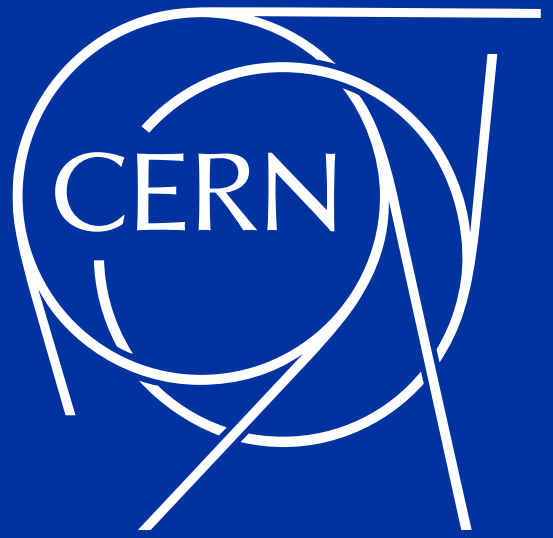

CERN

HPC / Slurm Clusters at CERN

Nils Høimyr, Ben Jones on behalf of the HPC team in IT/CD

High Performance Computing (HPC)

CERN local context

Motivation: *Address needs of parallel MPI applications and use cases that do not fit the standard batch High Throughput Computing (HTC) model*

SLURM MPI clusters as a complement to the HTCondor batch service

- **Theory and ATS sector main users**
  - Ref: workshop session 1 and session 2 in 2020 with presentations of applications and use cases
  - Restricted HPC service (KB0004975) and user community
- **Batch HTC under HTCondor (~400k cores) is the main compute service**
  - Worker nodes with up to 192 cores
  - A few “bigmem” nodes (1TB) for special use cases
  - Some GPU capacity (T4, V100, A100, H100)
- **For ML use cases: k8s & Kubeflow**

User Community

ATS	• Plasma Simulations for Linac 4 • Beam Simulations for LHC, CLIC, FCC… • Xtrack, PyOrbit etc • Picmc • Gdfdl (field calculations for RF cavities) • Field calculations (CST…) • Engineering (Ansys and Comsol)
TH	• Lattice QCD simulations
HSE	• Safety/Fire simulations (FDS, OpenFOAM)
EN	• CFD (Ansys-Fluent, OpenFOAM)

HPC MPI Clusters - Hardware

We have 4 Infiniband clusters, each on different Slurm partitions:

- **2x 72 nodes, with 2x Xeon® CPU E5-2630/20 cores (40HT), Infiniband FDR (partitions “inf-short” and “inf-long”**
- **72 nodes with 2x AMD EYPC 7302 32 cores, Infiniband EDR (partition “photon”)**
- **80 nodes with 2x Xeon® Gold 6442Y – 48 cores (96HT) Infiniband HDR (partition “muon”)**

