CERN

Building Scale-Out Storage Infrastructures with RADOS and Ceph

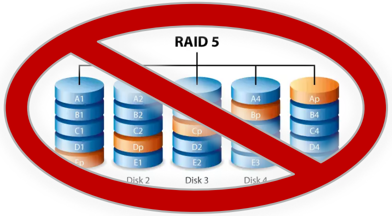
Dan van der Ster, CERN IT Storage Group

XLDB 2017: 10th Extremely Large Databases Conference

11 October 2017 | Clermont-Ferrand, France

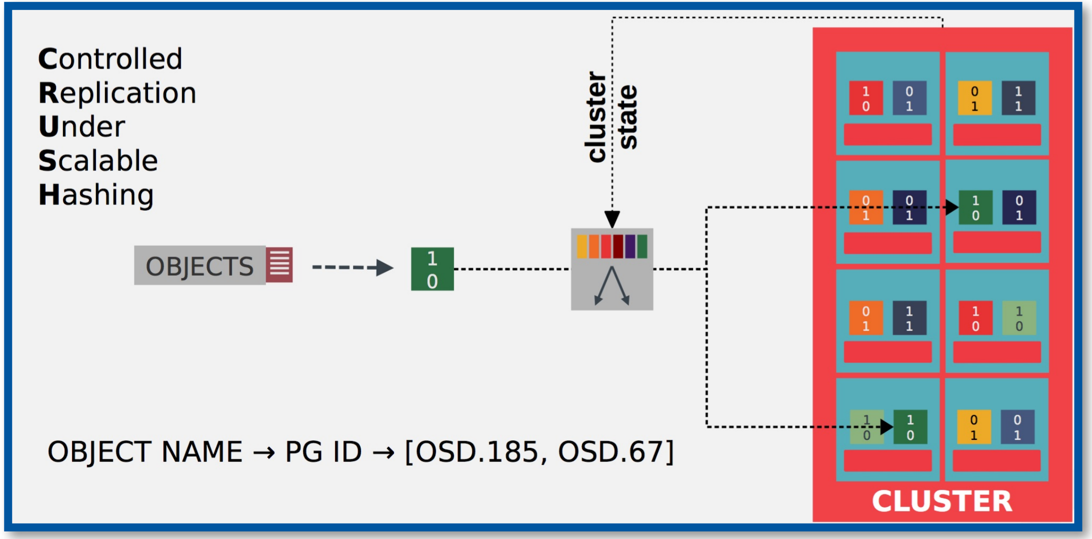
Modern, Software-Defined Storage

- **Reliable**: HA by moving IPs around is old fashioned – modern HA is built into the software. No SPOF, No special servers, and No RAID!
- **Infrastructure-aware**: naïve replication is not enough – need to place data across failure domains
- **Scale-out**: add or replace capacity/IOPS as needed without downtime
- **Low-cost**: use commodity hardware, spend money only where it matters
- **Flexible**: do you want high IOPS, low latency? Do you want cheap erasure-coded pools?
- **Object storage vs Filesystems**: need to support both modern and legacy applications



Why Ceph?

Object Storage with CRUSH



OBJECT NAME → PG ID → [OSD.185, OSD.67]

No namespace: objects placed according to storage topology, known by clients and servers

Fast: microseconds, even for very large clusters

Stable: minimal data movement when topology changes

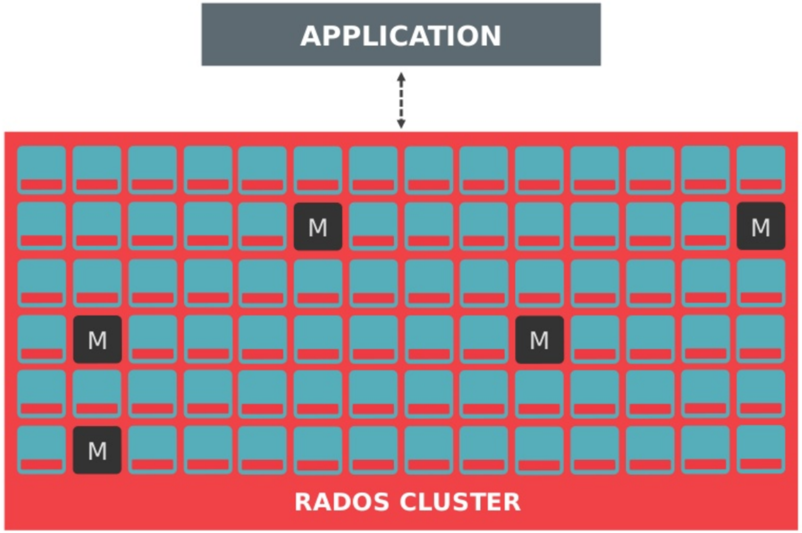
Reliable: object placement constrained by failure domains

