



redhat.

# Distributed File Storage in Multi-Tenant Clouds using CephFS

FOSDEM 2018

John Spray  
Software Engineer  
Ceph

Christian Schwede  
Software Engineer  
OpenStack Storage

# In this presentation

## Brief overview of key components

What is OpenStack Manila

## CephFS Native Driver

CephFS driver implementation (available since OpenStack Newton)

## NFS Ganesha Driver

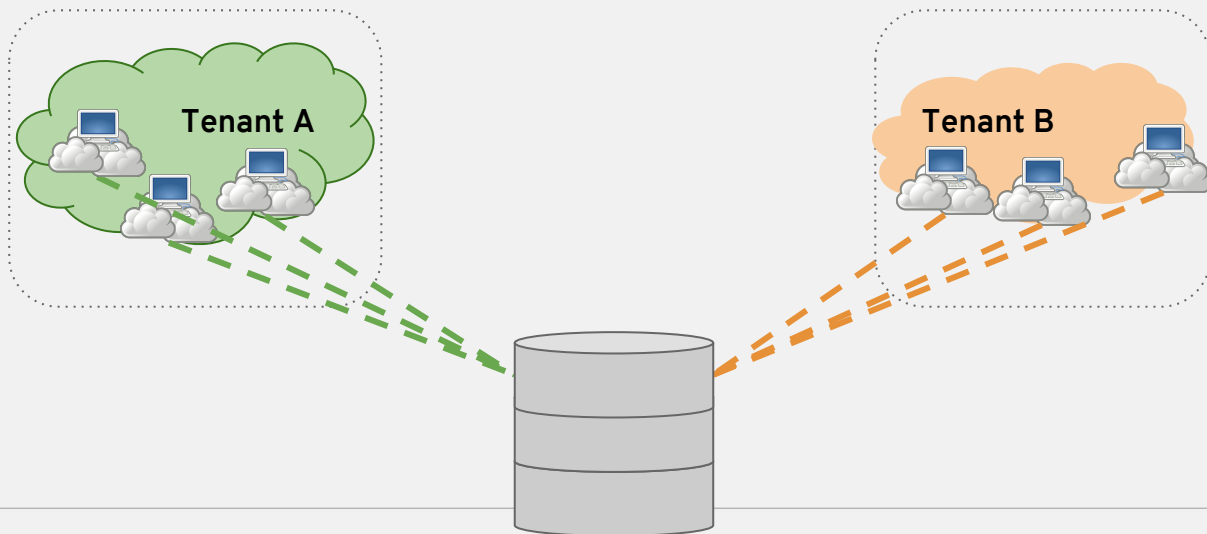
NFS backed with CephFS driver implementation (available since OpenStack Queens)

## Future work

OpenStack Queens and beyond

# What's the challenge?

- Want: a filesystem that is shared between multiple nodes
- At the same time: tenant aware
- Self-managed by the tenant admins



# How do we solve this?

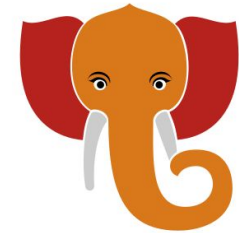


**MANILA**

*an OpenStack Community Project*



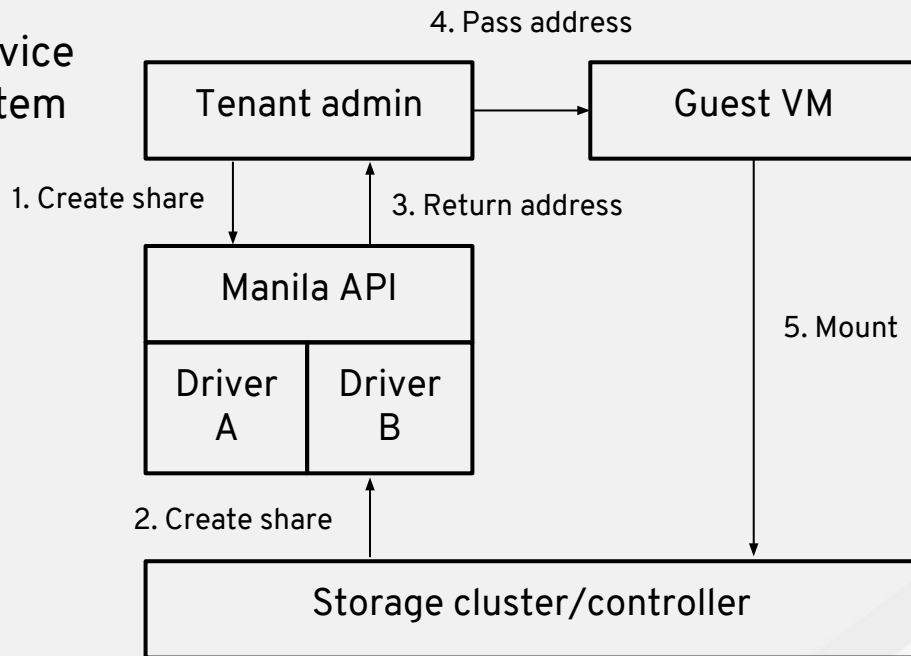
**ceph**



**GANESHA**

# OpenStack Manila

- OpenStack Shared Filesystems service
- APIs for tenants to request file system shares
- Support for several drivers
  - Proprietary
  - CephFS
  - “Generic” (NFS on Cinder)



# CephFS

OBJECT



**RGW**

S3 and Swift compatible object storage with object versioning, multi-site federation, and replication

BLOCK



**RBD**

A virtual block device with snapshots, copy-on-write clones, and multi-site replication

FILE



**CEPHFS**

A distributed POSIX file system with coherent caches and snapshots on any directory

