CHEP 2024

Ceph at CERN in the multi data centre era

October 24th, 2024

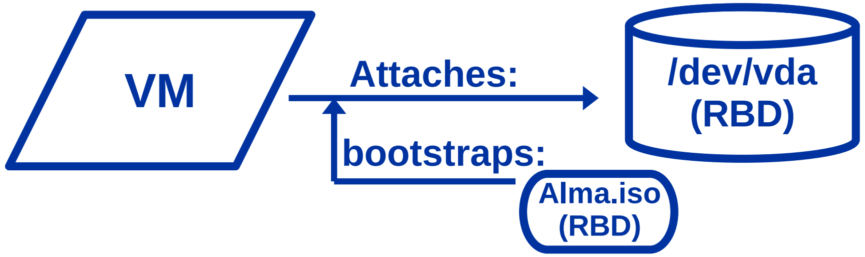
Zachary Goggin

Ceph at CERN:

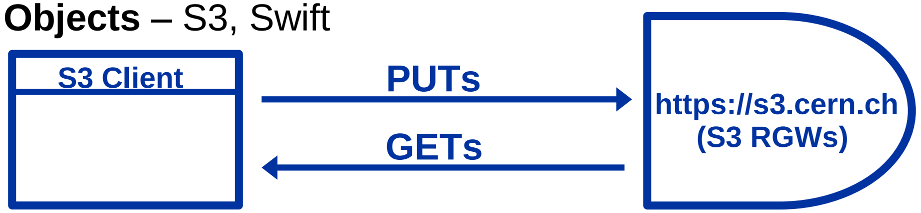
Ceph is a distributed storage platform:

- Provides 3 differing types of storage to end users
- Uses the same underlying “RADOS” object store under the hood

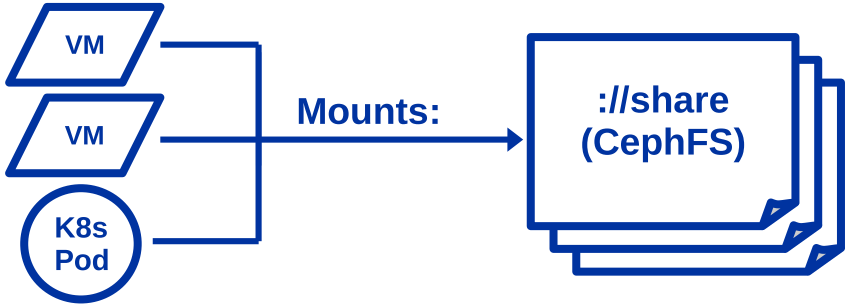
- **Block** – RBD, OpenStack Cinder Volumes / Glance Images



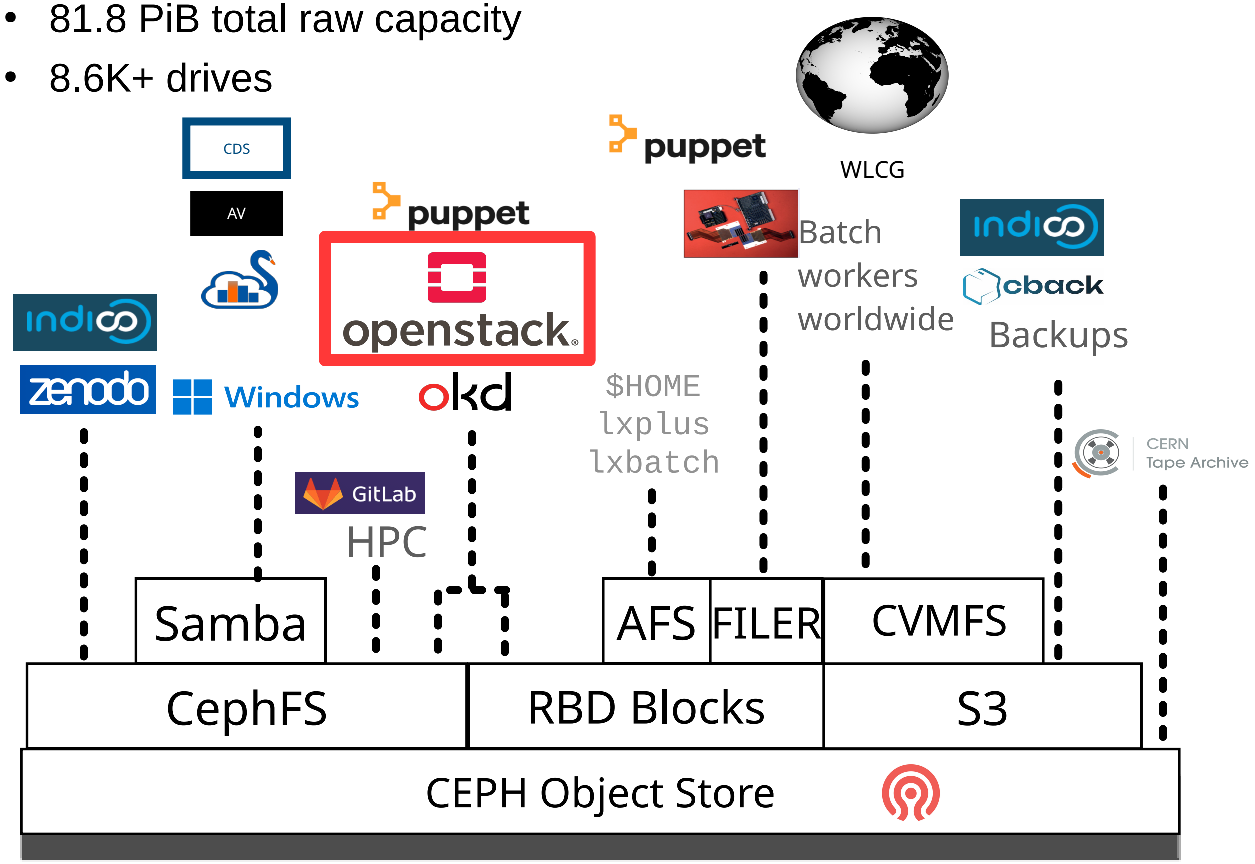
- **Objects** – S3, Swift



- **File System** – CephFS, Manila Shares



- 26 production clusters
- 81.8 PiB total raw capacity
- 8.6K+ drives

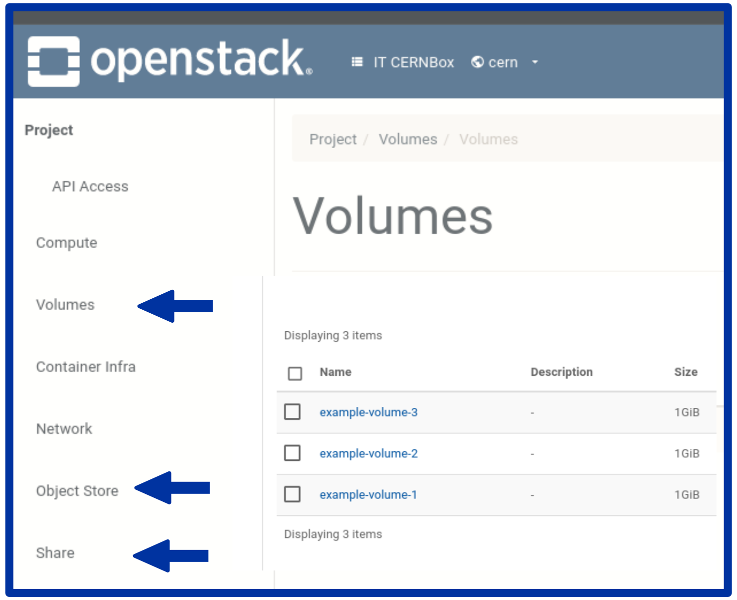


OpenStack at CERN:

Private Cloud for the entire Organization

- In production since July 2013
- Used to provision + life-cycle VM's / bare metal for services (including Ceph!)
- OS projects and quota are how we expose Ceph storage