Flexible: replication, erasure codes, complex placement schemes

RADOS

OSDs: 10s to 1000s in a cluster, Autonomous peering for IO and recovery

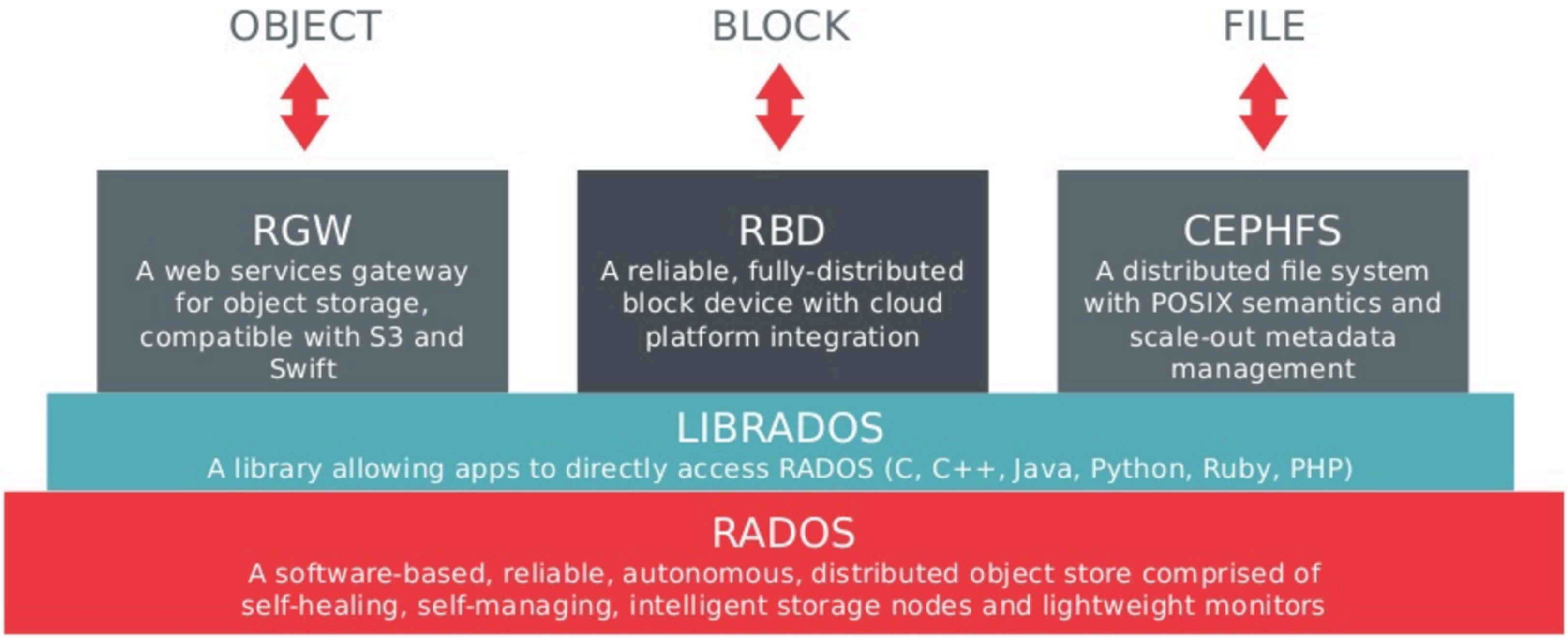
Mons: Quorum of k/v DBs that track the cluster state: *where are the OSDs? which CRUSH rules exist? which pools exist?…*



* RADOS makes bit/disk/host/network/… failures ~invisible, and enables organic evolution of the underlying hardware (growing/shrinking/replacement/…)

* *CRUSH is often cited as the key feature of Ceph – but RADOS makes it work in real life*

Ceph Open Source Storage



OBJECT

BLOCK

FILE

RGW
A web services gateway for object storage, compatible with S3 and Swift

RBD
A reliable, fully-distributed block device with cloud platform integration

CEPHFS
A distributed file system with POSIX semantics and scale-out metadata management

LIBRADOS
A library allowing apps to directly access RADOS (C, C++, Java, Python, Ruby, PHP)

RADOS
A software-based, reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes and lightweight monitors

Using Ceph

Using Ceph: RADOS

- Most users start with the `rados` CLI:
  - `get/put` objects in a pool, or run simple performance benchmarks
  - Useful for testing, not very useful for building an application
  - You *can* list a pool contents, but you *shouldn't*! RADOS is not indexed!!
- `librados` API:
  - rich api for read/write/modify, locking, watching, also a k/v store for each object
  - Bindings for most common languages. Good for writing your app!
- `libradosstriper` API:
  - rados deals with entire objects and the *best practise is to keep objects under ~10MB*.
  - `libradostriper` breaks single "objects" into several pieces for streaming to Ceph

- RADOS security is handled by *CephX* shared secrets granting "capabilities" on pools.
  - E.g. read/write, read-only, restrict to a an object prefix for multi-tenancy.