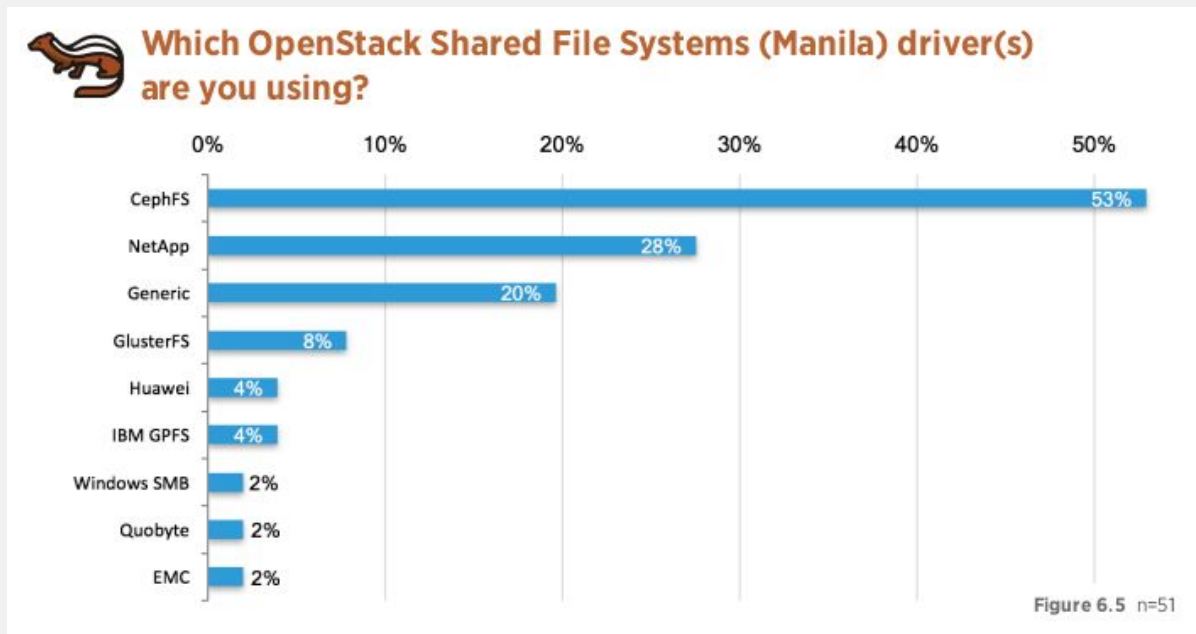
**LIBRADOS**

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

**RADOS**

A software-based, reliable, autonomic, distributed object store comprised of self-healing, self-managing, intelligent storage nodes (OSDs) and lightweight monitors (Mons)

# Why integrate CephFS with Manila/Openstack?



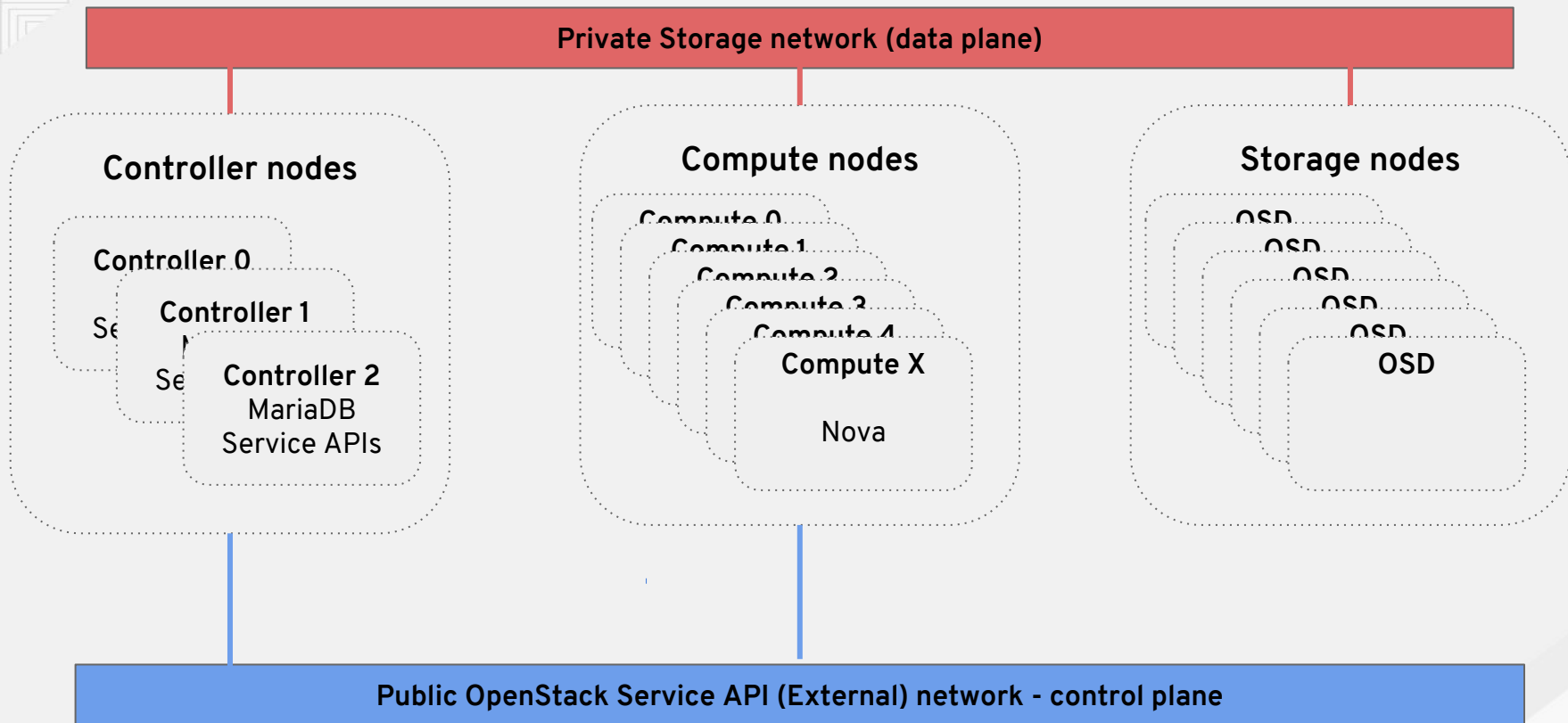
Most Openstack users are also running a Ceph cluster already

Open source storage solution

CephFS metadata scalability is ideally suited to cloud environments.


