




Present and Future of Ceph integration with OpenStack and k8s



Francesco Pantano
@fmount9
<https://fmount.me>

Agenda

- Ceph / OpenStack Integration Overview
- State of the art (ceph-ansible vs cephadm)
- cephadm based deployment
- Towards kubernetes (Rook/OpenStack in pods)
- Demo

The openstack project

OpenStack is a [free, open standard cloud computing](#) platform (**IaaS**).

The software platform consists of interrelated components that control diverse, multi-vendor hardware pools of processing, storage, and networking resources.



In July 2010, [Rackspace Hosting](#) and [NASA](#) announced an open-source cloud-software initiative known as OpenStack.

Mission statement: *"to produce the ubiquitous Open Source Cloud Computing platform that will meet the needs of public and private clouds regardless of size, by being simple to implement and massively scalable".*



BARBICAN



DESIGNATE



HEAT



KEYSTONE



IRONIC



CINDER



SWIFT



NOVA



MISTRAL



NEUTRON



OCTAVIA



GLANCE



KURYR



MANILA

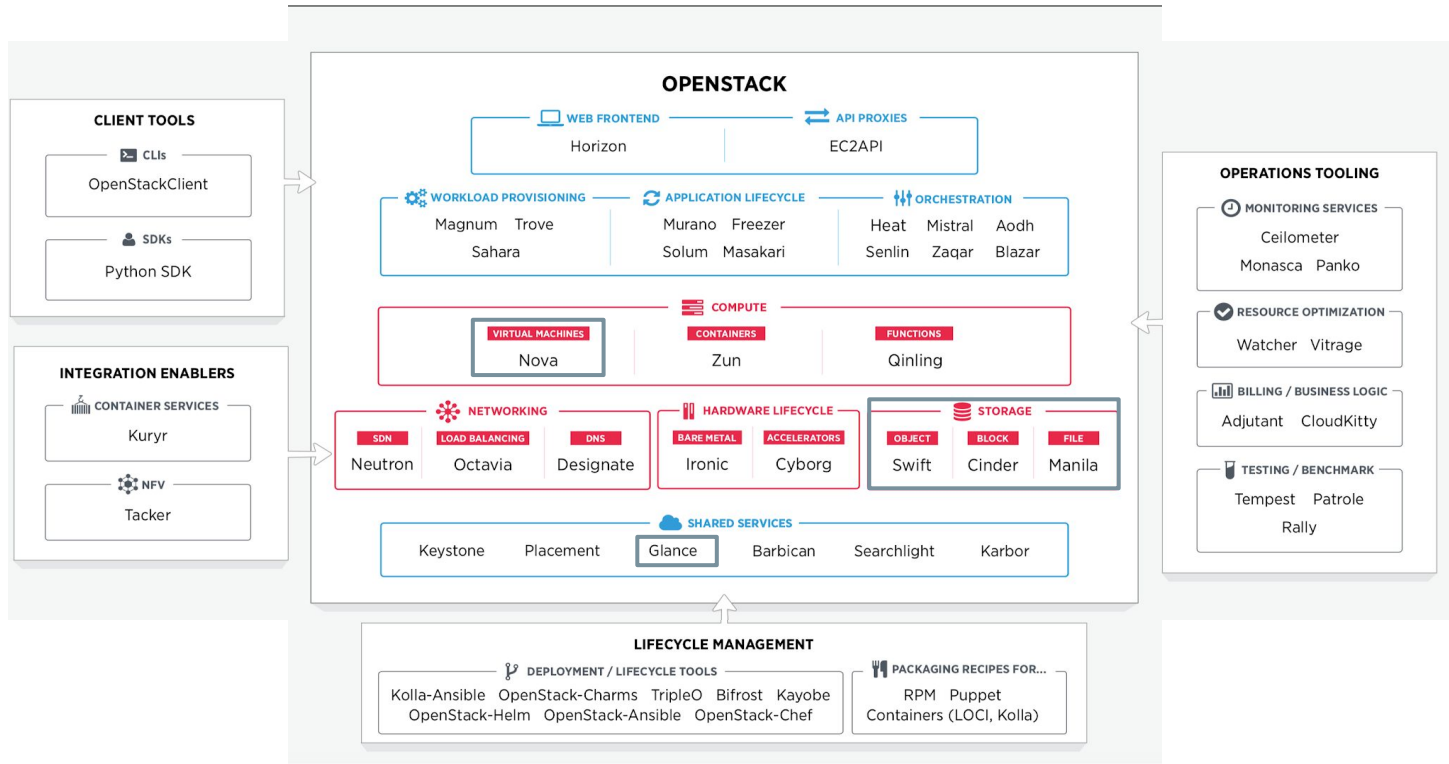


HORIZON



PLACEMENT

The openstack project






Ceph Integration with OpenStack

Why?

