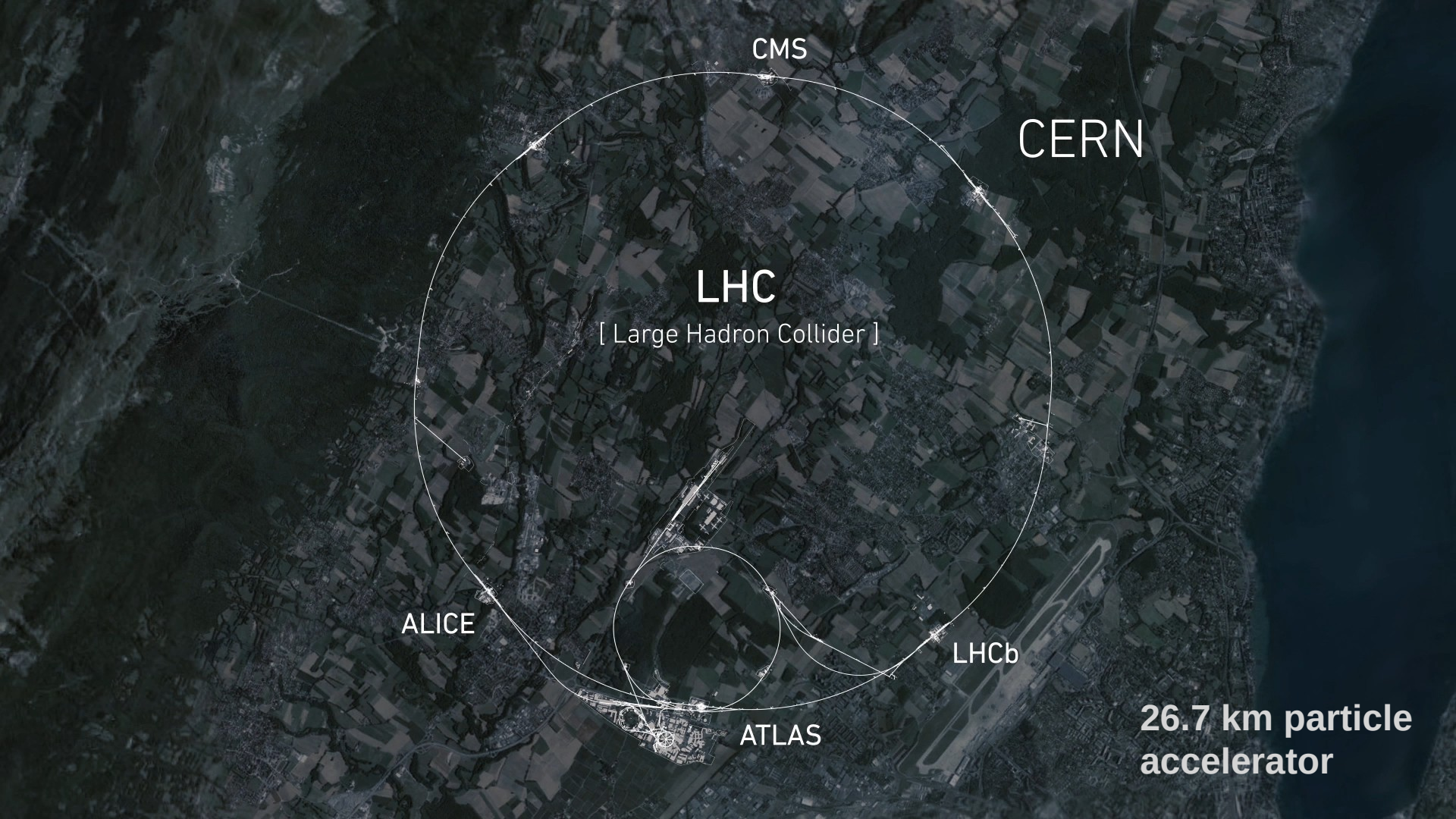

CERN


ceph
CERN
Ceph @ CERN
Julien COLLET | CERN IT | JTech Ceph Paris | 28 November 2018


CMS
CERN
LHC
[Large Hadron Collider]
ALICE
LHCb
ATLAS
26.7 km particle accelerator

O(10) GB/s - 50 PB / year

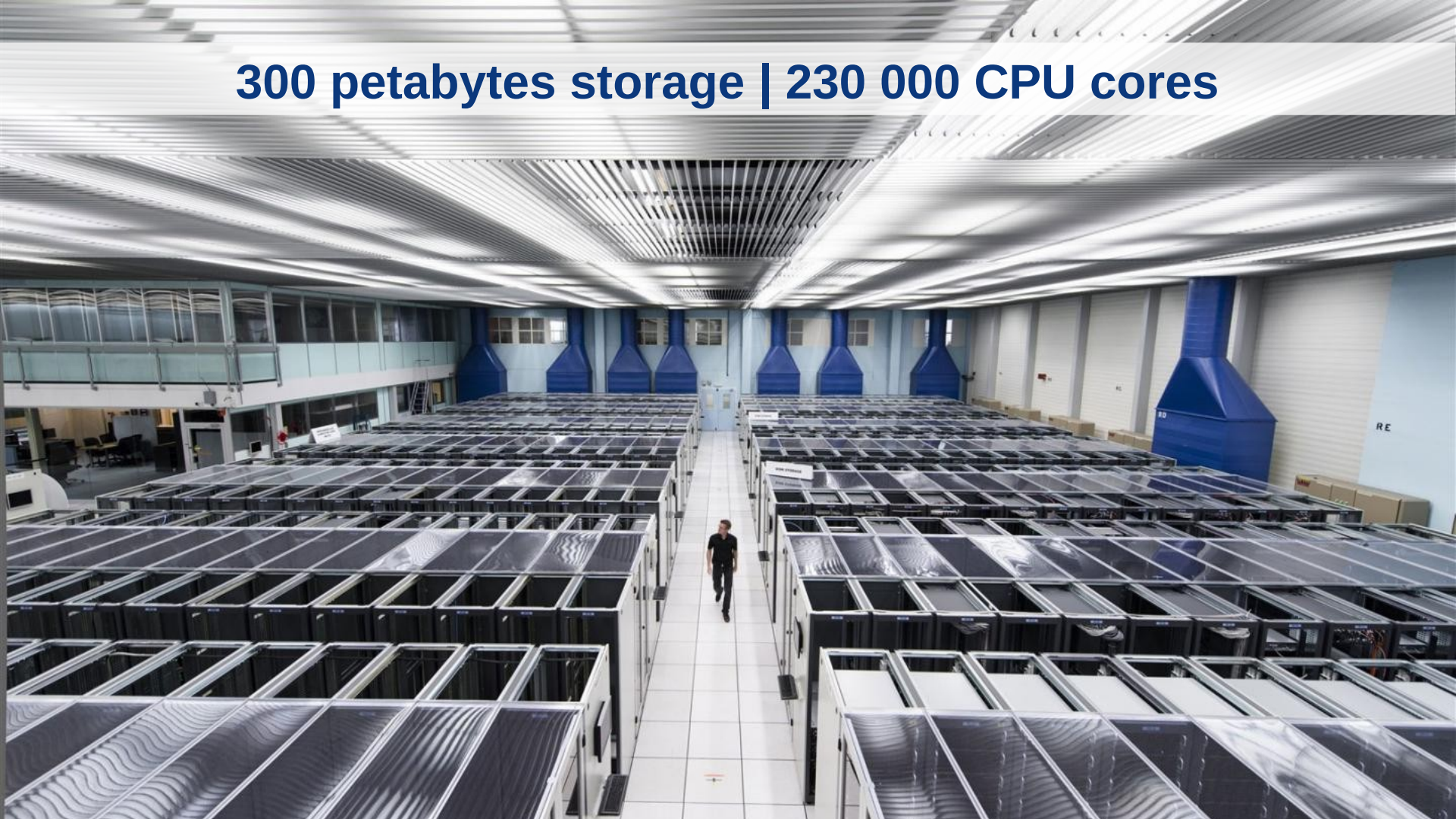
CMS
Compact Muon Solenoid]