<https://www.openstack.org/user-survey/survey-2017>

# Break-in: terms





# CephFS native driver\*

Since OpenStack Mitaka release and  jewel  
ceph

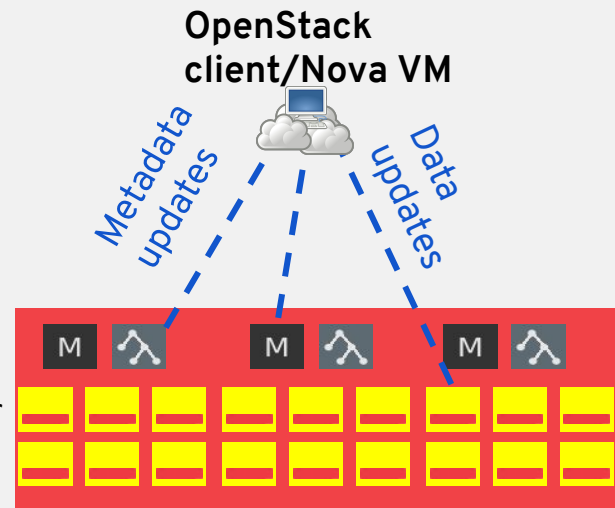
- \* for OpenStack private clouds, helps trusted Ceph clients use shares backed by CephFS backend through native CephFS protocol

# First approach: CephFS Native Driver

## Since Openstack Mitaka

- Best Performance
- Access to all CephFS features
- Simple deployment and implementation

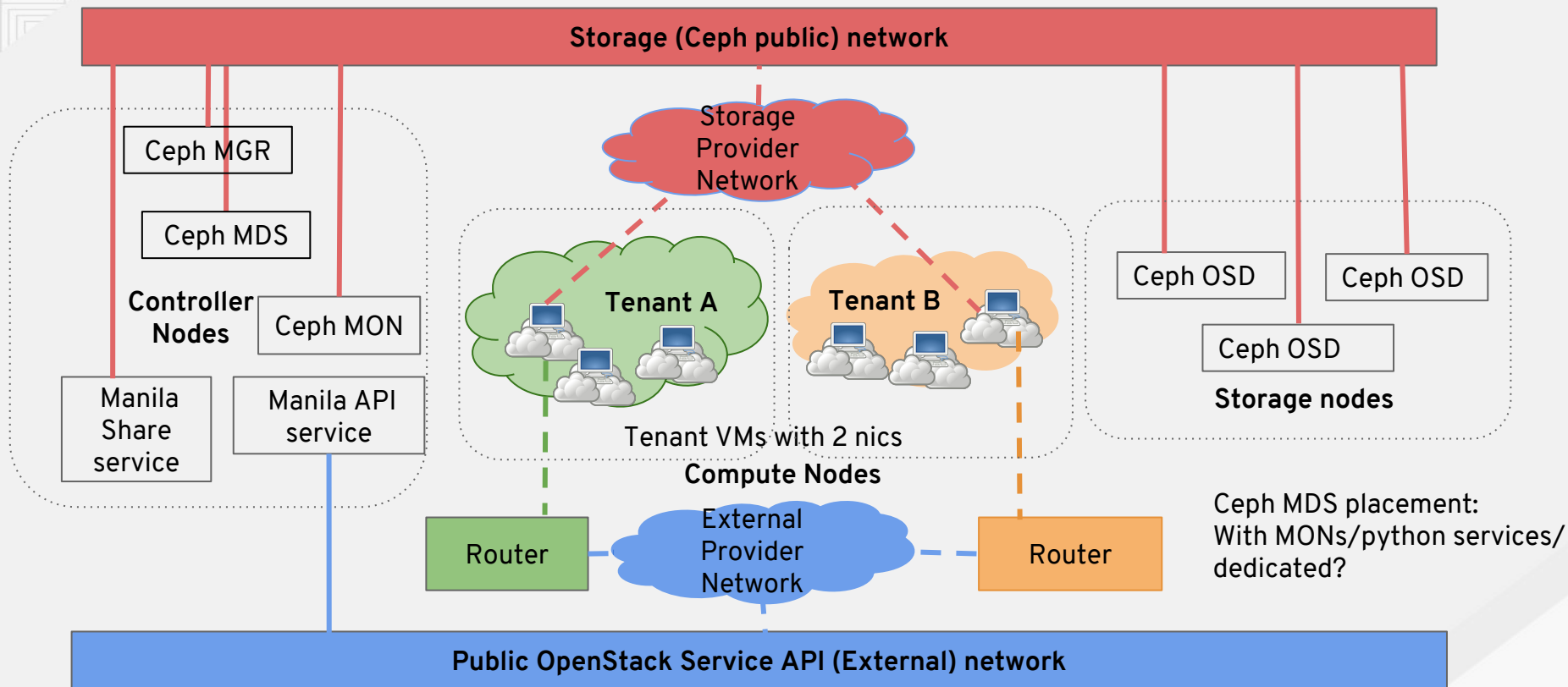
## Ceph server daemons



Manila on CephFS at CERN: The Short Way to Production by Arne Wiebalck

<https://www.openstack.org/videos/boston-2017/manila-on-cephfs-at-cern-the-short-way-to-production>

# CephFS native driver deployment



# CephFS Native Driver

## Pros

- Performance!
- Success stories, popular!
- Simple implementation.
- Makes HA relatively easy.

# CephFS Native Driver

## Cons

- User VMs have direct access to the storage network using ceph protocols.
- Needs client side cooperation.
- Share size quotas support only with Ceph FUSE clients
- Assumes trusted user VMs.
- Requires special client and key distribution.

