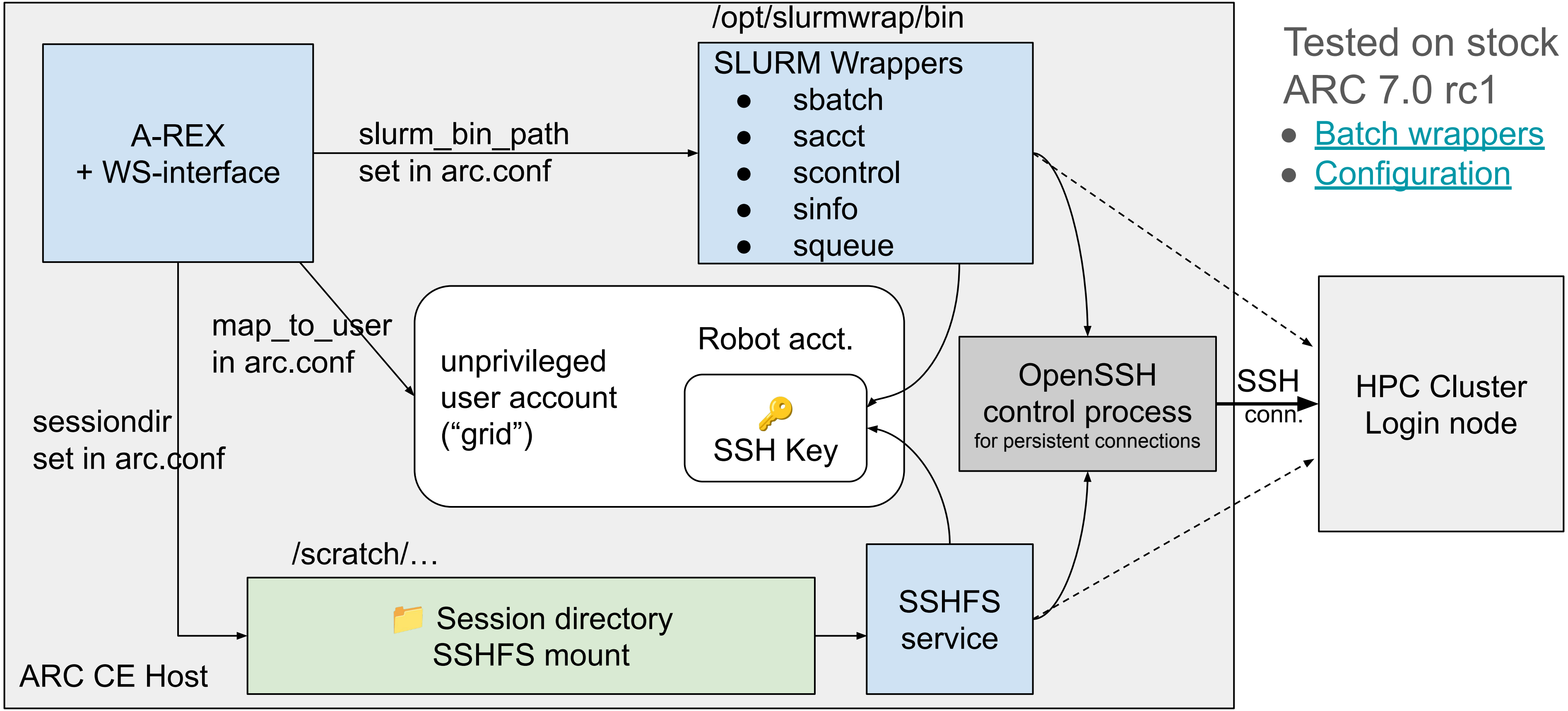
aCT – Expectations from ARC CE Local Environment

- CVMFS needs to be available in job context
  - Pilot sets up ATLAS environment from it

- Container runtime should be provided in job context
  - Pilot usually starts payload in a container
  - Automatic flavour detection – Apptainer/Singularity or Docker
  - Container images are located in CVMFS

- Data staging processed by ARC CE itself
  - ARC CE has credentials to access Rucio and external storage systems
  - Cache should be organized locally

ARC CE – Hosted outside of the HPC



Tested on stock ARC 7.0 rc1

- Batch wrappers
- Configuration

CVMFS

- Not provided centrally at HPC resource
  - Should be mounted in job context (unprivileged)

- Use cvmfsexec tool – different methods
  - Mount FUSE directly or via Apptainer/Singularity (LUMI only provides SingularityCE)
  - Should be patched to generate FUSE3 distribution (Puhti only provides FUSE3)

- Jobs cannot share cache
  - Local job scratch dir is wiped when job finishes
  - Use “Alien cache” option on shared scratch

