Centre de Calcul de l'Institut National de Physique Nucléaire et de Physique des Particules

Ceph storage with Openstack

Adrien GEORGET

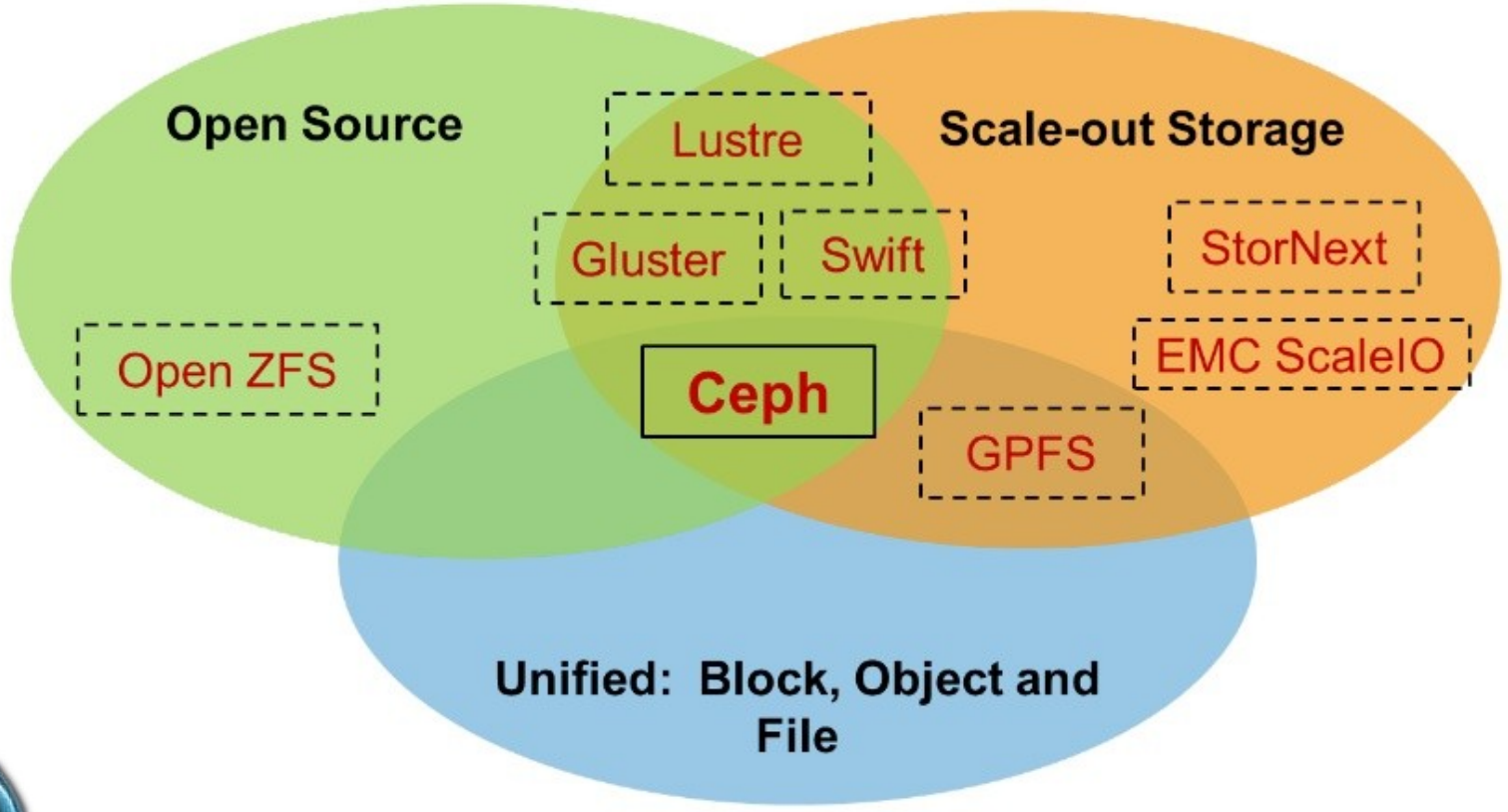
CCIN2P3

cnrs

Objectives

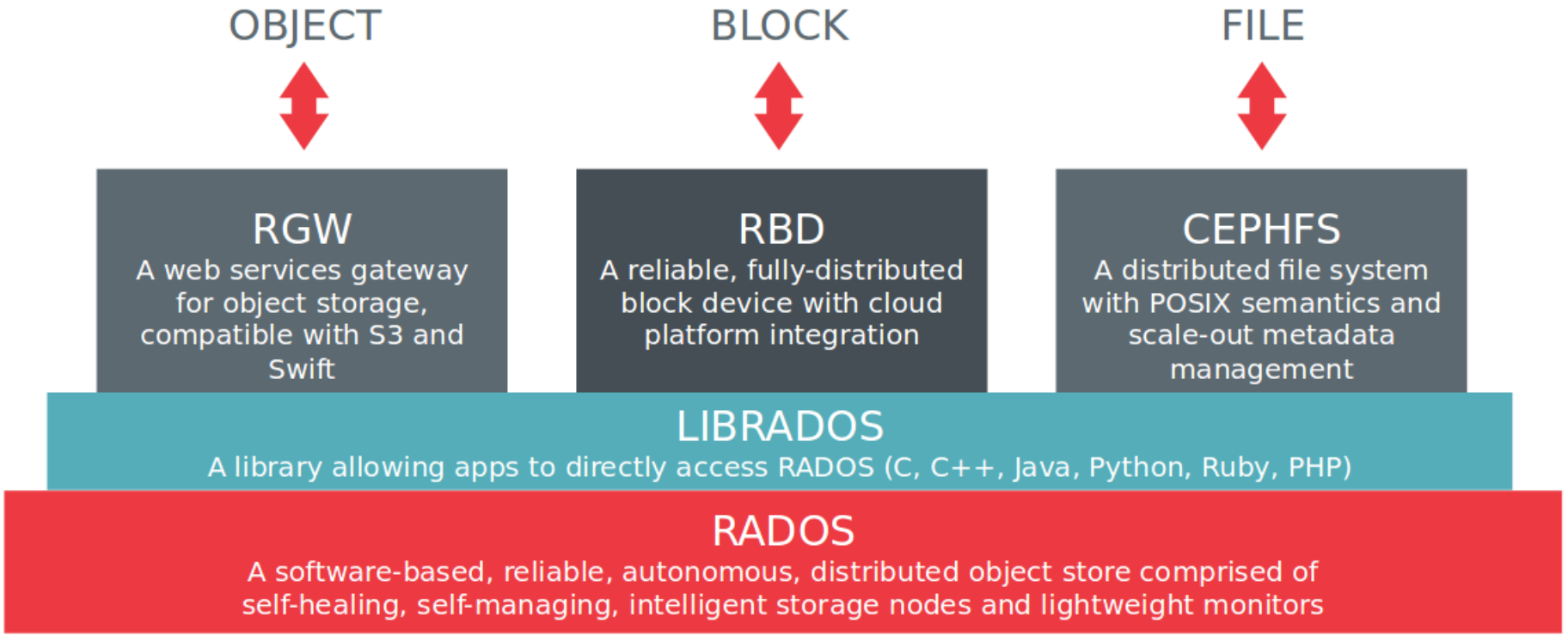
- Replace LVM storage backend for Openstack Cinder
- Provide storage on Openstack hypervisors to host VMs
- Enable Openstack Manila for shared FS ressources

