A new Ceph deployment using Cephadm at RAL

HEPiX Spring 2022

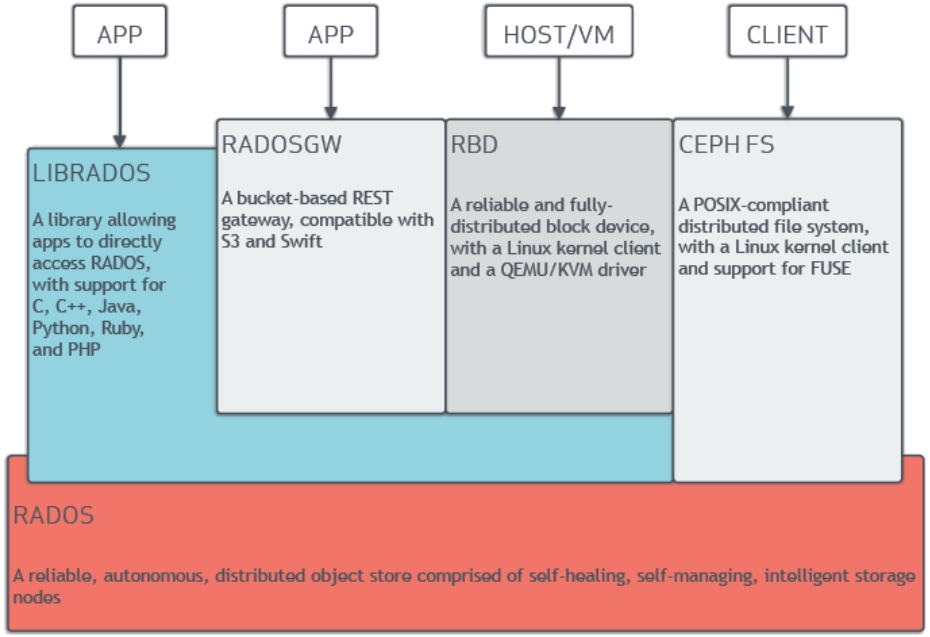
Kyle Pidgeon
Scientific Computing Department,
STFC UKRI

Ceph

- "*Ceph is an open-source software storage platform, implements object storage on a single distributed computer cluster, and provides 3-in-1 interfaces for object-, block- and file-level storage.*" – Wikipedia

- Scalable and reliable
  - Dynamic rebalancing and recovery

- Many plugins and drivers for integration with other technologies



Science and Technology Facilities Council

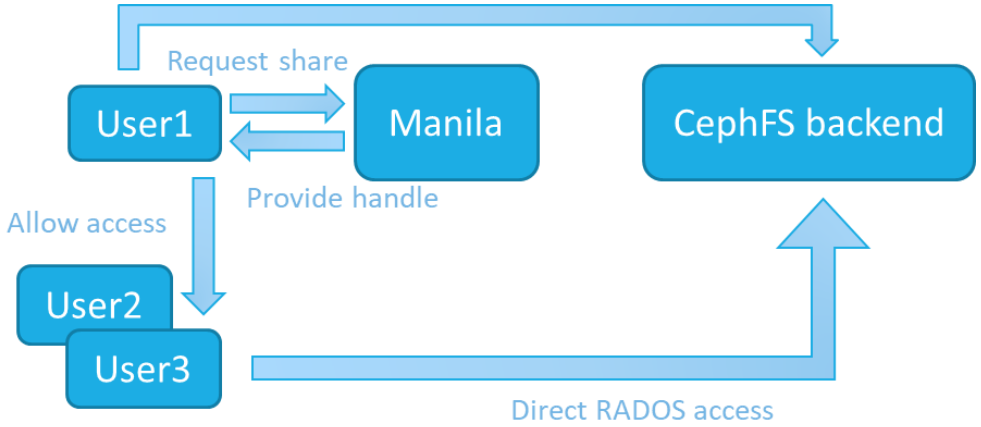
Scientific Computing

Ceph at RAL

- ECHO – Tier-1 storage
  - ~50PiB raw storage capacity
  - S3, XRootD, GridFTP
- Sirius
  - Block device storage for Cloud VMs
- Deneb
  - CephFS, used by several STFC Facilities e.g. ISIS and CLF

