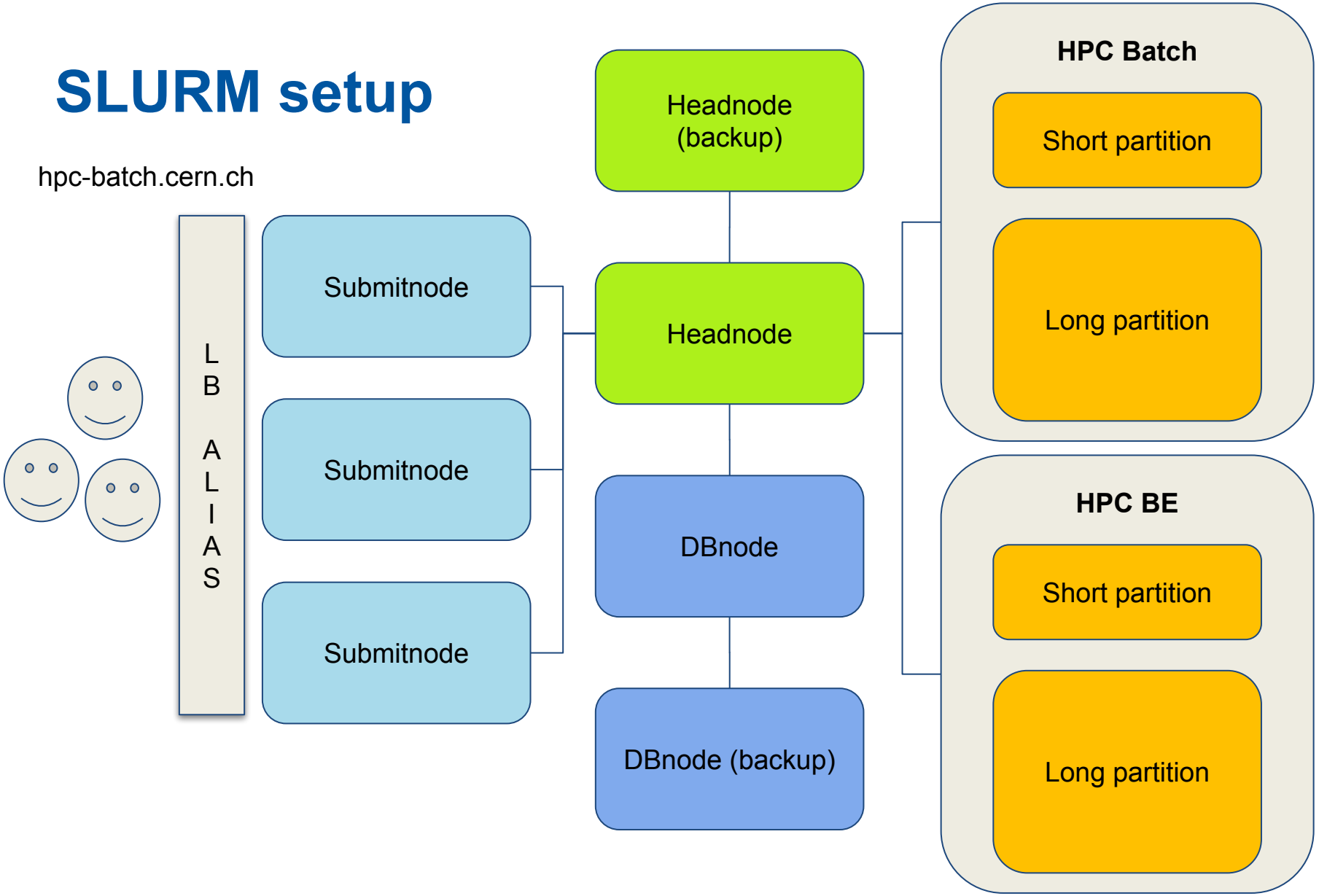
SLURM Clusters

- Batch
  - Roughly 100 nodes
  - Low-latency ethernet cards (~3.5us latency)
  - 2x 8-core CPUs, 32 hyperthreaded cores total
  - 128GB Memory
- Infiniband
  - Two 72-node clusters
  - InfiniBand interconnect (~1us latency)
  - 2x 10-core CPUs, 40 hyperthreaded cores total
  - 128GB Memory
  - **Hyperconverged** CephFS Cluster