Why Ceph ?

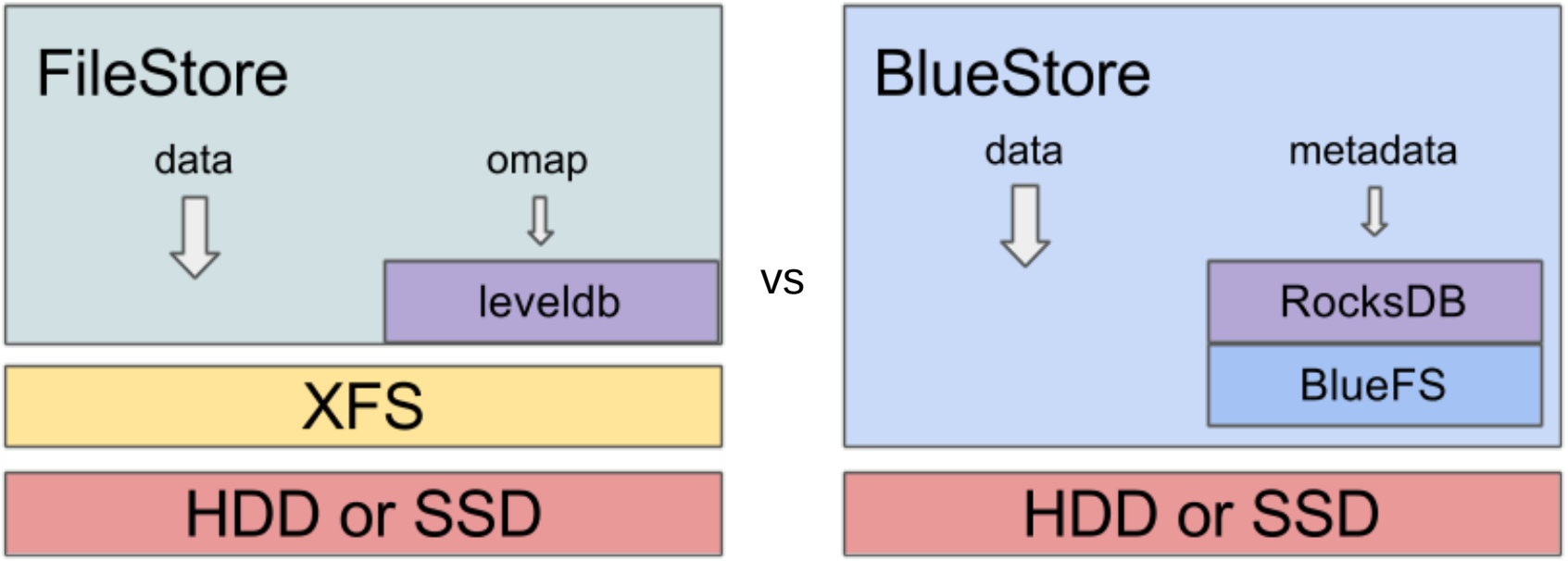

Open Source
Lustre
Scale-out Storage
Gluster
Swift
StorNext
Open ZFS
Ceph
EMC ScaleIO
GPFS
Unified: Block, Object and File

- Main features
  - Unified, distributed, and scalable storage solution
  - No single point of failure
  - Hardware agnostic
  - Self management (self healing, balancing, …)
  - Open source project with large community

