cafe24

GLOBAL
SUCCESS PARTNER.

Deployment Ceph Cluster Object Storage(RGW) using Rook Operator

윤찬열
Chanyeol yoon
(ycy1766@gmail.com)

Contents

Why Object Storages?

Why do we need object storage?

- ***Cafe24*** **has been operating public cloud services based on** ***openstack*** **from 4Q 2020.**
  - 2020 : rocky version base , manually ceph nautilus rbd backend
  - 2022: xena version base

- **For integrate various openstack components**
  - Backup Service : Freezer
  - Database Service : Trove

- **For various components that maintain the public cloud**
  - Monitoring Metric/Logging (non gnocchi)
  - Kubernetes Storages

- **For flexible infrastructure configuration, object service, not file service, has its strengths.**
  - Using Loadbalancer(Octavia)
  - Lightweight static web site

Why Object Storages?

Needed a new deployment method based on our existing Ceph Cluster operation experience.

- **Do stay up to date(at least with versions)** [1]
  - **Code Base Deployment/Upgrade**
    - Dev/Staging/Prod ENV Ceph cluster
      - For testing parameters
  - **Container Base Ceph Cluster**
    - To solve OS dependency and version issues

[1] https://www.youtube.com/watch?v=zJVoleSpSOk

How to deploy ceph ?

Various deployment ceph cluster

- **Ceph-volume (manually)**
  - Simply, But It cannot be distributed consistently.

- **Ceph-deploy**
  - Very Simply, But It cannot be distributed consistently.
  - These days ceph-deploy is no longer maintained, however, and doesn't even work with some newer distros like RHEL/CentOS 8.[1]
  - Last Commit Oct 2, 2020(https://github.com/ceph/ceph-deploy)

- **Ceph-ansible**
  - distributed consistently using ansible-playbook
  - Non-Container/Container-based deployment is possible. (playbook yaml; like a code)

- **CephADM**
  - Container-based deployment.
  - It doesn't offer a variety of components yet. (iSCSI gw,nfs ...)[1]

- **Rook Operator**
  - Deploy Ceph Cluster using Rook Operator
  - CRD Base
  - Very active project management
  - There is a disadvantage of having to run kubernets for the operator.

[1] https://ceph.io/en/news/blog/2020/introducing-cephadm/

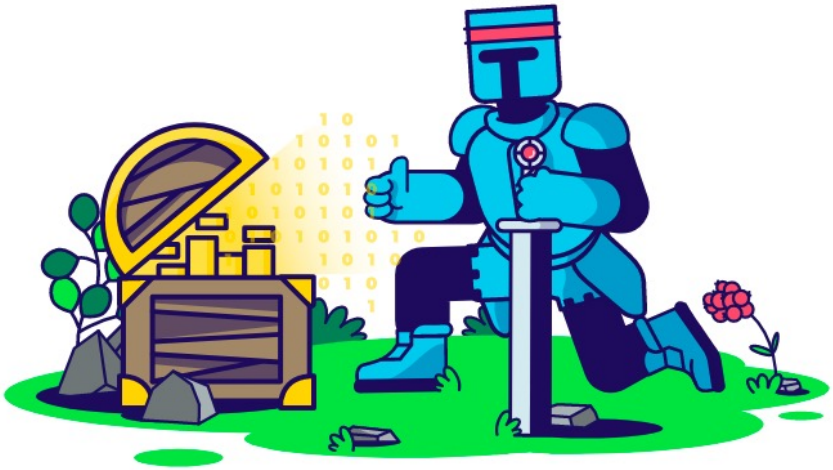
What is Rook?

What is Rook?

- Makes storage available inside your Kubernetes cluster
- **Self-managing, self-scaling, self-healing storage services**[1]
- Automates the tasks of a storage administrator: deployment, bootstrapping, configuration, provisioning, scaling, upgrading, migration, disaster recovery, monitoring, and resource management. [1]
- Kubernetes Operators and Custom Resource Definitions
    - Block Pool CRD, Cluster CRD, Object Store CRD, Object Bucket Claim …
- Automated management
    - Deployment, configuration, upgrades
- Open Source (Apache 2.0)
- **111 releases in 5 years**