- 1857 host hypervisors
- 10k bare metal hosts
- 14.6K virtual machines



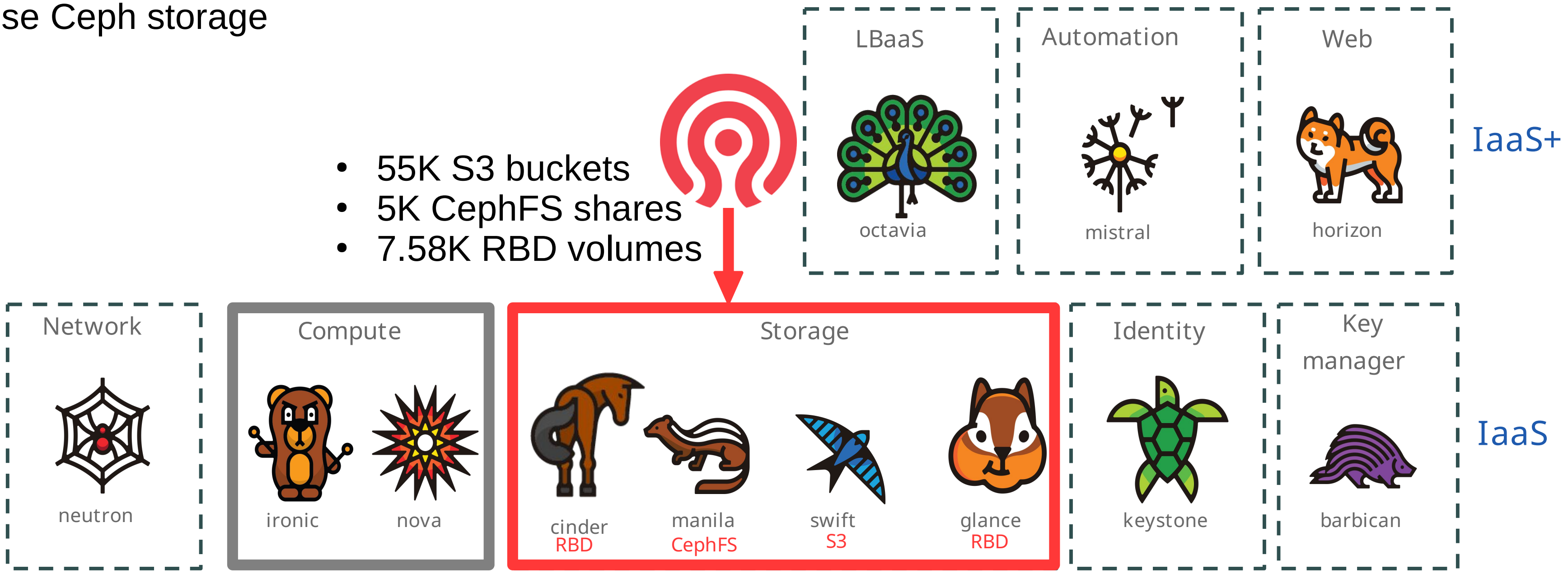
Volumes

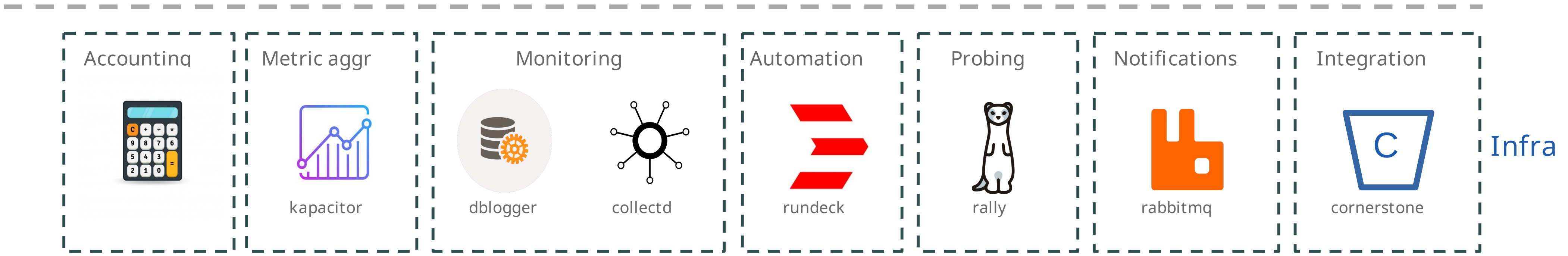
Displaying 3 items

☐	Name	Description	Size
☐	example-volume-3	-	1GiB
☐	example-volume-2	-	1GiB
☐	example-volume-1	-	1GiB

Displaying 3 items

- 55K S3 buckets
- 5K CephFS shares
- 7.58K RBD volumes



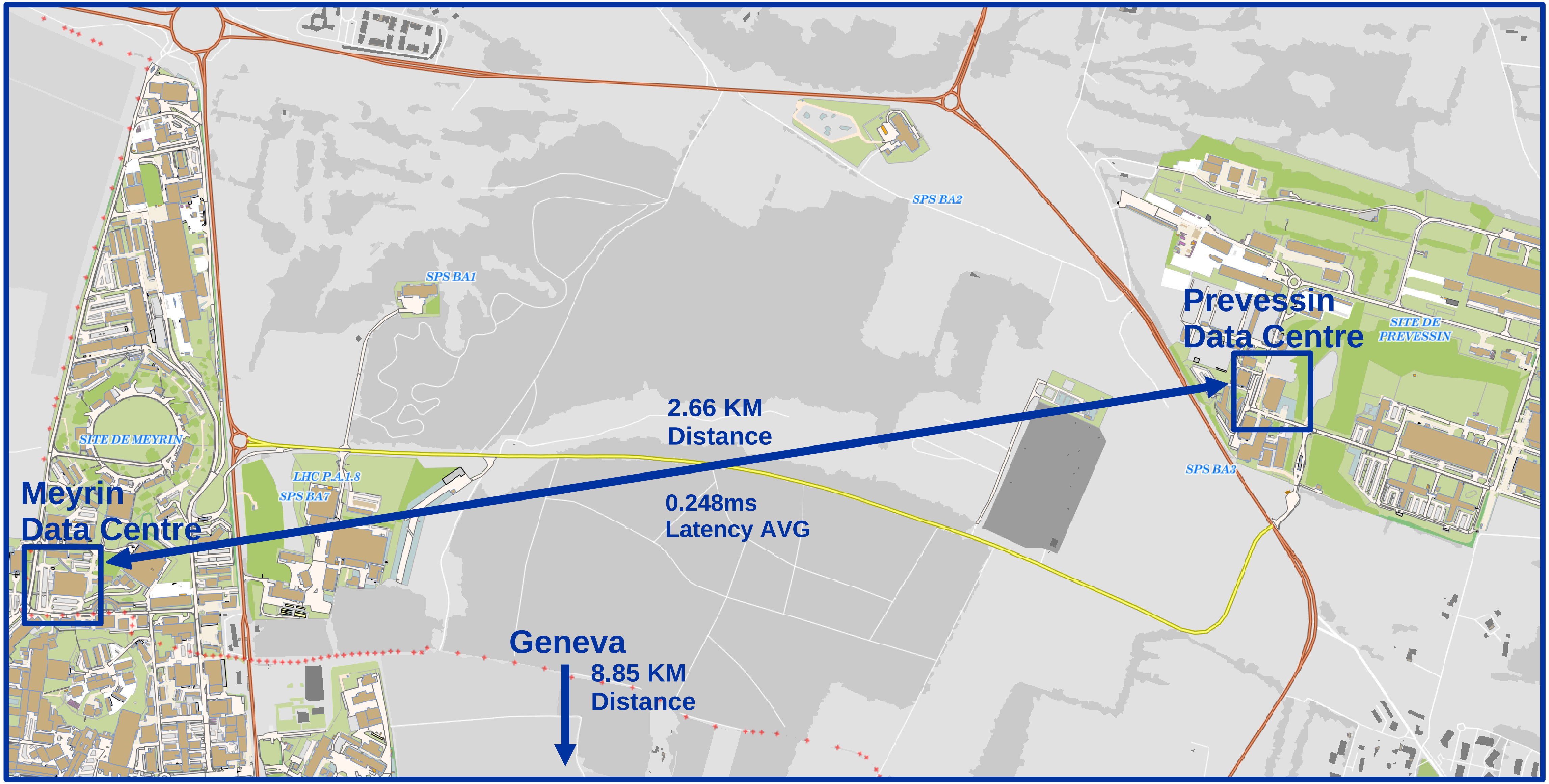


1

A whole new world.