# CephFS NFS driver\*

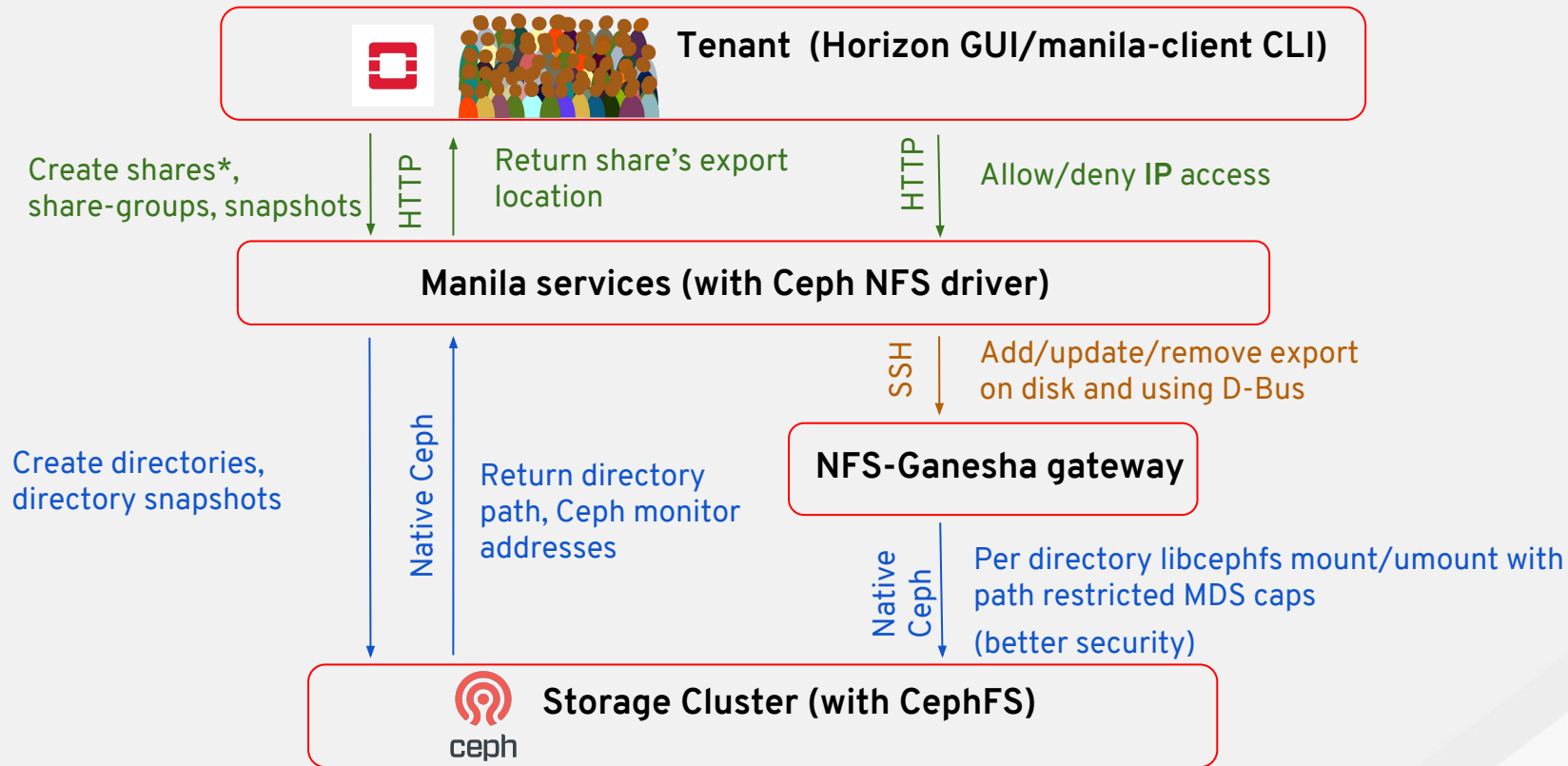
*Full debut in OpenStack Queens, with  luminous, NFS-Ganesha v2.54, Ceph Ansible 3.1.*

\* for OpenStack clouds, helps NFS clients use the CephFS backend via NFS-Ganesha gateways

# NFS Ganesha

- User-space NFSv2, NFSv3, NFSv4, NFSv4.1 and pNFS server
- Modular architecture: Pluggable File System Abstraction Layer allow for various storage backend (e.g. glusterfs, cephfs, gpfs, Lustre and more)
- Dynamic export/unexport/update with DBUS
- Can manage huge metadata caches
- Simple access for other user-space services (e.g. KRB5, NIS, LDAP)
- Open source