Storage Providers

- Stable
    - Ceph
- Alpha
    - Cassandra
    - NFS



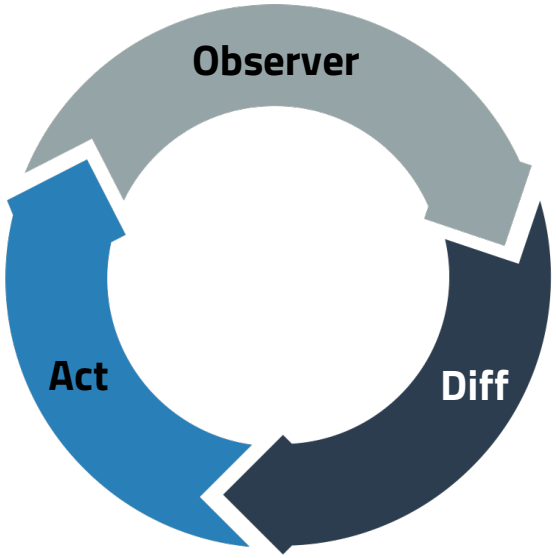
Watch	269	Star	9,138	Forks	2,099

[1] https://rook.io/

Rook Operator

Rook Operator

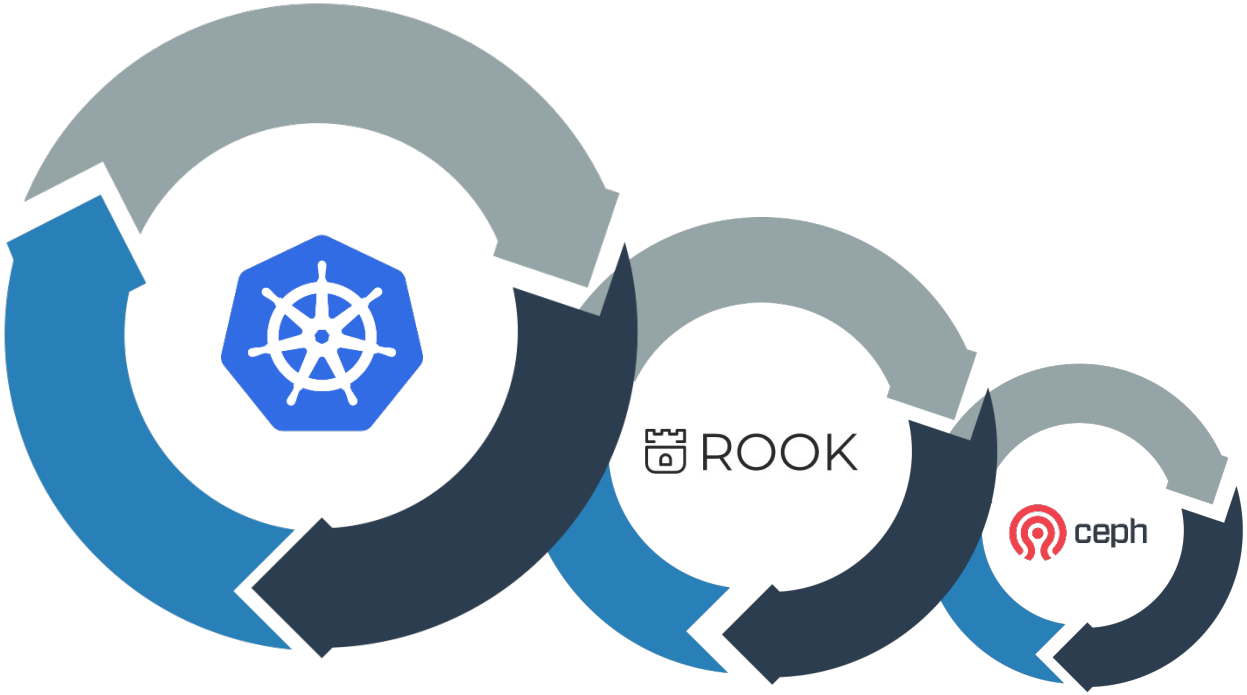
- It provides a more complicated reconciliation loop than the standard reconciliation loop of Kubernetes. So you can deploy and manage sensitive applications.
- So, in rook, ceph is deployed/managed through operator.
- The Rook operator automates configuration of storage components and monitors the cluster to ensure the storage remains available and healthy.[1]



[1] https://rook.io/docs/rook/v1.7/ceph-storage.html

Rook Operator

- Reconciliation loop occurs logically, and while maintaining the Ceph cluster, Rook Operator replaces human intervention and proceeds with deployment/management.
- Management by logically connecting multiple reclamation loops.