- Ceph is the most popular OpenStack storage backend
 - Make it easy for operators to use both
- There is an interest for Hyper Converged Infrastructure (HCI)
 - HCI = running storage + compute on the same node

Deployment tools/strategies

- Puppet-ceph 
-  Tripleo-ansible
- Kolla-ansible 
- ceph-ansible → cephadm



which is the Ceph attach rate in the OpenStack world



It is difficult to determine the exact attach rate of Ceph in the OpenStack world, as it depends on the specific deployment and the preferences of the organization. However, Ceph is widely regarded as a popular storage backend for OpenStack deployments, with many organizations choosing it for its scalability, reliability, and flexibility. In a survey conducted by the OpenStack Foundation in 2018, Ceph was the most commonly used storage backend for OpenStack, with a usage rate of 33%. It is likely that Ceph's attach rate in the OpenStack world has continued to grow since then.



how much the ceph integration in openstack is relevant



Ceph integration with OpenStack is very relevant, as it provides a highly scalable and reliable storage solution for OpenStack deployments. Ceph's distributed architecture allows it to scale horizontally, making it well-suited for OpenStack environments where there is a need to store large amounts of data. Ceph's integration with OpenStack also allows for the creation of distributed storage clusters, which can provide added resiliency and reliability for OpenStack deployments. In addition, Ceph's ability to support multiple storage protocols, such as block, object, and file storage, makes it a versatile storage solution for OpenStack environments. Overall, Ceph's integration with OpenStack is an important aspect of many OpenStack deployments, as it provides a robust and reliable storage solution.

Deployment workflows

ceph-ansible vs cephadm

ceph-ansible:

- The status of the cluster config is maintained at (ceph-)ansible level (vars defined and updated by **human** operators)
- Infrastructure playbooks are developed and are supposed to
 - Scale up/down the cluster nodes (e.g. shrink or add new services)
 - Perform updates/upgrades
 - Help operators on day2 operations



cephadm:

- The status of the cluster is maintained within the Ceph Orchestrator component
 - it continuously runs multiple checks and if the status of the cluster changes (e.g. a new osd is detected), it's able to take an action (e.g. run a new osd daemon)



Deployment workflow

Status of the integration with cephadm

Provision Networks

```
openstack overcloud network provision ...
```

Provision ALL nodes

```
openstack overcloud node provision ...
```

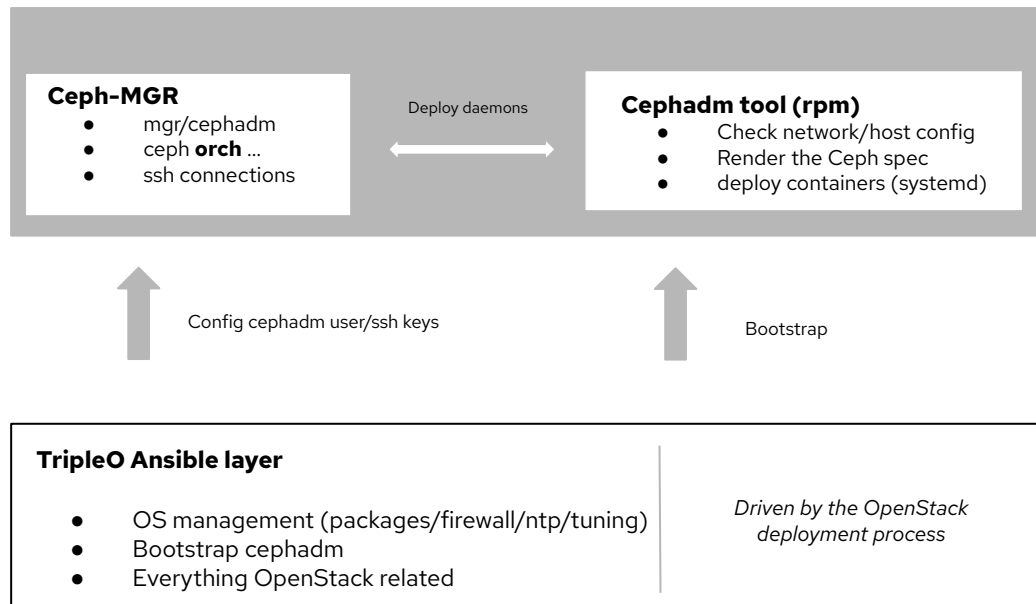
Deploy Base Ceph Cluster (RBD ready)

```
openstack overcloud ceph deploy ...
```