(Just down the road)

From one to two:



Prevessin Data Centre

Meyrin Data Centre

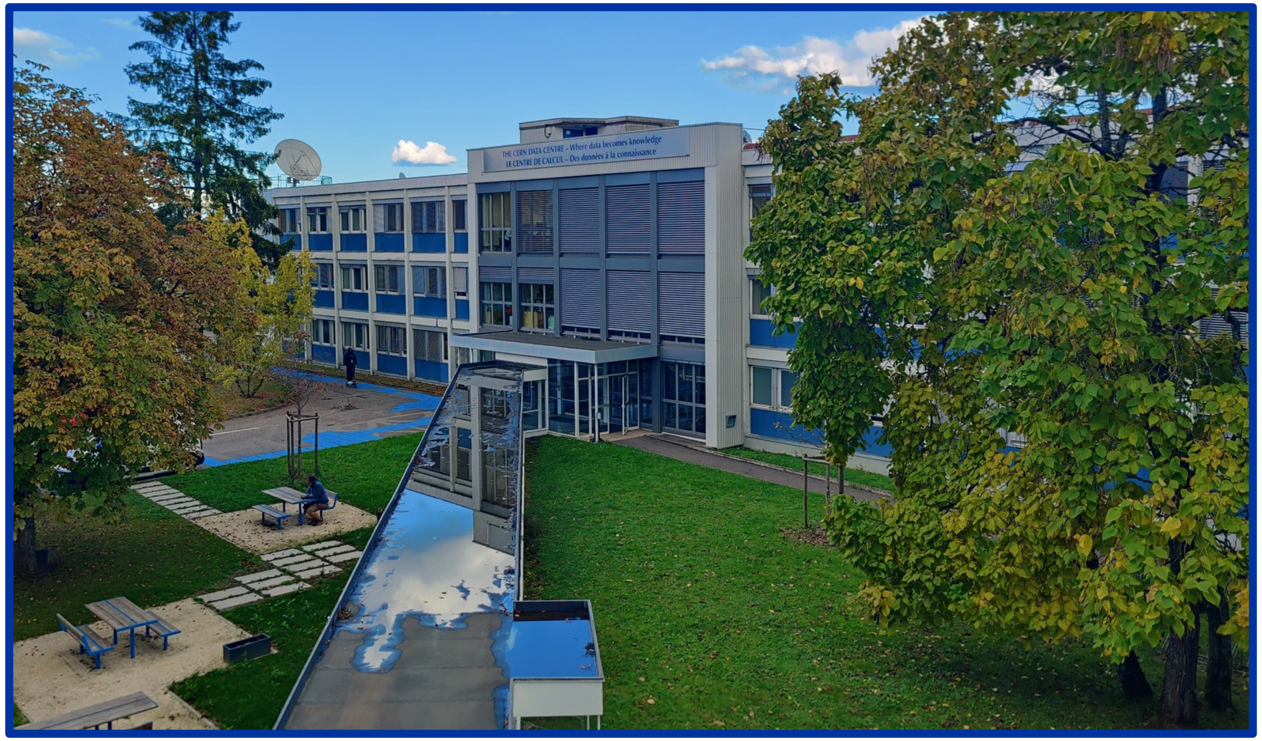
2.66 KM Distance

0.248ms Latency AVG

Geneva

8.85 KM Distance

A comparison:


THE CERN DATA CENTRE – Where data becomes knowledge
LE CENTRE DE CALCUL – Des données à la connaissance

Meyrin Data Centre (MDC):

- Built In the early 70s
- Operational power capacity nearing its limit
- Existing rack space for infrastructure expansion dwindling
- Limited space available in diesel-backed critical power area

Prevessin Data Centre (PDC):

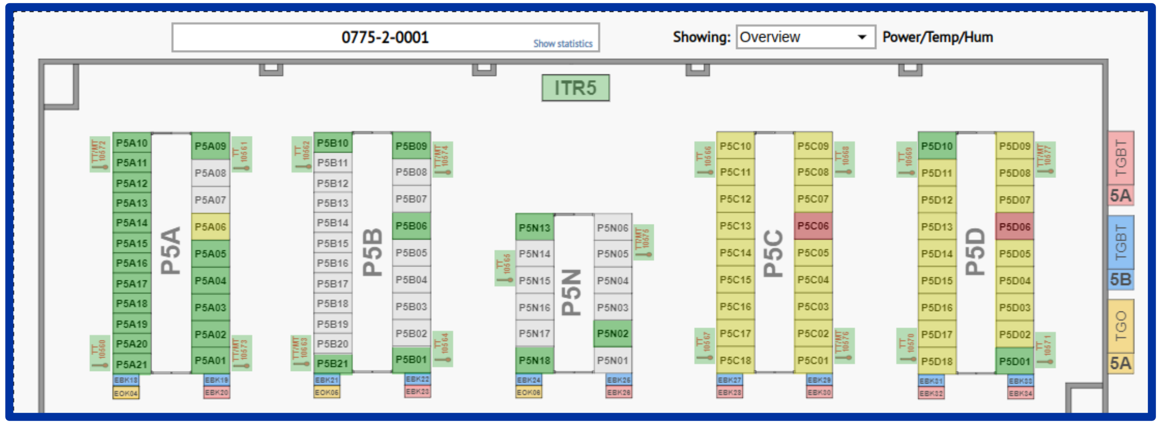
- Opened 23rd of Feb, 2024
- Built to cover the use case of:
  - HTC / Batch, experiment trigger system augmentation
- MDC limitations instigated plans for a second OS region
  - Ceph and OS go hand in hand, thus Ceph has a presence in PDC

PDC and Ceph:

Not a massive departure from what we already have and do…

Hardware Parameters	MDC (spinning cluster)	PDC (spinning cluster)
JBOD Size	2X24 HDD	1X60 HDD
Memory (per node)	251 GiB	251 GiB
NIC / Uplink throughput	1 x 25Gb/s 1 x 10Gb/s IPMI	1 x 25Gb/s 1 x 10Gb/s IPMI
Processor Model	AMD EPYC 7302	AMD EPYC 7402P
CRUSH Failure domain	Depends...	Rack (for now)

- PDC broadly mimics the services offered in MDC
- 5 production clusters (so far) offering flavours of block, object and file-system



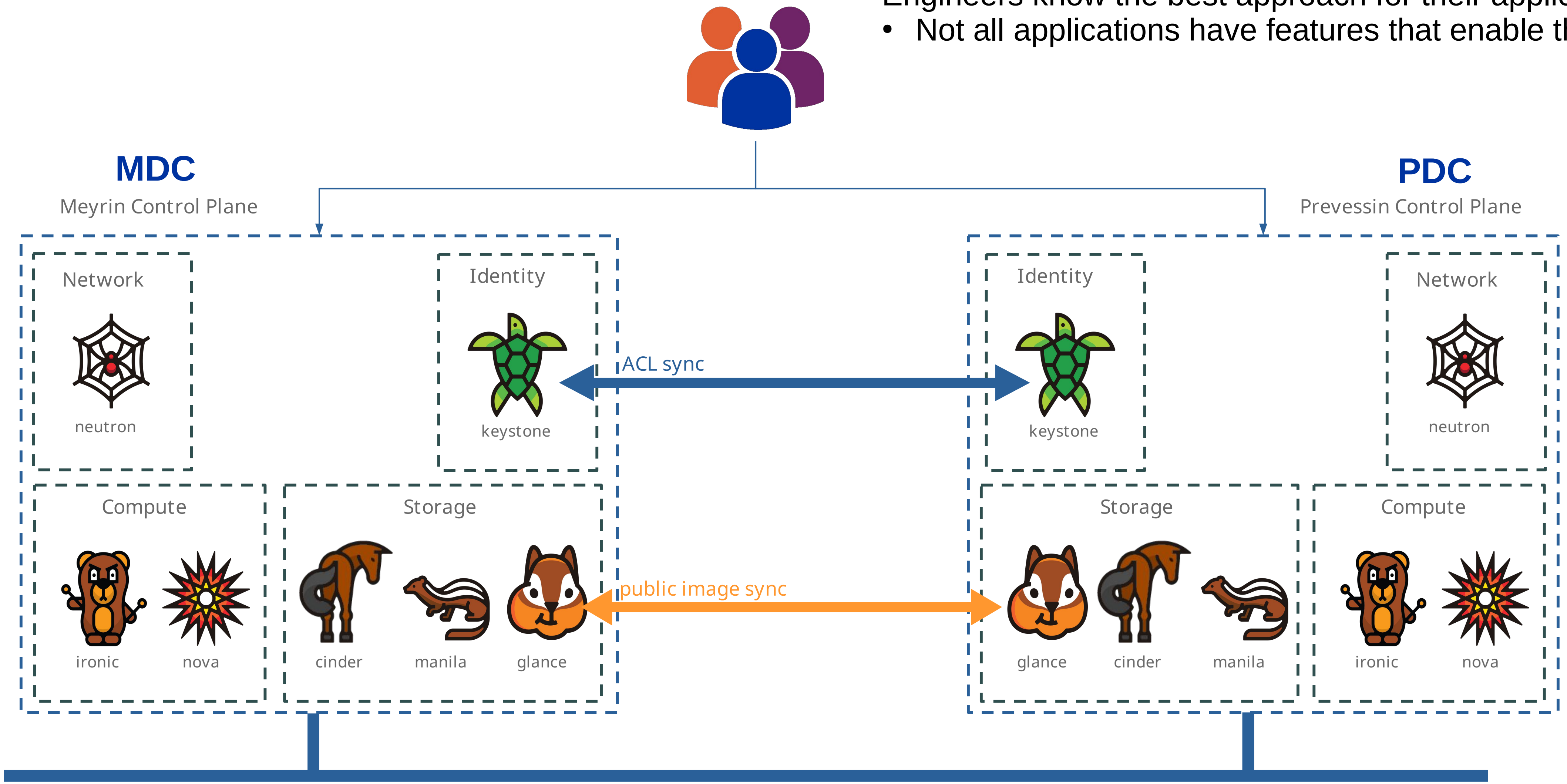
0775-2-0001

Showing: Overview

Deploying two regions:

From an application perspective:

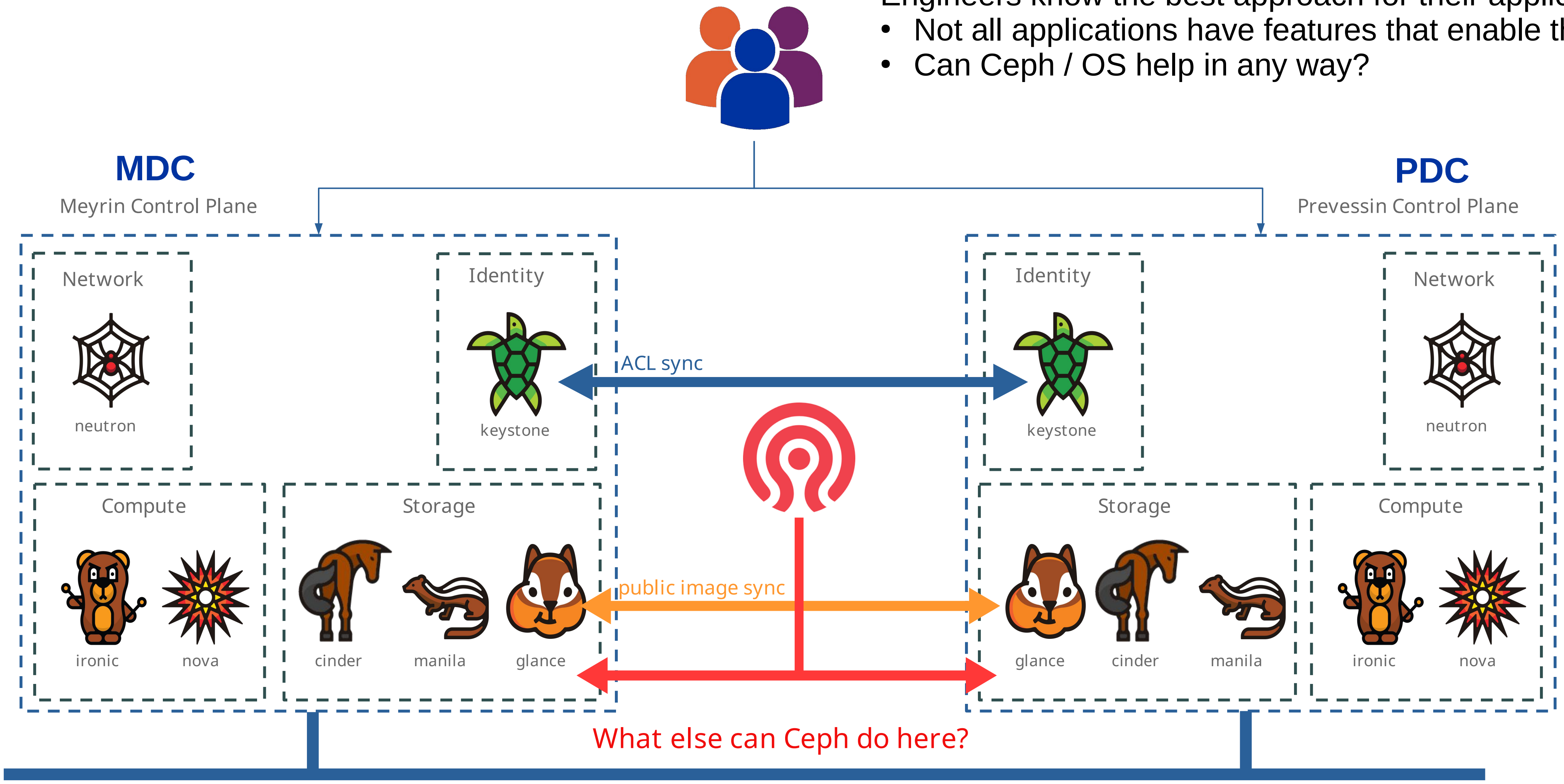
- Good designs can utilise two regions for BC/DR:
  - Engineers know the best approach for their application
    - Not all applications have features that enable this...



Deploying two regions:

From an application perspective:

- Good designs can utilise two regions for BC/DR:
  - Engineers know the best approach for their application
    - Not all applications have features that enable this...
    - Can Ceph / OS help in any way?



2

What else *can* Ceph do here?

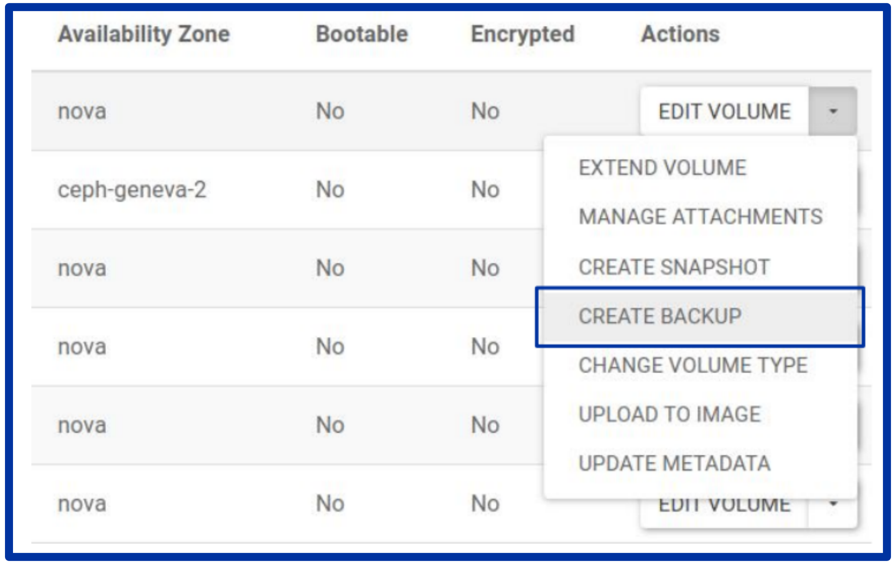
And what do we do?

Offsite backups for block storage:

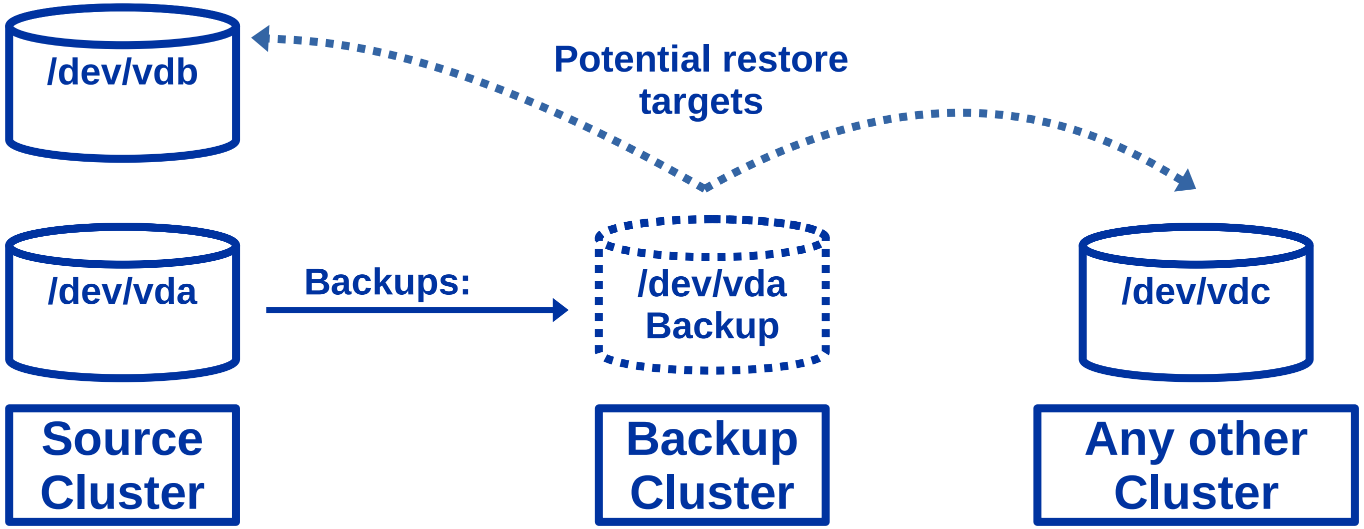
Ceph provides efficient tooling for RBD backups:

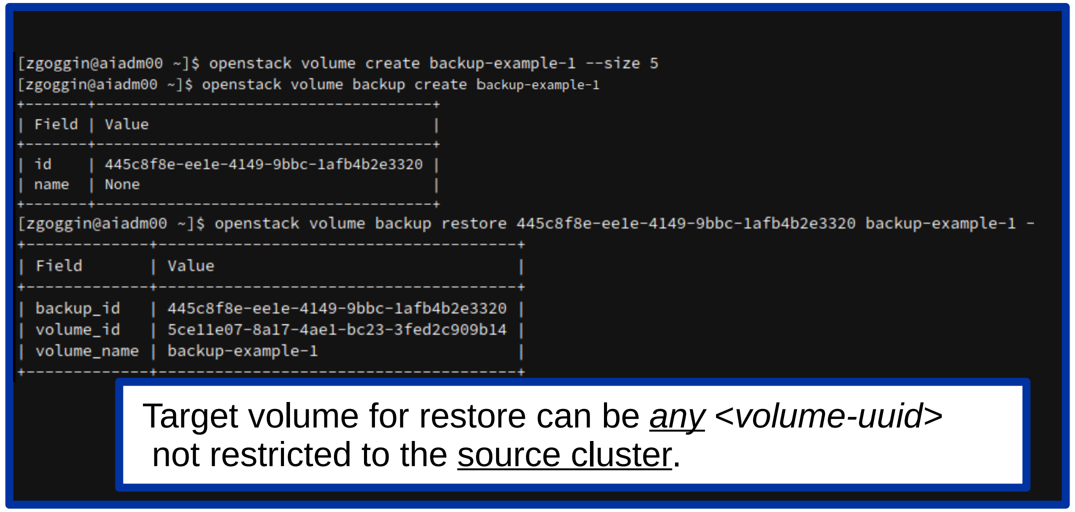
- Allows for *full* or *incremental* backups across clusters
- Location at rest can be a *different* Ceph cluster in a different region
- Full integration with OS: Fits well into our paradigm of “user driven”

- Two contending drivers:
  - RBD to RBD (**Good!**)
  - RBD to S3 (*Not so good...*)



Availability Zone	Bootable	Encrypted	Actions
nova	No	No	EDIT VOLUME
ceph-geneva-2	No	No	
nova	No	No	
nova	No	No	
nova	No	No	
nova	No	No	EDIT VOLUME





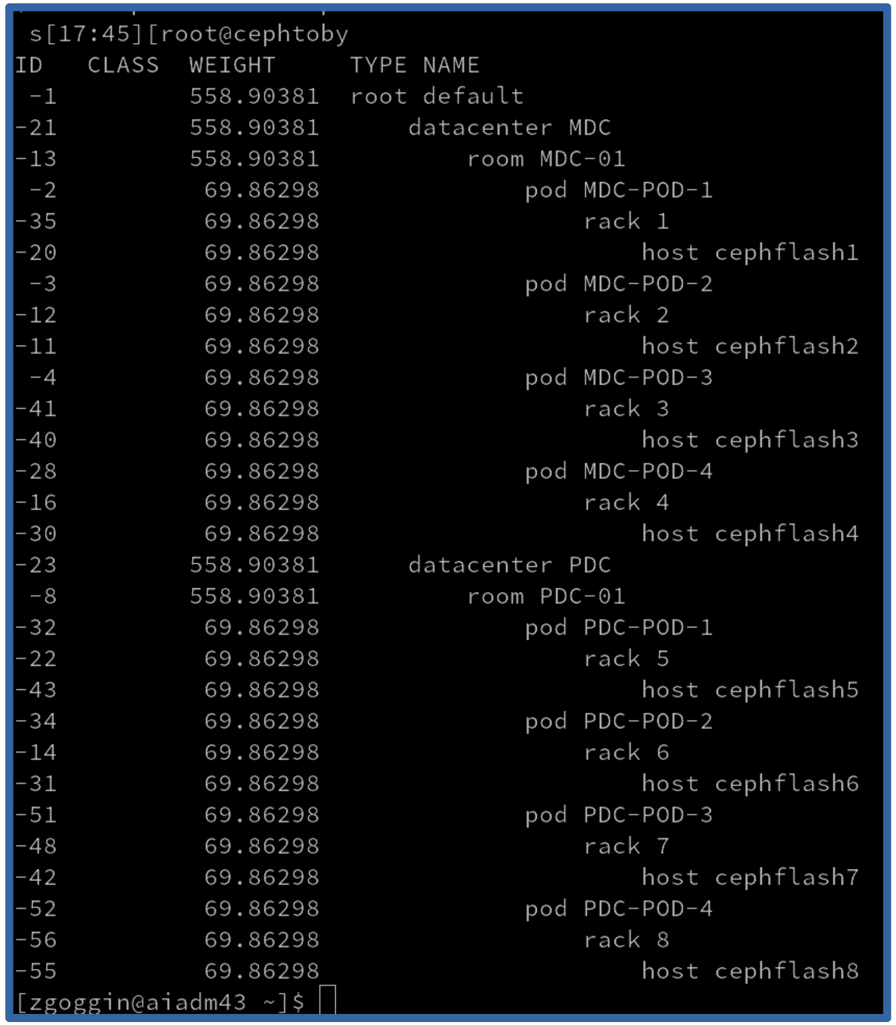
```
[zgoggin@aiadm00 ~]$ openstack volume create backup-example-1 --size 5
[zgoggin@aiadm00 ~]$ openstack volume backup create backup-example-1
+-------+--------------------------------------+
| Field | Value                                |
+-------+--------------------------------------+
| id    | 445c8f8e-ee1e-4149-9bbc-1afb4b2e3320 |
| name  | None                                 |
+-------+--------------------------------------+
[zgoggin@aiadm00 ~]$ openstack volume backup restore 445c8f8e-ee1e-4149-9bbc-1afb4b2e3320 backup-example-1 -
+-------------+--------------------------------------+
| Field       | Value                                |
+-------------+--------------------------------------+
| backup_id   | 445c8f8e-ee1e-4149-9bbc-1afb4b2e3320 |
| volume_id   | 5ce11e07-8a17-4ae1-bc23-3fed2c909b14 |
| volume_name | backup-example-1                     |
+-------------+--------------------------------------+
```

Target volume for restore can be *any* *<volume-uuid>* not restricted to the source cluster.

Cross region consistent storage:

A Ceph cluster can be “stretched” across two geographical points:

- Allows the survival of a cluster in the case of a site outage
- Possibility of “unstretching” cluster on total loss of a DC

