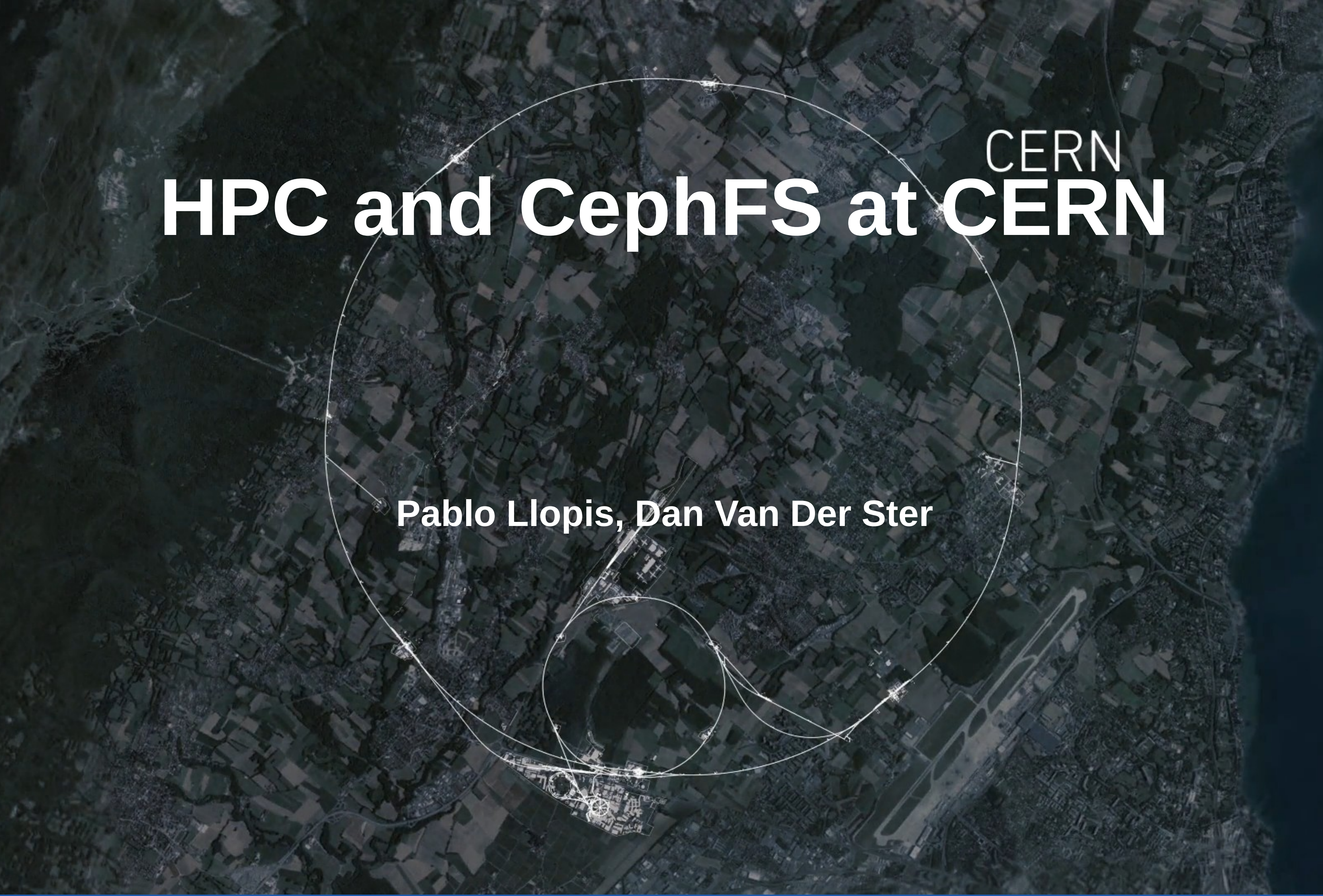