Deploy OSP &

```
openstack overcloud deploy --templates \  
... -e ..... /cephadm.yaml
```

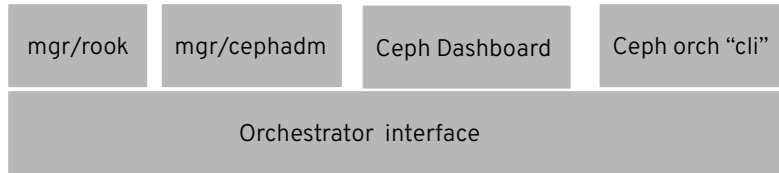
Finalize RHCS deploy Configures OSP



Demo: <https://asciinema.org/a/414971>

How do cephadm and orchestrator work?

- It uses ssh to connect to the nodes
- It doesn't rely on external configuration tools such as Ansible, Rook, and Salt.



Declarative approach

```
ceph orch apply rgw default '--placement=label:rgw  
count-per-host:1' --port=8080
```

```
---  
networks:  
- 1.2.3.0/24  
placement:  
  label:  
  - rgw  
  service_id: rgw.default  
  service_name: rgw.default  
  service_type: rgw  
spec:  
  rgw_frontend_port: 8080  
  rgw_realm: default  
  rgw_zone: default
```

Note:

Cephadm, within the Ceph codebase, represents the users endpoint and exposes CLI options to help operators have a better experience: this tool uses an interface through the orchestrator mgr module, which is responsible for managing the lifecycle of the Ceph components other than maintaining the status of the cluster

cephadm mini howto

1. Bootstrap the first node
 - `cephadm bootstrap --mon-ip *<mon-ip>*`
 - You now have one running Ceph Monitor and Manager
2. Distribute an **SSH keypair** to other nodes which cephadm can use
3. Add the other nodes and services by applying a spec
 - `ceph orch apply --in-file spec.yaml`

Day2 ops



Ceph Health Status

```
[admin@controller-0 ~]$ sudo cephadm shell -c /etc/ceph/ceph.conf -k /etc/ceph/ceph.client.admin.keyring -v /home/ceph-admin/specs:/specs
```

```
[ceph: root@controller-0 /]# ceph -s
```

```
cluster:
  id:          900f1b9f-df35-5d4f-969b-0096c5d78a93
  health: HEALTH_OK

services:
  mon: 1 daemons, quorum controller-0 (age 3d)
  mgr: controller-0.ujtfs(active, since 3d)
  osd: 16 osds: 16 up (since 3d), 16 in (since 3d)

data:
  pools:   7 pools, 177 pgs
  objects: 4 objects, 449 KiB
  usage:   216 MiB used, 512 GiB / 512 GiB avail
  pgs:    177 active+clean/
```

per-node daemons list

```
[ceph: root@controller-0 /]# ceph orch ps
```

NAME	HOST	PORTS	STATUS	REFRESHED	AGE	MEM	MEM LIM	VERSION	IMAGE ID	CONTAINER ID
crash.controller-0	controller-0		running (3d)	2m ago	3d	6405k	-	16.2.9-10469-g29e1fc17	f7e1b63ced56	14eafd7cd3d5
mgr.controller-0.ujtfs	controller-0	*:9283	running (3d)	2m ago	3d	543M	-	16.2.9-10469-g29e1fc17	f7e1b63ced56	64f1b0491b14
mon.controller-0	controller-0		running (3d)	2m ago	3d	441M	2048M	- 16.2.9-10469-g29e1fc17	f7e1b63ced56	db68aff25363
osd.0	cephstorage-0		running (3d)	2m ago	3d	77.3M	1262M	- 16.2.9-10469-g29e1fc17	f7e1b63ced56	87e05fd8043c
osd.1	cephstorage-0		running (3d)	2m ago	3d	78.7M	1262M	- 16.2.9-10469-g29e1fc17	f7e1b63ced56	0053aff39b80
...										
...										