[1] https://www.slideshare.net/chrismceniry/evolving-for-kubernetes

How to deploy rook ceph ?

Installation Rook(ceph)

- simple deployment

- CRD(CustomResourceDefinitions) deployment

```
# kubectl create  -f crds.yaml
```

- Namespace, RBAC deployment

```
# kubectl create  -f common.yaml
```

- Oerpator Config-Map, pod deployment

```
# kubectl create  -f operator.yaml
```

- Ceph Config-Map(;ceph.conf) deployment

```
# kubectl create  -f rook-config-override.yaml
```

- Ceph cluster deployment

```
# kubectl create  -f cluster.yaml
```

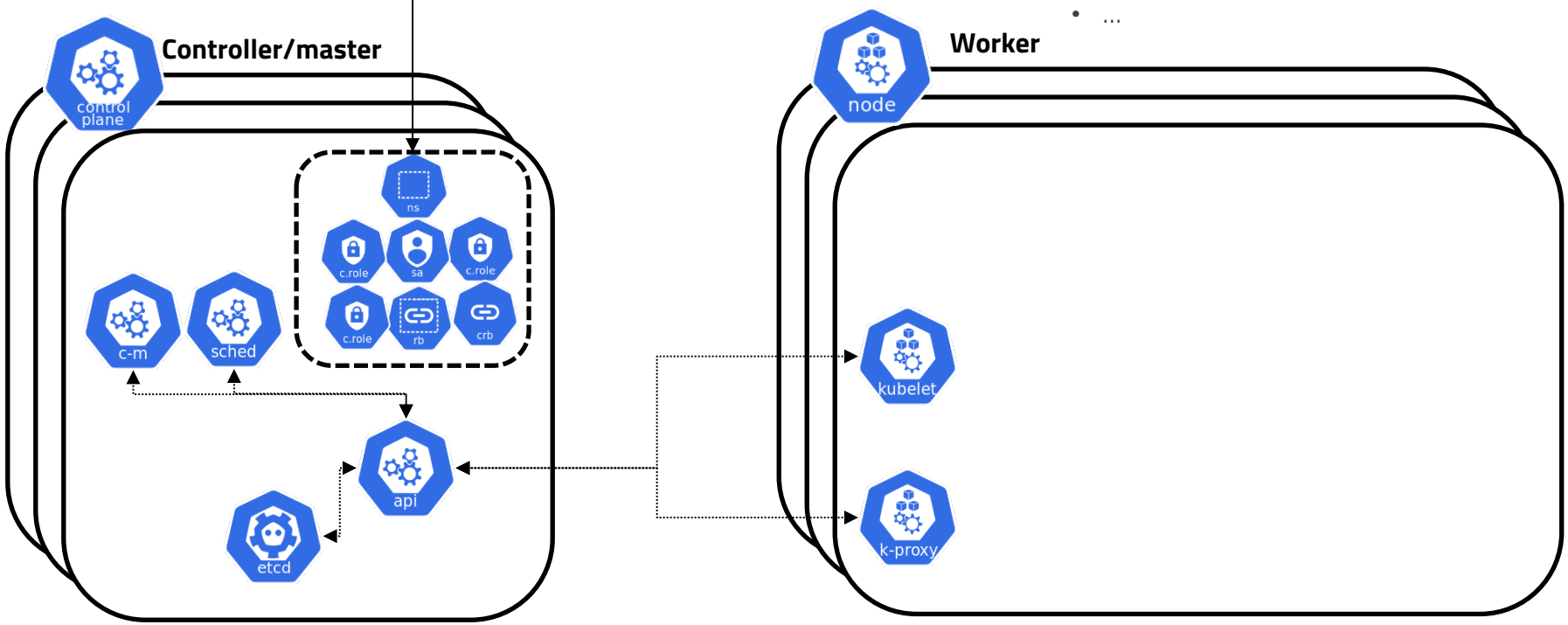
How to deploy rook ceph ?

