redhat.

Distributed File Storage in Multi-Tenant Clouds using CephFS

Openstack Vancouver 2018 May 23

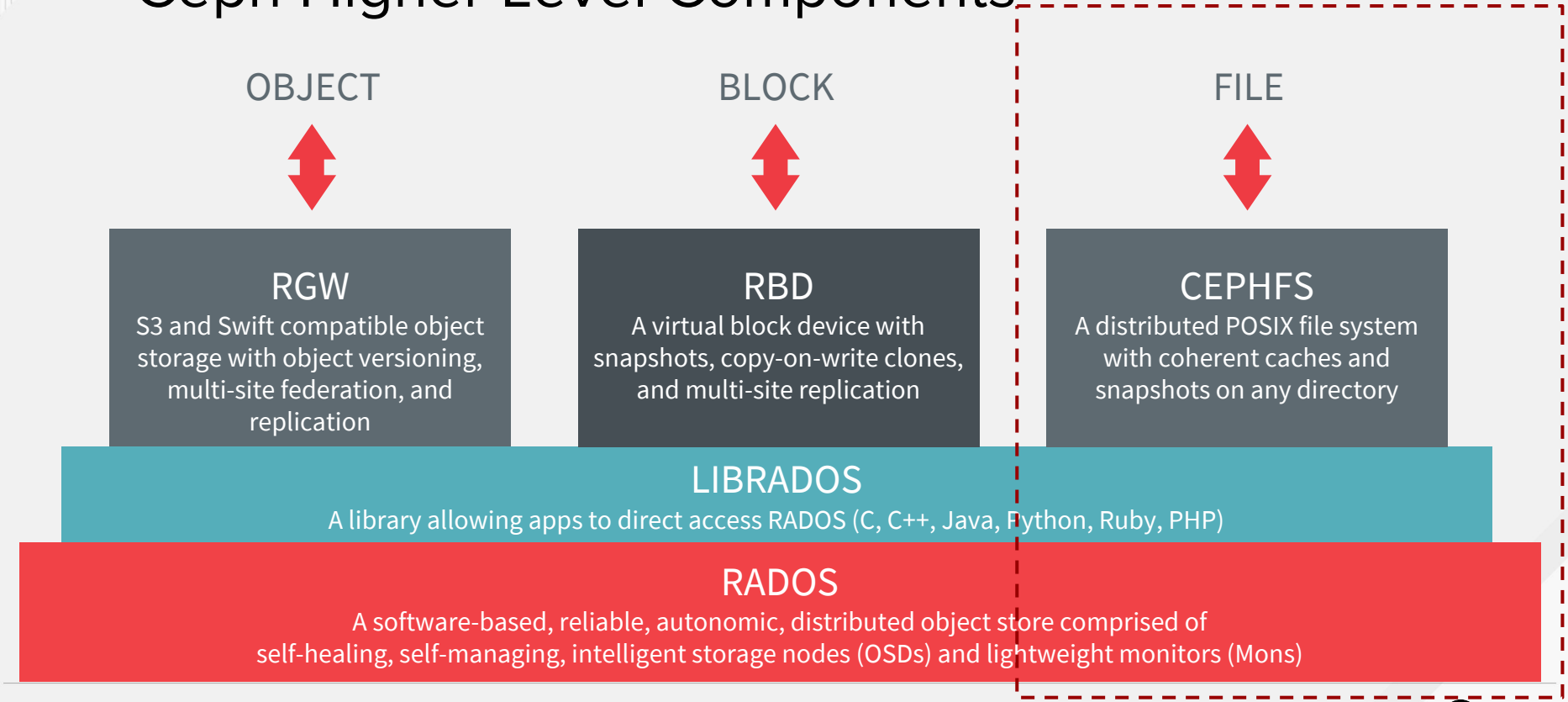
Patrick Donnelly
CephFS Engineer
Red Hat, Inc.

Tom Barron
Manila Engineer
Red Hat, Inc.

Ramana Raja
CephFS Engineer
Red Hat, Inc.