# CephFS NFS driver (in control plane)



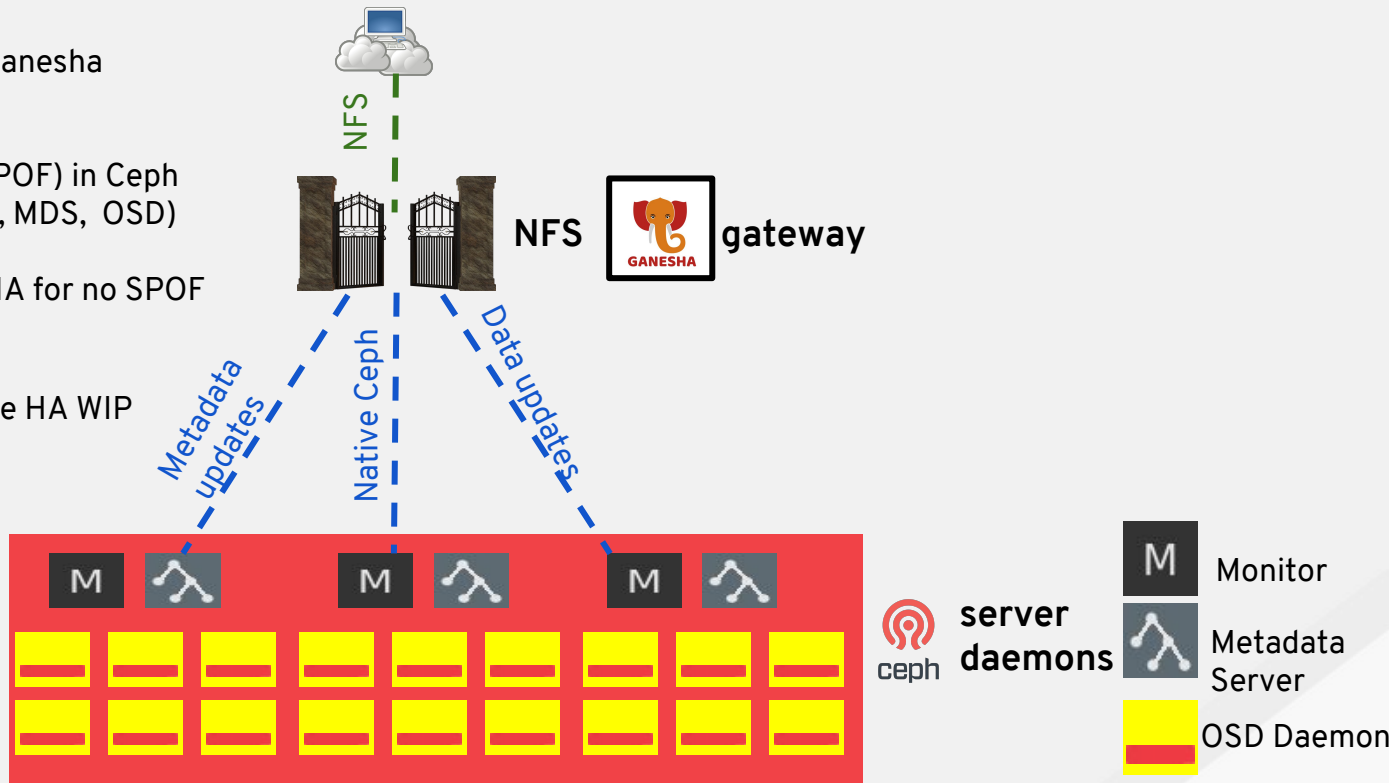
\* manila share = a CephFS dir + quota + unique RADOS name space



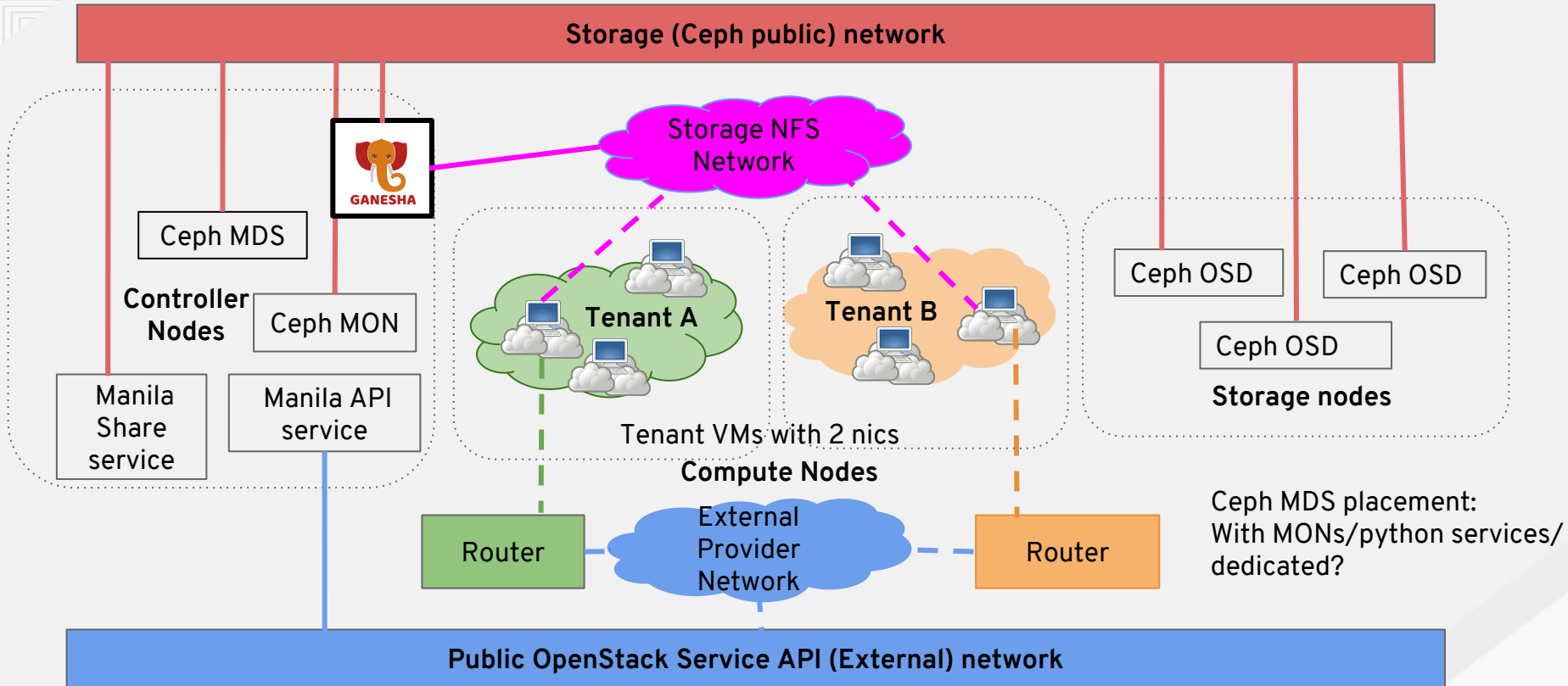
# CephFS NFS driver (in data plane)

