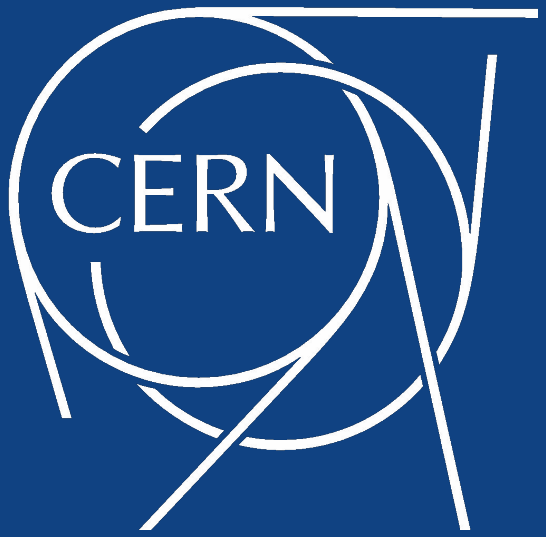

CERN

Integrating HPC into an agile and cloud-focused environment at CERN

Carolina Lindqvist, Pablo Llopis, Nils Høimyr, Dan van der Ster

Agenda

- HPC at CERN
- Challenges for HPC in an agile environment
- Challenges for HPC in cloud computing environments
  - Computing in the cloud
  - Parallel shared filesystems in the cloud
- Future work, our plans for our HPC infrastructure

High Performance Computing (HPC)

Applications and use cases that do not fit the standard batch High Throughput Computing (HTC) model. Typically parallel MPI applications

- 32-2000 cores for a single job
- Applications that scale well with parallelization
  - MPI application performance requires fast interconnects with low latency between nodes in a cluster
  - Stability of OS and environment critical
  - Some applications also require fast access to shared storage

User community

Beams

- gdfdl (field calculations for RF cavities)
- Plasma simulations for Linac 4, *Beam simulations for LHC, CLIC, FCC…*
- *PyOrbit*

Theoretical Physics

- Lattice QCD simulations

Health & Safety Engineering

- Safety/fire simulations (CFD)

Technology

- Picmc
- *Potentially also engineering (Ansys and Comsol)*

Engineering

- *CFD (Ansys-Fluent, OpenFOAM)*
- *Structural analysis (Ansys, LS-Dyna...)*

WLCG

- *Backfill with Grid jobs to increase cluster utilization (HTCondor-CE - GAHP - SLURM)*

Challenges of agile

Agile Methodologies

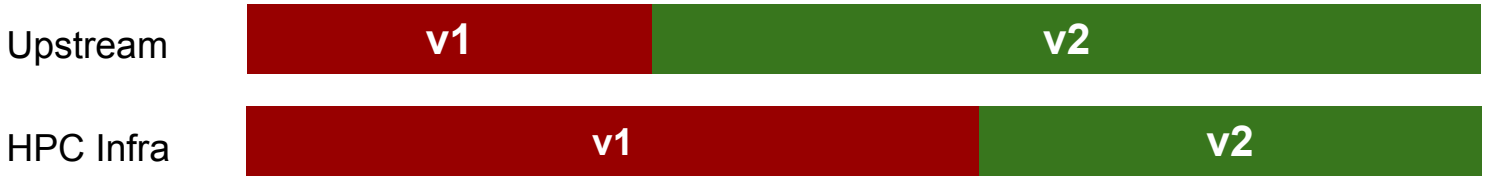
- High automation
- Frequent changes

HPC

- Long-running jobs (several weeks)
- Stability desired

Agile + HPC

- Keep high level of automation, frequent changes
- Separate testing and production environments
- Perform extensive testing before rolling out to production
- Rely on repository snapshotting to prevent external changes from sneaking into production without sufficient testing.