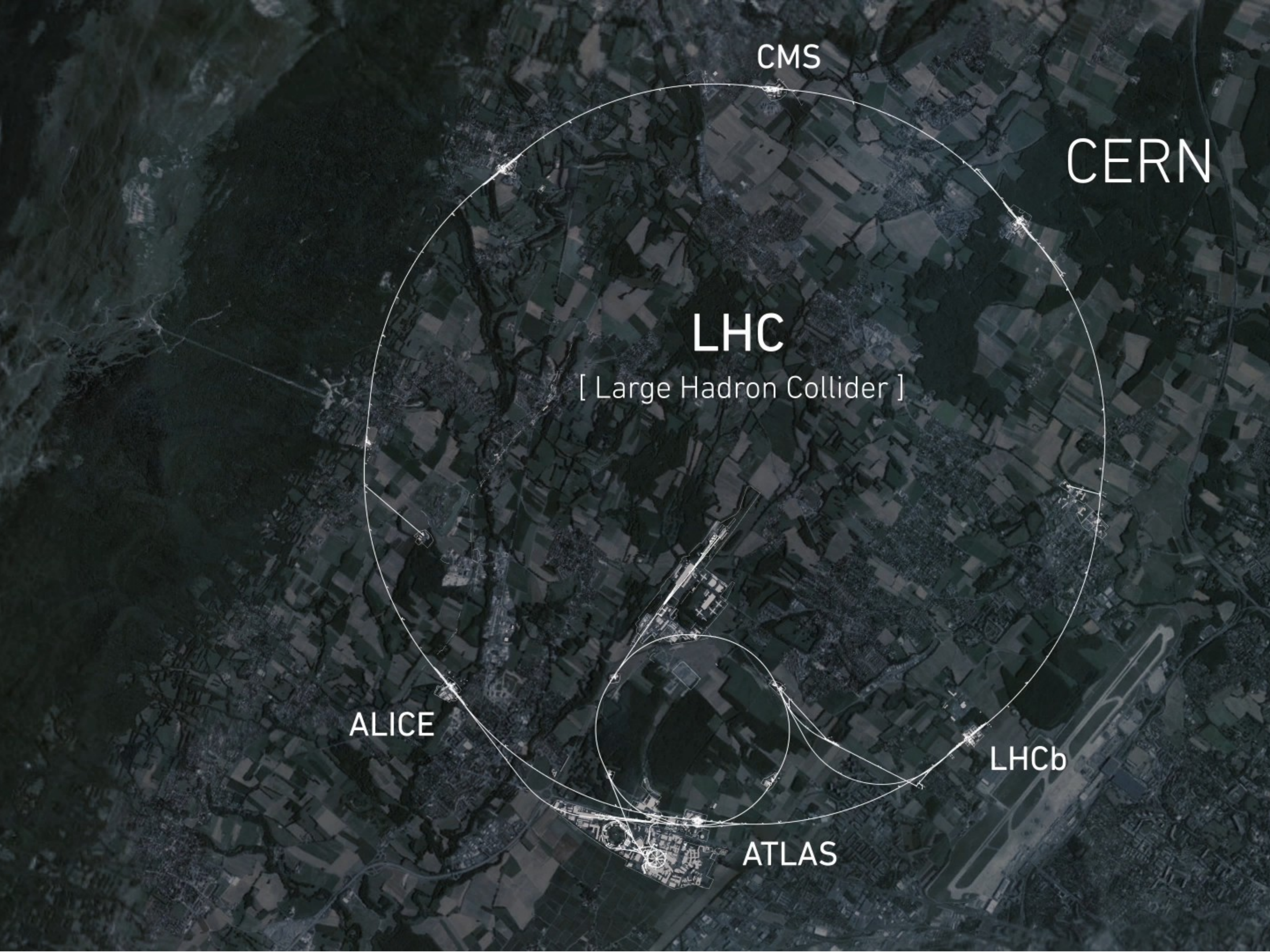
CERN