Using Ceph: RBD

- **R**ADOS **B**lock **D**evices
  - Virtual network block device that can be attached to a server remotely and used like a disk.
  - Thinly provisioned, resizable, snapshots, layering

- `librbd` for hypervisors such as `qemu-kvm`
- `krbd`, an rbd client built into Linux kernel

- Ceph RBD is the most commonly deployed OpenStack storage:
  - Glance image repository: allows to boot from network
  - Cinder volume service: attach extra storage devices to a running VM

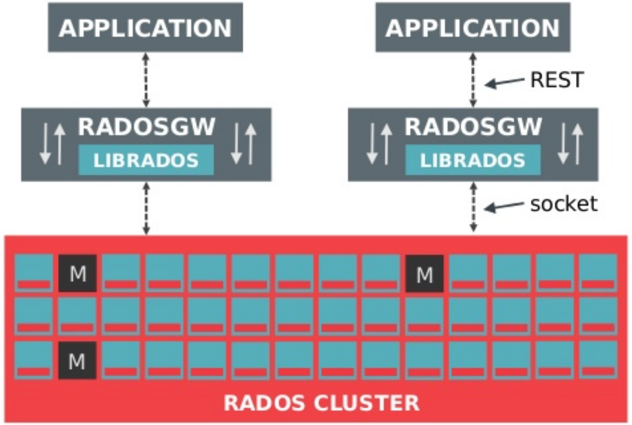
- `rbd-mirror`: asynchronously mirror a block device to a separate Ceph cluster for disaster recovery

Using Ceph: RGW

- **R**ADOS **G**ate**W**ay emulates S3/SWIFT APIs for Amazon-like object storage
  - Easily integrate with existing S3-compatible apps.
  - Enables cool things like presigned-URLs – securely grant time-limited access to objects/buckets.

- `rgw` daemons run on separate gateway nodes, translating S3/SWIFT into RADOS calls
  - Large S3 objects are broken into small RADOS objects
  - Security is handled by S3/SWIFT – users don't need cephx keys!

- S3 buckets are indexed, those indexes can grow!
  - rgw shards them once they grown above 100000 objects.
  - Multiple buckets are cheap – use several if you can!



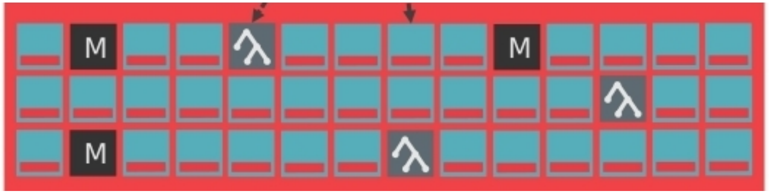
*New: try **librgw** to integrate rgw with your applications*

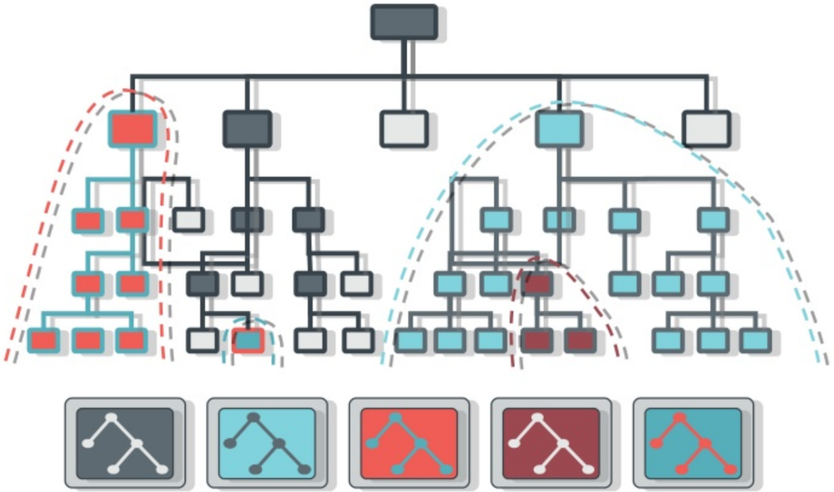
Using Ceph: CephFS

- **Ceph** **F**ile**S**ystem delivers full POSIX on top of RADOS
  - Kernel client: `mount -t ceph /cephfs`
  - FUSE client: `ceph-fuse /cephfs`
- **MDS** daemons handle the CephFS metadata
  - Several active daemons, hot/cold standbys

- CephFS Features:
  - POSIX user/group permissions & ACLs
  - Quotas, snapshots, configurable placement/striping layouts
  - Recursive statistics, recursive ctime

- **Multi-active metadata servers:**
  - MDS's dynamically rebalance the metadata
    - hot trees split to several MDSs, cold trees merged
  - Even single directories can be split across MDS's