A new cluster: Arided

- Backend for an OpenStack Manila service
  - ‘Shared Filesystem Service’
- Motivated by:
  - High demand for FS storage
  - Potential disruption if using e.g. Deneb
  - Ability to use faster SSD storage



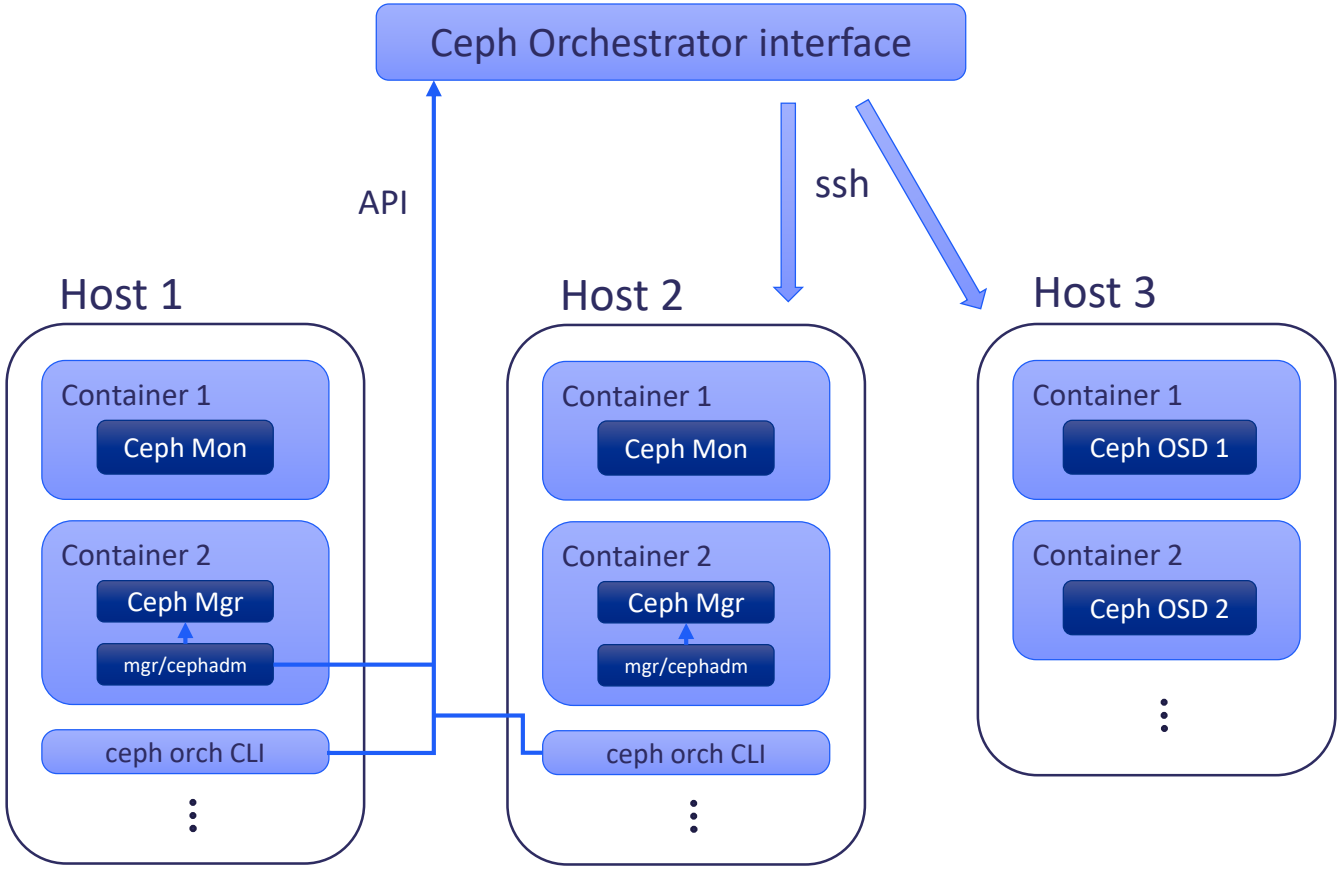
- Opportunity to try out new (to us) deployment methods
  - And a newer version of Ceph than is used on the current clusters (Octopus)