ALICE
[A Large Ion Collider Experiment]

ATLAS
[A Toroidal LHC ApparatuS]

LHCb
[Large Hadron Collider beauty]

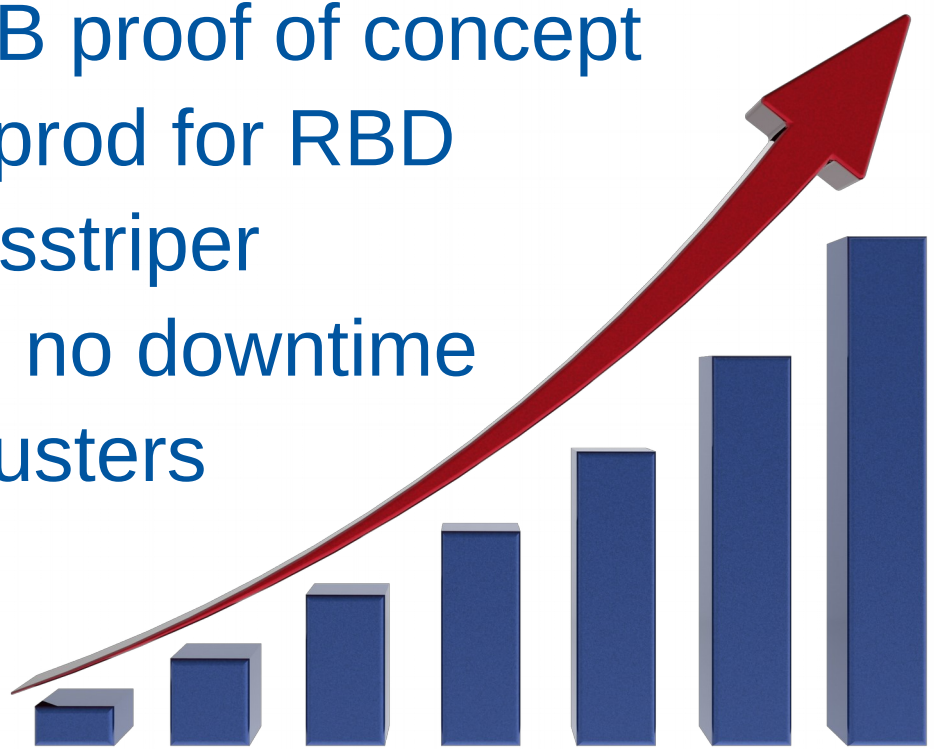
300 petabytes storage | 230 000 CPU cores



Background: Ceph at CERN

History

- March 2013: 300TB proof of concept
- Dec 2013: 3PB in prod for RBD
- 2014-15: EC, radosstriper
- 2016: 3PB to 6PB, no downtime
- 2017-18: 8 prod clusters



Ceph Clusters in CERN IT

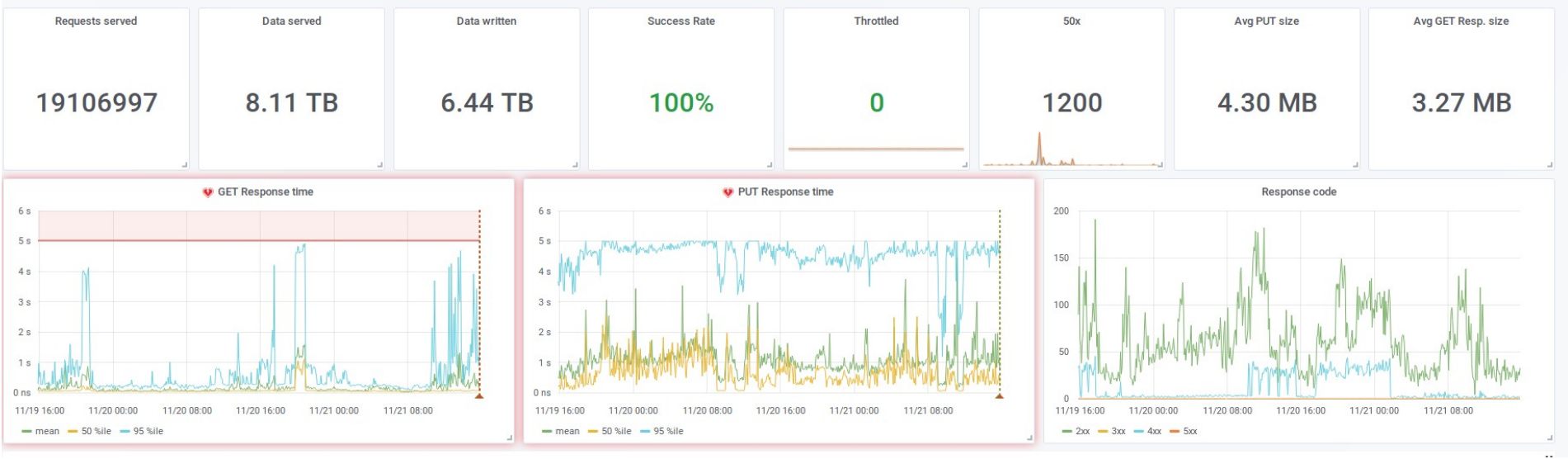
CERN Ceph Clusters		Size	Version
OpenStack Cinder/Glance	*Production*	6.2PB	luminous
	Satellite data centre (1000km away)	1.6PB	luminous
	Hyperconverged KVM+Ceph	16TB	luminous
CephFS (HPC+Manila)	*Production*	0.8PB	luminous
	Client Scale Testing	0.4PB	luminous
	Hyperconverged HPC+Ceph	0.4PB	luminous
CASTOR/XRootD	*Production*	4.9PB	luminous
	CERN Tape Archive	0.8TB	luminous
S3+SWIFT	*Production*	2.3PB	luminous