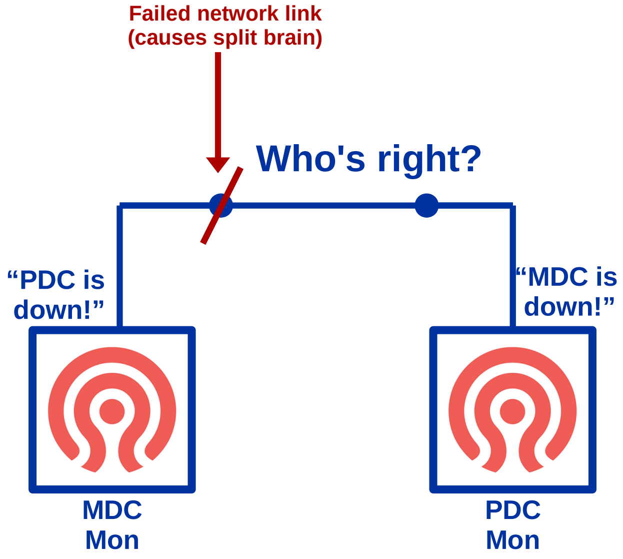


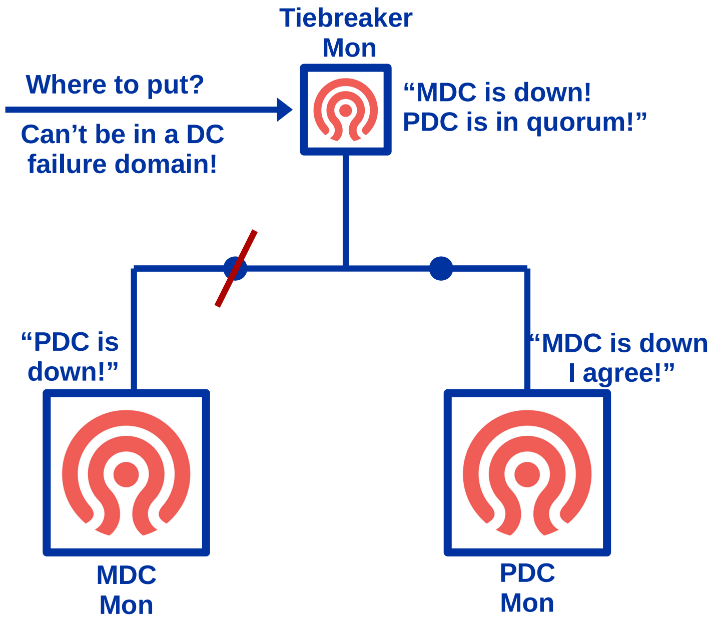
```
 s[17:45][root@cephtoby
ID   CLASS  WEIGHT     TYPE NAME
 -1         558.90381  root default
-21         558.90381      datacenter MDC
-13         558.90381          room MDC-01
 -2          69.86298              pod MDC-POD-1
-35          69.86298                  rack 1
-20          69.86298                      host cephflash1
 -3          69.86298              pod MDC-POD-2
-12          69.86298                  rack 2
-11          69.86298                      host cephflash2
 -4          69.86298              pod MDC-POD-3
-41          69.86298                  rack 3
-40          69.86298                      host cephflash3
-28          69.86298              pod MDC-POD-4
-16          69.86298                  rack 4
-30          69.86298                      host cephflash4
-23         558.90381      datacenter PDC
 -8         558.90381          room PDC-01
-32          69.86298              pod PDC-POD-1
-22          69.86298                  rack 5
-43          69.86298                      host cephflash5
-34          69.86298              pod PDC-POD-2
-14          69.86298                  rack 6
-31          69.86298                      host cephflash6
-51          69.86298              pod PDC-POD-3
-48          69.86298                  rack 7
-42          69.86298                      host cephflash7
-52          69.86298              pod PDC-POD-4
-56          69.86298                  rack 8
-55          69.86298                      host cephflash8
[zgoggin@aiadm43 ~]$
```

To consider:

- Cluster is a single, macro point of failure
- Latency plays a massive role in efficacy (0.248ms is in our favour)
    - Writes are **synchronous** across one Ceph cluster!

  Cost implications of redundancy
- How do you authoritatively decide when a site is “dead”?





Bucket policy to protect S3 backups:

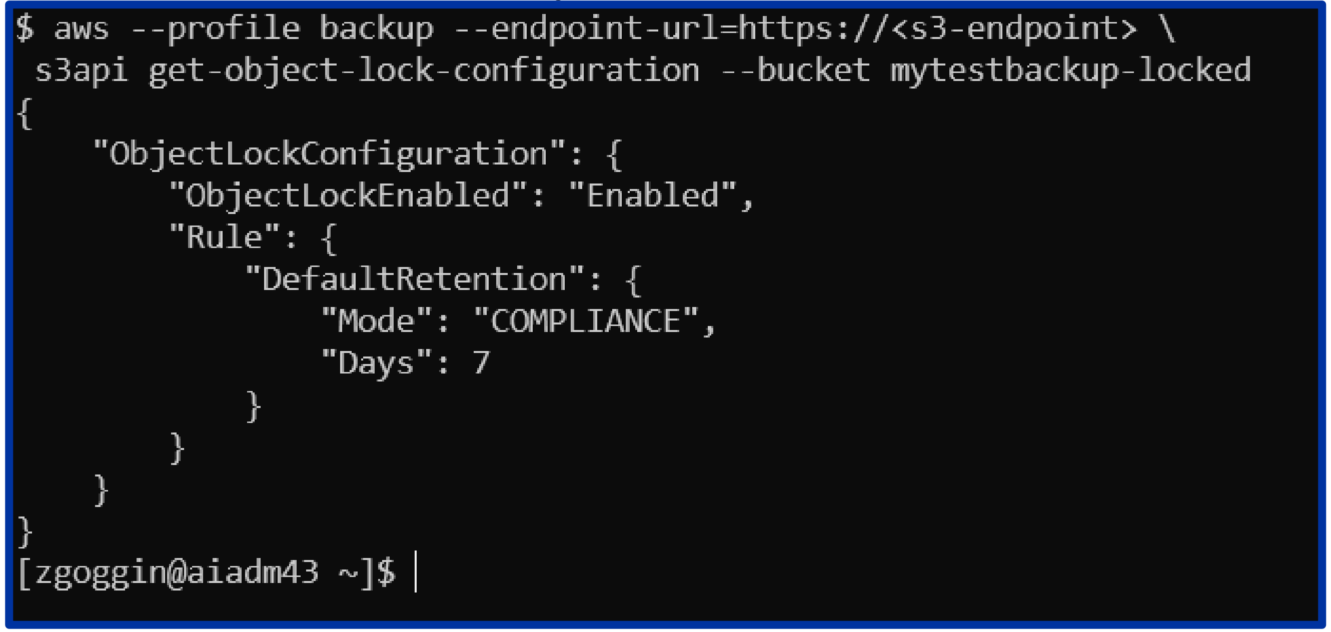
Lots of people use s3 as a backup endpoint:

- Bucket policies that can “protect” a backup bucket are useful

  Can **combine** two major S3 features to this end:

Object locks:

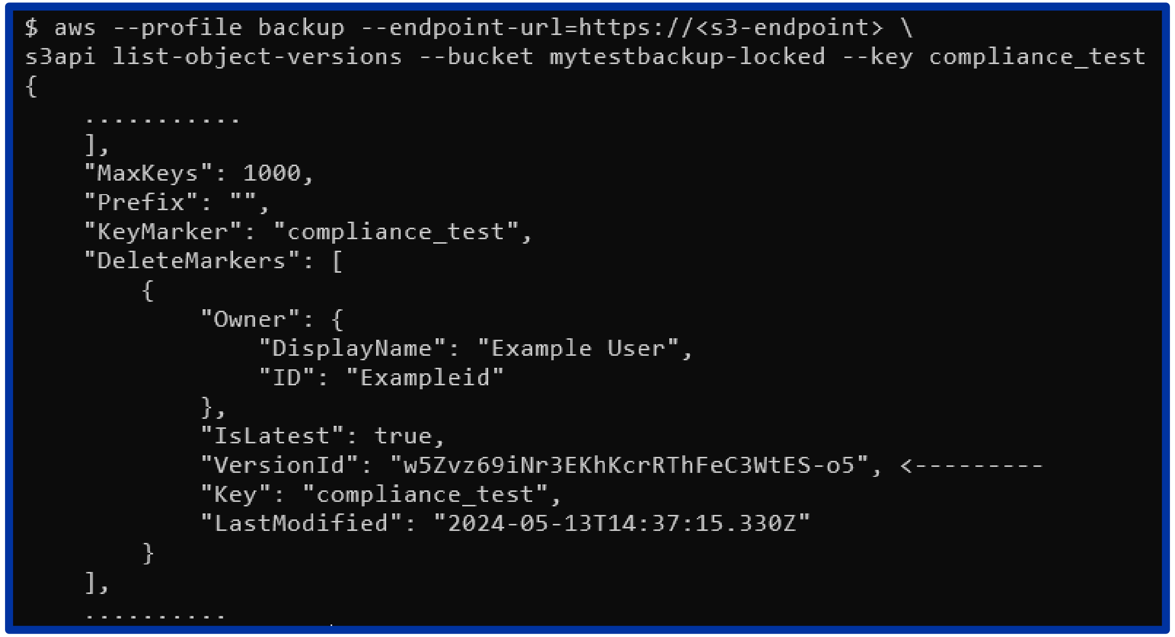
- Object locks provide granular permissions regarding object deletion
- Compliance mode, forces a grace period using on a <*retention-time*>
  - Objects deleted are “marked” but not acted upon until expiry
  - Cannot be overridden by the bucket owner or a administrator



```
$ aws --profile backup --endpoint-url=https://<s3-endpoint> \
 s3api get-object-lock-configuration --bucket mytestbackup-locked
{
    "ObjectLockConfiguration": {
        "ObjectLockEnabled": "Enabled",
        "Rule": {
            "DefaultRetention": {
                "Mode": "COMPLIANCE",
                "Days": 7
            }
        }
    }
}
[zgoggin@aiadm43 ~]$ 
```

Versioning:

- Allows for multiple versions of a specific object to exist wherein the current object is the newest version
  - Older versions are fetchable via a <*version-id*>
  - Stops attacks or mistakes that overwrite an object