- Architecture

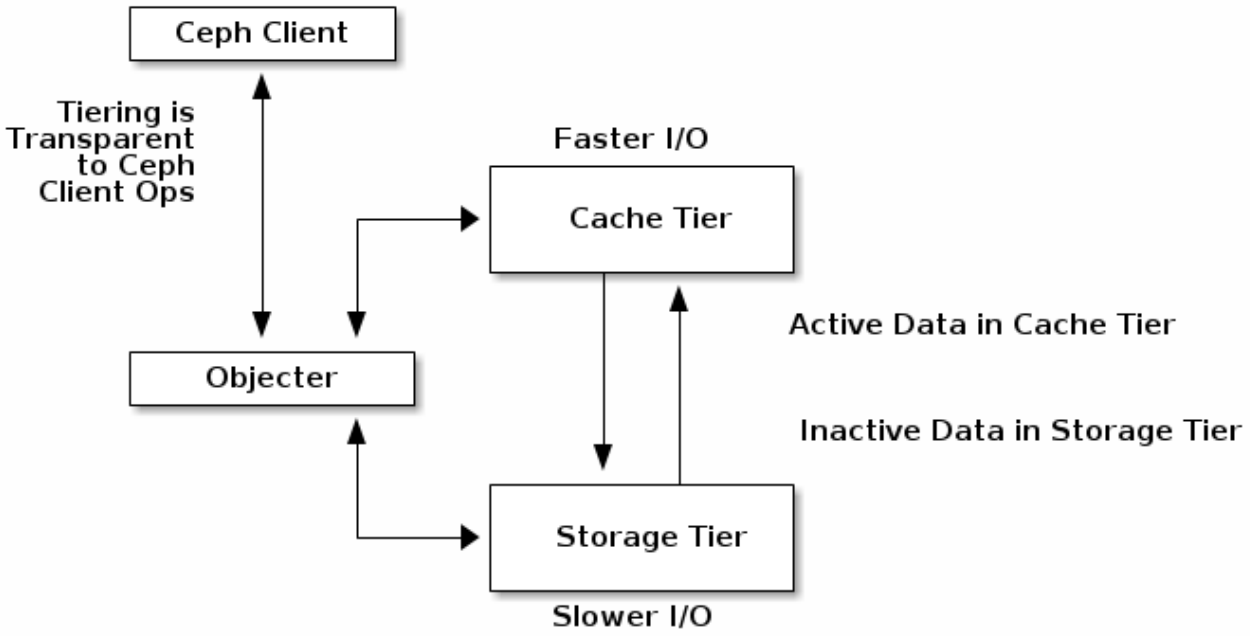


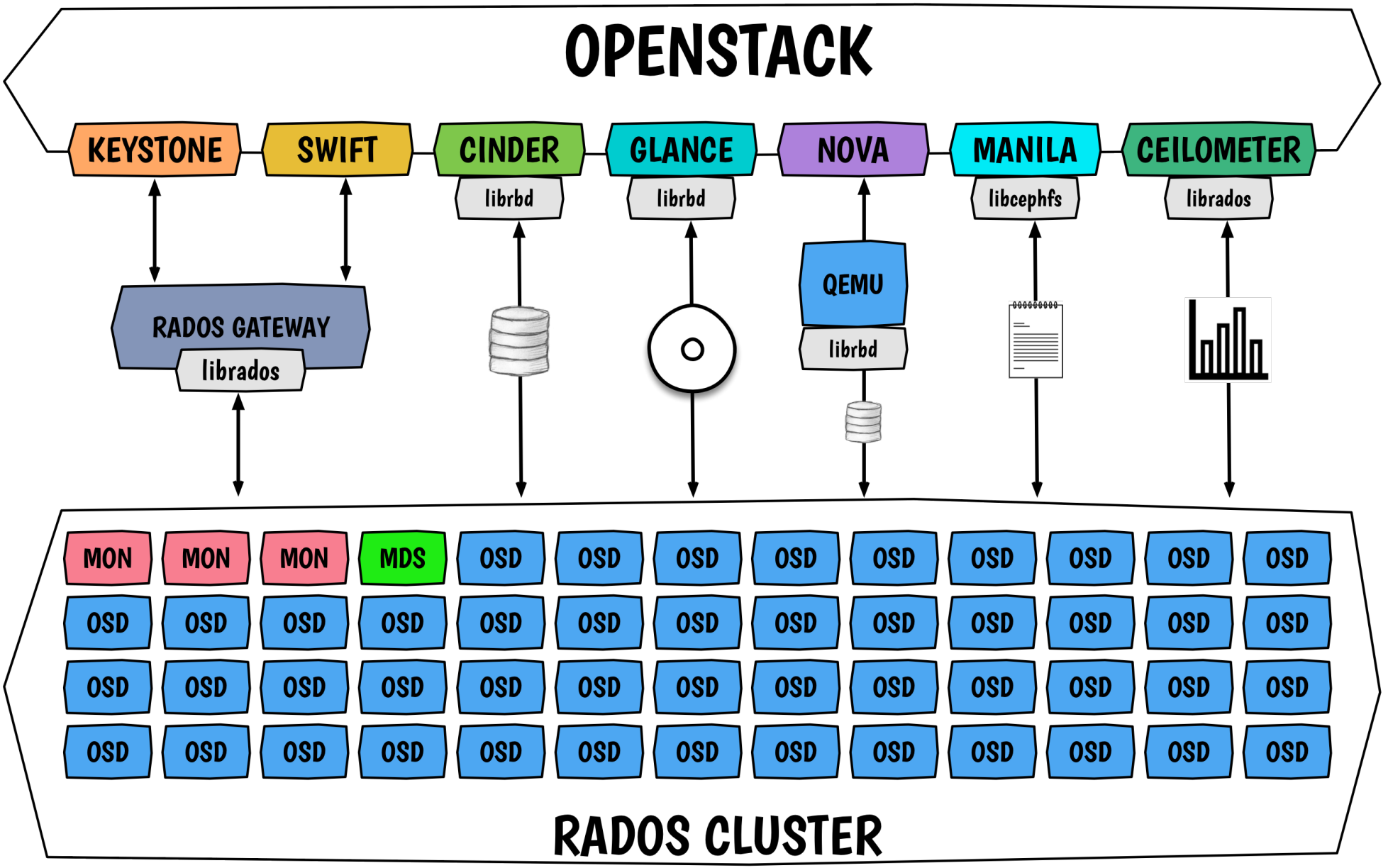
- Bluestore backend (since Luminous)