Science and Technology Facilities Council

Scientific Computing

Cephadm

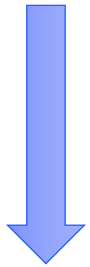
- Container orchestration for Ceph daemons
  - Can automate deployment, scaling, management of cluster
- Fully-featured
  - Implements all of Ceph's generic 'Orchestrator' features
  - Seems to be a primary focus of Ceph developers
- Minimal requirements
  - Python3, Docker, systemd, LVM


Ceph Orchestrator interface
API
ssh
Host 1
Host 2
Host 3
Container 1
Ceph Mon
Container 2
Ceph Mgr
mgr/cephadm
ceph orch CLI
Container 1
Ceph Mon
Container 2
Ceph Mgr
mgr/cephadm
ceph orch CLI
Container 1
Ceph OSD 1
Container 2
Ceph OSD 2

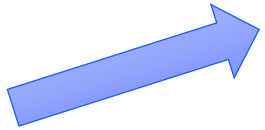
Cluster creation

(a test instance for now)

```
[root@ganesha1 ~]# cephadm check-host
podman|docker (/usr/bin/docker) is present
systemctl is present
lvcreate is present
Unit ntpd.service is enabled and running
Host looks OK
```



cephadm bootstrap



```
[ceph: root@ganesha1 /]# ceph status
  cluster:
    id:     d770062c-64b0-11ec-aec0-98039bcae034
    health: HEALTH_WARN
            OSD count 0 < osd_pool_default_size 3

  services:
    mon: 1 daemons, quorum ganesha1.nubes.rl.ac.uk (age 6m)
    mgr: ganesha1.nubes.rl.ac.uk.ccfibi(active, since 6m)
    osd: 0 osds: 0 up, 0 in

  data:
    pools:   0 pools, 0 pgs
    objects: 0 objects, 0 B
    usage:   0 B used, 0 B / 0 B avail
    pgs:
```

(Ceph version: Ceph Octopus 15.2.15)