HPC and CephFS at CERN

Pablo Llopis, Dan Van Der Ster





CERN

LHC

[Large Hadron Collider]

CERN ❤ CephFS

CephFS is a popular parallel shared filesystem for clouds.

Parallel, Consistent, Self-healing, Extremely scalable.

Heavily used at CERN (over 16 PB across various clusters), especially for 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)

CERN Ceph Clusters		**Size**	**Version**
OpenStack Cinder/Glance	*Production*	5.5PB	luminous
	Satellite data centre (1000km away)	1.6PB	luminous
	Hyperconverged KVM+Ceph	16TB	luminous
CephFS (HPC+Manila)	*Production*	0.8PB	luminous
	Client Scale Testing	0.4PB	luminous
	Hyperconverged HPC+Ceph	0.4PB	luminous
CASTOR/XRootD	*Production*	4.4PB	luminous
	CERN Tape Archive	0.8TB	luminous
S3+SWIFT	*Production*	2.3PB	luminous

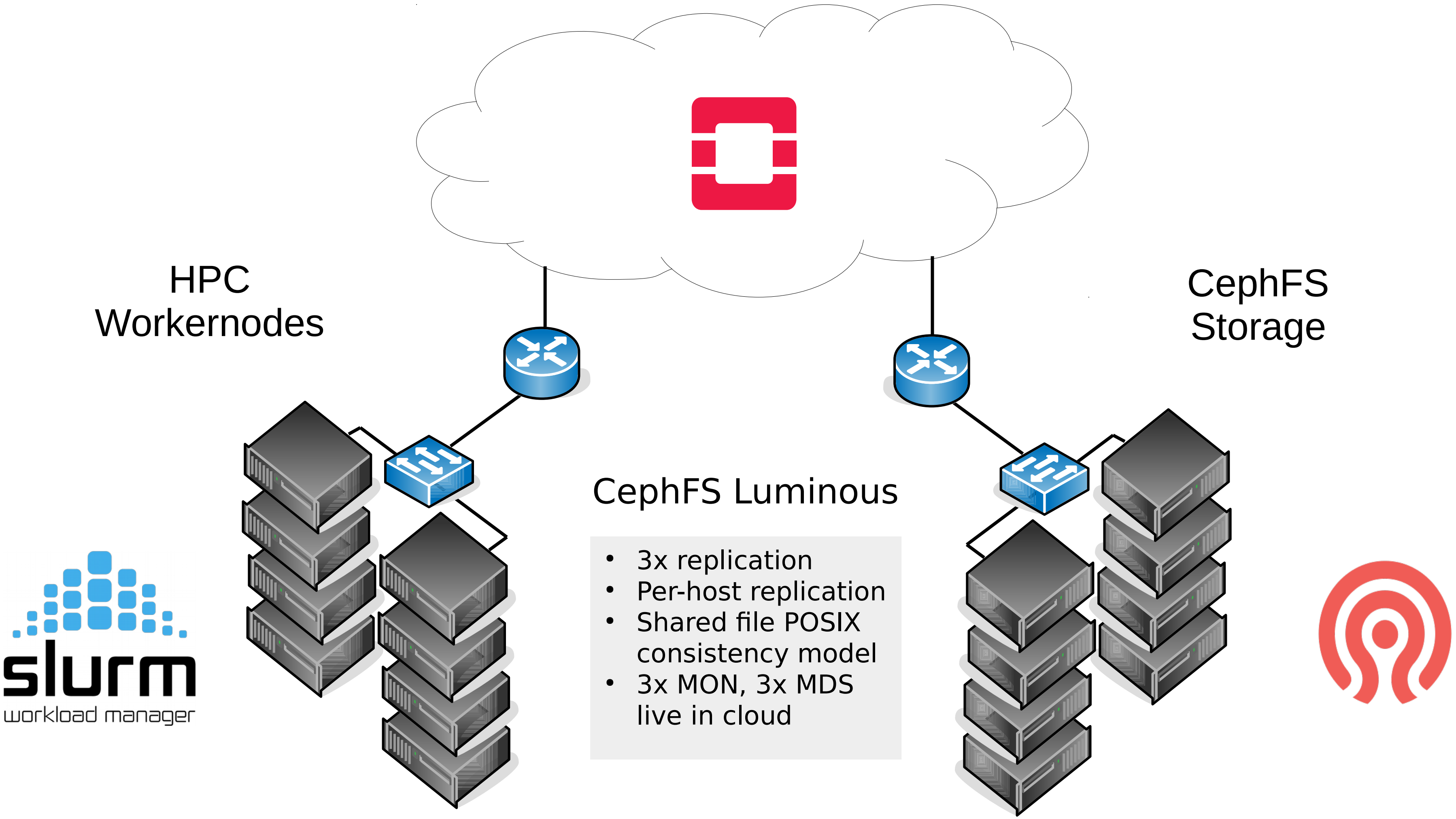
CERN ❤ CephFS

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)

Final score is a geometric mean of workload performance results.