List the enrolled Hosts

```
[ceph: root@controller-0 /]# ceph orch host ls
```

HOST	STATUS	ADDR	LABELS
cephstorage-0	192.168.24.31	osd	
cephstorage-1	192.168.24.9	osd	
cephstorage-2	192.168.24.18	osd	
cephstorage-3	192.168.24.46	osd	
controller-0	192.168.24.14	_admin	mon mgr
controller-1	192.168.24.30	mon	_admin mgr
controller-2	192.168.24.37	mon	_admin mgr

7 hosts in cluster

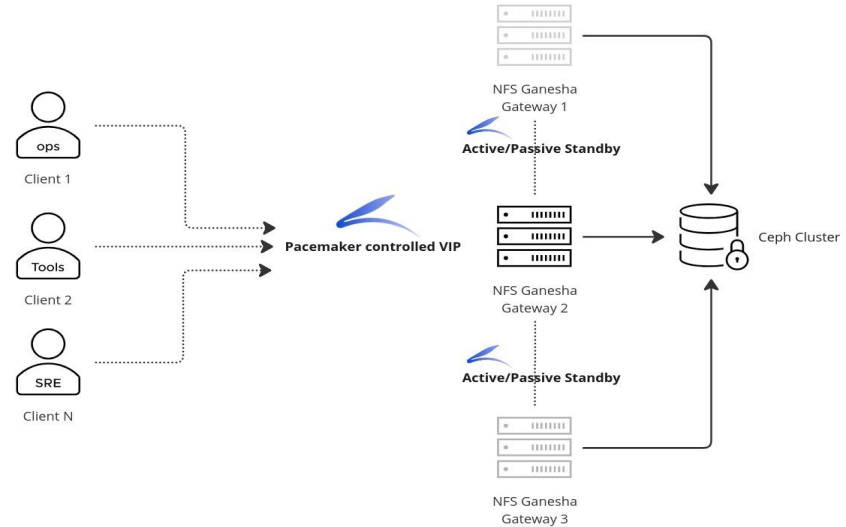
List the deployed services

```
[ceph: root@ceph-mon-0 /]# ceph orch ls
```

NAME	PORTS	RUNNING	REFRESHED	AGE	PLACEMENT
mgr	3/3	7m ago	3d		controller-0;controller-1;controller-2
mon	3/3	7m ago	3d		controller-0;controller-1;controller-2
osd	3/3	7m ago	3d		controller-0;controller-1;controller-2
rgw.rgw	3/3	?:8080	3d		controller-0;controller-1;controller-2

Example: Manila better HA for NFS Ganesha

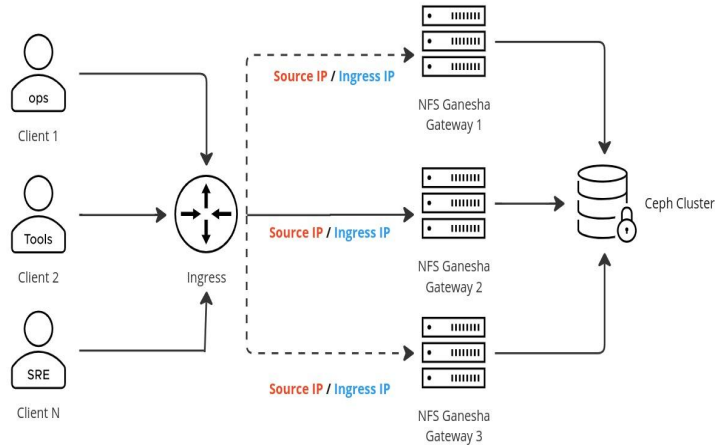
- Manila's CephFS driver sends **DBUS messages** to NFS-Ganesha
- DBUS messaging was scoped to privileged users - which meant that the NFS-Ganesha, and Manila's "share manager" process had to run in **privileged containers**
- DBUS socket was shared between the two containers making deployment inflexible and failovers complicated
- DBUS was "slower" - Ganesha export ID config had to be managed in the driver
- Service was deployed with ceph-ansible and offered as **active/passive** with the help of pacemaker



NFS Ganesha **Active-Passive** managed by **systemd** and **Pacemaker**

- Manila developed a **new driver interface** to interact with ceph mgr/orchestrator