Paradigm Collision

Traditional Cloud

- “Sparse” computing
- High throughput model
- Independent resources
- Communication through virtualized network
- Placement of computing resources meant for availability, not for performance

HPC

- Extremely sensitive to latency
- Benefits from dedicated resources
- High bandwidth requirements
- Placement affects latency
- Latency, latency, latency!

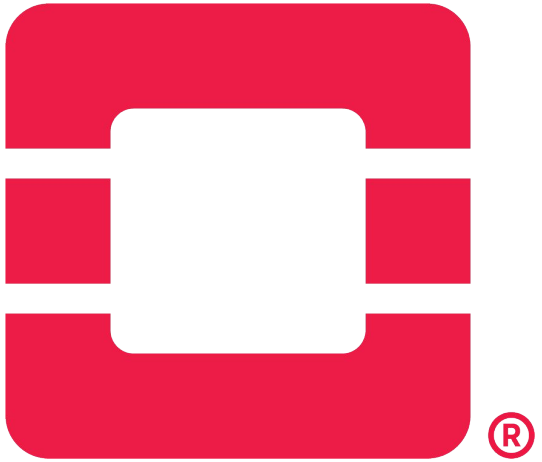
Paradigm Collision

Traditional Cloud

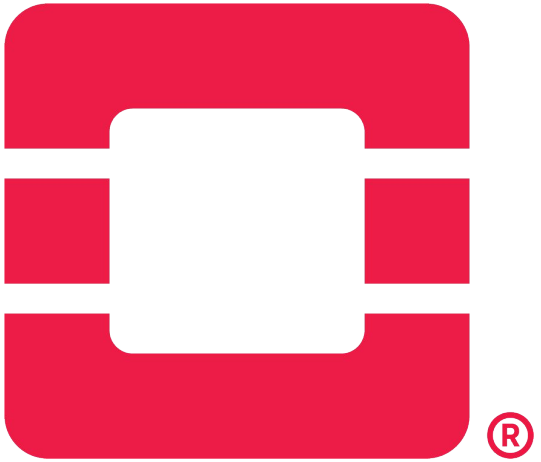
HPC

- High resource utilization
- Large scale
- Highly available
- Well known advantages for managing infrastructure more productively

HPC



CERN HPC runs on OpenStack Ironic



HPC

High Performance Computation

High Performance File I/O

HPC ❤ OpenStack

- OpenStack Ironic provides bare-metal provisioning
- Allows full access to raw resources without hypervisor isolation or overhead
  - No resource sharing among tenants
  - Bare-metal access means faster context switching, no hypercalls, less cache flushes, less overhead (latency!)
  - Bare-metal access means full PMU access
  - Possibility to optimize low-level BIOS and kernel settings
  - Take full advantage of fast interconnects such as Infiniband

HPC ❤ OpenStack

- However, granularity of resource placement is completely different
  - Cloud offers cluster-wide granularity. You do not know where your instance will be located
  - HPC needs switch-wide granularity
  - Latency!!

HPC ❤ OpenStack

Steps when deploying new nodes:

1. Create new bare metal instances
2. Gather network topology from resulting placement
   2.1. Infiniband provides topology auto discovery tools
3. Configure infrastructure with topology

HPC ❤ CephFS

CephFS is a popular parallel shared filesystem for clouds.

Parallel, Consistent, Self-healing, Extremely scalable.

Heavily used at CERN (over 13 PB across various clusters), especially on OpenStack.