Installation Rook(ceph)

- Deploy ns, rbac and crd with common.yaml, crds.yaml for rook-ceph-operator configuration

```
# kubectl create  -f crds.yaml
# kubectl create  -f common.yaml
```

- ceph-client-crd
- ceph-cluster-crd
- ceph-filesystem-crd
- ceph-fs-mirror-crd
- ceph-nfs-crd
- ceph-object-multisite-crd
- ceph-object-store-crd
- ceph-object-store-user-crd
- ...

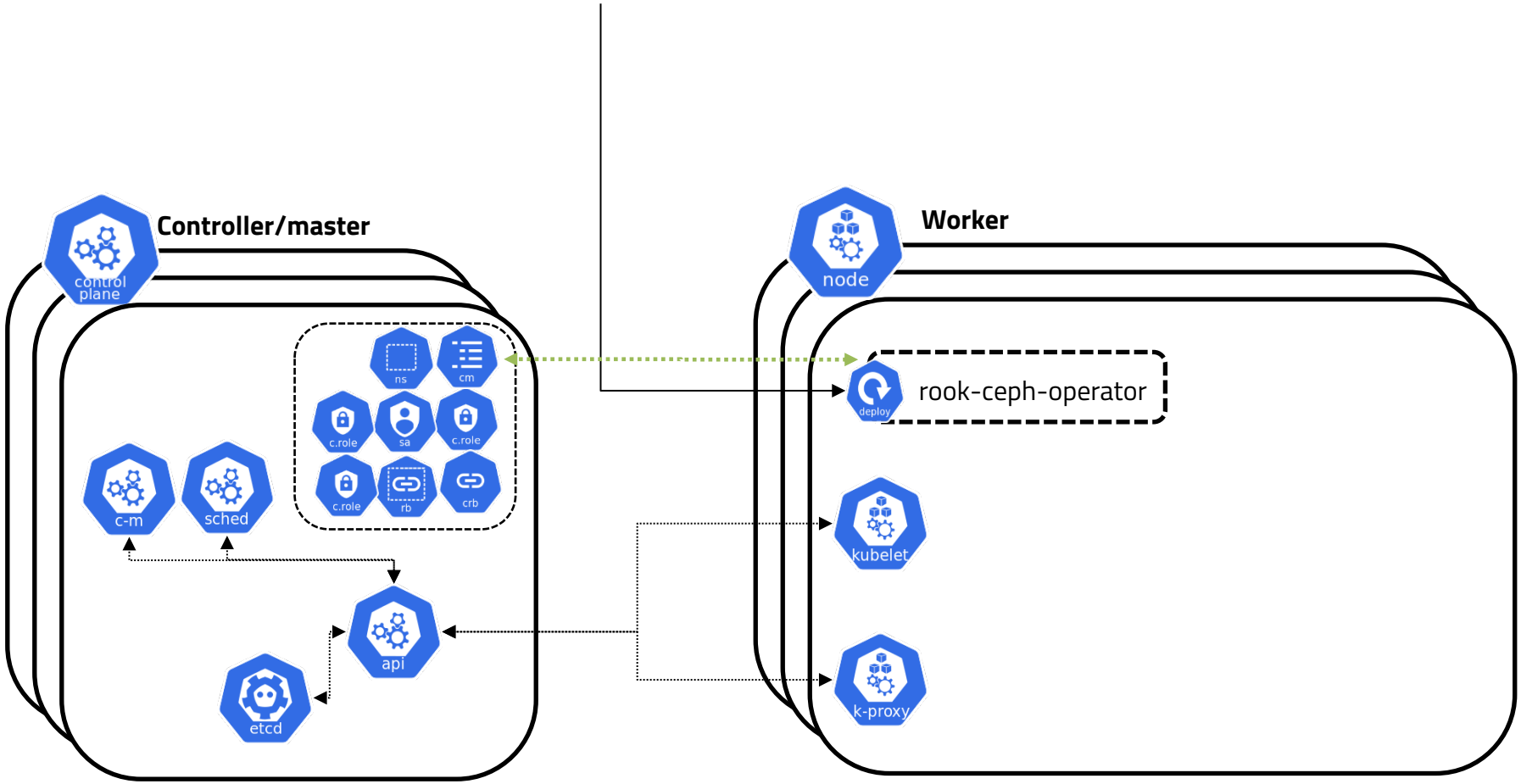


How to deploy rook ceph ?