Ceph Clusters in CERN IT

- Standard Ceph node architecture
  - 16-core Xeon / 64-128GB RAM
  - 24x 6TB HDDs
  - 4x 240GB SSDs (journal/rocksdb)

RGW

S3 @ CERN


Requests served
19106997
Data served
8.11 TB
Data written
6.44 TB
Success Rate
100%
Throttled
0
50x
1200
Avg PUT size
4.30 MB
Avg GET Resp. size
3.27 MB
GET Response time
6 s
5 s
4 s
3 s
2 s
1 s
0 ns
11/19 16:00
11/20 00:00
11/20 08:00
11/20 16:00
11/21 00:00
11/21 08:00
mean
50 %ile
95 %ile
PUT Response time
6 s
5 s
4 s
3 s
2 s
1 s
0 ns
11/19 16:00
11/20 00:00
11/20 08:00
11/20 16:00
11/21 00:00
11/21 08:00
mean
50 %ile
95 %ile
Response code
200
150
100
50
0
11/19 16:00
11/20 00:00
11/20 08:00
11/20 16:00
11/21 00:00
11/21 08:00
2xx
3xx
4xx
5xx

S3 @ CERN

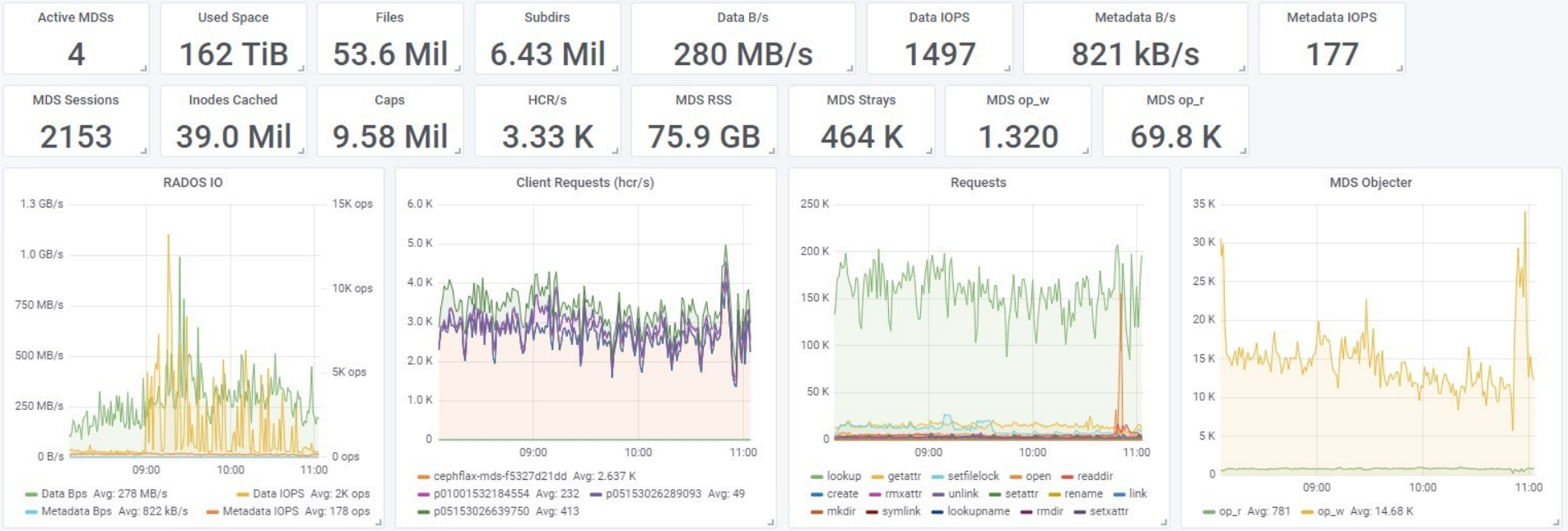
- Running as a production service since 2018
  - Use-cases: IT applications, http accessible object storage, backups, …
  - ATLAS + LHC@home using it for physics data already since 2016

- Basic architecture
  - s3.cern.ch is load-balanced across 10 VMs running Traefik/RGW
  - 4+2 erasure coding for data, 3x replication on HDDs for bucket indexes
  - Local authentication for most users, recently enabled OpenStack Keystone

- Current issues:
  - Bucket index performance is critical: moving to a 3x SSD pool
  - Automatic index re-sharding is still scary. By default, we use 32 shards per bucket
  - Multi-region: experimenting with rgw-mirror and cloudsync

CephFS

CephFS


Active MDSs
4
Used Space
162 TiB
Files
53.6 Mil
Subdirs
6.43 Mil
Data B/s
280 MB/s
Data IOPS
1497
Metadata B/s
821 kB/s
Metadata IOPS
177
MDS Sessions
2153
Inodes Cached
39.0 Mil
Caps
9.58 Mil
HCR/s
3.33 K
MDS RSS
75.9 GB
MDS Strays
464 K
MDS op_w
1.320
MDS op_r
69.8 K
RADOS IO
1.3 GB/s
1.0 GB/s
750 MB/s
500 MB/s
250 MB/s
0 B/s
15K ops
10K ops
5K ops
0 ops
09:00
10:00
11:00
Data Bps Avg: 278 MB/s
Data IOPS Avg: 2K ops
Metadata Bps Avg: 822 kB/s
Metadata IOPS Avg: 178 ops
Client Requests (hcr/s)
6.0 K
5.0 K
4.0 K
3.0 K
2.0 K
1.0 K
0
09:00
10:00
11:00
cephflax-mds-f5327d21dd Avg: 2.637 K
p01001532184554 Avg: 232
p05153026289093 Avg: 49
p05153026639750 Avg: 413
Requests
250 K
200 K
150 K
100 K
50 K
0
09:00
10:00
11:00
lookup
getattr
setfilelock
open
readdir
create
rmxattr
unlink
setattr
rename
link
mkdir
symlink
lookupname
rmdir
setxattr
MDS Objecter
35 K
30 K
25 K
20 K
15 K
10 K
5 K
0
09:00
10:00
11:00
op_r Avg: 781
op_w Avg: 14.68 K

CephFS

- HPC scratch area in production for more than 2 years
  - Extensive benchmarking w/IO-500 (see next slides)
- OpenStack Manila in production during 2018
  - Users create NFS-like shares for their general NAS use-cases
  - Authenticated using cephx to small number of clients
- Mostly using ceph-fuse for compatibility and needed features
  - Now in CentOS 7.6 the ceph kernel driver is working

HPC on CephFS?