- Data written directly to raw device, no more underlying FS or dedicated journal device
- Key/value database (RocksDB) for metadata
- 2-3X performance boost
- Inline compression, full data checksums, … (optional)

- Cache Tiering
  - Better I/O perf using fast storage (SSD, NVMe, ...)
  - Each Tier is a rados pool (with replication or Erasure Code)
  - Same conf for RBD, RGW or CephFS
  - Workload dependant, wisely sizing/tuning needed




OPENSTACK
KEYSTONE
SWIFT
CINDER
GLANCE
NOVA
MANILA
CEILOMETER
librbd
librbd
libcephfs
librados
RADOS GATEWAY
librados
QEMU
librbd
MON
MON
MON
MDS
OSD
RADOS CLUSTER

Which OpenStack block storage (Cinder) drivers are in use?

Cinder drivers also remained relatively constant, with Ceph RDB up 8 points and both LVM and NetApp up 3 points.

Just a handful of respondents indicated IBM Storwize, Huawei, HDS, IBM GPFS, Dell EqualLogic, IBM XIV/DS800, Windows Server 2012, Nexenta, SAN/Solaris, HP LeftHand, XenAPI Storage Manager, Sheepdog and IBM NAS.

Among the largest clouds with 1,000 or more cores, Ceph RDB is still dominant, but not used by the majority, while other block storage drivers were also popular.

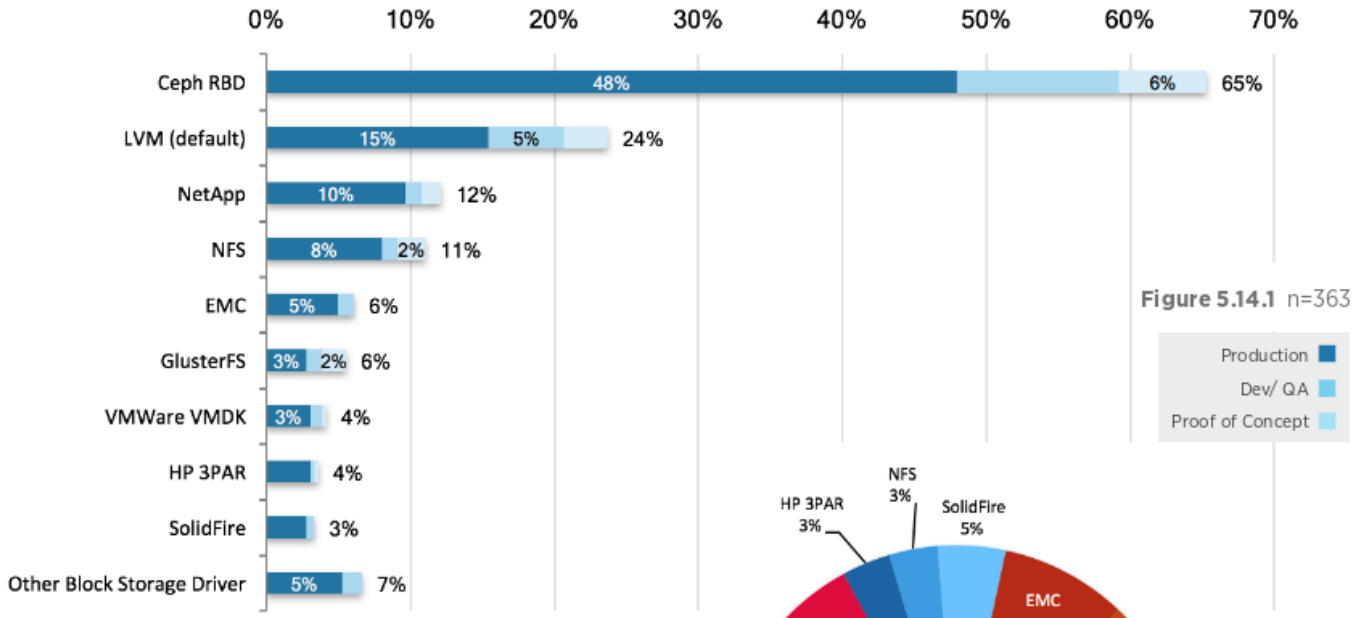


Figure 5.14.1 n=363

Among the largest clouds with 1,000 or more cores, Ceph RDB is still dominant, but not used by the majority, while other block storage drivers were also popular.