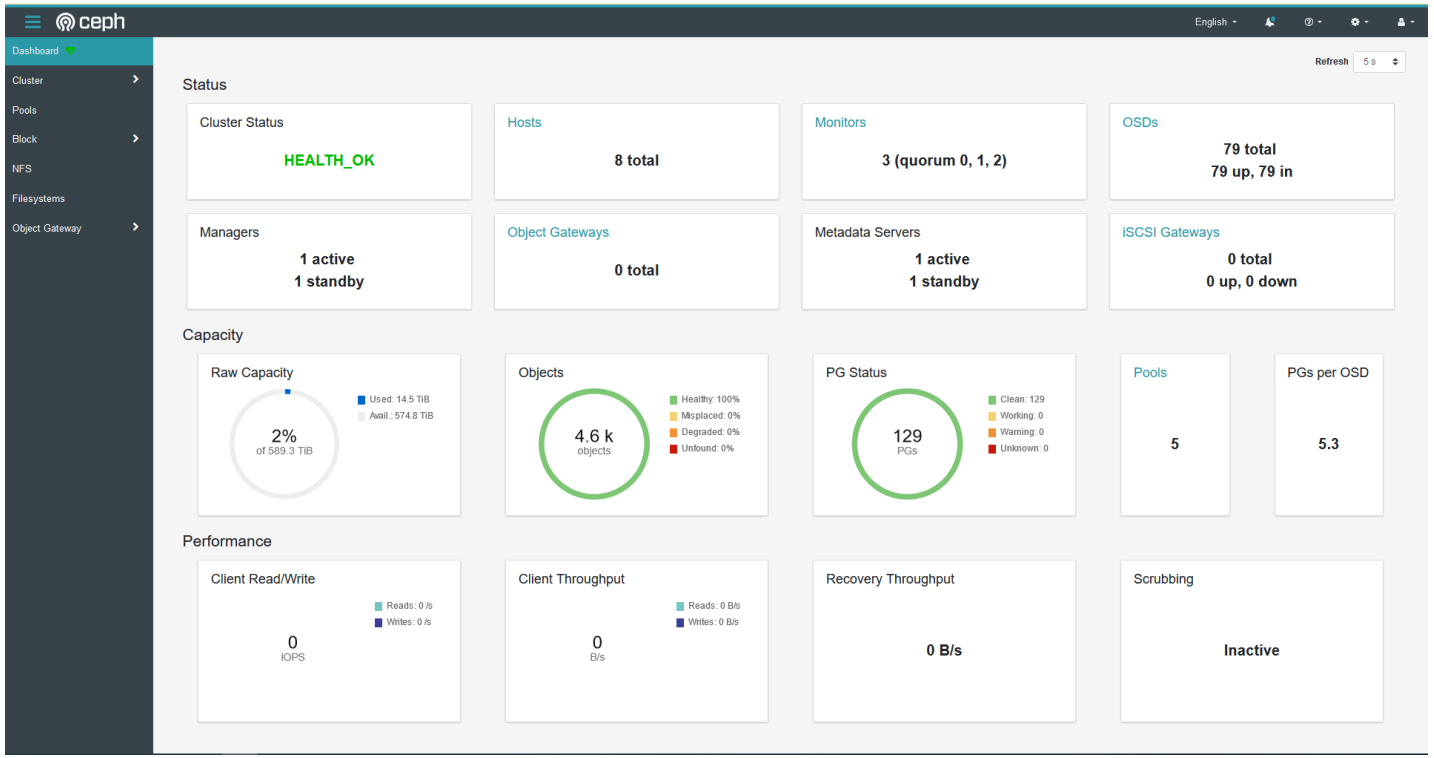
‘Day 2’ procedures

- Addition of hosts to cluster
  - Copy Ceph’s pub key and
    ceph orch host add <host>
- Service placement specs
  - Example uses HDDs for primary data storage and SSDs for metadata DBs

```
service_type: osd
service_id: osd_spec_default
placement:
  host_pattern: '*'
data_devices:
  rotational: 1
  size: '7TB:'
  model: 'HGST HUS728T8TAL'
db_devices:
  rotational: 0
  size: ':2TB'
  model: 'Dell Express Flash PM1725b 1.6TB AIC'
```

Monitoring

- Full monitoring stack deployed by default
- Can perform ops via Dashboard (interacts via Orchestrator)



Science and Technology Facilities Council

Scientific Computing

Cluster admin

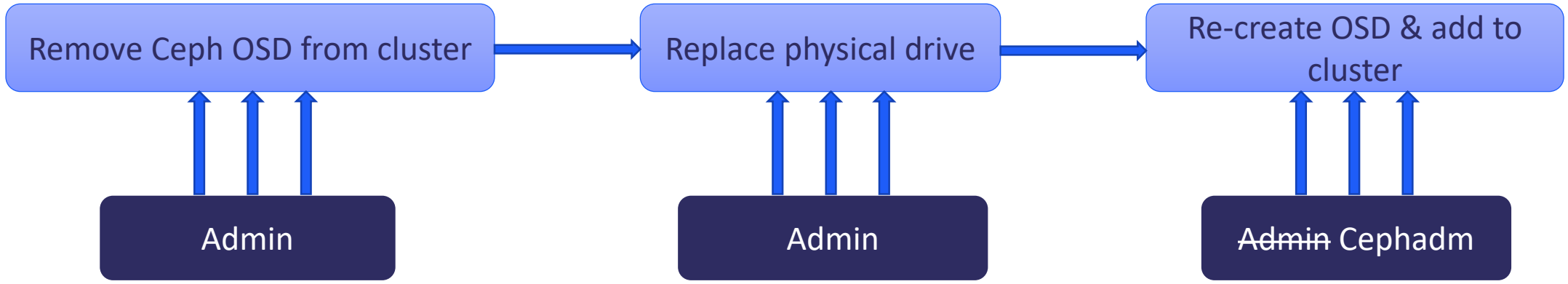
- Orchestrator API centralises a lot of routine ops
  - Daemon management from a single admin node

```
[ceph: root@ganesha1 /]# ceph orch daemon restart osd.0
Scheduled to restart osd.0 on host 'sn117.nubes.rl.ac.uk'
```