All nodes with shared CephFS file system `/hpcscratch`

HPC – Software and OS

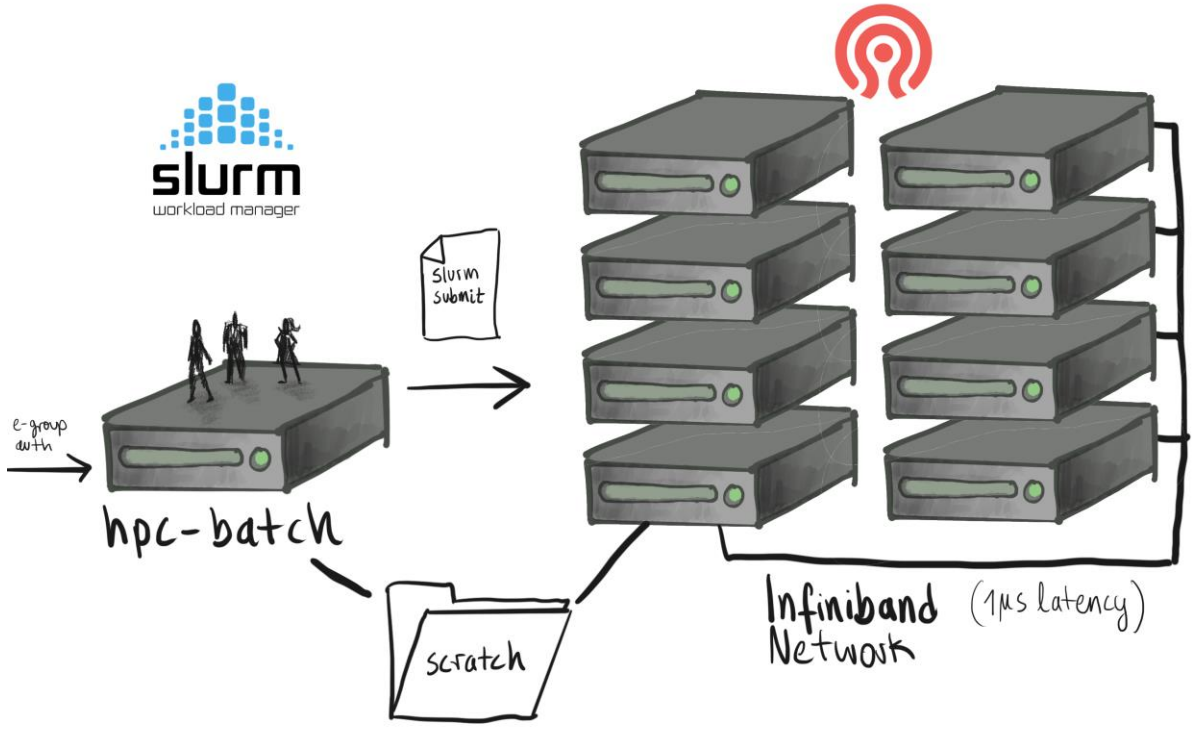
- **Clusters now running EL9 Linux (RHEL 9.5)**
- **Slurm 23.02.07**
- **MPI versions via modules**
  - OpenMPI 316 and 411 (4.1.6 now also available)
  - Mvapich 2.3
- **Same software and OS packages and LxPlus and LxBatch for compatibility**
  - Also CVMFS mounts etc
  - Try to maintain the same compute environment across these platforms to provide consistency to users

HPC Batch cluster – user environment

- **Login / submit node: “hpc-batch.cern.ch”**
  - Users’ home and scratch directories on the /hpcscratch file system (CephFS)
  - AFS and EOS are available, similar to LxPlus
  - Applications on AFS or CVMFS, (also local or EOS)
  - EOS for data copy and project storage
- **SLURM for HPC scheduling**
  - Jobs typically run unauthenticated (run times up to several weeks)
  - Submission with Kerberos token supported via Auks, for copy back to EOS

Submit node

- **Users compile their jobs against the MPI distribution they choose using the appropriate module ie:**
  `module load mpi/openmpi/4.1.6`
- **Users launch their jobs, check job status, cancel jobs…**
- **Similar to LxPlus, but reserved for HPC**
- **Separate system necessary due to need for shared scratch**



HPC – Slurm partitions and queues

Partition Name	Max run time	Main users
inf-short	5 days	ATS, HSE, engineering
inf-long	21 days	ATS, HSE, engineering
photon	10 days	ATS, TH
phodev	2 hours	ATS, TH
muon	10 days	ATS, TH
mudev	2 hours	ATS, TH