- CERN is mostly a high *throughput* computing lab:
  - Embarrassingly parallel workloads, tolerant to relaxed consistency

- Several HPC corners exist within our lab:
  - Beam simulations, accelerator physics, plasma simulations, computation fluid dynamics, QCD, ASIC design …
  - Require full POSIX, read-after-write consistency, and parallel IO

HPC Storage with CephFS

- CERN's approach is to build HPC clusters with commodity parts:
  - Software-defined HPC rather than an expensive integrated system
  - Compute side is solved with HTCondor/SLURM
  - Typical HPC storage is not very attractive (missing expertise + budget)

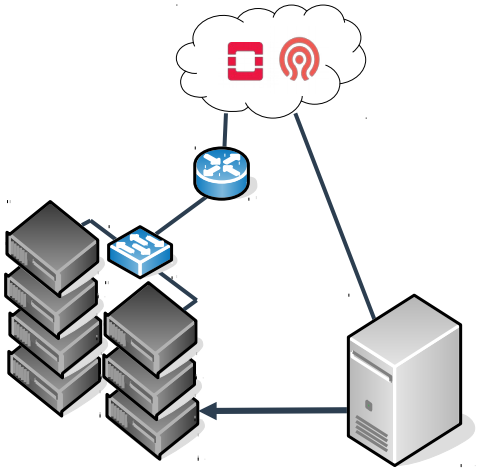
HPC Worker nodes

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

Operated since mid-2016:

~500 client nodes

~1PB CephFS (RAW)



CephFS on BlueStore

- 3x replication
- Per-host replication
- MDS as near as possible
- Hyperconvergence...

IO-500

- “The goal of the IO-500 is simple: to improve parallel file systems by ensuring that sites publish results of both "hero" and "anti-hero" runs and by sharing the tuning and configuration they applied to achieve those results.”

- **Latest** results:
  - Bandwidth: **2.83** GB/s (1.94x)
  - IOPS: **20.16** kIOPS (15.23x)
  - Score total: **7.56** (3.9x)

Pablo Llopis: HPC and CephFS at CERN, *SuperComputing’18*

19	Queen Mary; University Of London	Apocrita	E8	GPFS	10	240	zip	9.79	4.32	22.21
20	Clemson University	Palmetto	Dell	BeeGFS	48	48	zip	8.46	2.93	24.41
21	CERN	Bytecollider		CephFS	64	64	zip	7.56	2.83	20.16
22	SNL	Serrano	IBM	Spectrum	16	160		4.25*	0.65	27.98*

CephFS scale-test

- Client scale testing with k8s (10,000 cephfs clients…!)
- 2 benchmarks: idle + busy clients
- 3-node test Ceph cluster
- Goals:
  - Verify CSI CephFS implementation
  - Verify the Manila Provisioner implementation
  - Verify the Ceph MDS with 10k clients
- See Ricardo Rocha's Talk @ OpenStack Summit 2018

Wed 14, 11:50am - 12:30pm CityCube Berlin - Level 1 - Hall A2

Dynamic Storage Provisioning of Manila/CephFS Shares on Kubernetes

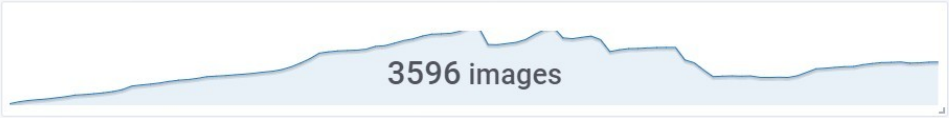
Container Infrastructure

CephFS performance tips

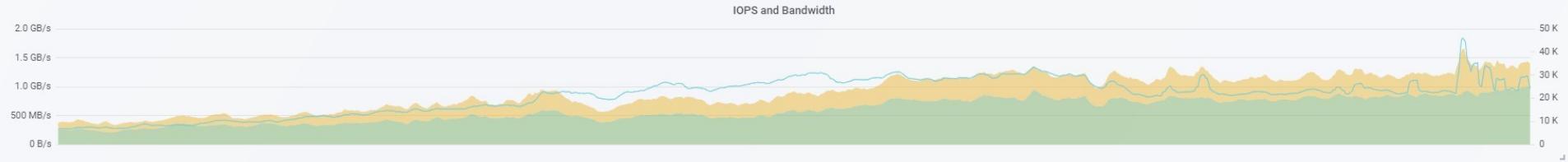
- Run several active MDSs: since v12.2.10 the md balancing is working well
- Use flash-only OSDs for the cephfs_metadata pool
- Locate the MDSs as nearby to the OSDs as possible (to reduce latency)
- Give as much RAM as possible to the MDSs – caching inodes and caps is crucial

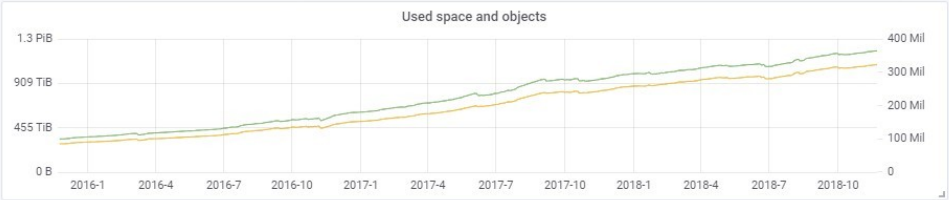
RBD

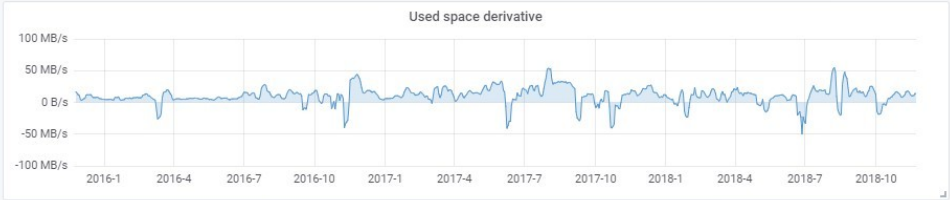
RBD @ CERN


3596 images


5895 volumes


IOPS and Bandwidth
2.0 GB/s
1.5 GB/s
1.0 GB/s
500 MB/s
0 B/s
50 K
40 K
30 K
20 K
10 K
0


Used space and objects
1.3 PiB
909 TiB
455 TiB
0 B
400 Mil
300 Mil
200 Mil
100 Mil
0
2016-1
2016-4
2016-7
2016-10
2017-1
2017-4
2017-7
2017-10
2018-1
2018-4
2018-7
2018-10


Used space derivative
100 MB/s
50 MB/s
0 B/s
-50 MB/s
-100 MB/s
2016-1
2016-4
2016-7
2016-10
2017-1
2017-4
2017-7
2017-10
2018-1
2018-4
2018-7
2018-10