Installation Rook(ceph)

- Deploy configmap with operaator.yaml and operator and their settings
- Stating reconciliation loop

```
# kubectl create  -f operator.yaml
```

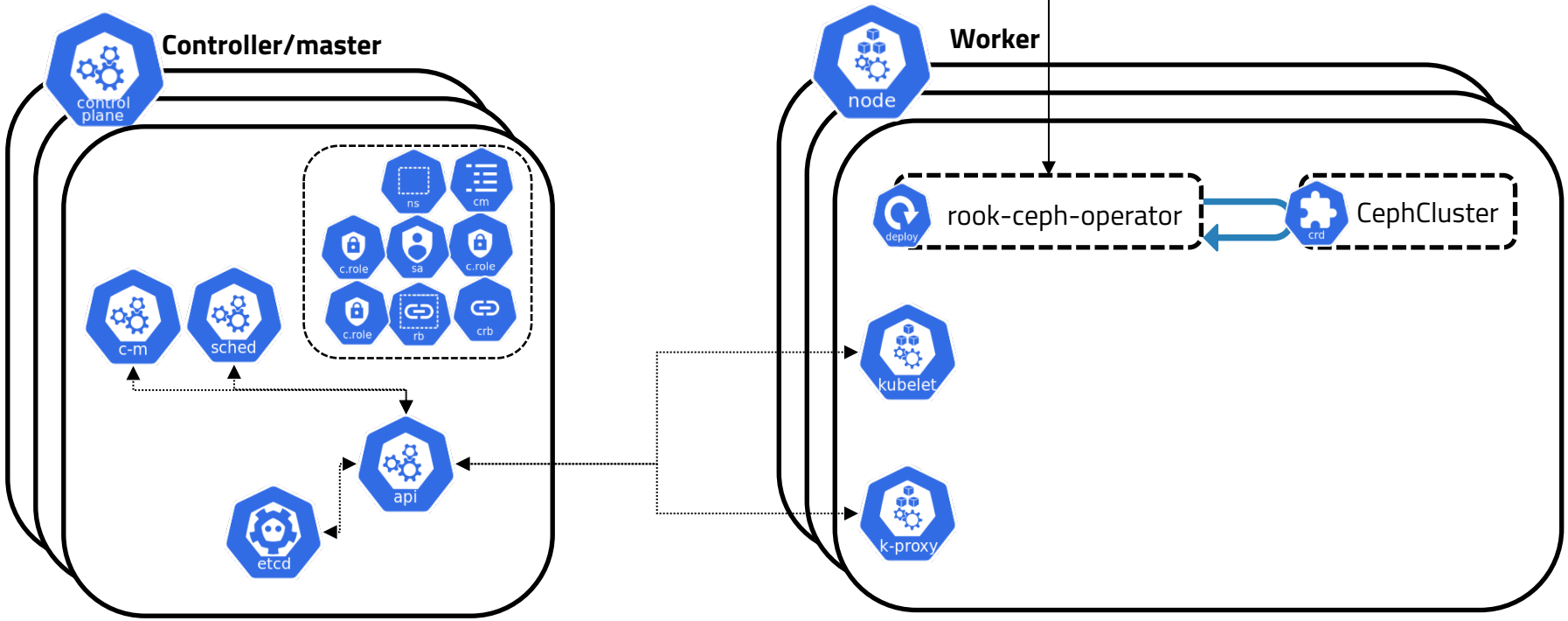


How to deploy rook ceph ?

Installation Rook(ceph)

- Deploy rook-config-override cm to be used as ceph.conf and apply it to CephCluster CRD through cluster.yaml

```
# kubectl create  -f rook-config-override.yaml
# kubectl create  -f cluster.yaml
```



How to deploy rook ceph ?