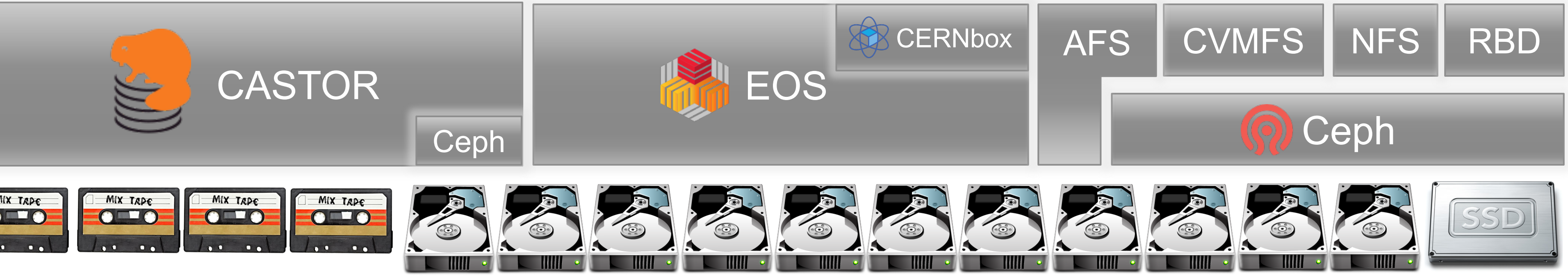




Ceph @ CERN Ops Experience

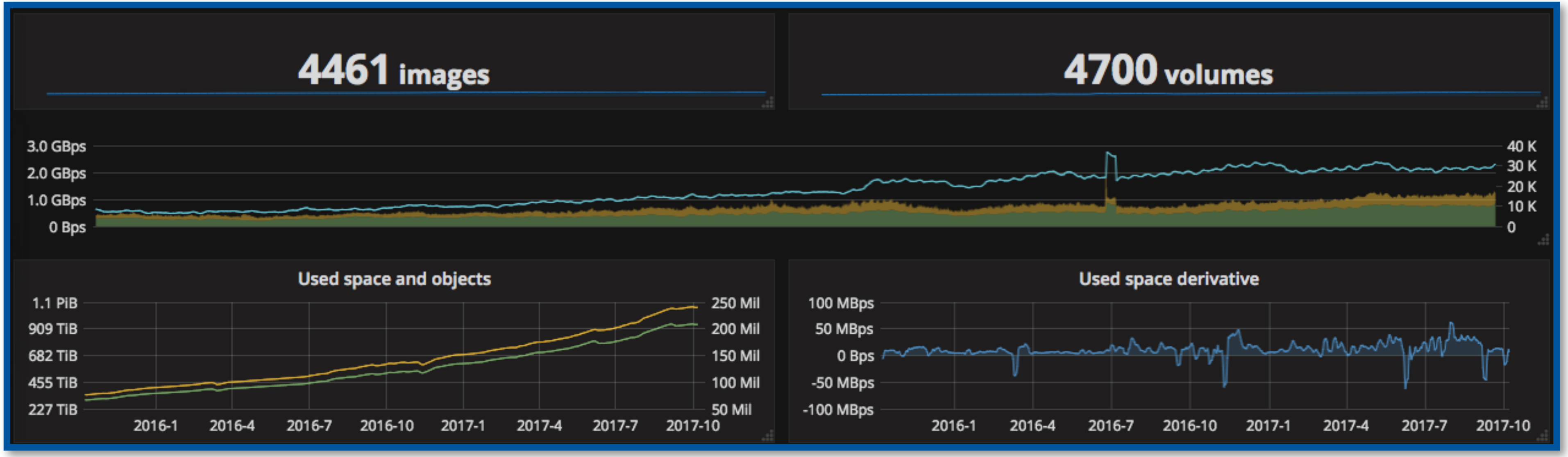
Storage for Particle Physics and CERN

- Huge data requirements (>200PB now, +50PB per year)
- Worldwide LHC Grid *standards* for accessing and moving data
  - GridFTP, Xrootd to access data, FTS to move data, SRM to manage data