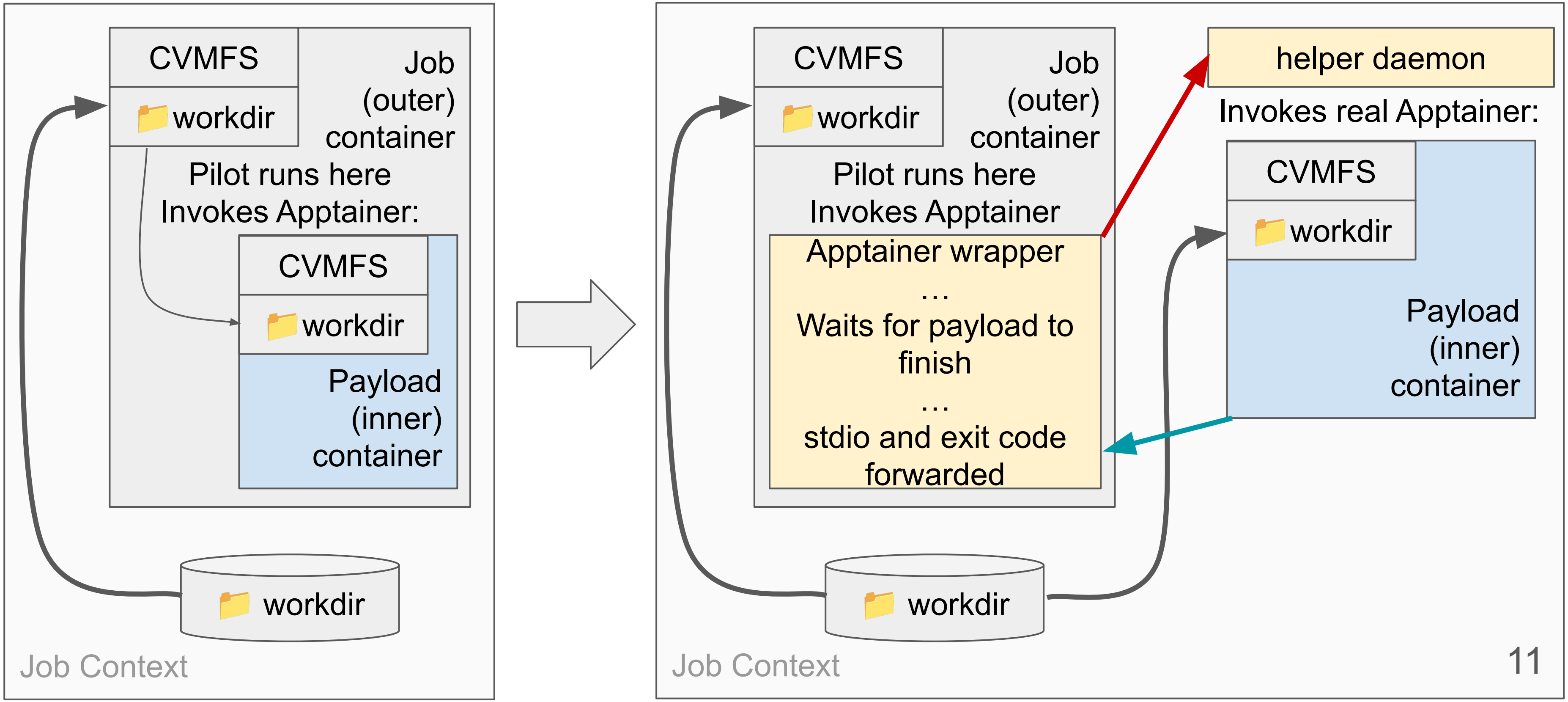
Containers

- System-provided privileged (suid) Apptainer/Singularity installation cannot run containers directly from FUSE mount
  - Cannot use container images / filesystems in user-mounted CVMFS
  - Wrapper script to rebuild SIF image from CVMFS and/or public registries and store locally
    - Can be cached for future use by other jobs

- Containers cannot be nested
  - Rely on unprivileged user namespaces in Linux – feature disabled by HPC vendor
  - If pilot is run in a container, then it cannot execute payload
    - Run pilot outside with FUSE-mounted CVMFS – possible on Puhti, but not LUMI

  - Need a solution to un-nest containers to run on LUMI
    - LUMI prohibits mounting FUSE directly, only inside a container
    - Need to run pilot AND payload in containers to have CVMFS

Un-nesting containers – universal approach WIP



CVMFS

workdir

Job (outer) container

Pilot runs here

Invokes Apptainer:

