LUMI onboarding update

Tomas Lindén[1], Gianfranco Sciacca[2], Ievgen Sliusar[3]

HIP[1], UniBE[2], UiO[3]

27.02.2025

NLCG meeting, 27th of February 2025, hybrid

Contents

CSCs current national HPC-resources

- **Puhti**
  - dual Intel Xeon Gold 6230 CPUs with 20 cores each
  - *Puhti has 3760 cores with local disk + the GPU nodes*
    - M-IO 48 nodes 192 GB RAM and 1490 GiB local disk
    - L-IO 40 nodes 384 GB RAM and 3600 GiB local disk
    - BM-IO 6 nodes 1.5 TB GB RAM and 5960 GiB local disk
  - No native CVMFS installation
  - Has fuse3, cvmfsexec mode 1 and singcvmfs works
  - Suggested by Sebastian to be easier to use than LUMI
  - Used by levgen for developing the tools to run on LUMI
- **Mahti**
  - dual AMD Rome 7H12 SMT CPUs with 64 cores each, 256 GB RAM
  - interactive partition: 4 nodes with a 3.5 TB NVMe drive
  - *small partition: 56 nodes (7168 cores) each with a 3.5 TB NVMe drive*
  - No native CVMFS installation
  - Has fuse3, cvmfsexec mode 1 and singcvmfs works
  - *Recommended for I/O jobs since the 14th of February 2025*

CSCs next national HPC-resource

Roihu

- Roihu is expected to be taken into use in January or February 2026
- Roihu will replace Puhti and Mahti
- BullSequana XH3000 system
- 486 nodes, dual AMD Turin 192 core CPUs (186 624 cores)
- GPU: 132 nodes with a total of 528 Nvidia GH200 GPUs
- GPU nodes for visualization
- Special bigmem nodes
- **Nodes will have ”small and slow” SSDs for OS- and CVMFS use**

LUMI activites

CSC on native CVMFS installation on LUMI

- LUMI will be in use well into 2027
- CSC dropped the CSCS Alps RADOS blockdevice CVMFS solution
- CSC is now planning an Alien cache CVMFS installation on LUMI-F
- The installation schedule is unknown

Current LUMI HEP development projects:

- 11.10.2025, 2006540 LUMI pilot, CSC project on cPouta for ARC-CEs, etc
- 01.03.2026, 462000245: Exploring the Use of AMD GPUs for High-Performance
- 12.03.2026, 462000247, LUMI-as-a-Service - actual LUMI project Computing in the CMS Reconstruction

LUMI activites

LUMI Pilot Project, CSC cPouta OpenStack project

- This project hosts the ARC CEs snowarc and arc2lumi and a squid
- Ievgen has patched cvmfsexec for fuse3, `https://github.com/cvmfs/cvmfsexec/pull/94`
- On Puhti and Mahti cvmfsexec method 1 and singcvmfs works
- snowarc was used for Puhti submission with ssh and sshfs using the fuse3 patch and cvmfsexec method 1
- ATLAS has run more than 1,5 k SIM-jobs in this way
- Puhti usage planned to be moved to Mahti
  - On Mahti there is more CPU cores in terms of NVMe equipped nodes

LUMI activites

LUMI as a service, LUMI project

- LUMI has no native CVMFS installation, nor any local disks
- The Lustre project scratch area is 50 TB
- On LUMI only singcvmfs works with stock cvmfsexec
  - Running containers from inside CVMFS does not work with singcvmfs
- The cvmfsexec fuse3 patch enables cvmfsexec method 1 also on LUMI
- Ievgen has made a tool fapptainer for unnesting nested containers for singcvmfs
- arc2lumi ssh submission to LUMI with sshfs and singcvmfs with fapptainer is used currently
  - ATLAS HammerCloud and MC-jobs run on LUMI in this way
  - CMS ETF/SAM testjobs run on LUMI in this way
  - CMS MC-validation jobs will be run
- Issues with the setup are fixed as usage increases