Installation Rook(ceph)

- Ceph cluster deployment

```
# kubectl create  -f cluster.yaml
```

```yaml
apiVersion: ceph.rook.io/v1
kind: CephCluster
metadata:
  name: rook-ceph
  namespace: rook-ceph
spec:
  cephVersion:
    image: quay.io/ceph/ceph:v16.2.6
  dataDirHostPath: /var/lib/rook
  mon:
    count: 3
    allowMultiplePerNode: false
  mgr:
    count: 1
    modules:
      - name: pg_autoscaler
        enabled: true
  crashCollector:
    disable: false
  storage:
    useAllNodes: false
    useAllDevices: false
    nodes:
      - name: node-a
        devices:
          - name: "sdb"
      - name: node-b
        devices:
          - name: "sdb"
      - name: node-c
        devices:
          - name: "sdb"
  network:
    provider: host
    #provider: multus
```

Ceph Version (image tag)

Ceph Mon Setting

Ceph MGR Setting

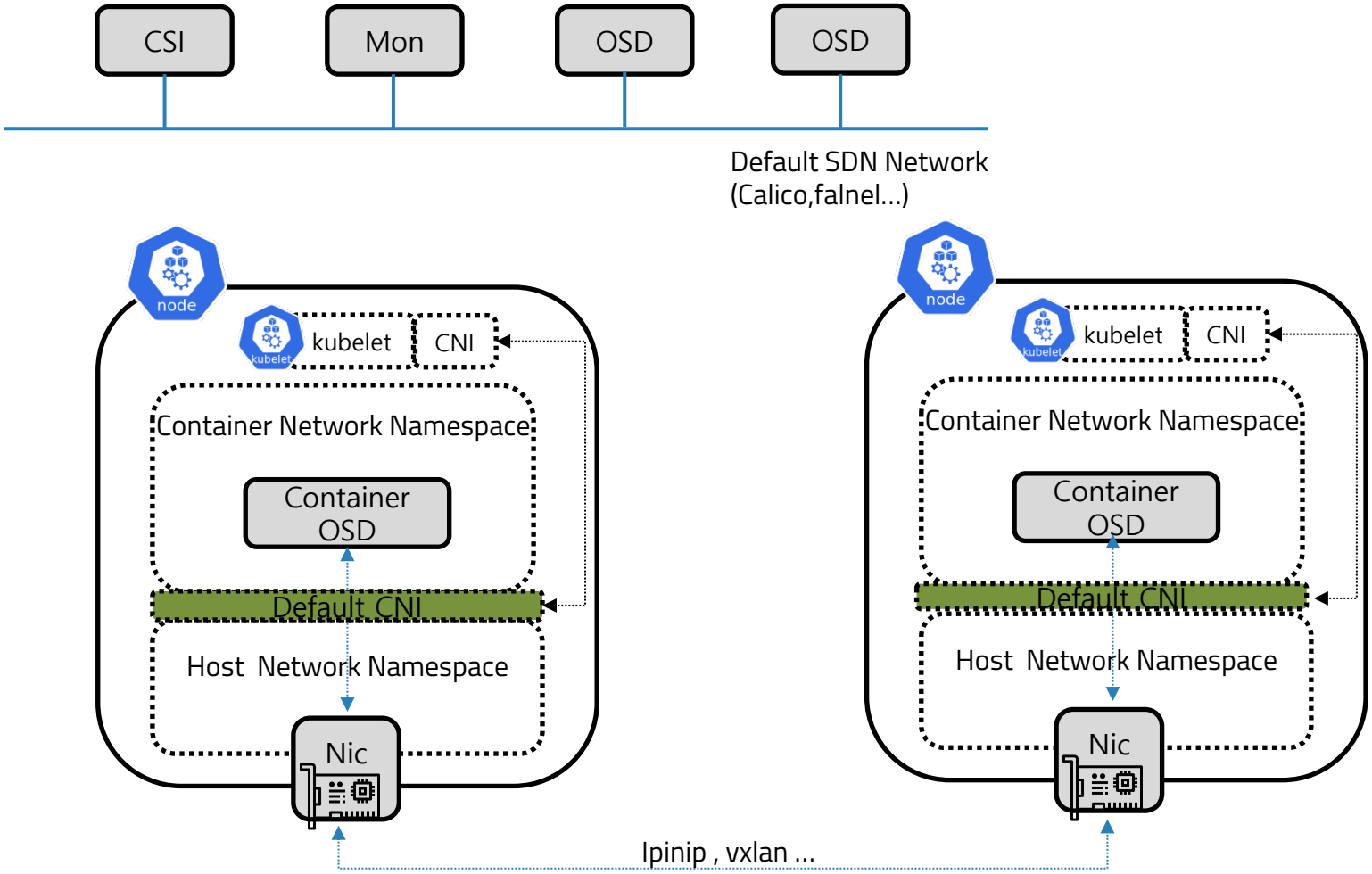
Ceph OSD Setting

Ceph Network Setting

How to deploy rook ceph ?