Initial design: Cloud mountpoints



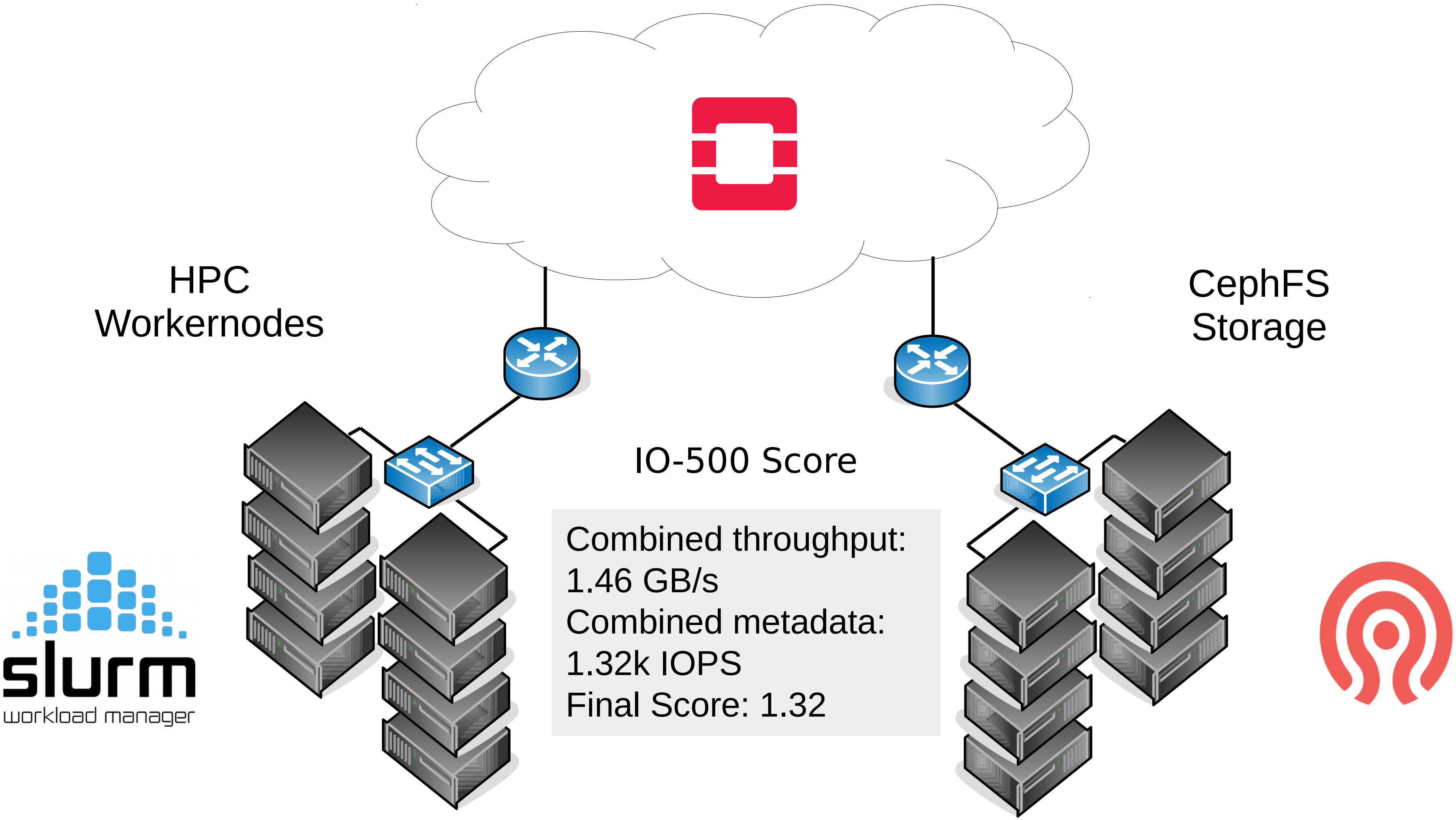
HPC
Workernodes

CephFS
Storage

CephFS Luminous

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

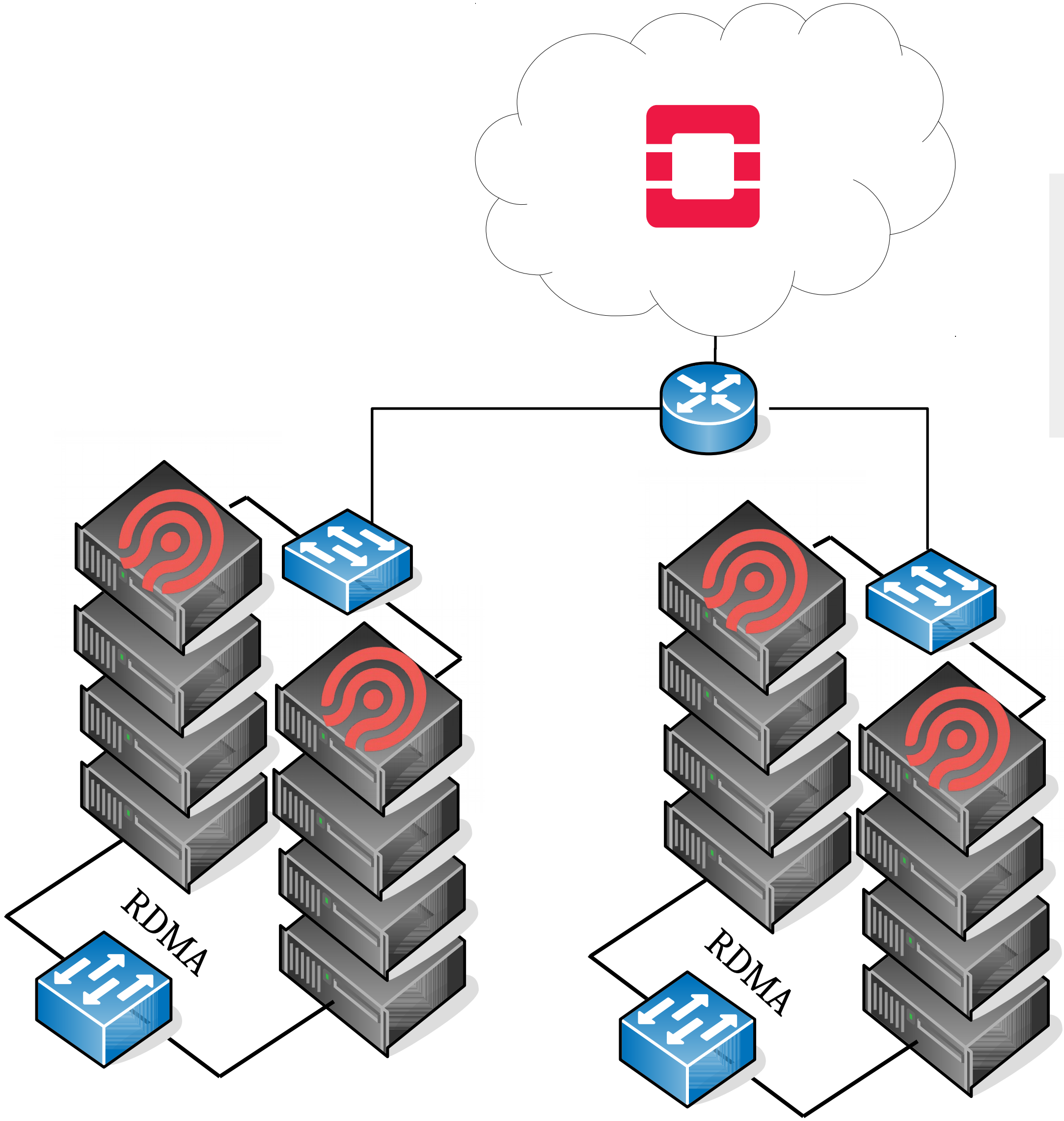
Initial design: Cloud mountpoints