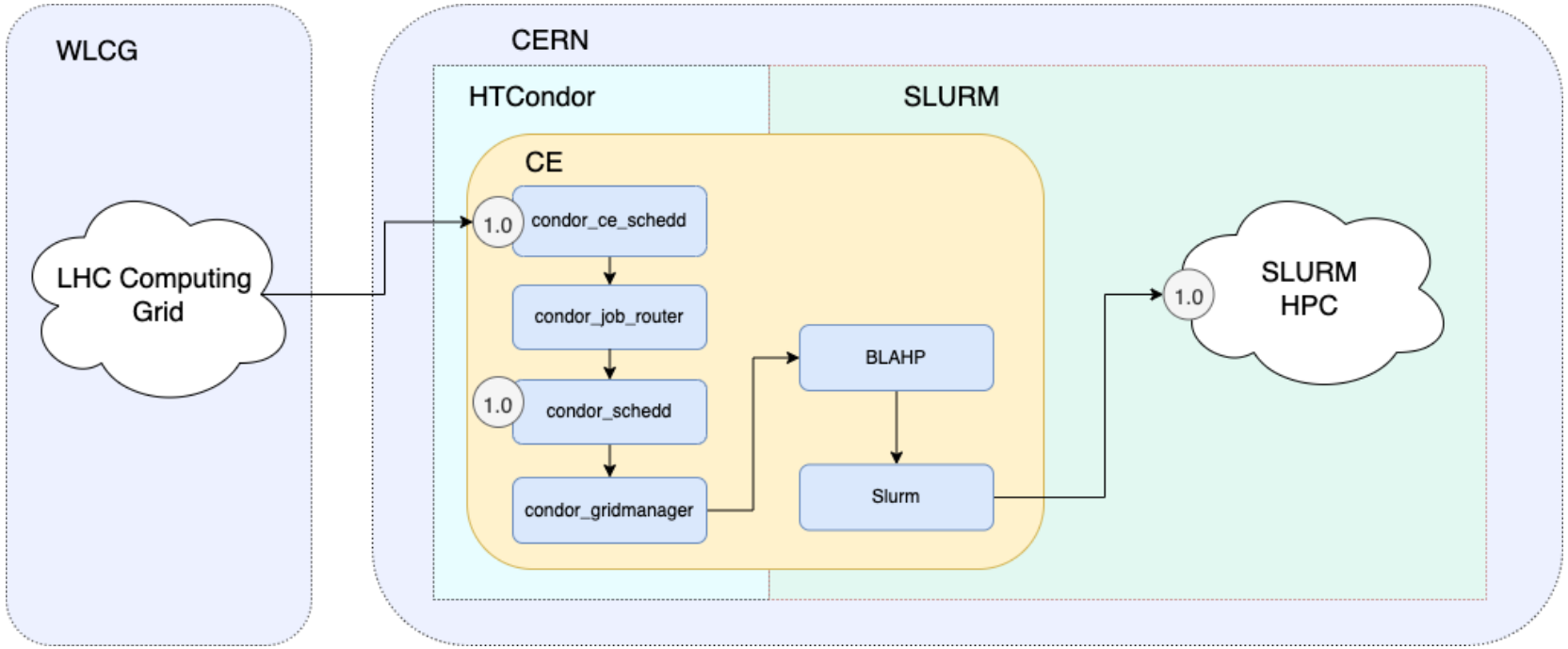
CephFS Scratch File System

- **Home directories for users:** `/hpcscratch/user`
- **Project areas:** `/hpcscratch/project`
- **Slurm run-time bitmap:** `/hpcscratch/statesavelocation`
- **The shared filesystem is located on the Ceph cluster “Jim” managed by the Ceph team in IT-SD**
- **TH/QCD also have a CephFS mount:** `/hpcqcd`

cf CephFS documentation and relevant Storage talk

HPC Backfill

- To Maximize use of HPC resources, nodes not allocated to multi-node MPI user jobs are backfilled with WLCG grid jobs
  - htcondor-ce “cehpc.cern.ch”
  - User HPC job will preempt backfill job (SIGCONT & SIGTERM ; 5 mins ; SIGKILL)



External HPC Pilots

- **Cloud partition (2021)**
  - 100 workernodes on Azure added s a Slurm partition, fully integrated in the cluster (including the shared file systems)
  - Implemented Express Route to Azure/Amsterdam and batch/cloud integration on the previous CERN MS Azure contract
- **EuroHPC VEGA:**
  - Benchmark/dev grant during 2021 used by a few TH and ATS users
  - Procedure with ssh keys and external accounts barrier to entry for regular CERN users.
    - (Should be noted that users in ATS & TH who regularly run on HPC centres are more used to these technical details)