OpenStack client/Nova VM

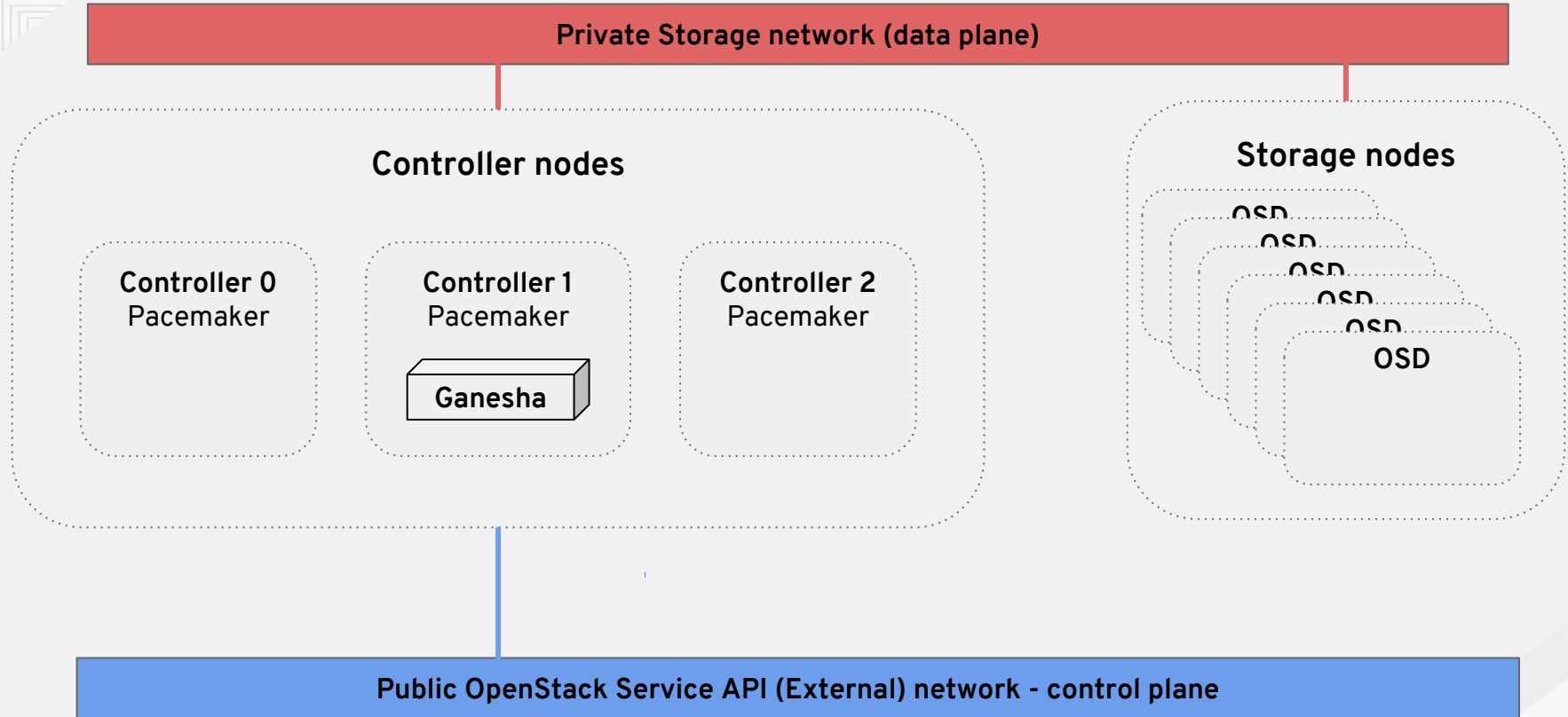
- Clients connected to NFS-Ganesha gateway. Better security.
- No single point of failure (SPOF) in Ceph storage cluster (HA of MON, MDS, OSD)
- NFS-Ganesha needs to be HA for no SPOF in data plane.
- NFS-Ganesha active/passive HA WIP (Pacemaker/Corosync)



# CephFS NFS driver deployment



# OOO, Pacemaker, containers, and Ganesha



# Current CephFS NFS Driver

## Pros

- Security: isolates user VMs from ceph public network and its daemons.
- Familiar NFS semantics, access control, and end user operations.
- Large base of clients who can now use Ceph storage for file shares without doing anything different.
  - NFS supported out of the box, doesn't need any specific drivers
- Path separation in the backend storage and network policy (enforced by neutron security rules on a dedicated StorageNFS network) provide multi-tenancy support.

# Current CephFS NFS Driver

## Cons

- Ganesha is a “man in the middle” in the data path and a potential performance bottleneck.
- HA using the controller node pacemaker cluster impacts our ability to scale
- As does the (current) inability to run ganesha active-active, and
- We’d like to be able to spawn ganesha services on demand, per-tenant, as required rather than statically launching them at cloud deployment time.

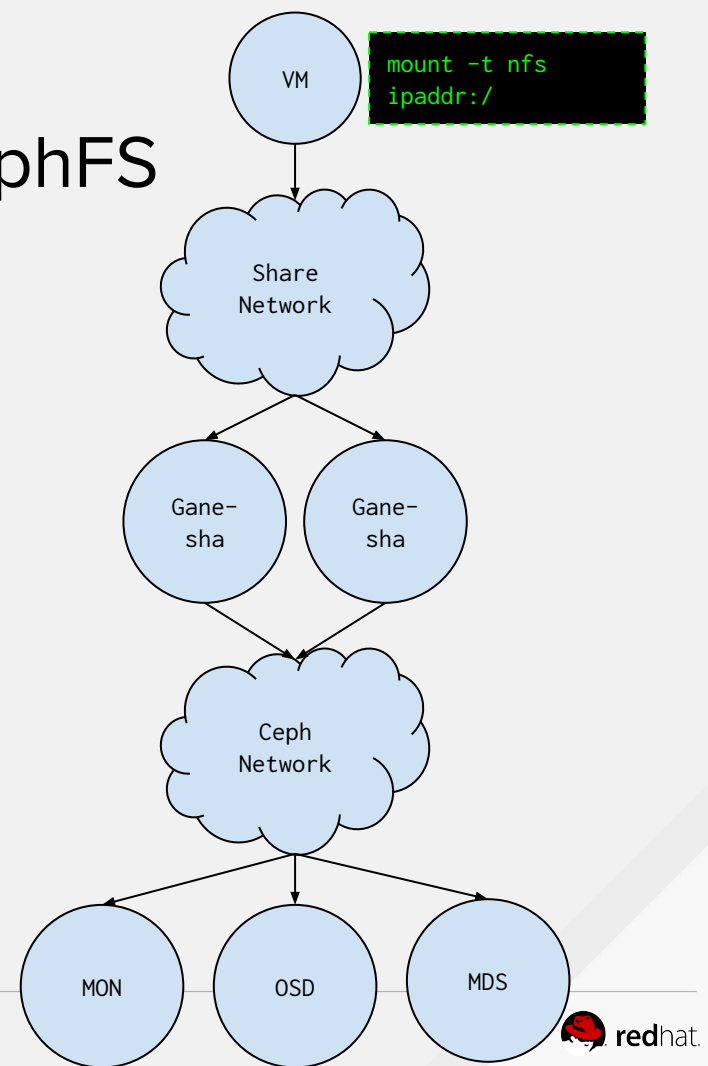


What lies ahead ...