- Creating a CephFS: ceph fs volume create <fs name>
  - Automatically deploys metadata servers (MDS)

- Cephadm automates some time-consuming procedures
  - Cluster upgrades
  - Example: storage additions/replacements

Disk replacements



Ceph OSD = a daemon that interfaces with an *Object Storage Device*

Science and Technology Facilities Council

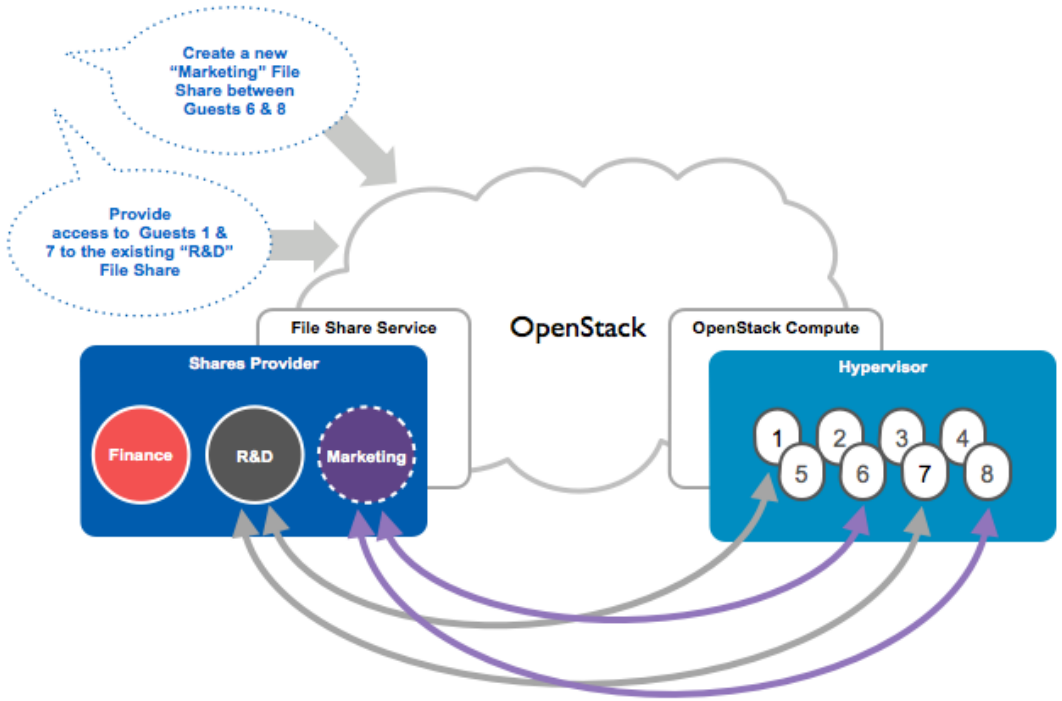
Scientific Computing

Potential downsides

- Performance (but probably not)
  - If degraded, probably compensated for by SSDs
- Scalability
  - Questions surrounding cephadm's scalability
  - Seems to have developer attention

Integration with Manila

- Manila will run on the STFC Cloud OpenStack
  - Brings file shares into the same ecosystem as compute etc.
- Specialised driver interfaces with CephFS
  - Creates CephFS ‘subvolumes’ for shares etc.
- Shares are mountable via Ceph FUSE



Science and Technology Facilities Council

Scientific Computing

Summary

- Developed test instance of a file shares service
  - Made use of Cephadm for the Ceph backend
- Cephadm…
  - …allowed rapid deployment of a Ceph cluster
  - …can simplify cluster admin procedures
- Intending to use for production instance
  - (If anyone has experience running a cephadm deployed/managed cluster in production, it'd be great to hear about it)

Questions?

Science and Technology Facilities Council

Scientific Computing