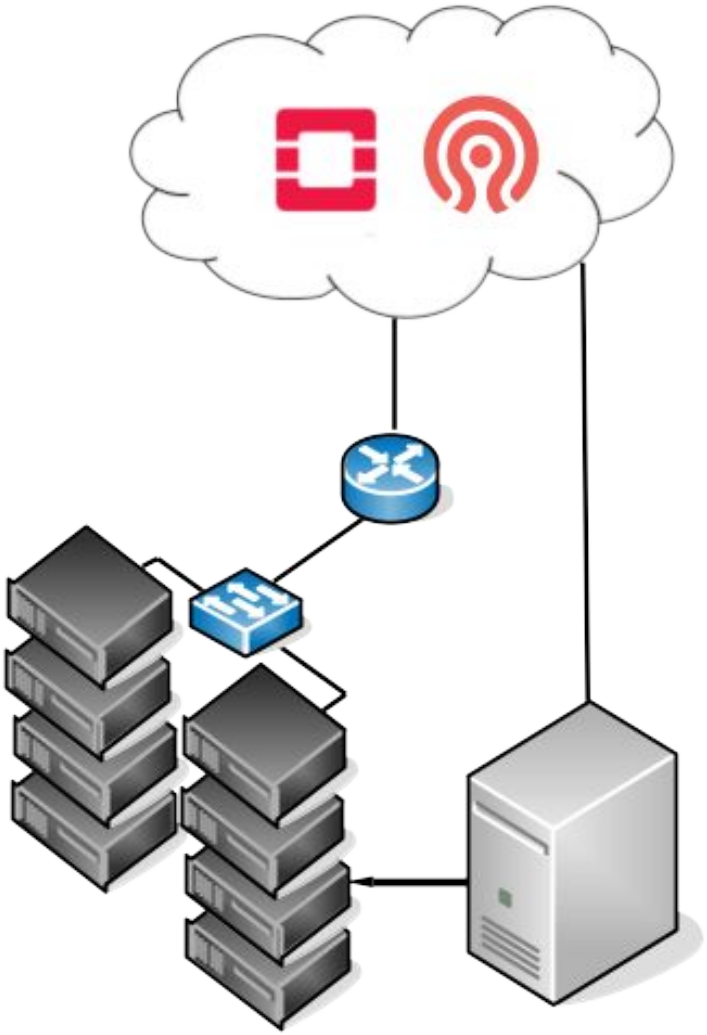
Benchmarking: IO500 score

- IOR easy (throughput; independent parallel file I/O)
- IOR hard (throughput; shared parallel file I/O)
- Mdtest easy (metadata; independent metadata I/O)
- Mdtest hard (metadata, shared directory metadata I/O)

HPC ❤ CephFS

HPC
Worker nodes

- Intel Xeon E5 2630 v3
- 128GB Memory 1600Mhz
- RAID 10 SATA HDDs
- Low-latency Chelsio T520-LL-CR
- Communication iWARP/RDMA CM



CephFS Jewel

- 3x replication
- Per-host replication
- Shared file POSIX consistency model
- Mon, MDS live in cloud