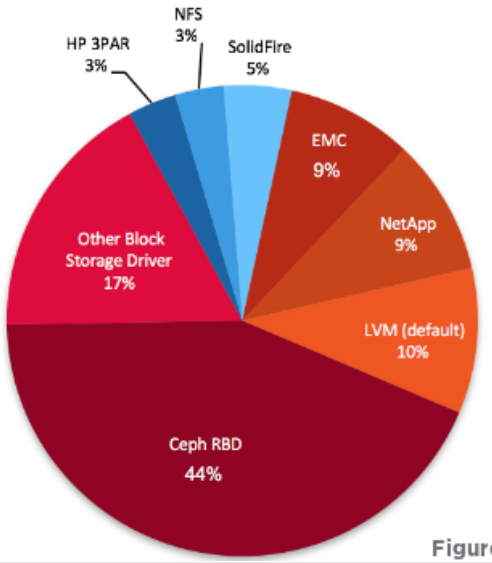


Figure 5.14.2 n=88

https://www.openstack.org/assets/survey/April2017SurveyReport.pdf

- Cinder implementation (in PRD since 01/2018)
  - CentOS 7.4
  - Ceph Luminous 12.2.2
  - 2x replication for cache and 3x RBD pools
  - Dedicated pools/disks for volumes

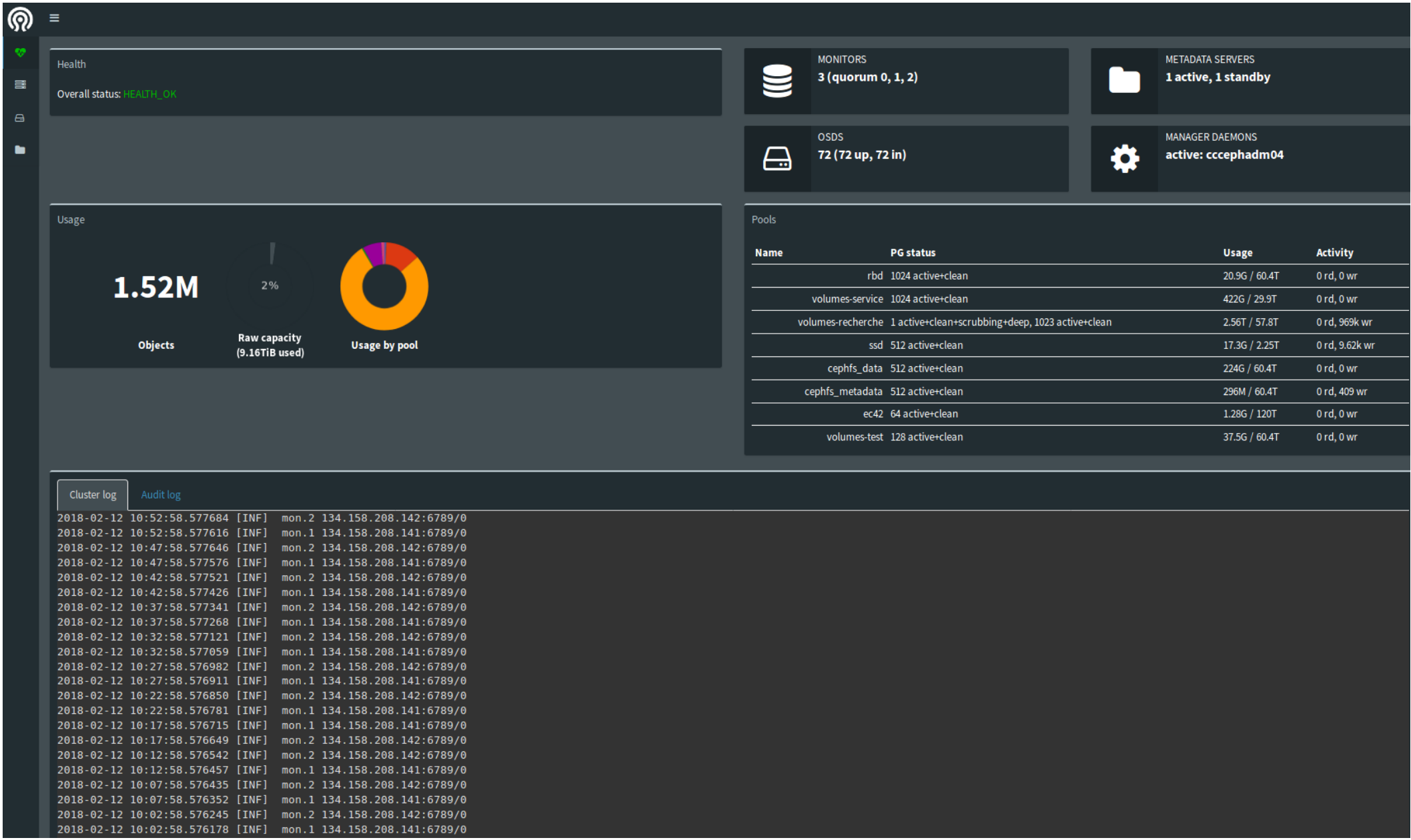
- Provides volumes on demand for Vms with various QoS
  - 30TB Volumes-service with cache tier SSD (2.5TB)
  - 60TB volumes-research
  - 60TB test (RBD, cephfs)