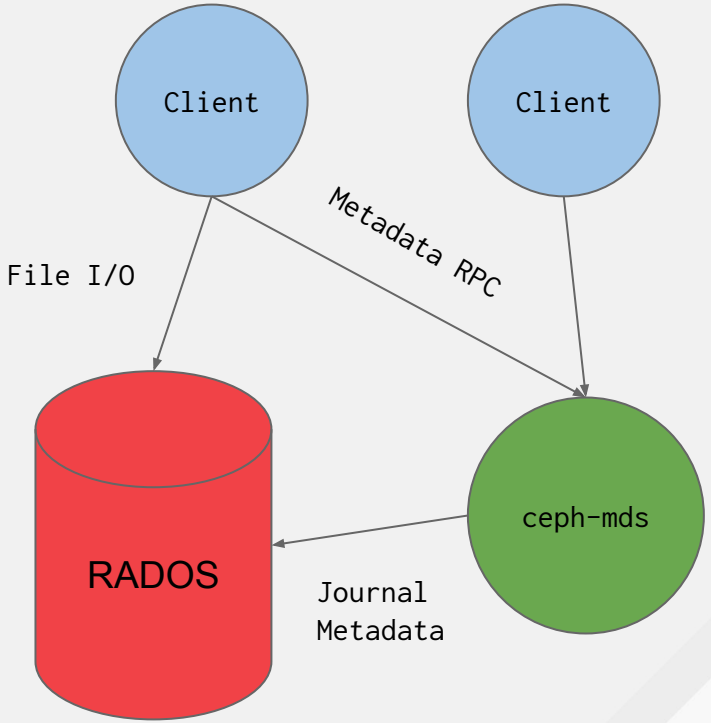
How do we solve this?