IO-500 Score

Combined throughput:
1.46 GB/s
Combined metadata:
1.32k IOPS
Final Score: 1.32

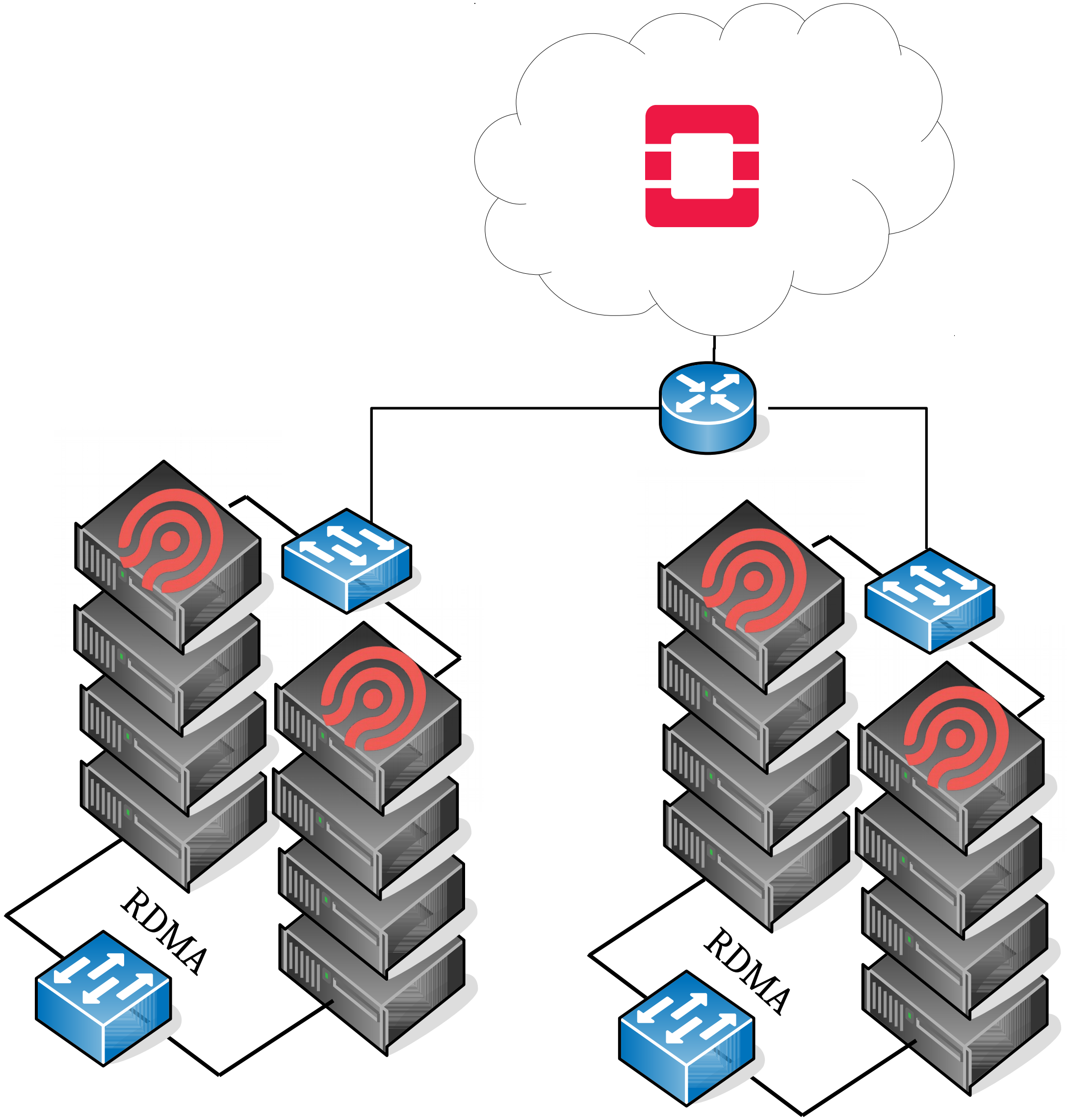
IOR independent I/O 2.5GB/s



CephFS Luminous

- Shared file POSIX consistency model
- **Hyperconverged OSDs**