LUMI activites

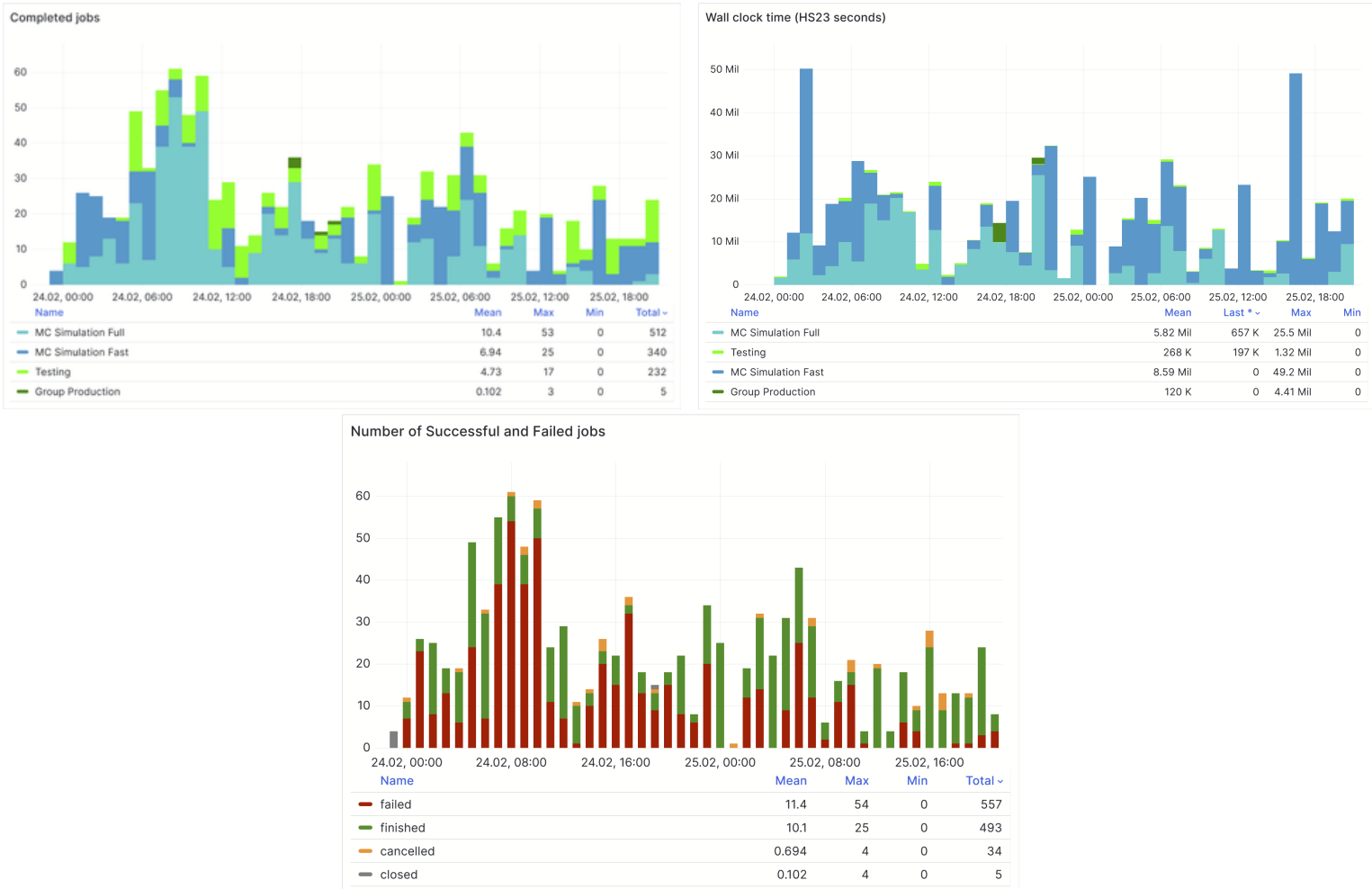

Figure: ATLAS jobs run on LUMI.