Ceph Higher Level Components

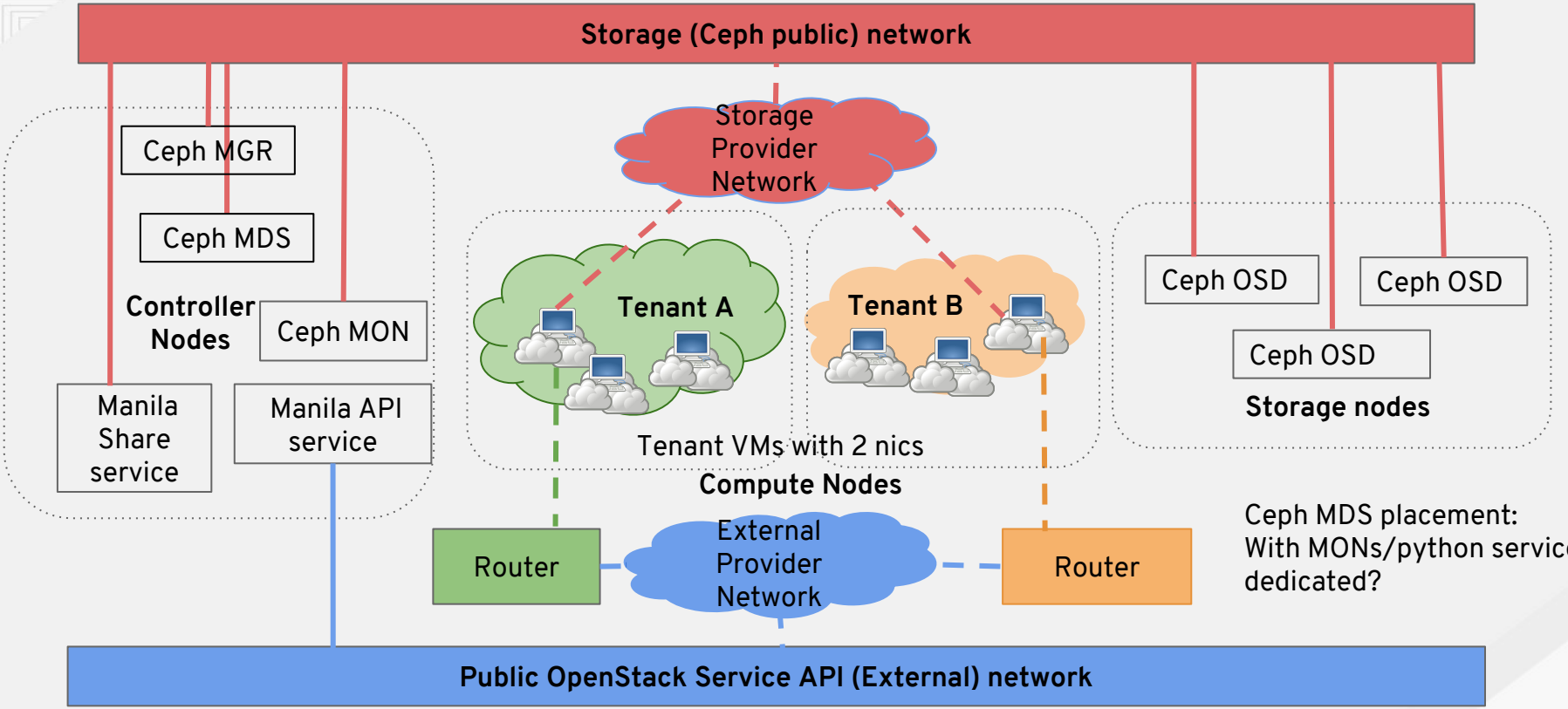


What is CephFS?

- CephFS is a POSIX-compatible distributed file system!
- 3 moving parts: the MDS(s), the clients, and RADOS.
- Files, directories, and other metadata are stored in RADOS. The MDS has no local state.
- Mount using:
  - FUSE: `ceph-fuse ...`
  - Kernel: `mount -t ceph ...`
- Coherent caching across clients. MDS issues inode capabilities which enforce synchronous or buffered writes.
- Clients access data directly via RADOS.