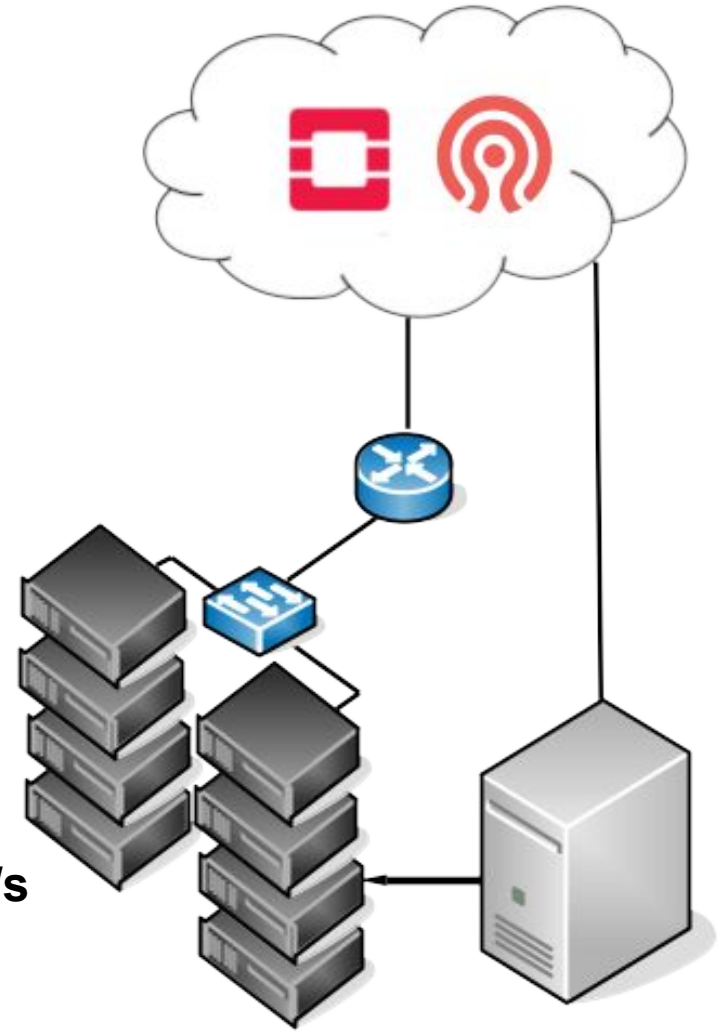
Legacy
Bare-metal
provisioning

VMs on OpenStack

HPC ❤ CephFS

HPC
Worker nodes

- Intel Xeon E5 2630 v3
- 128GB Memory 1600Mhz
- RAID 10 SATA HDDs
- Low-latency Chelsio T520-LL-CR
- Communication iWARP/RDMA CM



CephFS Jewel

- 3x replication
- Per-host replication
- Shared file POSIX consistency model
- Mon, MDS live in cloud

Legacy
Bare-metal
provisioning

VMs on OpenStack

IO500 SCORE:
Throughput: 1.74 GB/s
Metadata: 3.47k IOPS
Final Score: 2.46

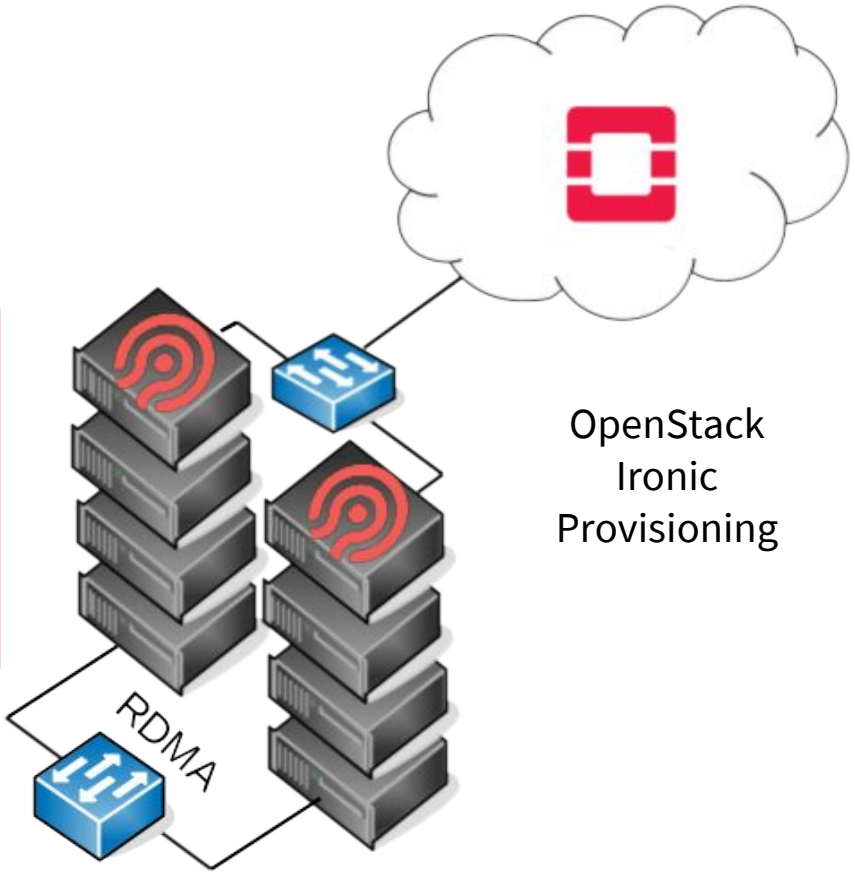
HPC ❤ CephFS

Openstack Pike + CephFS Luminous

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 Luminous 12.2.5
- Network-local
- Pinned MDS
- OSDs on compute nodes
- 2x replication
- Rack-aware replication
- Lazy I/O relaxed POSIX