SLURM setup

hpc-batch.cern.ch



SLURM Additions

- Hyperconverged CephFS cluster
  - Ceph OSDs run on compute nodes
  - Also MDS servers run on dedicated nodes (no user jobs run on these)
  - Large fast local storage O(100TB)
  - Runs on 10GB ethernet though)
  - Infiniband support for CephFS is very limited, no latency gains due to OSDs having high I/O latency anyway.
- HTCondorCE-Slurm gateway
  - Users can condor_submit to a CE machine
  - Job is run on SLURM

HPC ❤ CephFS

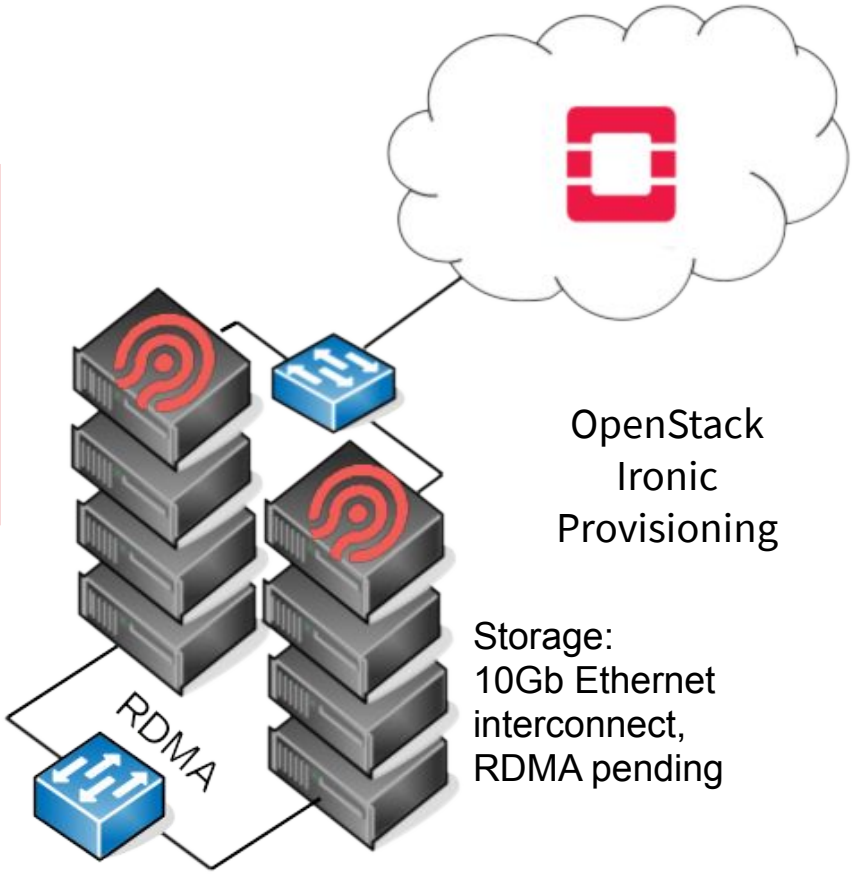
Hyperconverged
Compute + Storage

- Intel Xeon E5 2630 v4
- 128GB 2400Mhz 18ASF2G72PDZ-2G3B1
- 4x 960GB Intel S3520 SATA3
- RDMA Interconnect (compute)
- Mellanox MT27500 ConnectX-3 56Gb/FDR
- 10Gb Ethernet (storage)

- CephFS
- Network-local
- Pinned MDS
- OSDs on compute nodes
- 2x replication
- Rack-aware replication
- (Lazy I/O relaxed POSIX)

IO500 SCORE:
Throughput: 3.77 GB/s
Metadata: 8.20k IOPS
Best Score: 5.56
(On 10Gb Ethernet)
Summer 2018

Openstack + CephFS



OpenStack
Ironic
Provisioning

Storage:
10Gb Ethernet
interconnect,
RDMA pending