CephFS native driver deployment

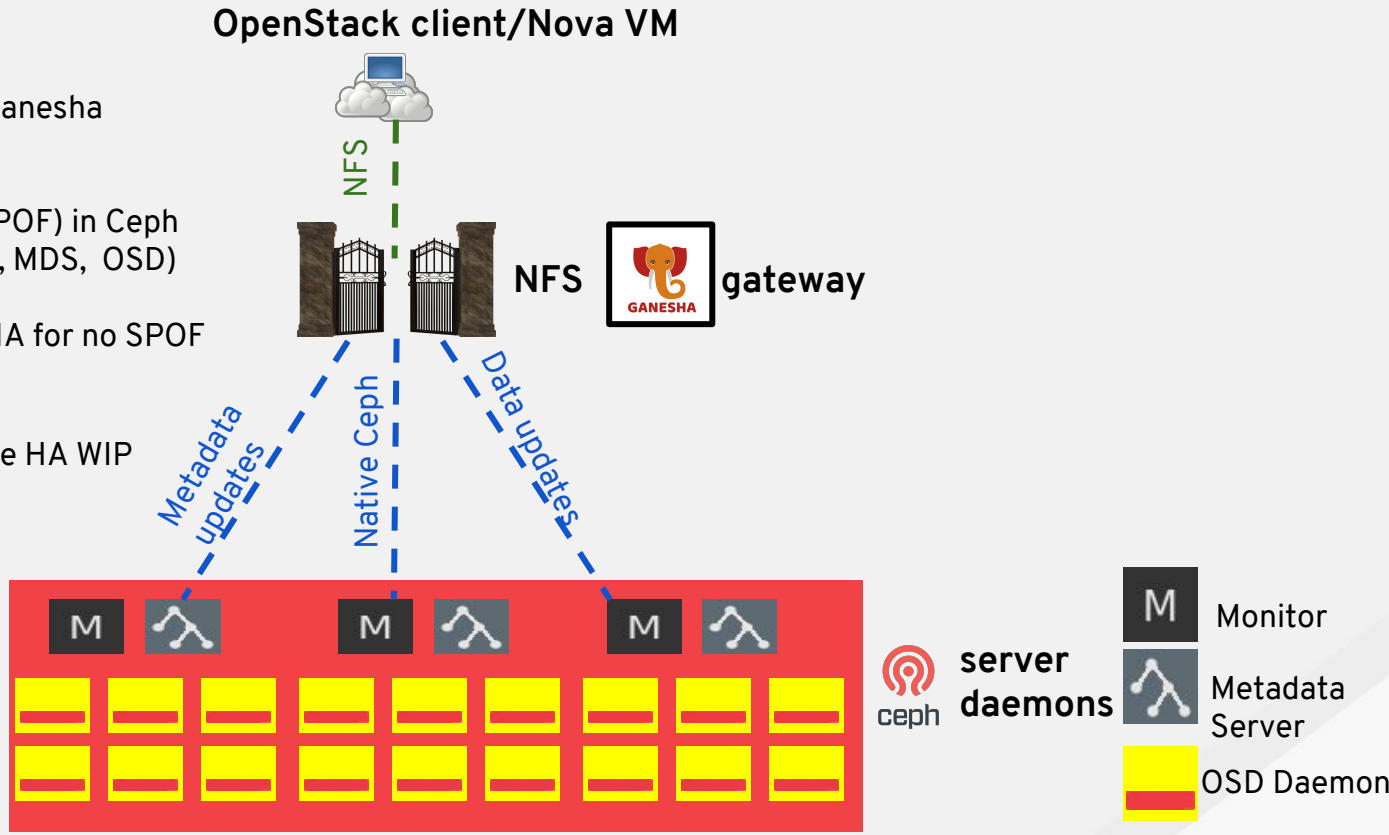


Ceph MDS placement:
With MONs/python services/
dedicated?

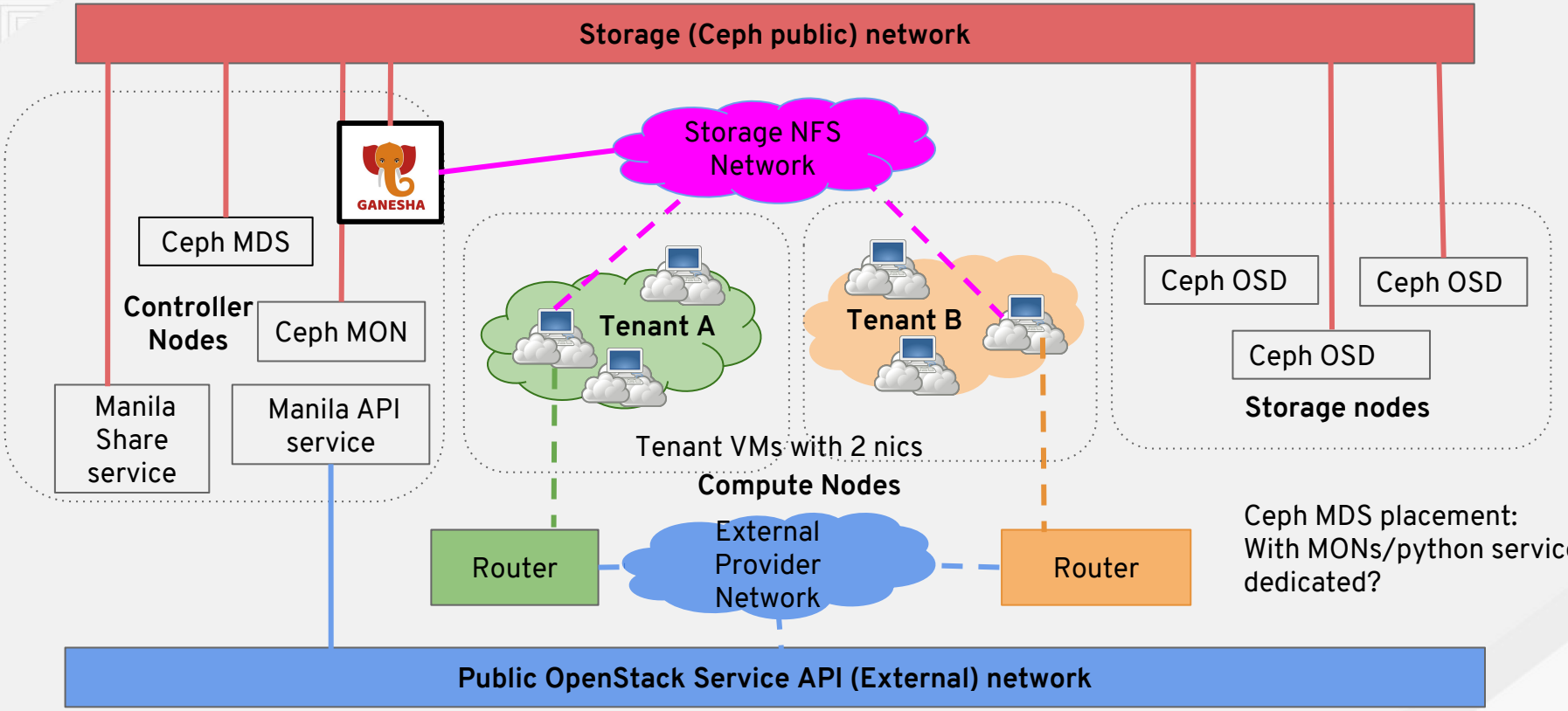
https://docs.openstack.org/manila/ocata/devref/cephfs_native_driver.html