- ~40 volumes research, ~30 volumes services

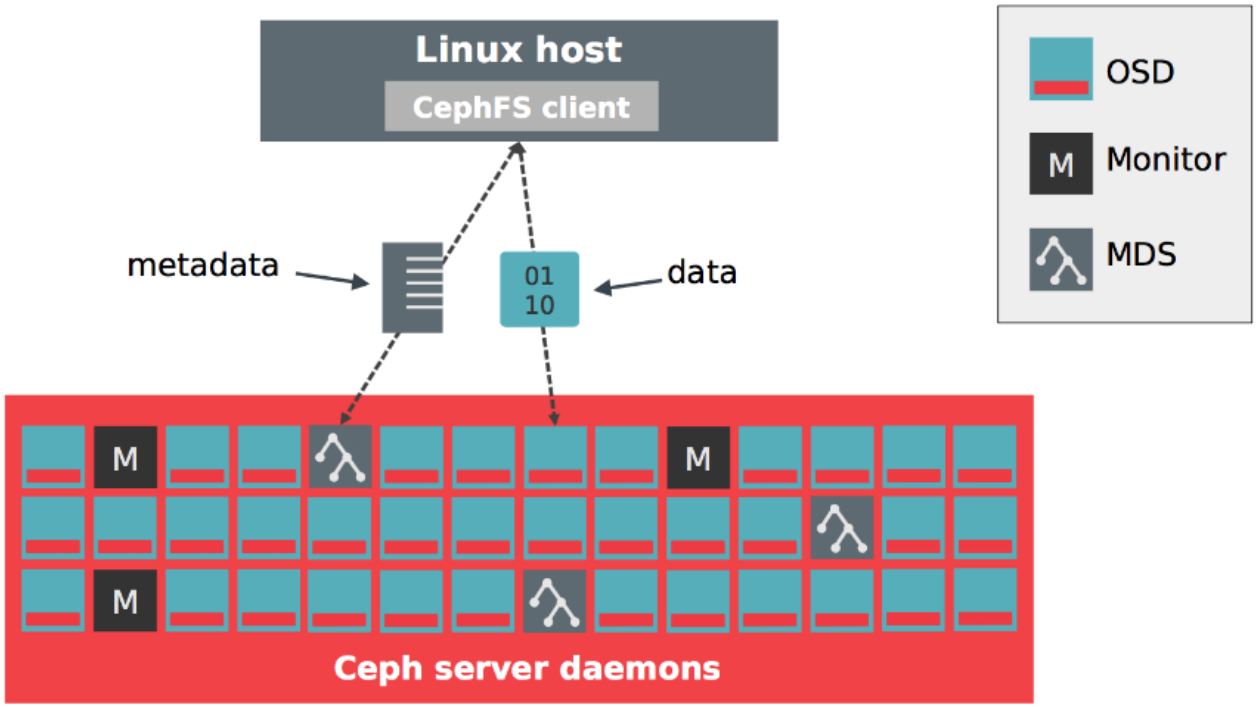
- Ceph Cinder cluster

Monitors servers (x3)	
Model	Dell R430
Processor	Intel(R) Xeon(R) CPU E5-2609 v4
Memory	32 GB
OSD servers (x6)	
Model	Dell R730XD
Processor	Intel(R) Xeon(R) CPU E5-2620 v4
Memory	64GB
Storage	10x 8TB SAS Nearline 2x 400GB SSD Write Intensive
Network	10Gbps interface

Ceph & Openstack Cinder



- Testing CephFS to host Openstack VMs
  - CentOS 7.4
  - Ceph Luminous 12.2.2
  - 2x replication for cache and 3x for data/metadatas pools
  - 2 active-active MDS (metadata servers) and 1 standby