Example: Manila better HA for NFS Ganesha



With PROXY / Without PROXY protocol enabled

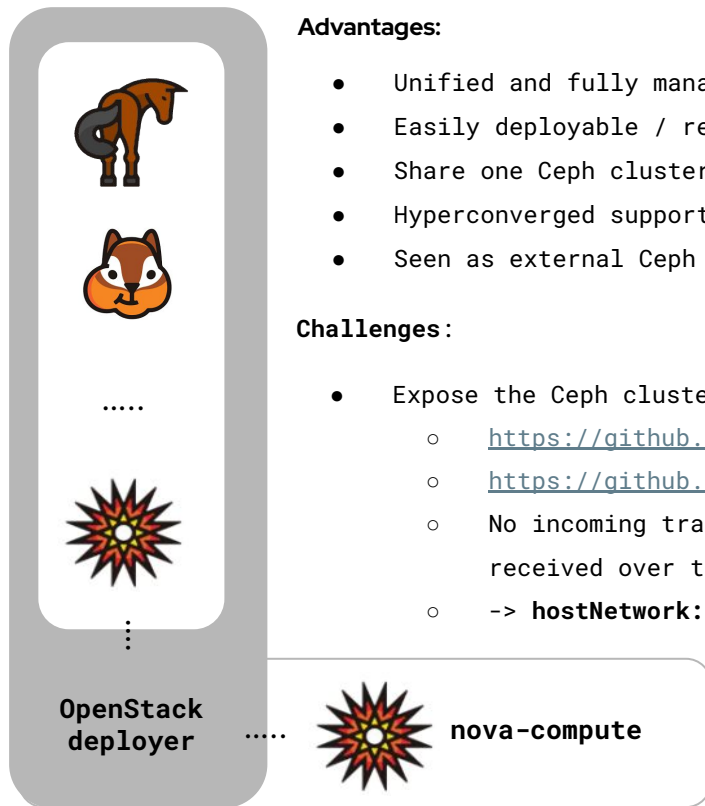
Cephadm ingress service

- Ingress service == HAProxy (tcp mode) + keepalived
- Ingress service allows using a VIP as a single entrypoint across the cluster with H/A awareness (IP managed by keepalived)
- Load-distribution across NFS Ganesha servers (via HAProxy)
- Clustered active/active NFS service offering faster recovery time and better scalability
- Manila's Ceph driver can now use the ceph mgr interface to manage exports which hugely simplifies the driver (no dbus anymore)

Limitations:

- Client restrictions are not available (yet) - HAProxy terminates client connections preventing Ganesha from knowing the end client - client restrictions are a huge part of OpenStack NFS usage
- Ingress adds an additional hop and a possible performance drain

Towards kubernetes (Rook/OpenStack in pods)



Advantages:

- Unified and fully managed by kubernetes
- Easily deployable / reproducible
- Share one Ceph cluster between OpenStack and Kubernetes
- Hyperconverged support is provided out of the box
- Seen as external Ceph from OpenStack perspective

Challenges:

- Expose the Ceph cluster outside: Multus service
 - <https://github.com/rook/rook/issues/10410>
 - <https://github.com/rook/rook/issues/9488>
 - No incoming traffic to the mons, mgr, and rgw can be received over the **multus network**.
 - -> **hostNetwork: true**



Fully managed by k8s but keeping the workload outside

Demo

<https://asciinema.org/a/555694>

Additional Resources



- Cephadm
 - <https://docs.ceph.com/en/latest/cephadm/>
- Notes about Rook/crc/multus
 - <https://gist.github.com/fmount/24b9fb2cfaff9dc813b8211414cb4de0>
- Notes about Rook/minikube with host networking
 - <https://gist.github.com/fmount/13f56ac2310b5013810424f0bc6f350b>
- TripleO
 - https://docs.openstack.org/project-deploy-guide/tripleo-docs/latest/features/deployed_ceph.html
- TripleO Standalone Container based deployment
 - <https://docs.openstack.org/project-deploy-guide/tripleo-docs/latest/deployment/standalone.html>
- What is OpenStack
 - <https://en.wikipedia.org/wiki/OpenStack>

Contacts

- Mailing List:
 - <https://lists.openstack.org/cgi-bin/mailman/listinfo/openstack-discuss>
- IRC Channel (oftc):
 - #openstack-dev
 - #tripleo
 - #openstack-manila
- Email
 - fpantano@redhat.com

THANK YOU