CephFS NFS driver (in data plane)

- Clients connected to NFS-Ganesha gateway. Better security.
- No single point of failure (SPOF) in Ceph storage cluster (HA of MON, MDS, OSD)
- NFS-Ganesha needs to be HA for no SPOF in data plane.
- NFS-Ganesha active/passive HA WIP (Pacemaker/Corosync)

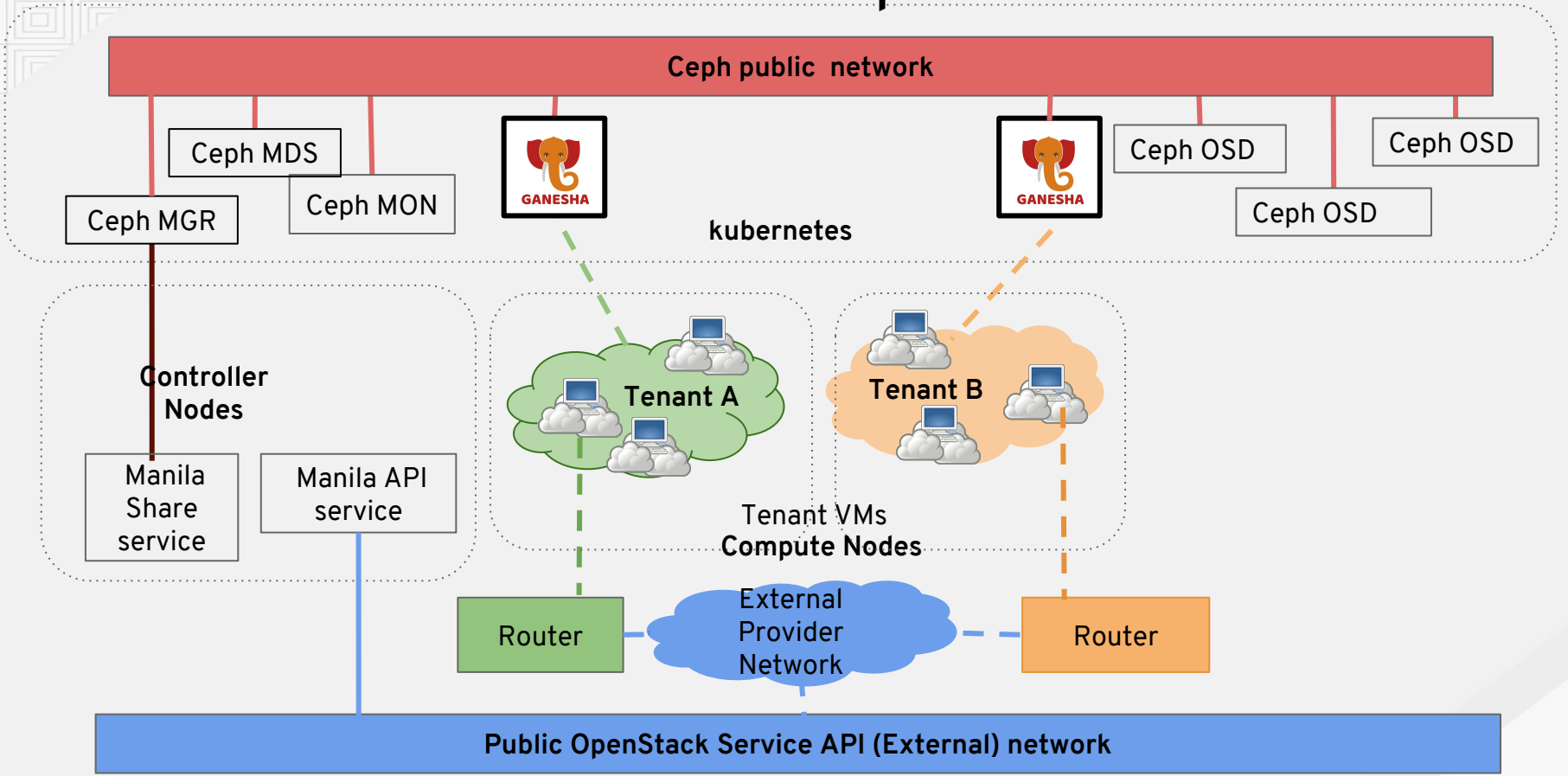


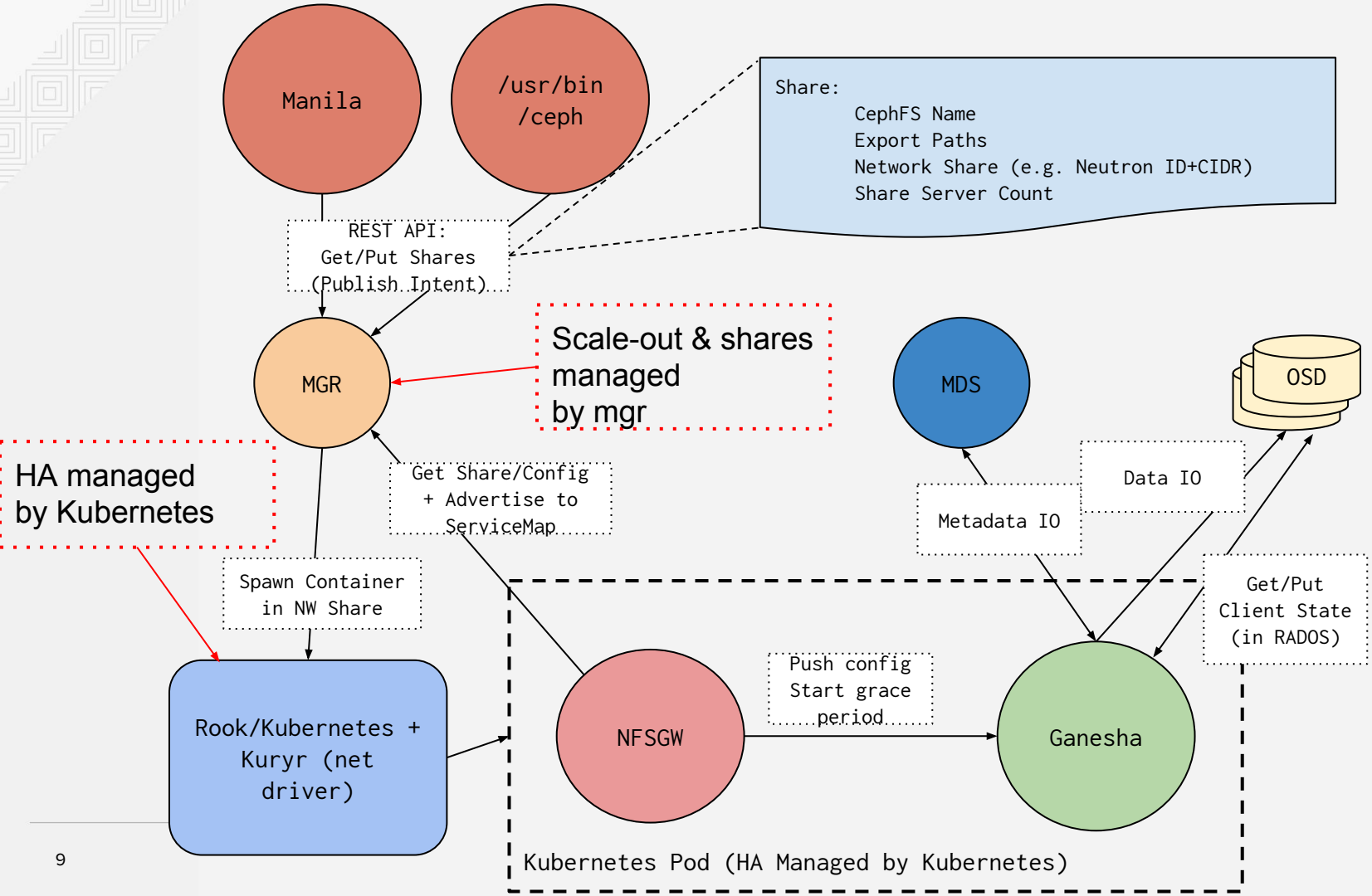
CephFS NFS driver deployment



https://docs.openstack.org/manila/queens/admin/cephfs_driver.html

Future: Ganesha per Tenant