RBD @ CERN

- OpenStack RBD remains the biggest use-case of Ceph at CERN
- Steady growth of ~1PB per year (raw space)
- Expanding the cluster to >6PB

- Ongoing work:
  - Just finished an expanded rbd trash feature: Delete a 50TB image in 1s
  - **Working on understanding/improving librbd performance**
    - **Thanks to CERN Openlab/Rackspace**

Benchmarking ceph/rbd

Rackspace Collaboration

Benchmarking a Ceph cluster[1]:

- Raw disk baseline performance: fio
- Ceph storage level performance: rados bench
- Block device performance: fio (librbd) and rbd bench

- Metrics: IOPS, Disk IO and CPU utilizations, latency

- Single-node cluster to avoid network latency impacts on performance

[1]https://github.com/colletj/rbdperfscripts

Rackspace Collaboration

Benchmarking a Ceph cluster:

Starting point: raw fio results give 85.1 kIOPS (SSD) and 232 IOPS (HDD), what is RADOS performance ? (4k random sync writes)

SSD: bluestore ssd

HDD: bluestore ssd

MIX: bluestore data:hdd db:ssd

FS: filestore data:hdd journal:ssd

Rackspace Collaboration

Benchmarking a Ceph cluster:

Starting point: raw fio results give 85.1 kIOPS (SSD) and 232 IOPS (HDD), what is RADOS performance ? (4k random sync writes)

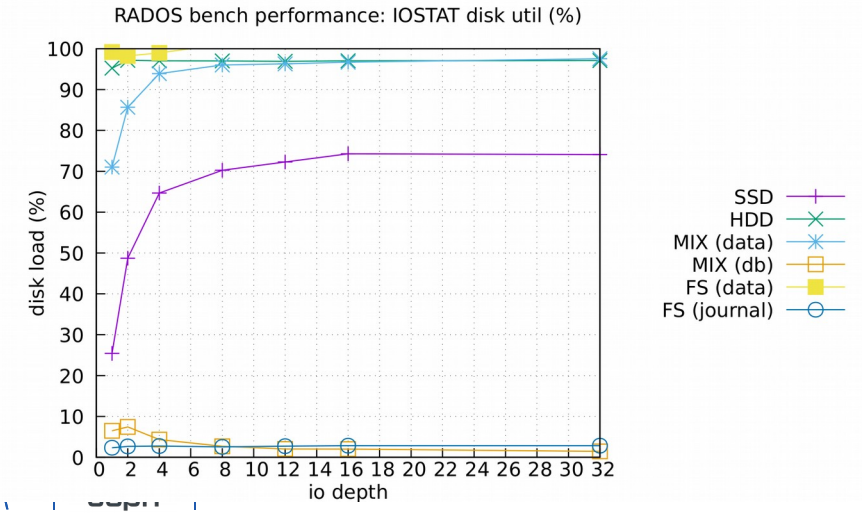
SSD: bluestore ssd

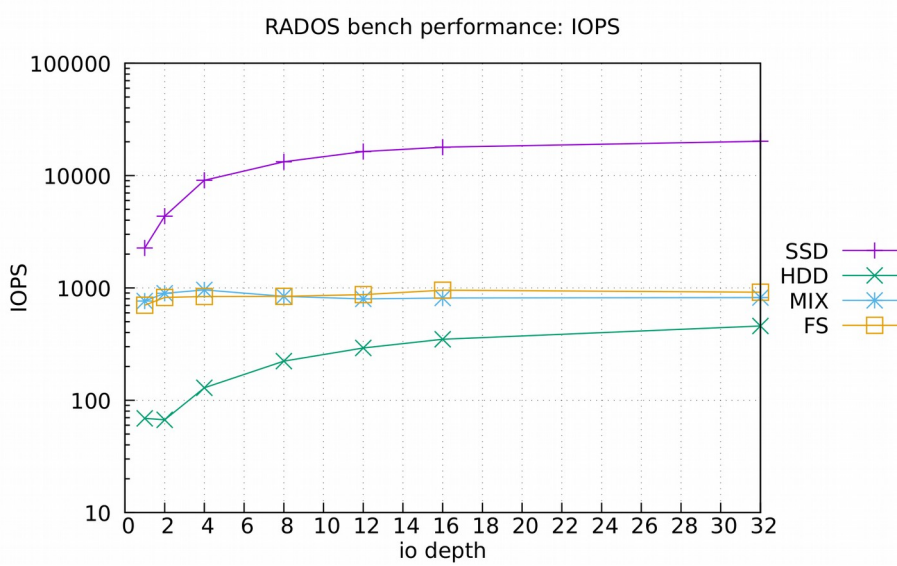
HDD: bluestore ssd

MIX: bluestore data:hdd db:ssd

FS: filestore data:hdd journal:ssd

- Mixed configurations: SSD not-stressed, HDD remaining the bottleneck with a 100% I/O util

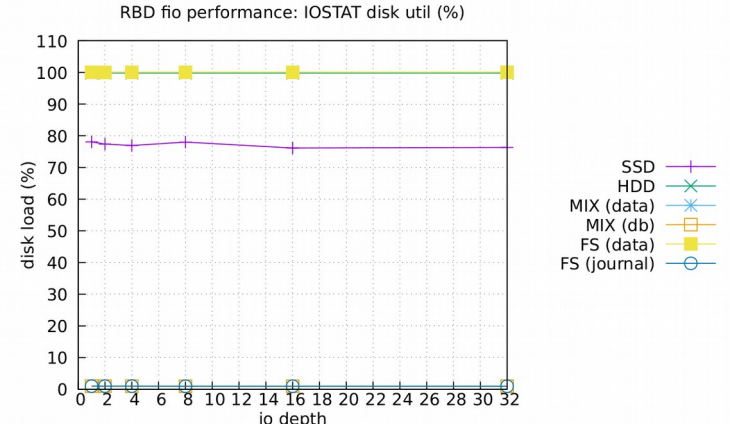




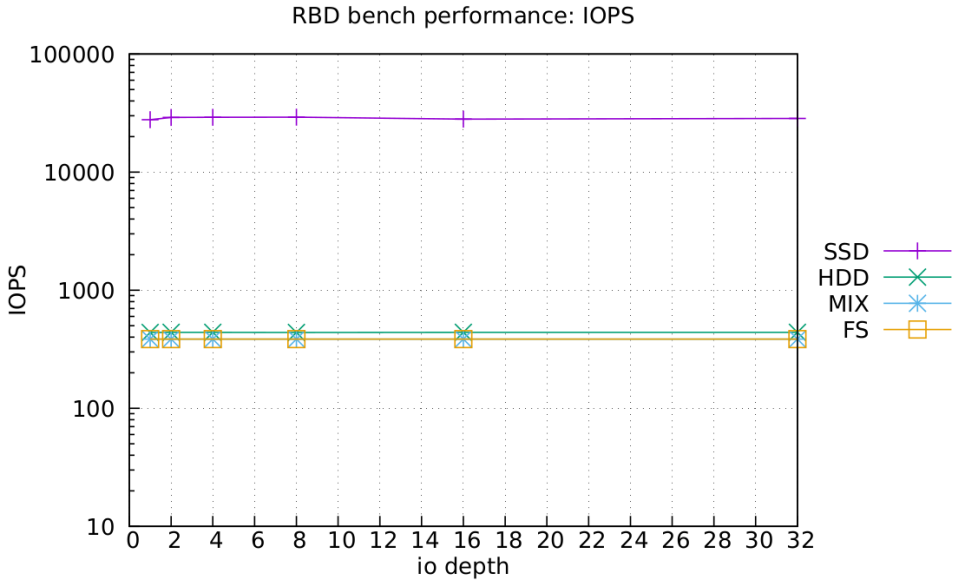
Rackspace Collaboration

Benchmarking a Ceph cluster:

Starting point: raw fio results give 85.1 kIOPS (SSD) and 232 IOPS (HDD), what is RBD performance ? (4k random sync writes)



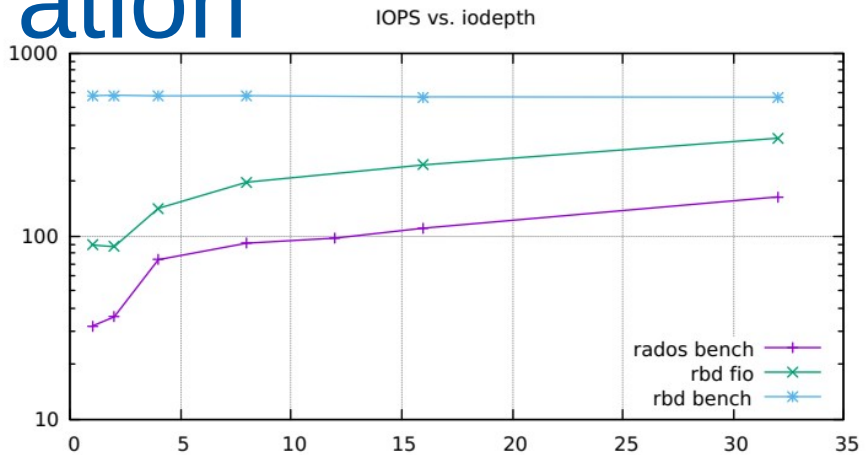
- Mixed configurations roughly equivalent, unable to stress the SSD…
- As SSDs are never I/O bound, recommended to use partitioning (multiple OSDs per SSDs)

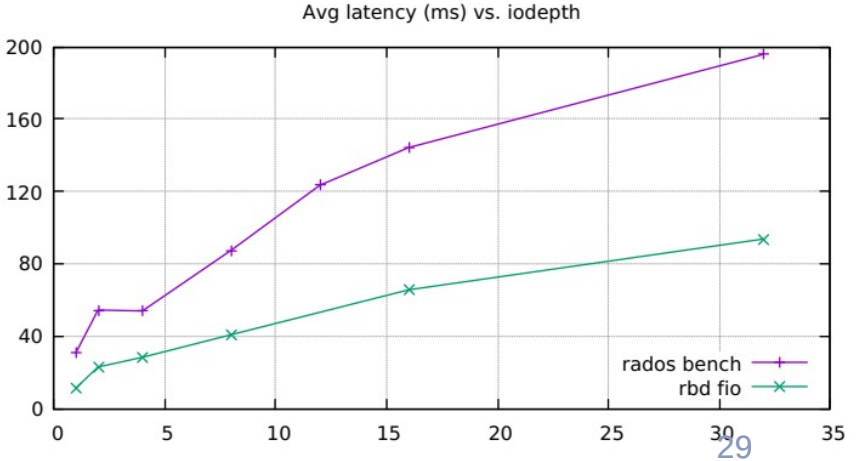


Rackspace Collaboration

OSD-side caching with dm-cache

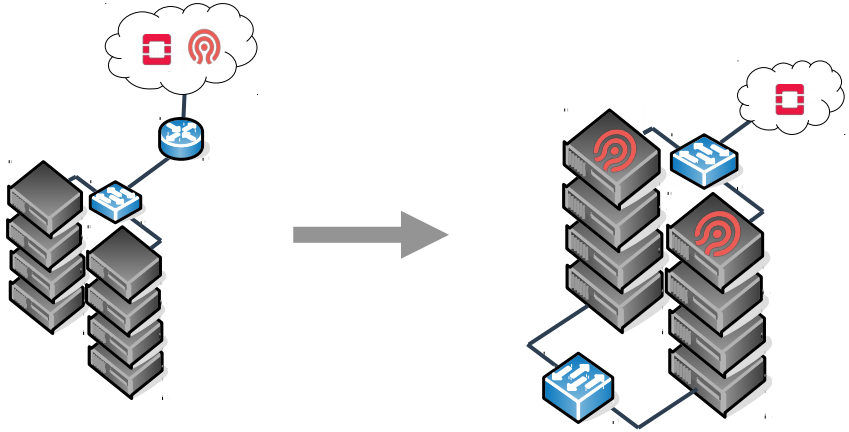
- Slightly better performance at the RBD bench level compared to BlueStore OSD but not a silver bullet either: *dm-writecache?*
- *All flash ?*





Hyperconverged OpenStack+RBD

- Cluster: 22 nodes
  - 16-core Xeon (SMT disabled)
  - 128GB of memory
  - 16x 960GB SSDs



- Configuration
  - Memory: 64GB for the VMs, 32GB for Ceph, rest for overheads
  - 14 OSDs per node

Plan

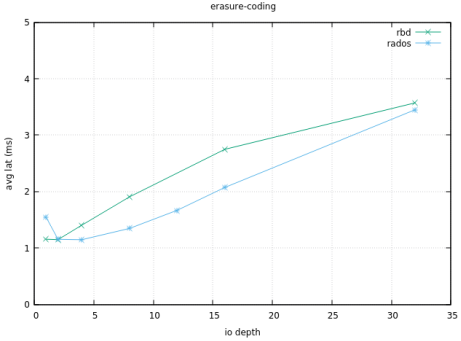
- Build it
- Internal perf tests
- Disaster tests
- Develop ops procedures
- Invite early adopters