CVMFS

workdir

Payload (inner) container

workdir

Job Context

CVMFS

workdir

Job (outer) container

Pilot runs here

Invokes Apptainer

Apptainer wrapper

…

Waits for payload to finish

…

stdio and exit code forwarded

helper daemon

Invokes real Apptainer:

CVMFS

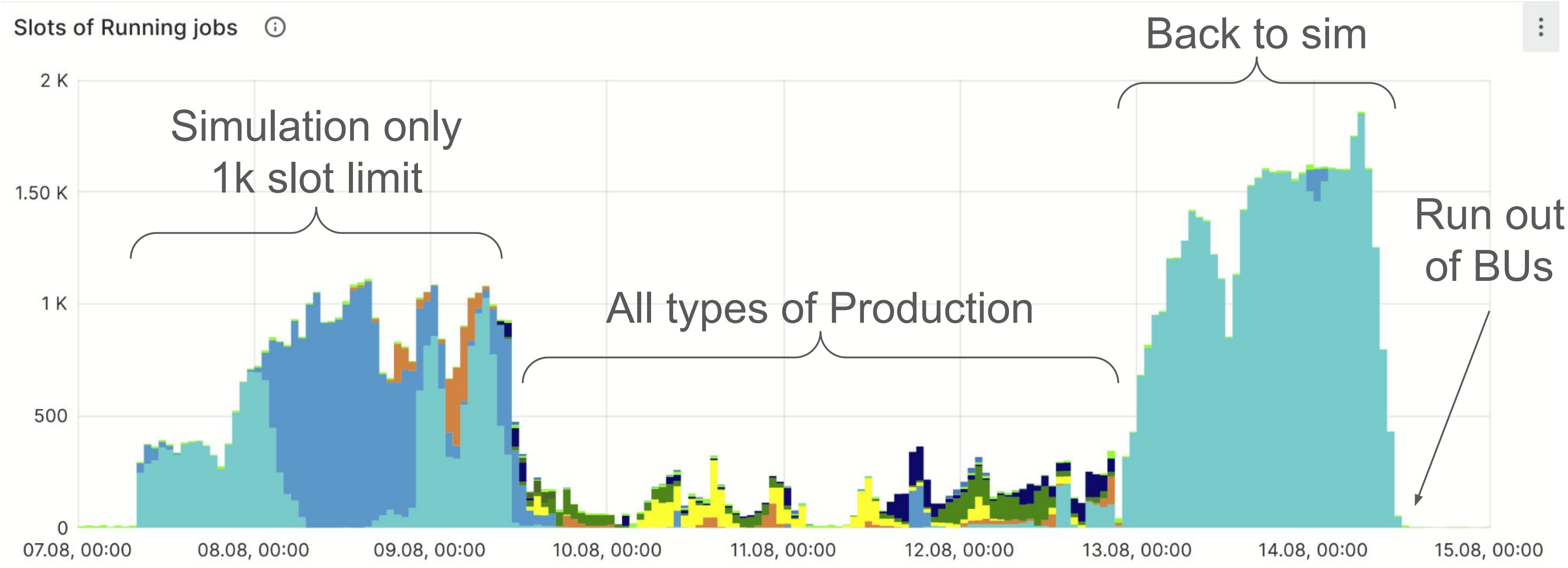
workdir

Payload (inner) container

workdir

Job Context

Testing on Puhti HPC: 7-15 Aug 2024



Simulation only
1k slot limit

All types of Production

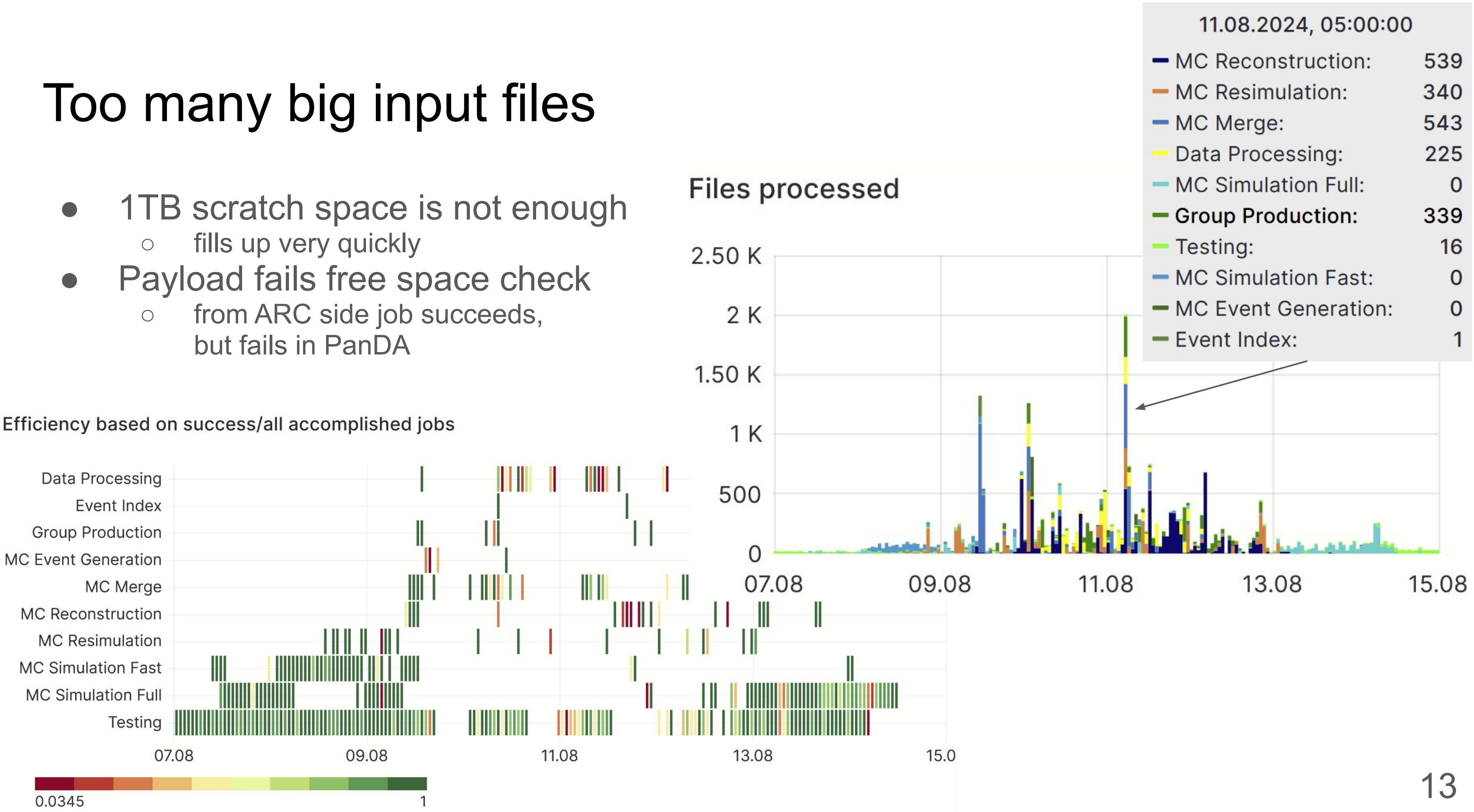
Back to sim

Run out
of BUs

[ATLAS Grafana Dashboard]

Too many big input files

- 1TB scratch space is not enough
  - fills up very quickly
- Payload fails free space check
  - from ARC side job succeeds, but fails in PanDA



Data Staging

- Scratch space provided by HPC is limited and cannot be extended
  - 1TB for Puhti, 50TB for LUMI per project
  - Used for
    - container cache
    - CVMFS “alien cache”
    - job scratch directory

- Object Store alternative – future plans
  - Provided by CSC, also can be any other cloud service (GCP, AWS, …)
  - Not a POSIX filesystem
  - Accessible from outside of HPC, S3-compatible access protocol
  - Can be extended to 150TB on demand
  - Can be mounted via FUSE driver: s3fs-fuse

Summary – Generalizing the approach

- ARC CE can be put in front of HPC to process ATLAS production workloads
  - Single CE can submit to single HPC remotely, but can utilize various queues
  - Standard ARC distribution with tailored configuration

- “Ingredients” to be provided inside HPC

CVMFS	PoC usable with FUSE mount or via container runtime
Container runtime	PoC usable with image cache, un-nesting tool in development
Data staging	ad-hoc works, needs more testing

- Next steps
  - Complete development of un-nesting tool
  - Replicate setup on LUMI and document all the recipes
  - Make more tests on real ATLAS workloads (not only MC sim)
  - Try running CMS workloads via the same ARC CE

Conclusions

- Generic approach to run ATLAS workloads on HPC seems feasible
  - The only requirement is functioning container system Apptainer/Singularity w/FUSE support
  - … and ARC CE installation with reliable SSH access to the HPC

- Performance varies depending on workload type
  - MC Simulation runs smoothly and takes all available slots
  - More data-demanding tasks require improved data staging

- Data staging needs to be addressed
  - Try using S3-compatible Object Store provided by HPC to be more scalable
  - s3fs-fuse can be used for mounting from ARC CE and from job context (unprivileged)