# Next Step: Integrated NFS Gateway in Ceph to export CephFS

- Ganesha becomes an integrated NFS Gateway to the Ceph file system.
  - Targets deployments beyond Openstack which need a gateway client to the storage network (e.g. standalone appliance, kerberos, openstack, etc.)
  - Provides an alternative and stable client to avoid legacy kernels or FUSE.
  - Gateways/secures access to the storage cluster.
  - Overlays Ganesha potential enhancements (e.g. Kerberos)

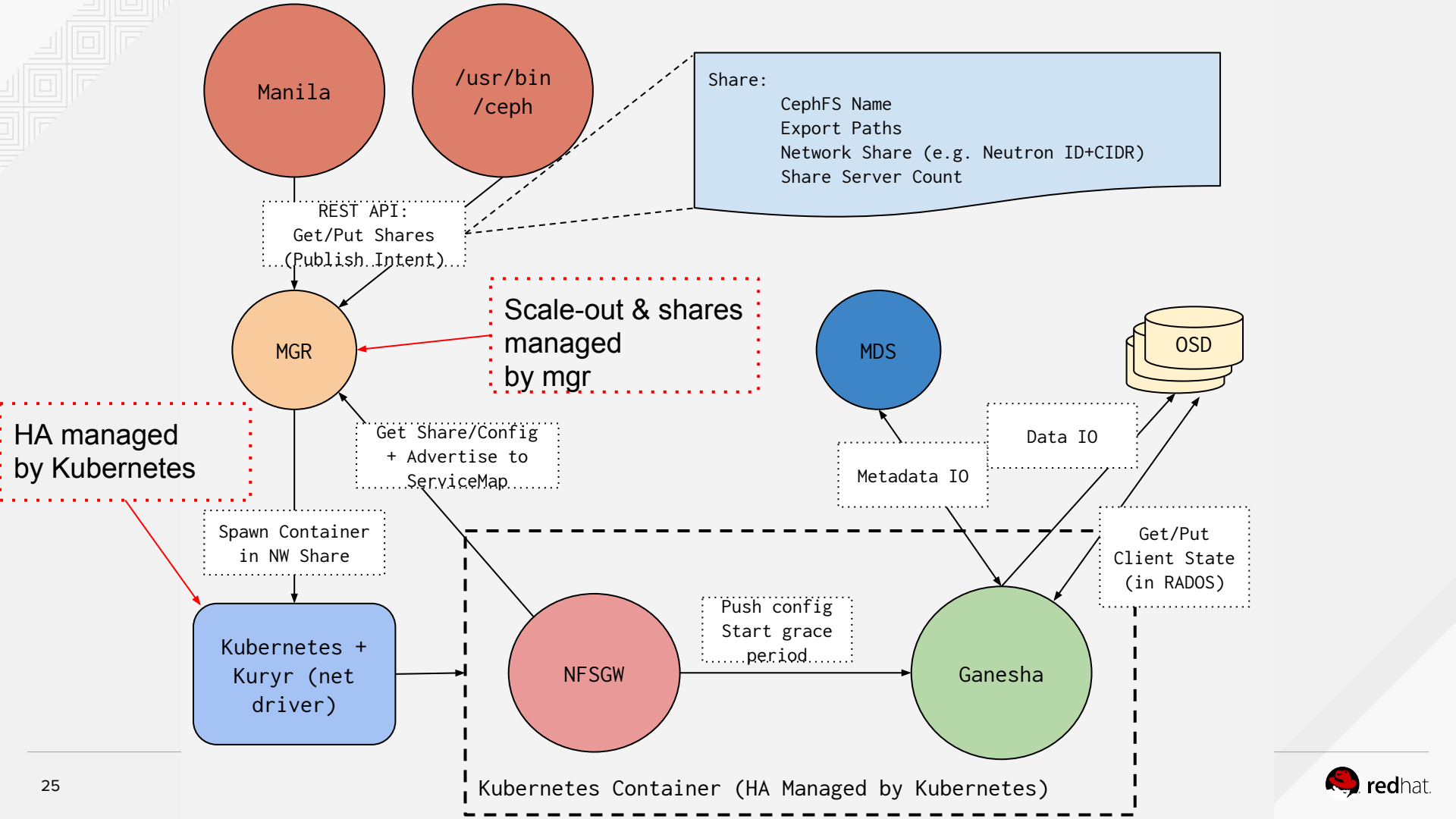
See also John Spray's talk at Openstack in Apr 2016:  
<https://www.youtube.com/watch?v=vt4XUQWetq0&t=1335>