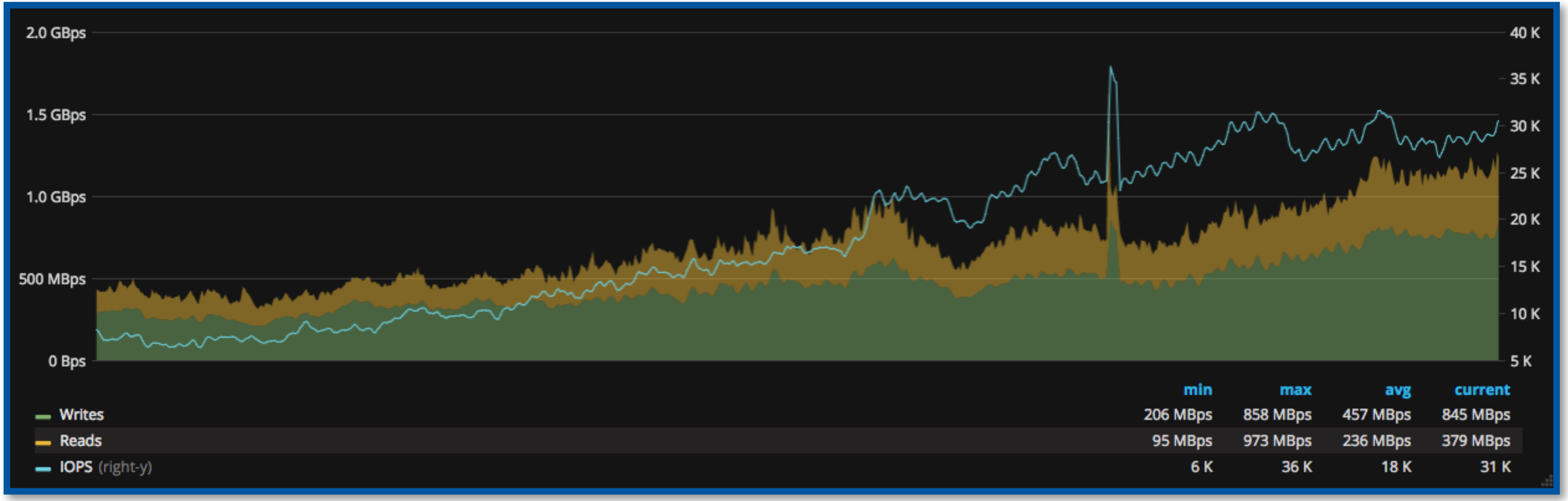
- Not just physics: we also operate a pretty standard IT infrastructure – largely based around OpenStack – for our ~10000 users.
- Ceph plays a large role for the cloud infrastructure, and a growing role for physics.

OpenStack Glance + Cinder



- OpenStack is Ceph’s killer app, usage grew by 4x in 2 years.
- Very stable, very few incidents in 3 years operations.
  - Zero issues related to data durability or corruption.

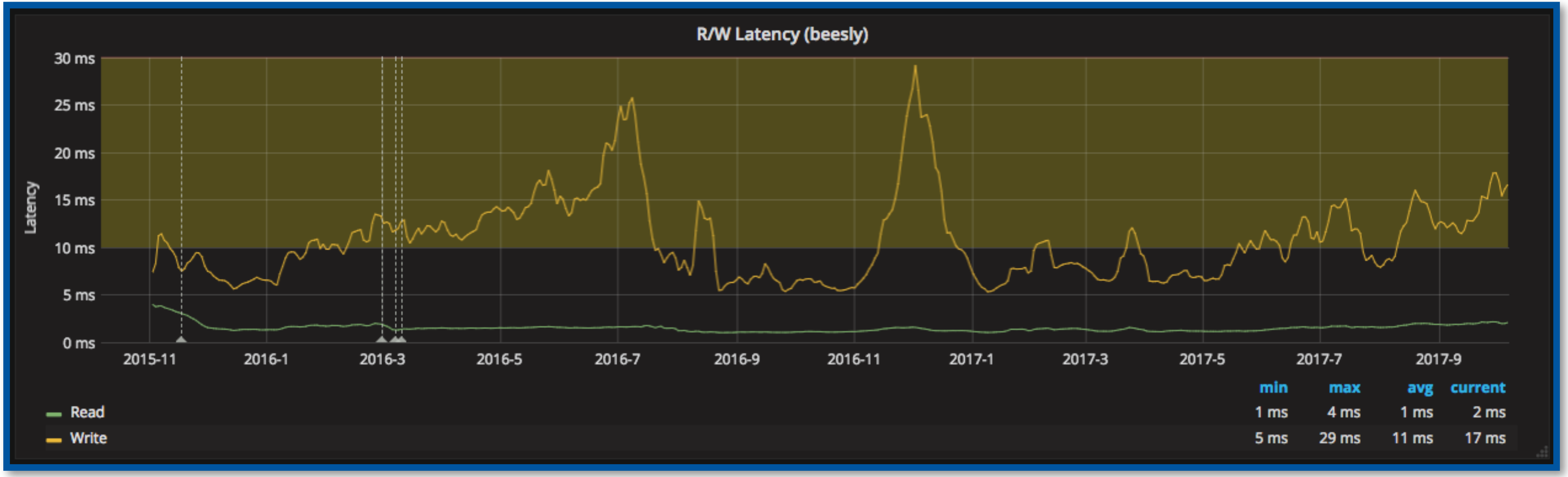
OpenStack Glance + Cinder



	min	max	avg	current
🟩 Writes	206 MBps	858 MBps	457 MBps	845 MBps
🟨 Reads	95 MBps	973 MBps	236 MBps	379 MBps
🟦 IOPS (right-y)	6 K	36 K	18 K	31 K

- From ~300MBps to ~1.2GBps block IO and from ~6000 to ~31000 IOPS.