OpenStack
Ironic
Provisioning

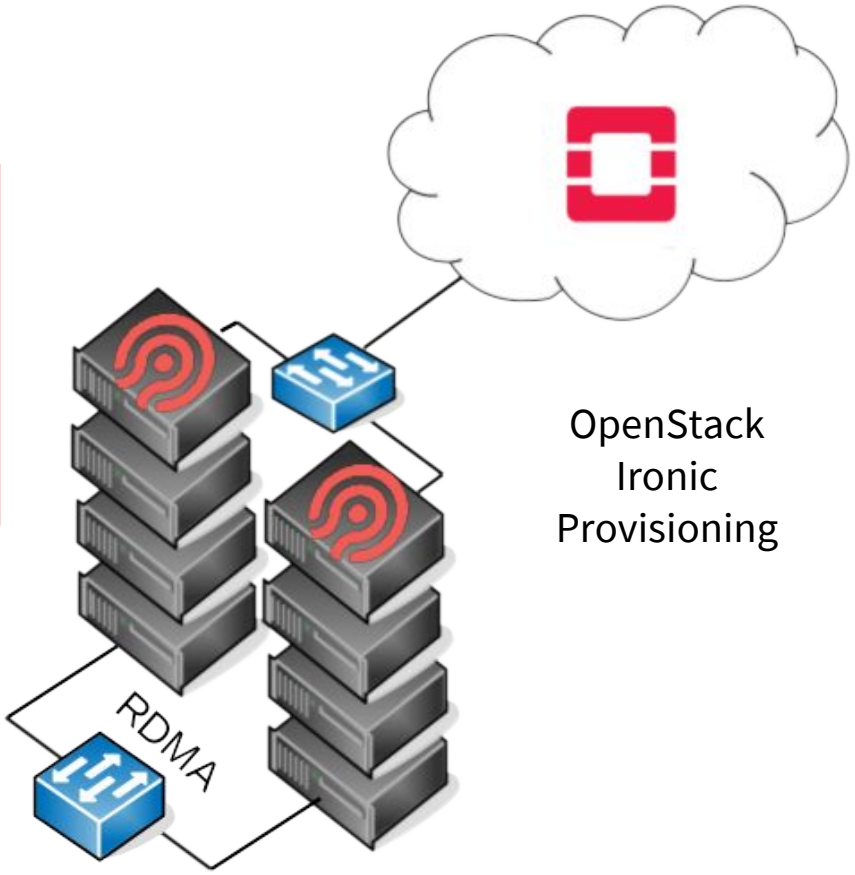
HPC ❤ CephFS

Openstack Pike + CephFS Luminous

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 Luminous 12.2.5
- Network-local
- Pinned MDS
- OSDs on compute nodes
- 2x replication
- Rack-aware replication
- Lazy I/O relaxed POSIX