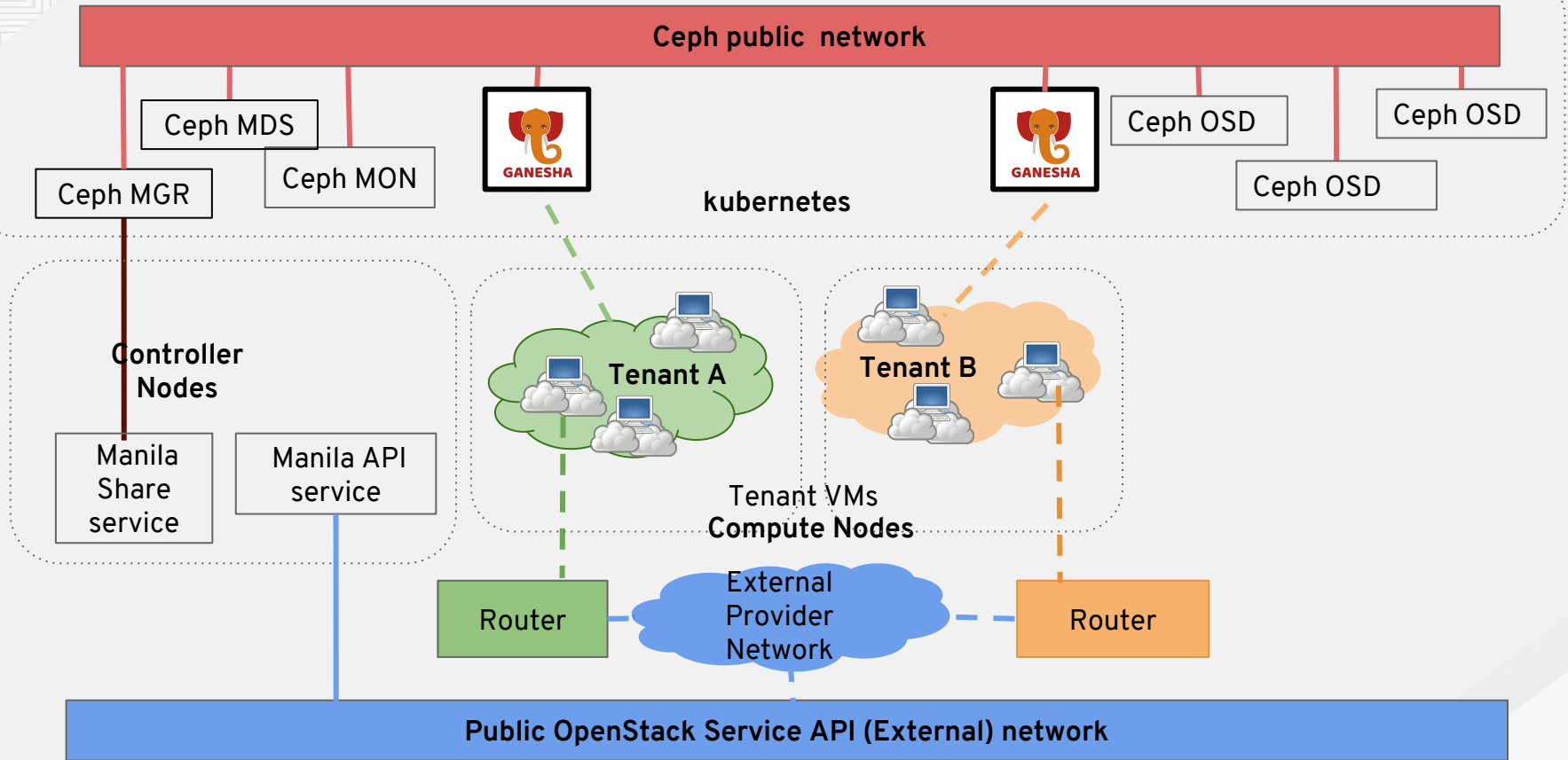
# HA and Scale-Out

- **High Availability**
  - **Kubernetes** managed Ganesha container
    - Container life-cycle and resurrection not managed by Ceph.
    - **ceph-mgr** creates shares and launches containers through Kubernetes
- **Scale-Out (avoid Single Point of Failure)**
  - **ceph-mgr** creates multiple Ganesha containers for a share.
  - (Potentially) **Kubernetes** load balancer allows for automatic multiplexing between Ganesha containers via a single service IP.





# Future: trivial to have Ganesha per Tenant



# Challenges and Lingering Technical Details

- How to recover Ganesha state in the MDS during failover (opened files; delegations)
  - One solution: put all Ganesha shares into grace during failover to prevent lock/capability theft. (Heavy weight approach)
  - → Preserve CephFS capabilities for takeover on a timeout; introduce sticky client IDs
    - Need a mechanism to indicate to CephFS that state reclamation by the client is complete.
    - Need to handle cold (re)start of the Ceph cluster where state held by the client (Ganesha) was lost by the MDS cluster (need to put entire Ganesha cluster in grace while state is recovered).

# Further future

- Performance:
  - Exploit **MDS scale-out**
  - **NFS delegations**
  - **pNFS**
- Container environments:
  - Implementing **Kubernetes** Persistent Volume Claims
  - Re-using underlying NFS/networking model
  - Perhaps even re-use Manila itself **outside of OpenStack**

# Thanks !

John Spray  
[jspray@redhat.com](mailto:jspray@redhat.com)

Christian Schwede  
[cschwede@redhat.com](mailto:cschwede@redhat.com)

# Links

- Openstack Talks
  - Sage Weil, “The State of Ceph, Manila, and Containers in OpenStack”, OpenStack Tokyo Summit 2015: <https://www.youtube.com/watch?v=dNTCBouMaAU>
  - John Spray, “CephFS as a service with OpenStack Manila”, OpenStack Austin Summit 2016: <https://www.youtube.com/watch?v=vt4XUQWetg0>
  - Ramana Raja, Tom Barron, Victoria Martinez de la Cruz, “CephFS Backed NFS Share Service for Multi-Tenant Clouds”, OpenStack Boston 2017: <https://www.youtube.com/watch?v=BmDv-iQLv8c>
  - Patrick Donnelly, “Large-scale Stability and Performance of the Ceph File System”, Vault 2017: <https://docs.google.com/presentation/d/1X13IVeEtQUc2QRJ1zuzibJEUhHg0cdZcBYdiMzOOqLY>
  - Sage Weil et al., “Ceph: A Scalable, High-Performance Distributed File System”:  
<https://dl.acm.org/citation.cfm?id=1298485> Sage Weil et al., “Panel Experiences Scaling File Storage with CephFS and OpenStack” <https://www.youtube.com/watch?v=IPhKEi3aRPg>
- CERN
  - Storage Overview - <http://cern.ch/go/976X>
  - Cloud Overview - <http://cern.ch/go/6HID>
  - Blog - <http://openstack-in-production.blogspot.fr>