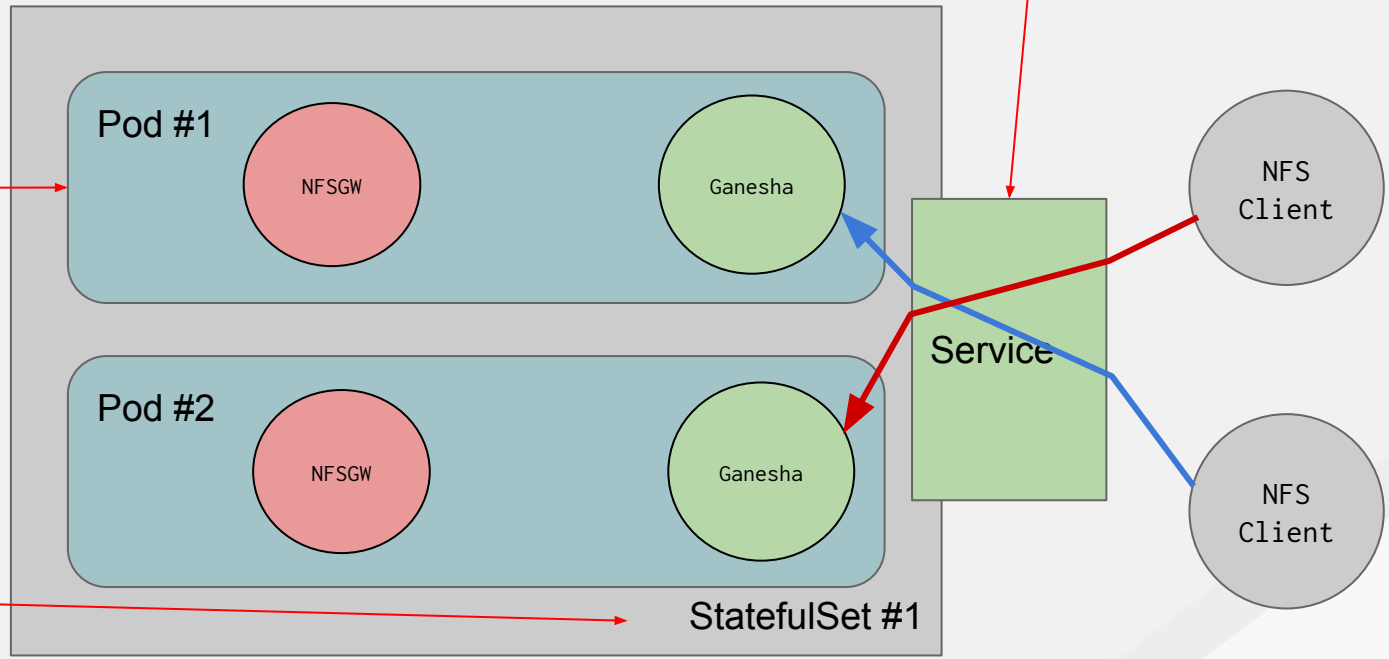



Manila
/usr/bin
/ceph
Share:
CephFS Name
Export Paths
Network Share (e.g. Neutron ID+CIDR)
Share Server Count
REST API:
Get/Put Shares
(Publish Intent)
MGR
Scale-out & shares
managed
by mgr
MDS
OSD
HA managed
by Kubernetes
Get Share/Config
+ Advertise to
ServiceMap
Metadata IO
Data IO
Spawn Container
in NW Share
Get/Put
Client State
(in RADOS)
Rook/Kubernetes +
Kuryr (net
driver)
NFSGW
Push config
Start grace
period
Ganesha
Kubernetes Pod (HA Managed by Kubernetes)

Managing Scale-Out: “CephFSShareNamespace”

One IP to access NFS Cluster for tenant; only via tenant network

# of pods equal to scale-out for “CephFSShareNamespace”; dynamically grow/shrink??



Stable network identifiers for pods!

Thanks!

Patrick Donnelly

pdonnell@redhat.com

Thanks to the CephFS team: John Spray, Greg Farnum, Zheng Yan, Ramana Raja, Doug Fuller, Jeff Layton, and Brett Niver.

Homepage: http://ceph.com/

Mailing lists/IRC: http://ceph.com/IRC/