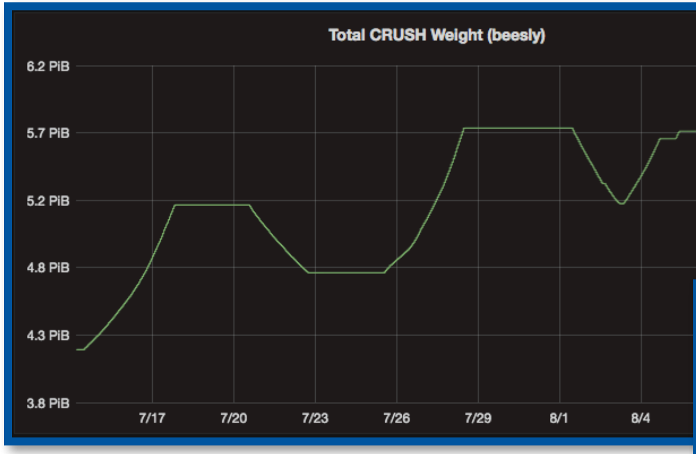
OpenStack Glance + Cinder



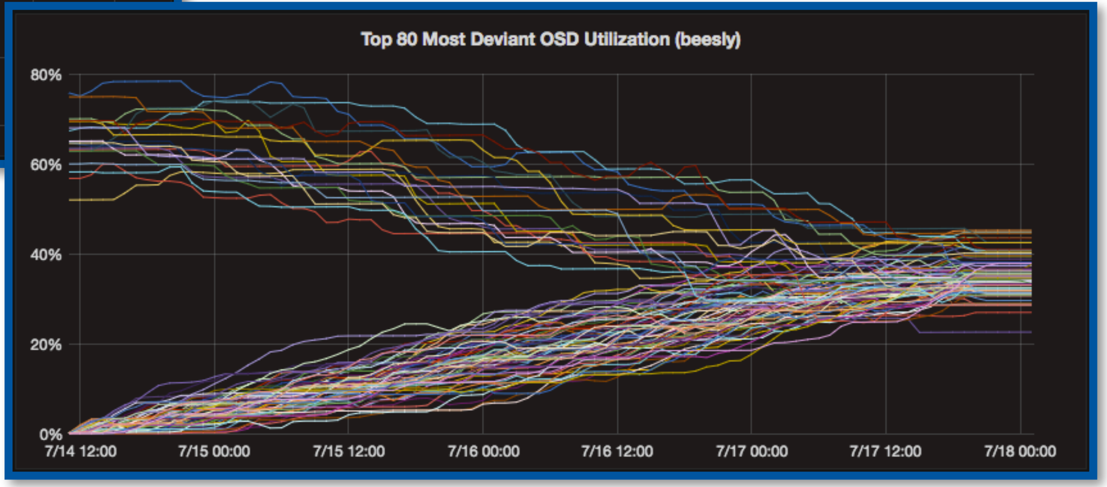
	min	max	avg	current
🟩 Read	1 ms	4 ms	1 ms	2 ms
🟨 Write	5 ms	29 ms	11 ms	17 ms

- Goal latency is <10ms for a 4kB write.
- We maintain the latency through new hardware, tuning and software improvements.

OpenStack Hardware Replacement



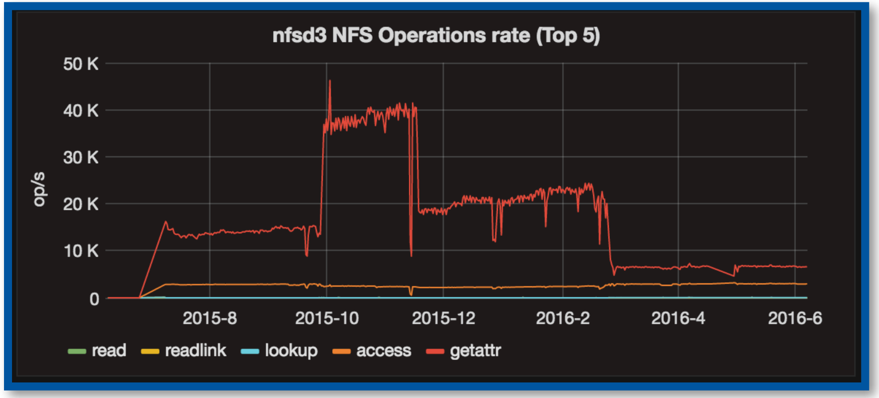
Fully replaced 3PB of block storage with 6PB new hardware over several weeks, transparent to users.



NFS on RBD

- ~60TB across 28 servers:
- OpenStack VM + RBD
- CentOS 7 with ZFS for DR

- *Not highly-available, but…*
- cheap, thinly provisioned, resizable, trivial to add new filers