- **Hyperconverged MDSs**
- **2x replication**

Hyperconverged Architecture



CephFS Luminous

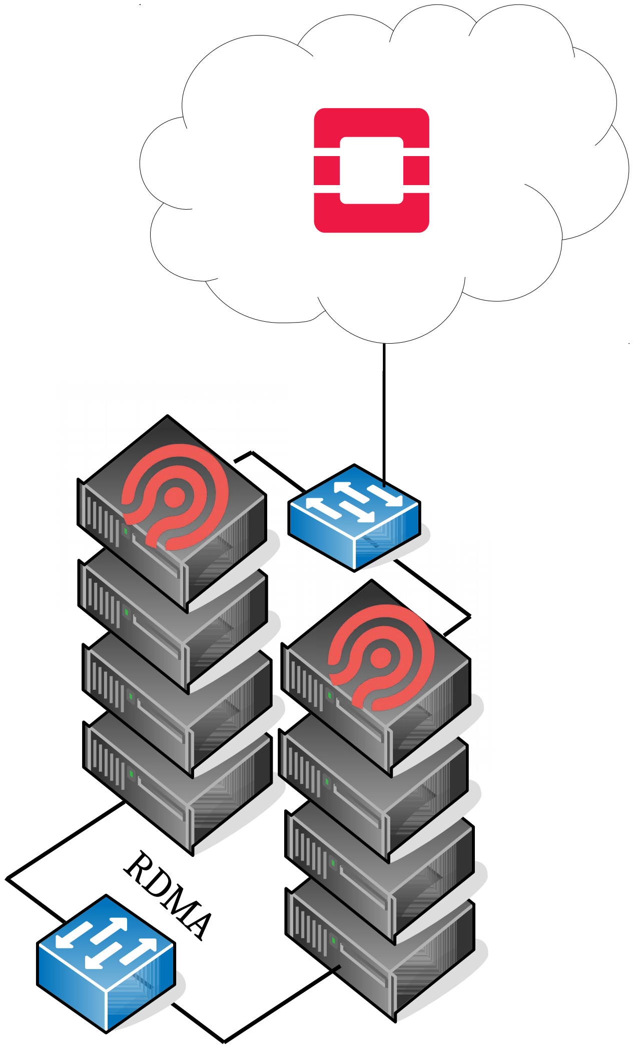
- Shared file POSIX consistency model
- **Hyperconverged OSDs**
- **Hyperconverged MDSs**
- **2x replication**