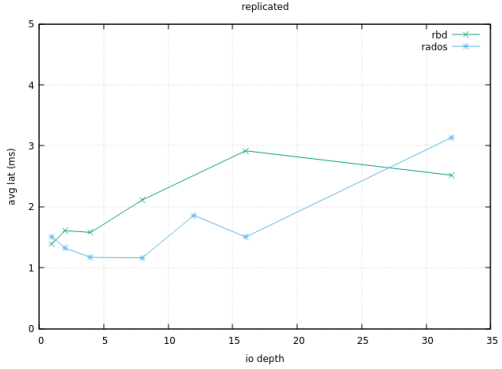
CERN | ceph | rackspace

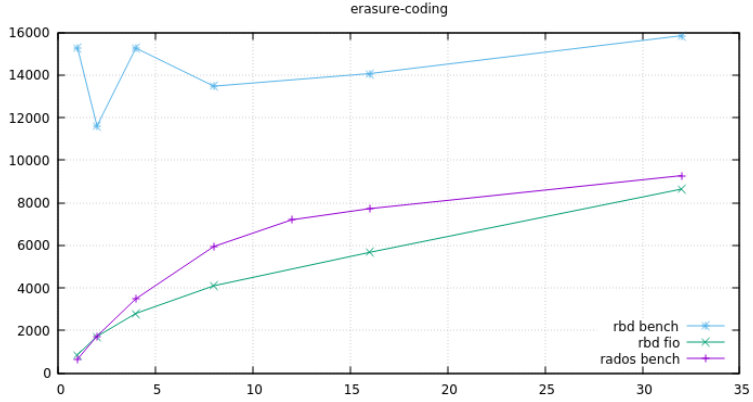
Rackspace Collaboration

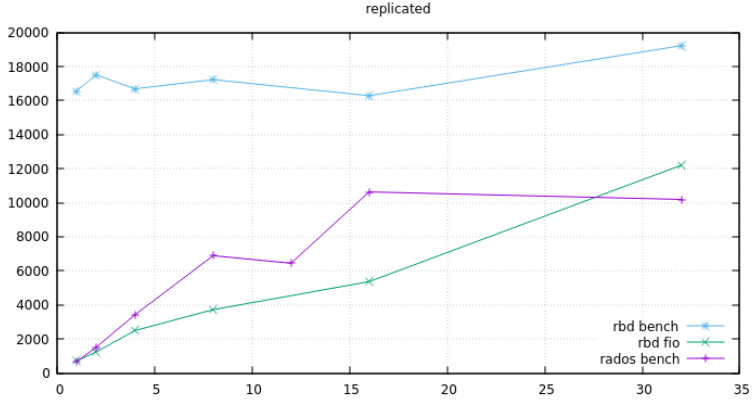
Hyperconverged platform performance

- Benchmarking a hyperconverged cluster to bring Ceph/rbd to other CERN use-cases
- Outperforms by far OSD-side caching alternatives









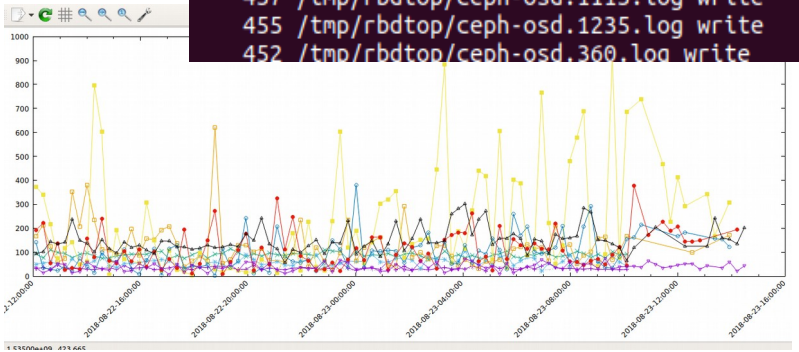
Monitoring ceph/rbd performance

Ceph/rbd top v1

Identifying busiest OSDs/RBD Images

- List all OSDs on the host
- Activate logging for the OSDs for a given time
- Extract read/write
- Sort/Filter by most active
- Generate a report/plot

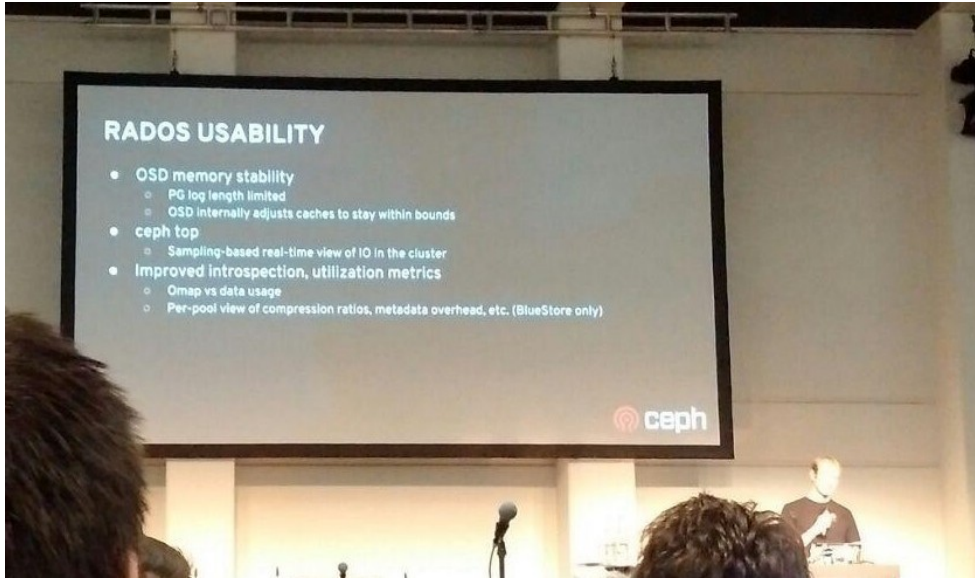
```
[2018-08-22 12:12:13/rbdtop] Logs collected, parsing
[2018-08-22 12:12:13/rbdtop] logfile is /tmp/rbdtop/ceph-osd.[0-9]*.log
[2018-08-22 12:12:13/rbdtop] OSD operation summary (1117 active images):
    807 /tmp/rbdtop/ceph-osd.357.log write
    640 /tmp/rbdtop/ceph-osd.1238.log write
    586 /tmp/rbdtop/ceph-osd.1246.log write
    492 /tmp/rbdtop/ceph-osd.1243.log write
    492 /tmp/rbdtop/ceph-osd.361.log write
    457 /tmp/rbdtop/ceph-osd.1113.log write
    455 /tmp/rbdtop/ceph-osd.1235.log write
    452 /tmp/rbdtop/ceph-osd.360.log write
```



Ceph/rbd top v2

Integrating top as a Ceph feature (work-in-progress)

- Feature announced at Ceph Day Berlin on 13.11.2018
- Will help operators to identify “hot” clients/images
- Still work-in-progress