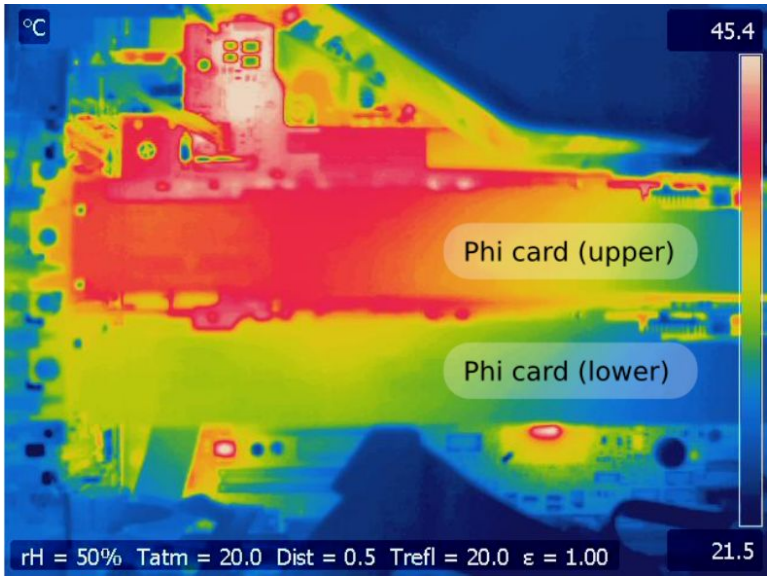
OpenStack
Ironic
Provisioning

IO500 SCORE:
Throughput: 3.77 GB/s
Metadata: 8.20k IOPS
Best Score: 5.56

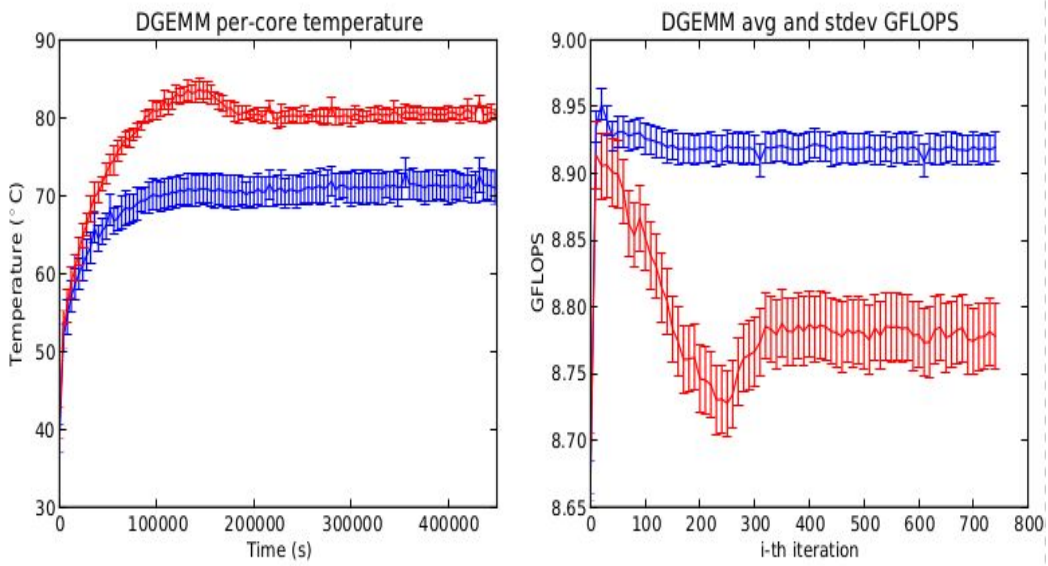
Detailed info on numbers in the following contribution:
https://indico.cern.ch/event/587955/contributions/2936868/

Future Work

- Increase resource utilization
- Increase workload power and performance efficiency
- Improve data gathering and analysis of HPC workloads



Minimizing Thermal Variation Across System Components, Zhang et al., IPDPS 2015



Understanding the dynamic nature of modern processors in preparation for exascale computing, Llopis et al. ANL Talk.

Highlights

- Fully integrate HPC in CERN's OpenStack environment
- Safe integration of HPC with agile methodologies
- We use Ironic bare-metal provisioning and re-engineer our tools and deployment methodology to adapt to a dynamic network topology
- We perform several optimizations to the CephFS instance, creating a hyperconverged compute+storage solution for high performance I/O
- Parallel file I/O speedups of 3x

Questions and discussion