IO-500 Score

Combined throughput: 2.13 GB/s
Combined metadata: 6.52k IOPS
Final Score: 3.73

IOR independent I/O 3.7 GB/s

Hyperconverged Architecture



CephFS Luminous

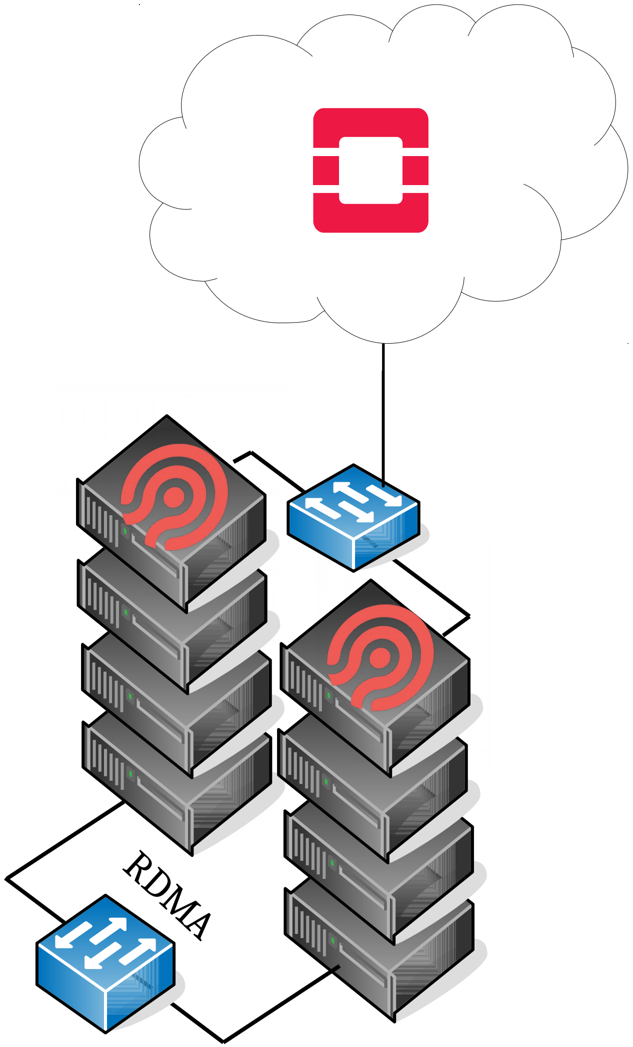
- Hyperconverged OSDs
- Hyperconverged MDSs
- 2x replication
- Shared file POSIX consistency model
- **No routing in the data path**

IO-500 Score

Combined throughput: 2.54 GB/s
Combined metadata: 6.55k IOPS
Final Score: 4.08

IOR independent I/O 4.01 GB/s

Hyperconverged Architecture



CephFS Luminous

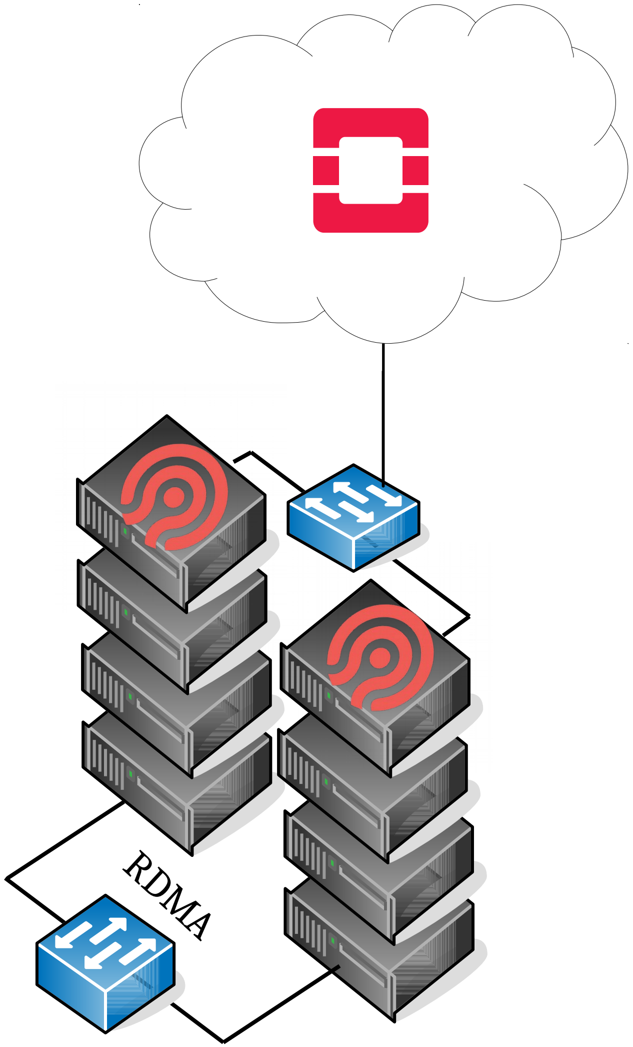
- Hyperconverged OSDs
- Hyperconverged MDSs
- 2x replication
- Shared file POSIX consistency model
- 3x MON, 3x MDS live in cloud
- No routing in the data path
- **MDS pinning**