Example: ~25 puppet masters reading node configurations at up to 40kHz

puppet labs

GitLab

Platform Computing an IBM Company

BOINC

OPENSHIFT

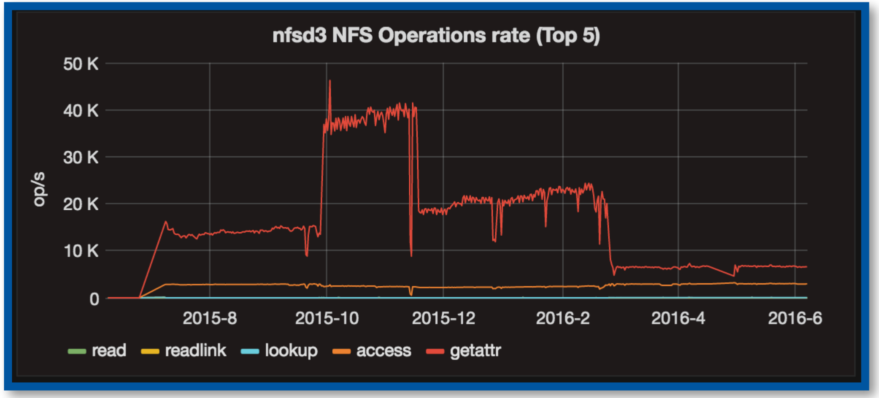
collaborate with TWiki®

elastic

NFS on RBD

- ~60TB across 28 servers:
- OpenStack VM + RBD
- CentOS 7 with ZFS for DR

- *Not highly-available, but…*
- cheap, thinly provisioned, resizable, trivial to add new filers



Example: ~25 puppet masters reading node configurations at up to 40kHz

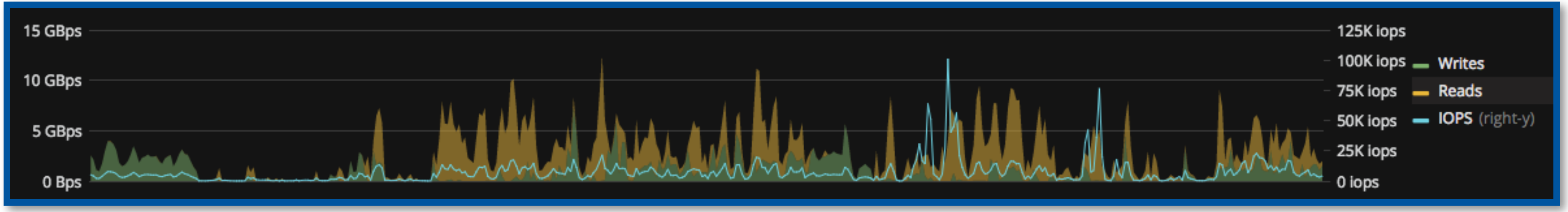
Migration to CephFS ongoing!

CephFS for HPC

- CERN is mostly a high *throughput* computing lab:
  - Embarrassingly parallel workloads, quite tolerant to relaxed consistency.
- Several HPC corners exist within our lab:
  - Beam simulations, accelerator physics, plasma simualtions, computation fluid dynamics, QCD …
  - Require full POSIX, read-after-write consistency, and parallel IO
- ~100 HPC nodes accessing ~1PB of CephFS since mid-2016:
  - Few bugs found, quite stable, but for perf++, extent locking and/or O_LAZY needs some dev attention.
- With our NFS→CephFS project + HPC on CephFS, we'll be getting more practical experience during 2018.

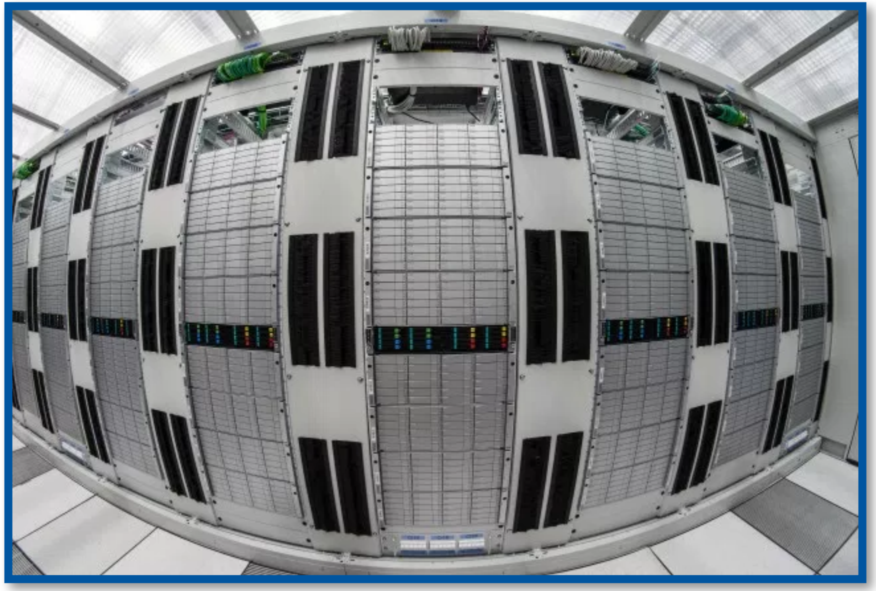
Ceph for Physics Data

- CASTOR: CERN Tape Archive System
  - Files sent to disk, then CASTOR pushes those to tape.
- 2PB disk buffer now implemented in RADOS.
- Contributed `libradosstriper` to Ceph
  - fast parallel streaming, and to keep object sizes small.



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/

Ceph Deployment Tips

One or many clusters?