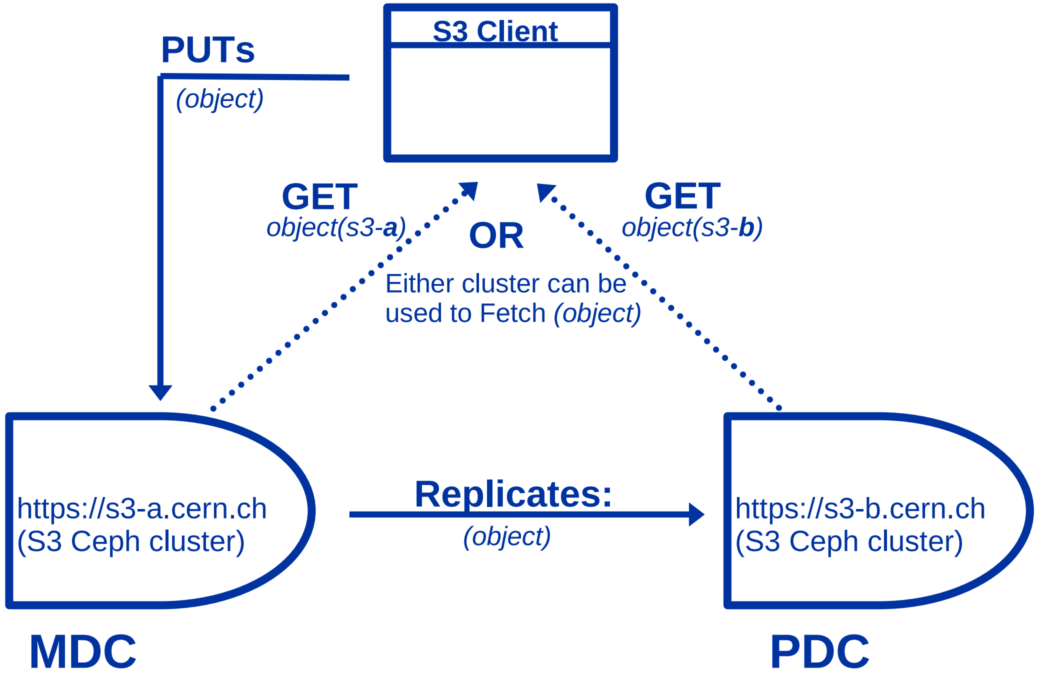


```
$ aws --profile backup --endpoint-url=https://<s3-endpoint> \
s3api list-object-versions --bucket mytestbackup-locked --key compliance_test
{
    ...........
    ],
    "MaxKeys": 1000,
    "Prefix": "",
    "KeyMarker": "compliance_test",
    "DeleteMarkers": [
        {
            "Owner": {
                "DisplayName": "Example User",
                "ID": "Exampleid"
            },
            "IsLatest": true,
            "VersionId": "w5Zvz69iNr3EKhKcrRThFeC3WtES-o5", <---------
            "Key": "compliance_test",
            "LastModified": "2024-05-13T14:37:15.330Z"
        }
    ],
    ..........
```

S3 Multisite across two regions:

Muitisite is where two (or more) Clusters are mirrored:

- Writes on one bucket are replicated to another cluster
- Multisite “Zones” have policies that control this behaviour



S3 Client

PUTs
(object)

GET
object(s3-**a**)

OR

GET
object(s3-**b**)

Either cluster can be used to Fetch *(object)*

https://s3-a.cern.ch
(S3 Ceph cluster)

Replicates:
(object)

https://s3-b.cern.ch
(S3 Ceph cluster)

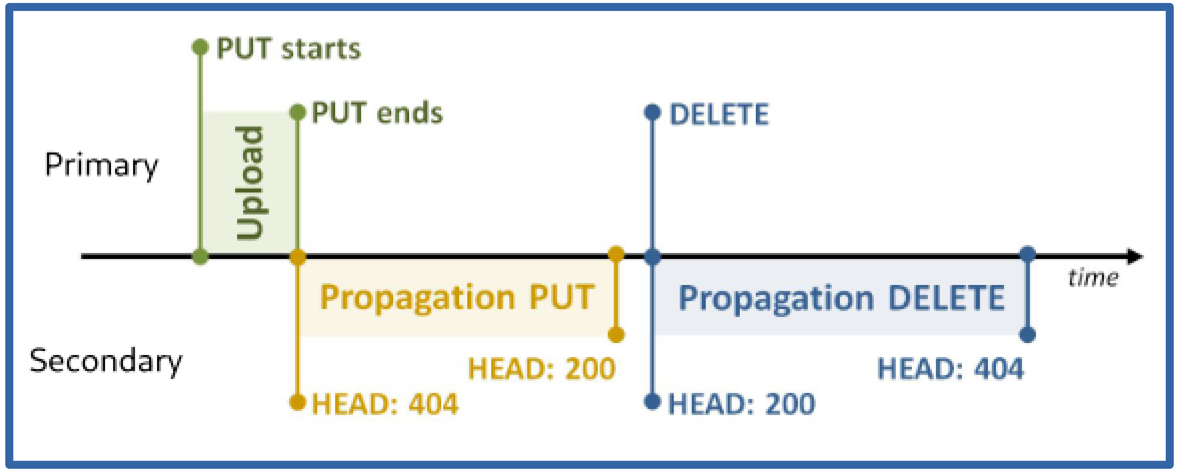
MDC

PDC

To consider:

- Multisite comes with an **Intrinsic** replication delay
- Replications are **always asynchronous** (two clusters, not one!)
- Range for delay varies...

	Inital PUT	Propagation PUT	Propagation DELETE
Mean	1.15s	7.75s	10.80s
Stdev	0.01	2.70	2.00



3 Conclusion.

The verdict:

Ceph has numerous features that can make good use of a redundant data centre

- RBD, S3, CephFS all have coverage (sometimes by the same solution)
- Your mileage may vary requisite to your site/deployments needs…

This talk only focused on what we are actively using, not discussed:

- CephFS Snapshots + Mirroring
- OpenStack Manila backup drivers
- External S3 providers (AWS, Glacier, etc.)
- And certainly others…

Addendum:

4-5 December 2024 | #Cephalocon
CERN | Geneva, Switzerland

https://ceph.io/en/community/events/2024/cephalocon-2024/

Thanks for your time!

zachary.goggin@cern.ch

Backup slides:

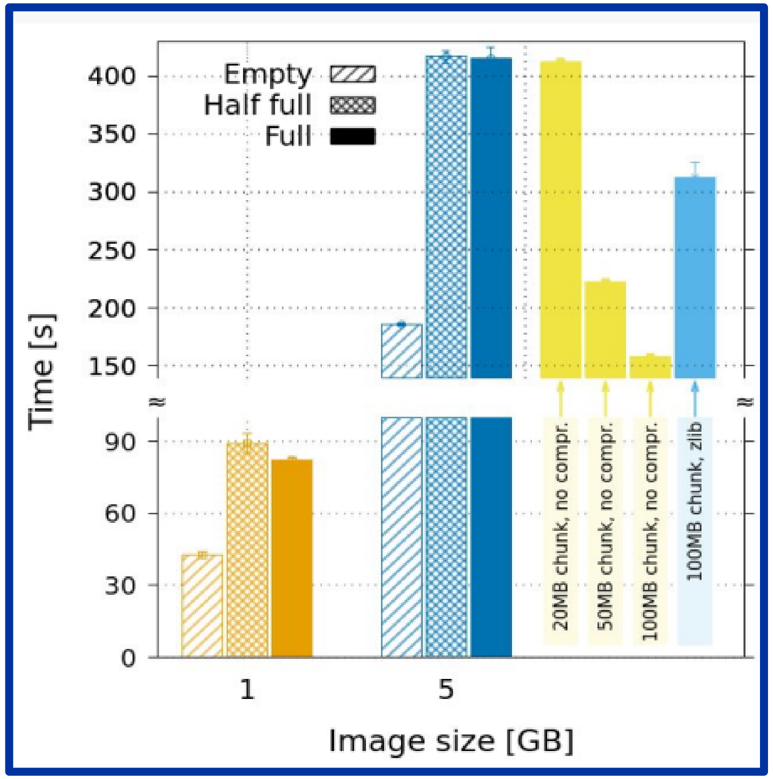
RBD Backups: Benchmarking

Only real interest here is in dissuading people from using rbd -> s3. Otherwise will spend too long talking here...