IO-500 Score

Combined throughput: 2.72 GB/s
Combined metadata: 8.43k IOPS
Final Score: 4.79

mdtest independent I/O went from 12 kIOPS to 26 kIOPS

Hyperconverged Architecture



CephFS Luminous 12.2.9

- Hyperconverged OSDs
- Hyperconverged MDSs
- 2x replication
- Shared file POSIX consistency model
- 3x MON, 3x MDS live in cloud
- No routing in the data path
- **3 local MDSs**
- **Replace ceph-fuse by kernel client**

IO-500 Score

Combined throughput: 2.82 GB/s
Combined metadata: 14.49k IOPS
Final Score: 6.39

Hyperconverged Architecture

File-level locking during collective I/O operations

```
[RESULT] BW   phase 1            ior_easy_write                4.757 GB/s
[RESULT] BW   phase 2            ior_hard_write                0.838 GB/s

[RESULT] BW   phase 3             ior_easy_read                7.562 GB/s : time 270.85
seconds
[RESULT] BW   phase 4             ior_hard_read                2.104 GB/s : time  43.94
seconds
[RESULT] IOPS phase 1         mdtest_easy_write                9.137 kiops : time 200.90
seconds
[RESULT] IOPS phase 2         mdtest_hard_write                5.709 kiops : time 227.21
seconds
[RESULT] IOPS phase 3                      find              146.550 kiops : time  17.47
seconds
[RESULT] IOPS phase 4          mdtest_easy_stat               58.724 kiops : time  25.93
seconds
[RESULT] IOPS phase 5          mdtest_hard_stat               27.526 kiops : time  49.22
seconds
[RESULT] IOPS phase 6        mdtest_easy_delete                5.392 kiops : time 239.88
seconds
[RESULT] IOPS phase 7          mdtest_hard_read                6.330 kiops : time 204.64
seconds
[RESULT] IOPS phase 8        mdtest_hard_delete                4.617 kiops : time 279.74
seconds
[SCORE] Bandwidth 2.82241 GB/s : IOPS 14.4933 kiops : TOTAL 40.90603
```

LAZY IO in CephFS is implemented for the FUSE client

In commit *c6d0c0* developed by *ukernel*

Merged in *master* but *master* is very far from what we are running in production

Patch is easily ported across versions

LAZY IO in CephFS is implemented for the FUSE client

In commit # developed by ukernel

Merged in master but master is very far from what we are running in production

Patch is easily ported across versions

```
[RESULT] BW   phase 1            ior_easy_write                4.01 GB/s
[RESULT] BW   phase 2            ior_hard_write                3.97 GB/s
```

File locking and LAZY_IO

Network locality: brought services closer together to reduce latency and increase throughput. Tuned crushmap.

MDS: MDS pinning vs Automatic scale-out and balancing.

Client implementation: Kernel has higher performance, ceph-fuse allows lazy I/O semantics.

CephFS knobs we turned

Storage and Interconnect: Every workload was running on SATA SSDs and 10GbE.
We would expect a big performance boost from upgrading HW to NVMe and enabling RDMA.

Improve metadata locality: We usually only run 3 MDSs, so we could create a crush rule that moves the metadatapool closer to the MDSs.

System tuning. Kernel parameters, buffer/queue sizes, etc.

Future work

Questions and discussion