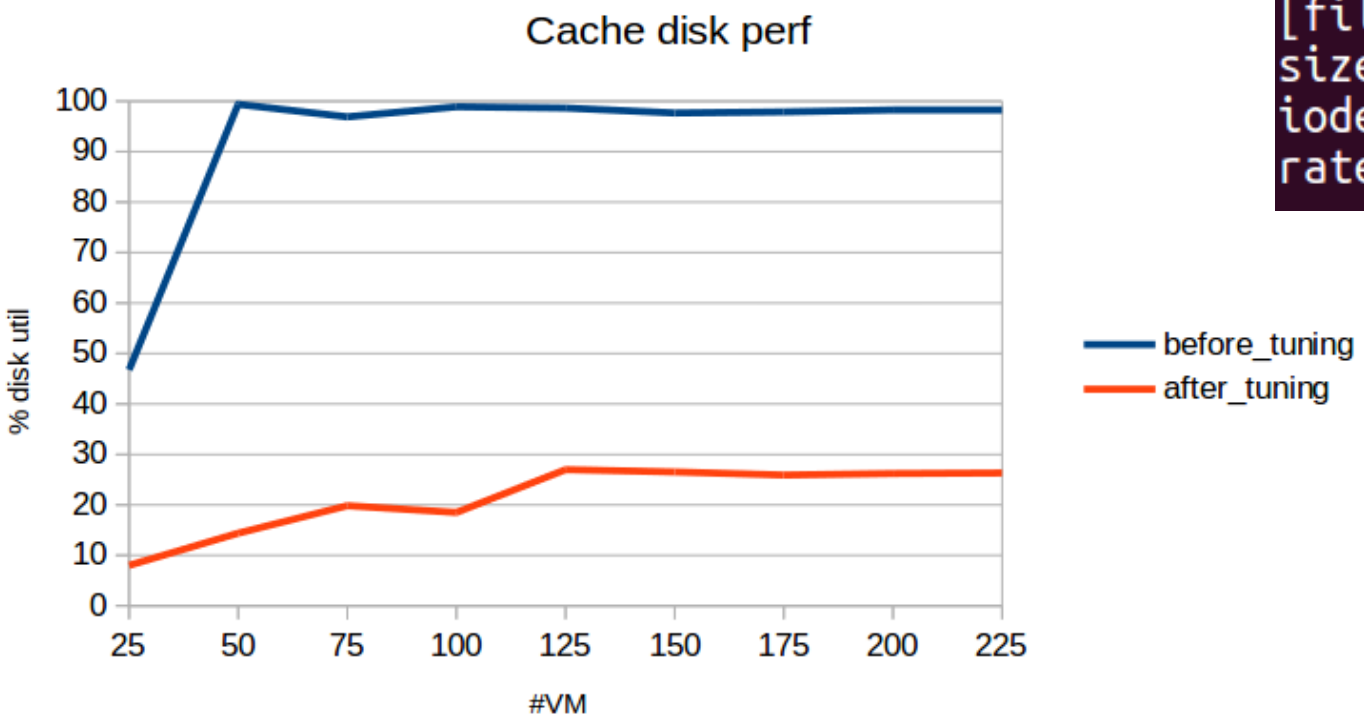
- CephFS implementation

Monitors servers (x3)	
Model	Dell R430
Processor	Intel(R) Xeon(R) CPU E5-2623 v4
Memory	32 GB
OSD servers (x2)	
Model	Dell R630 / MD3460
Processor	Intel(R) Xeon(R) CPU E5-2623 v4
Memory	32GB
Storage	7x 8TB SAS Nearline 5x 800GB SSD Write Intensive
Network	10Gbps interface

- Benchmark cache & tuning
  - Workload 80% randwrite / 20 % randread
  - 4k blocs, 8G size
  - 15min bench

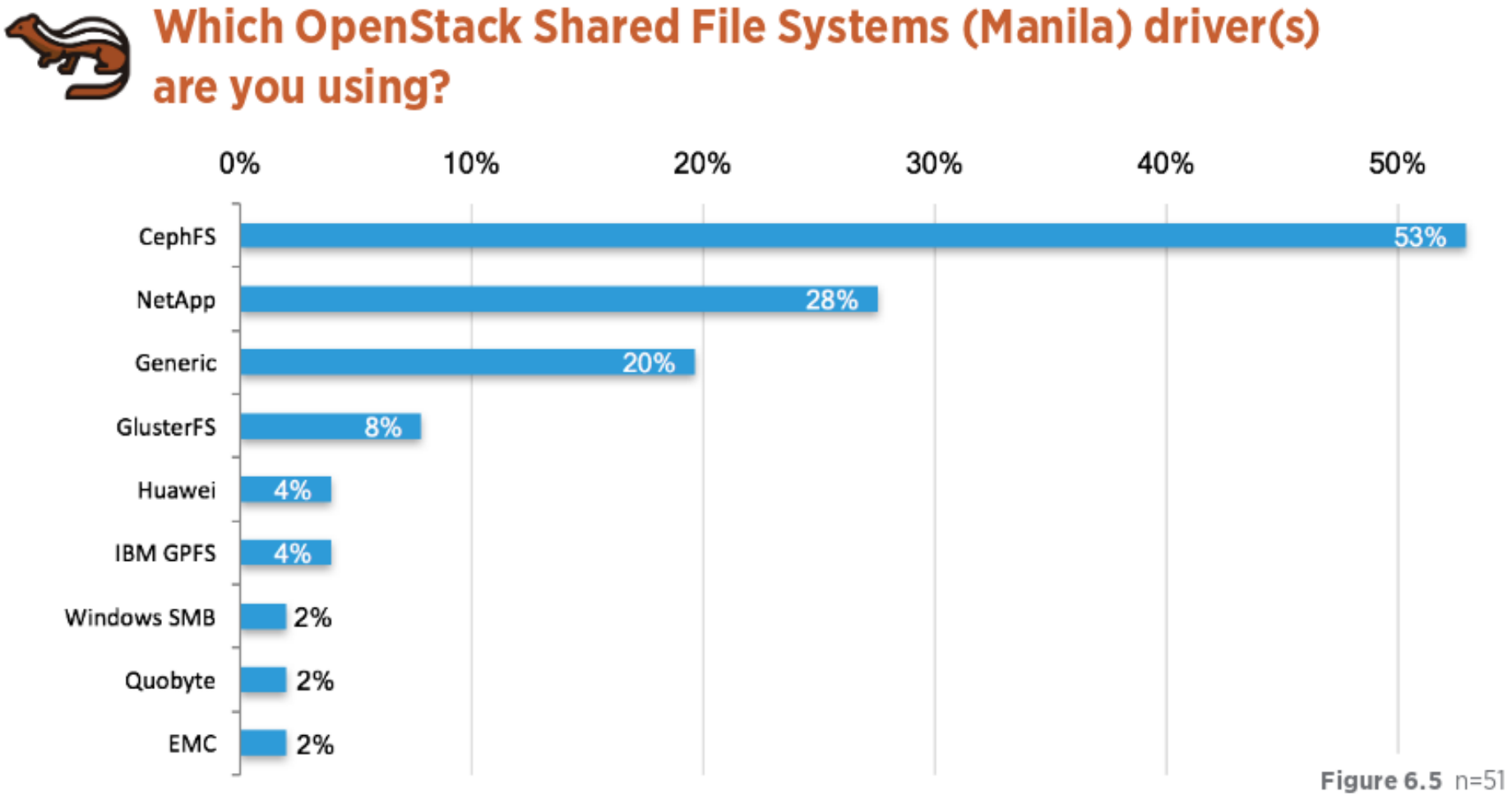
```
#
[global]
directory=/data/cloudio-randrw
rw=randrw
rwmixread=20
rwmixwrite=80
bs=4k
direct=0
time_based=1
runtime=900

[file1]
size=30G
iodepth=4
rate_iops=7,28
```