In production

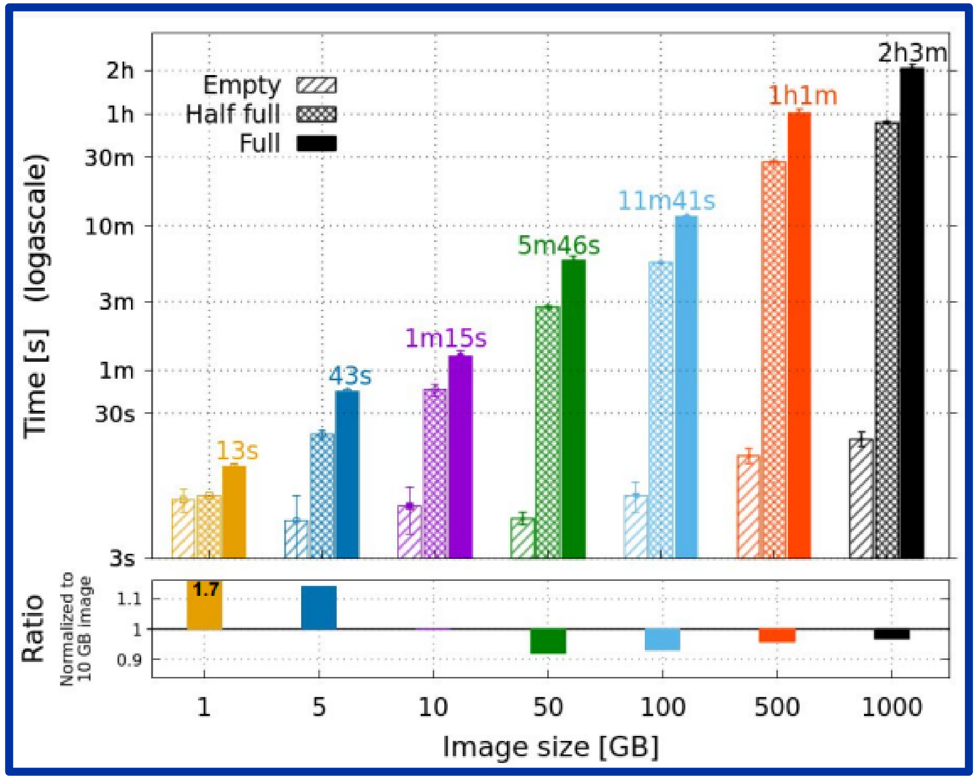
RBD -> S3: Full Backup

- In our testing, painfully slow.
- *Some* tuning parameters (*compression type*, *s3 object size*, *total pool connections*) but no real improvements…
- Incremental support is likewise, pretty bad:
  - Driver reads full source and backup to gen diff



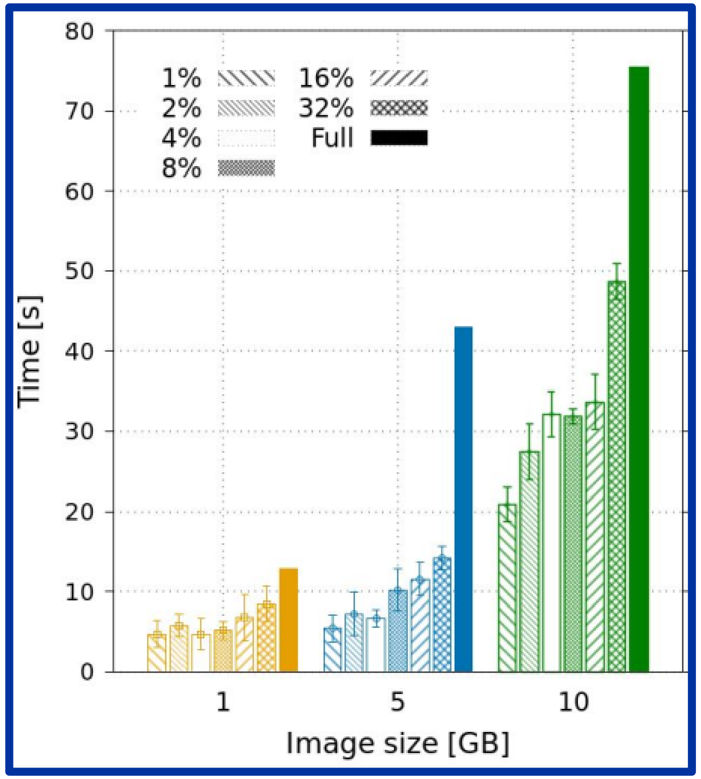
RBD -> RBD: Full Backup

- *Significantly* better write performance. Reliance on *librbd.*
- Good performance out of the box! (~120MB/s per backup)
- Write speed is sustained as volume size increases
- Scales well with large numbers of concurrent backups
- Utilization of volume has an impact on time to conclusion



RBD -> RBD: Incremental

- Based on a diff of the previous snapshot and the current image state. Backup is effectively:
- Source `*rbd export diff* ` *to export snapshot diff*
- *Target* ` *rbd import diff* ` *to merge diff into backup*
- Leverages ceph block optimizations for speedup:
  - ` *Fast-diff, exclusive-lock, object-map* `



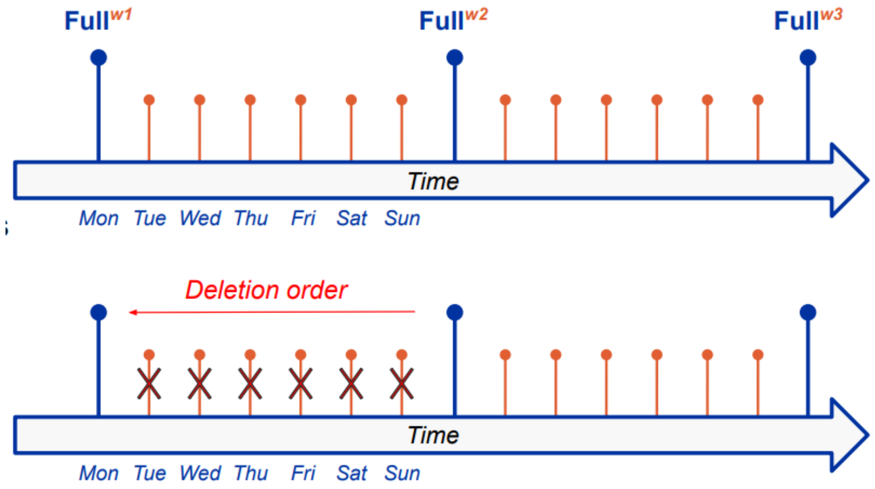
RBD Backups: Caveats

Restores:

- Restore is *allways* a full restore, no concept of a differential restore in either driver
- Restore speed with the RBD -> RBD driver is comparable to backup speeds (120MB/s per restore)

Deletion with Incrementals:

- Deletion cannot occur oldest -> newest
- Better to make full backups, spaced out, each with their own incrementals, rather then one “long” backup



Consistency:

- Having a backup does not mean your data is “safe”! it just means you have a copy of it at a given time under certain circumstance:

Bad.

<u>Inconsistant:</u>

- Backup occurs while block volume is “live”,
- Contents may change! Backup may not be readable or even usable after restore.

RBD backups on their own are here → <u>Crash consistant:</u>

- Point in time consistent backup, of all blocks in a given volume, outstanding IO may not be captured, but existing blocks will not change.
- Typically “good enough” for most applications

Ideal for databases or state concerned apps → <u>Appliccation consistant:</u>

- Pending IO transactions are flushed to disk, presumes application stops /reads/writes via *fsfreeze* or other mechanisms before and after backup starts.

Getting better.

CHEP is for reporting on WORK DONE, not for providing explicit documentation / tutorials on how a concept works.
We are not writing CEPH documentation

Ceph at CERN: clusters of note

Cluster Application	Cluster medium type	Size (Raw)	Release Version
RBD (OpenStack Cinder/Glance krbd)	HDD’s (Replica 3)	9.7 PiB	Pacific
^	Full-flash (4+2 EC)	392 TiB	Pacific
CephFS (OpenStack Manila – K8/OKD PVs, HPC)	HDD’s (Replica 3)	4.2 PiB	Pacific
^	Full-flash (Replica 3)	1.1 PiB	Pacific
RGW (S3 + Swift)	HDD’s (4+2 EC)	4.2 PiB	Pacific
Backup and Preservation (S3/RBD)	HDD’s (4+2 EC)	24 PiB	Pacific
RADOS CERN Tape Archive (CTA) Tape DB, Disk Buffer and repacking	full flash (4+2 EC)	220 TiB	Quincy

Clearly not an exhaustive list.

- Largely using Pacific in production
  - Slowly upgrading to Quincy, newer clusters go straight to 17.x

20 of our 26 clusters are in our primary datacenter

  - More on that later.

Could flip this and aggregate size for each storage paradigm

+ No of clusters providing each service. Maybe simpler to read?