LUMI activites

Figure: CMS ETF/SAM tests run 25.02.2025 on LUMI.

LUMI allocations

- LUMI projects are self contained and cannot include other CSC resources
- The ARC CE:s are hosted on CSC:s cPouta in a non LUMI project
- LUMI resources are allocated proportional to the investments (EU 50 %, FI 25 %, rest of the consortium 25 %)
- LUMI allocations `https://lumi-supercomputer.eu/get-started/`
- LUMI allocations in Finland `https://lumi-supercomputer.eu/get-started-2021/users-in-finland/`

"Regular Access mode, for resource applications with a case to enable progress of science in the domains covered. These applications are expected to be able to justify the need for large allocations in terms of compute time, data storage and support resources because they are significantly contributing to the progress of science in their domain.
middle size projects, duration fixed 12 months
max 10 million core-h, 200,000 GPU-h, 570 000 TiB-h"

Moving to LUMI production usage

Possible scaling limitations:

- Can more I/O-intensive jobs than MC-job be run?
- Memory usage increases due to injob CVMFS
- - Can be mitigated by requesting more RAM
- LUMI Lustre has limits of 2 M files (scratch) and 1 M files (flash)
- - These might be worked around with blockdevices formatted as ext3
- ARC CE ssh and sshfs scaling limit is currently unknown
- - Place the ARC CE close to the HPC to minimize latency
- - Mitigation is to run as few and as large jobs as possible
- So far the ARC CE VM resources 3 cores, 4 GB RAM, 80 GB disk have been enough

Moving to LUMI production usage

Moving to LUMI production usage:

- Where to host needed services, near LUMI or further away?
- ARC ssh(fs) submission allows only one HPC and one VO per CE
- Could CSC host some services?
  - It remains to see if CSC installs CVMFS on LUMI?
  - A CSC installed ARC CE could server multiple VOs or multiple allocations (from different countries)
  - ARC cache(s)
  - Squid cache(s)
- LUMI Allocations need to be applied for (always max 12 months).
  - Could cross experiment allocations be negotiated from CSC?
- A LUMI robot account is needed for each allocation.
- Documentation and maintenance of the setup

LUMI AI Factory

LUMI successor:

From `https://indico.cern.ch/event/1466360/contributions/6325288/attachments/3007776/5302233/CSC-update-300125.pdf`

- combined HPC + AI + QC platform
  - AI-optimised supercomputer LUMI-AI
  - AI Factory service center
  - Experimental QC-AI platform LUMI-IQ
- CSC (Finland) with Czechia, Denmark, Estonia, Norway and Poland
- Total budget over 600 MEUR

Will this new environment be supportive of HEP workloads?

Will this new environment have node local storage?

Summary

- Ievgen has made great work with the new tools
- ATLAS and CMS MC-jobs can run on LUMI using the new tools
- Other workflows will be tested
- Other HPCs can be used in a similar way if needed
- We will submit an abstract to ACAT about this

Links to the software and documentation

- Ievgens ATLAS qualification talk has more explanations and details: https://docs.google.com/presentation/d/1XqqZvw9-9Lvn0mtnLBzBNqMDtI_J3hoeuT566BZ4tlw/edit?usp=sharing
- fuse3 patch: https://github.com/cvmfs/cvmfsexec/pull/94
- Batch wrappers: https://docs.google.com/document/d/10i03EldpBgYkUzc2ruGgS43JhGJagvgaKGjdEZ_YePA/edit?usp=sharing
- ARC and sshfs: https://docs.google.com/document/d/1J-fg8AEEjXUEMyw_MIHhJ9RAEzX8JZgzWel1Pn9jZQ4/edit?tab=t.0
- fapptainer: https://source.coderefinery.org/slu/fapptainer