- Openstack Manila
  - FSaaS for Openstack VMs
  - Management & provisioning of file shares
  - Client restrictions, quotas, ...

- Why CephFS with Manila ?
  - Openstack cluster already include a Ceph cluster (Cinder)
  - CephFS driver Up for Manila
  - Easy to deploy (set up and working in few hours)

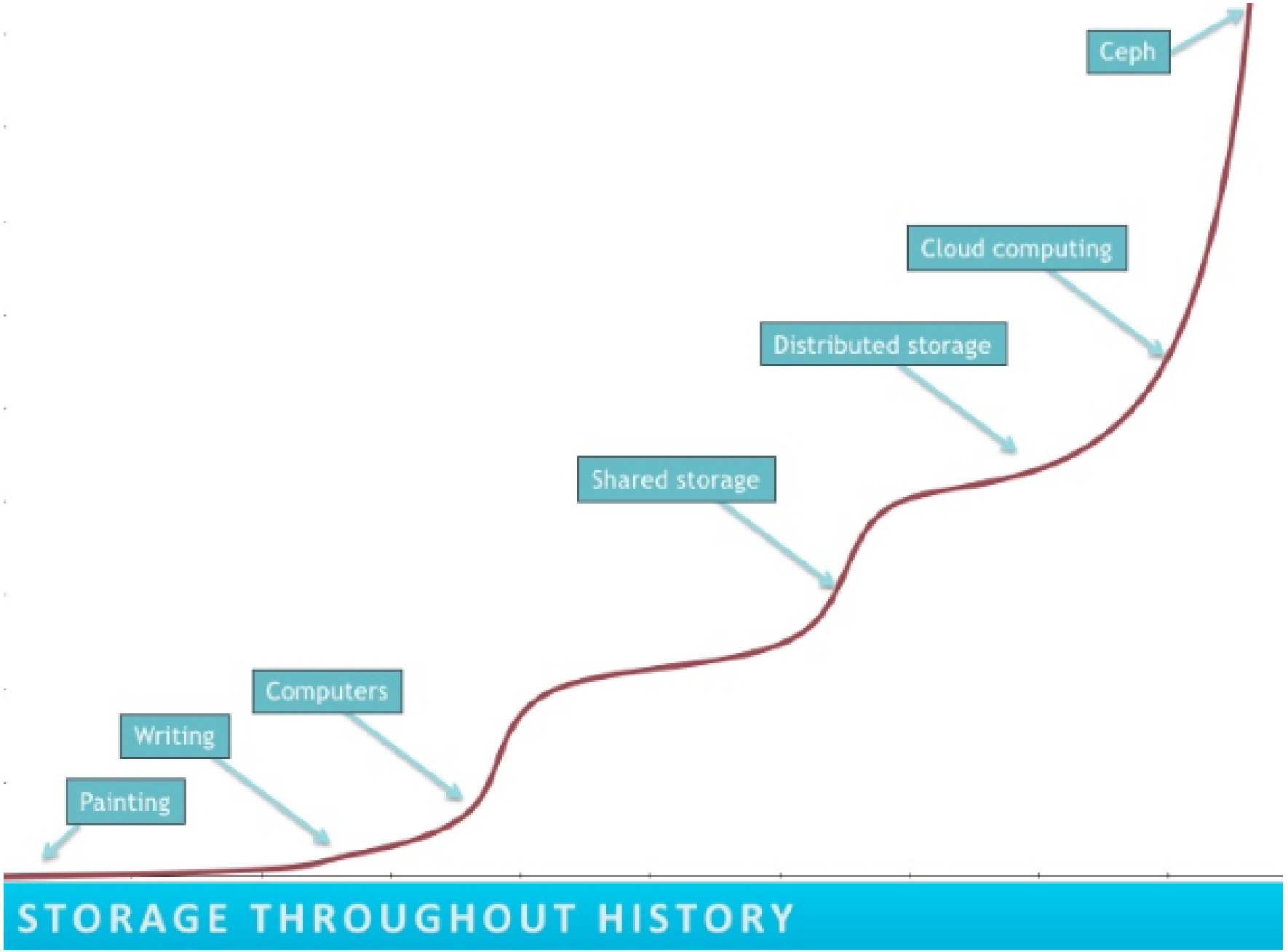


https://www.openstack.org/assets/survey/April2017SurveyReport.pdf

What’s next ?

- Erasure coding vs replication
- Enable Manila in production
- Improve logs collecting/monitoring
- RGW testbed (LSST object storage tests)

Questions ?


Ceph
Cloud computing
Distributed storage
Shared storage
Computers
Writing
Painting
STORAGE THROUGHOUT HISTORY