Ceph/rbd top v2

Integrating top as a Ceph feature

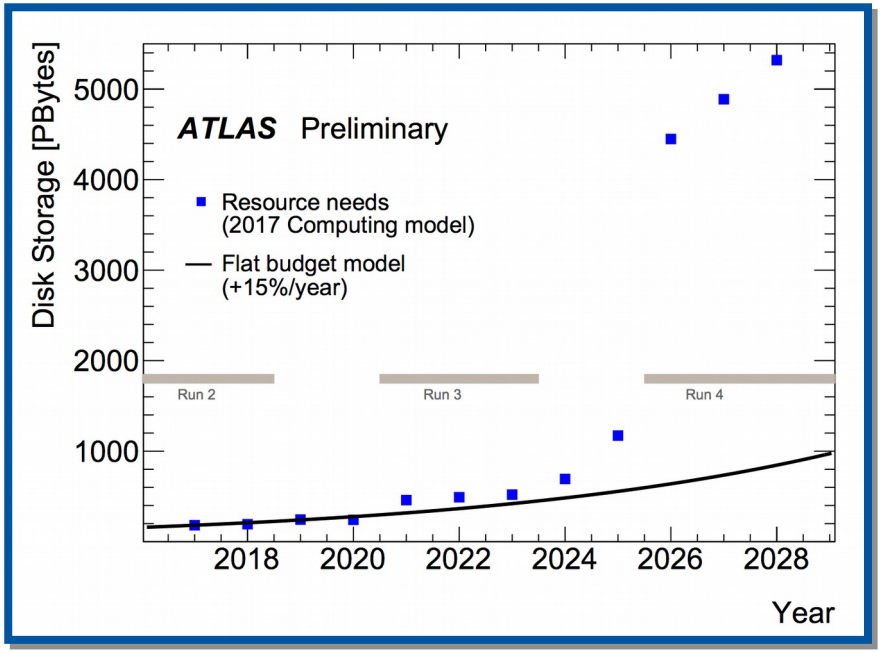
- Mgr module issuing requests to OSDs to collect perf metrics
- Python interface to add/remove requests and get query results
- Group by object prefix (rbd image name)

```
maha:~/ceph/ceph/build% ceph mgr module enable osd_perf_query
maha:~/ceph/ceph/build% ceph osd perf query add client_id
added query client_id with id 0
0
maha:~/ceph/ceph/build% ceph osd perf query add rbd_image_id
added query rbd_image_id with id 1
1
maha:~/ceph/ceph/build% for i in 1 2 3; do rbd bench --io-type write --rbd-cache=false --io-s
bench  type write io_size 4096 io_threads 16 bytes 409600 pattern random
  SEC       OPS   OPS/SEC   BYTES/SEC
elapsed:     0  ops:      100  ops/sec:   499.99  bytes/sec: 2047978.70
bench  type write io_size 4096 io_threads 16 bytes 409600 pattern random
  SEC       OPS   OPS/SEC   BYTES/SEC
elapsed:     0  ops:      100  ops/sec:   543.47  bytes/sec: 2226063.81
bench  type write io_size 4096 io_threads 16 bytes 409600 pattern random
  SEC       OPS   OPS/SEC   BYTES/SEC
elapsed:     0  ops:      100  ops/sec:   595.23  bytes/sec: 2438069.87
maha:~/ceph/ceph/build% ceph osd perf counters get 0
counters for query with id 0
+---------------+-----------+----------+-------------+------------+---------------+---------
|   client_id   | write_ops | read_ops | write_bytes | read_bytes | write_latency  | read_lat
+---------------+-----------+----------+-------------+------------+---------------+---------
| client.164136 |    107    |    24    |  409600/107 |   366/24   | 2618503617/107 | 11166778
| client.164140 |    107    |    24    |  409600/107 |   366/24   | 2833574010/107 | 14450420
| client.164159 |    107    |    24    |  409600/107 |   366/24   | 2357477064/107 | 11717881
+---------------+-----------+----------+-------------+------------+---------------+---------
maha:~/ceph/ceph/build% ceph osd perf counters get 1
counters for query with id 1
+---------+---------------+-----------+----------+-------------+------------+----------------
| pool_id |  rbd image_id | write_ops | read_ops | write_bytes | read_bytes | write_latency
+---------+---------------+-----------+----------+-------------+------------+----------------
|    3    | 1e6157e263d9e |    100    |    0     |  409600/100 |    0/0     | 2548654136/100
|    3    | 1e64961688492 |    100    |    0     |  409600/100 |    0/0     | 2742848291/100
|    3    | 1e6526e037e21 |    100    |    0     |  409600/100 |    0/0     | 2289681719/100
+---------+---------------+-----------+----------+-------------+------------+----------------
```

Future…

HEP Computing for the 2020s

- Run-2 (2015-18):
  ~50-80PB/year
- Run-3 (2020-23): ~150PB/year
- Run-4: ~600PB/year?!



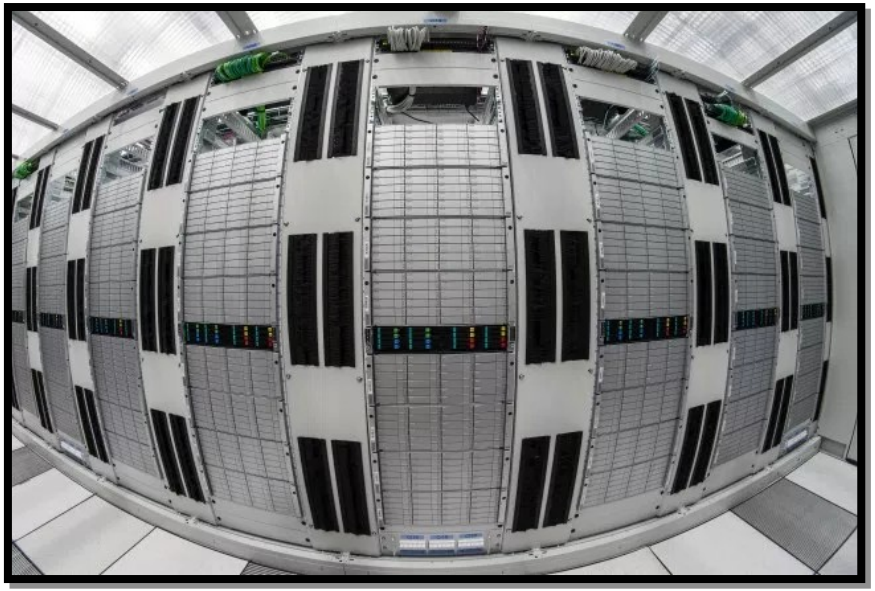
https://arxiv.org/abs/1712.06982

CERN

ceph

Scale Testing

- *Bigbang* scale tests mutually benefitting CERN & Ceph project
- *Bigbang I:* 30PB, 7200 OSDs, Ceph hammer. Found several *osdmap* limitations
- *Bigbang II:* Similar size, Ceph jewel. Scalability limited by OSD-MON traffic. Led to dev of *ceph-mgr*.
- *Bigbang III:* 65PB, 10800 OSDs.



https://ceph.com/community/new-luminous-scalability/

Thanks!

Merci!


CERN