- *Can I host all of my applications in one single Ceph cluster?*
- **Yes!** RGW + RBD + CephFS, all in one cluster…indeed this is technically possible
  - We can use separate pools for each use-case
- **But no!**
  - Quality-of-service concerns:
    - Ceph does not (yet) offer pool-level QoS – intensive applications can drown out the others
    - Latency vs Throughput: RBD is latency-sensitive – you probably don't want mix RBD hardware with your high throughput Big Science RADOS disks?
  - Client compatibility impracticalities:
    - RBD clients (VMs) have very long uptimes. This can lead to upgrade inconvenience, if you want to enable new incompatible Ceph features. You can upgrade, but not enabled new features!

NVMe, SSD, HDD

Double write penalty?
1 write, at least 6 seeks

- ***Where do I need flash? Where did all my IOPS go?***
- The story of one 4kB write (Ceph v10 with XFS FileStore):
  - Client calculates the 3 replica placement [4,1,3], then sends the 4kB object to osd.4 across the network
  - osd.4 writes and flushes to a journal device or file; osd.4 also writes buffered to an XFS filesystem
  - osd.4 dispatches the 4kB write to osd.1 and osd.3; osd.1 and osd.3 do like osd.4 above
  - Client sees the write acknowledge after all three replicas have the 4kB written and flushed (to the journals!).
- Ceph OSD Filestore journal: write ahead log, easily accelerated by flash
- Ceph v12 includes a new OSD implementation – BlueStore – that improves several of these double-write concerns.
  - RocksDB and it's write-ahead-log can profit from flash.
- Ceph v12 let's you easily build SSD/HDD pools, with CRUSH rules based on device types
- Ceph has a native cache tiering feature: my advice is to avoid this.

Two replicas

- ***I can't afford 3 replicas. Can I get away with 2x?***

- Consider the following:
  - OSDs *A* & *B* share a placement group. We allow writes when at least one is up.
  1. *A* up; *B* is down: *A* accepts some writes.
  2. *B* is restored: *B* starts replaying the writes he missed.
  3. While *B* is recovering, *A* goes down.
  - At this point, the placement group becomes *inactive*, objects are *unfound*, IO stops.

- Be safe: use 3x replicas, require min up OSDs = 2.

- Erasure coding lets us save money without losing durability!

Erasure Coding

- *Erasure coding looks great, can I save loads of money by using a 25+2 profile?*

- Things to consider:
  - EC splits objects into smaller pieces, amplifying IOPS
  - Long tail of latency: clients have to wait for the slowest OSD
  - Updating objects is expensive (full rewrite of the object)
  - Might be CPU intensive if you’re doing high throughput.

- In practise, start with k=4,m=2. Maybe 8+3.

The Ceph Open Source Project

Community

- The Ceph open source project has a large and growing user/dev community
  - 71 organisations contributed to *luminous*
  - 271 individual committers

- Much more on ceph.com:
  - ceph-users, ceph-devel mailing lists
  - Ceph Days events scheduled globally
  - Tech talks online to learn about Ceph,
  - Developer monthly meeting to propose a project

RED HAT
SUSE
X-ION GMBH
RELIANCE JIO INFOCOMM LTD.
CANONICAL
ISTUARY INNOVATION GROUP
EISOO
ZTE
DIGITALOCEAN
EAYUN
ALIBABA
VIVA64
XSKY
WALMART LABS
MIRANTIS
QUANTUM CORPORATION
NEUNN
FLIPKART
UNAFFILIATED
SPECTRA LOGIC
BUDGETENERGIE
TRENDY TECH
WANDA FFAN
T2CLOUD
INSPUR
MELLANOX
DIGIWARE
CLOUDIN
WIDO 420N
UMCLOUD
CHINA MOBILE
LINARO
LIYAN
LETV
GODADDY
GENTOO
H3C
INKTANK
SALESFORCE
INTEL
DEUTSCHE TELEKOM
ORACLE
TERA
IBM
DREAMHOST
CLOUDWATT
UNITEDSTACK
CERN
KYLIN CLOUD
VIRTUOZZO
SYNESIS
IGALIA
EASYSTACK
TENCENT
BIGTERA
OVH
XTAO CO. LTD.
SANDISK
ESPOL
EBAY
FAU
CATALYSTS
99CLOUD INC
SK TELECOM

Governance

cisco
FUJITSU
CERN
redhat
42 on.com
intel
SUSE
CANONICAL
SanDisk

- The Ceph Open Source project is governed by a group of individuals and organizations that are making large commitments and long-term strategic bets on Ceph. Announced in October 2015, this initiative serves to increase contributions and streamline participation through the leadership, mentoring, and assistance of our board members.

Summary

Summary

- Ceph has many APIs, so you need to plan your applications carefully
  - librados vs. block storage vs. S3 vs. CephFS
- CERN is operating Ceph at scale
  - OpenStack + CASTOR + CephFS/HPC + S3
- Ceph is reliable and scalable, but you need to plan your deployments carefully
  - Single vs. multi-tenant clusters? Flash vs. HDDs? Replication vs. Erasure coding

CERN