Rook CephCluster Networking Methods

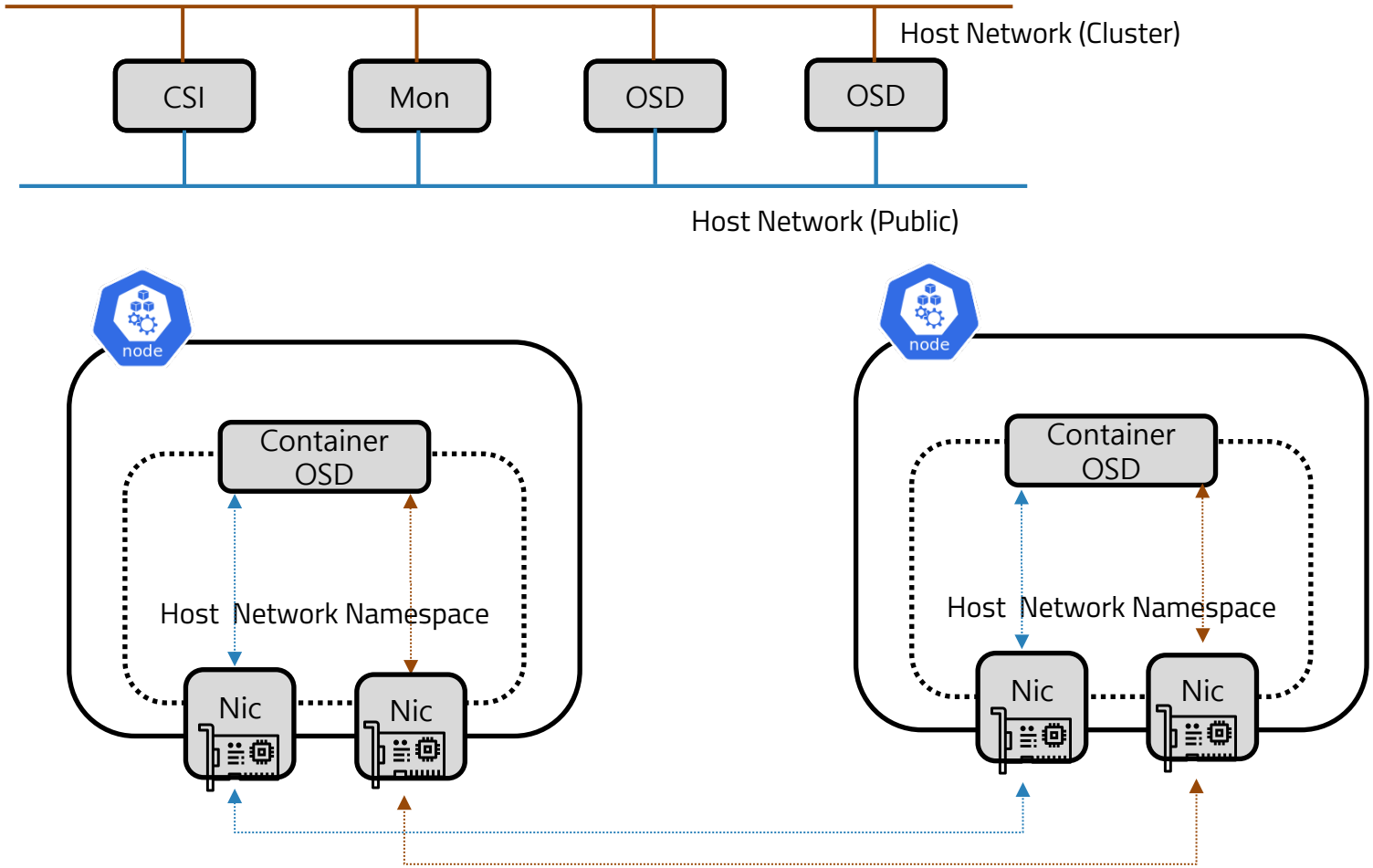
- Traditional pod networking - single network interface - default SDN
  - Security
  - Low Performance



How to deploy rook ceph ?

Rook CephCluster Networking Methods

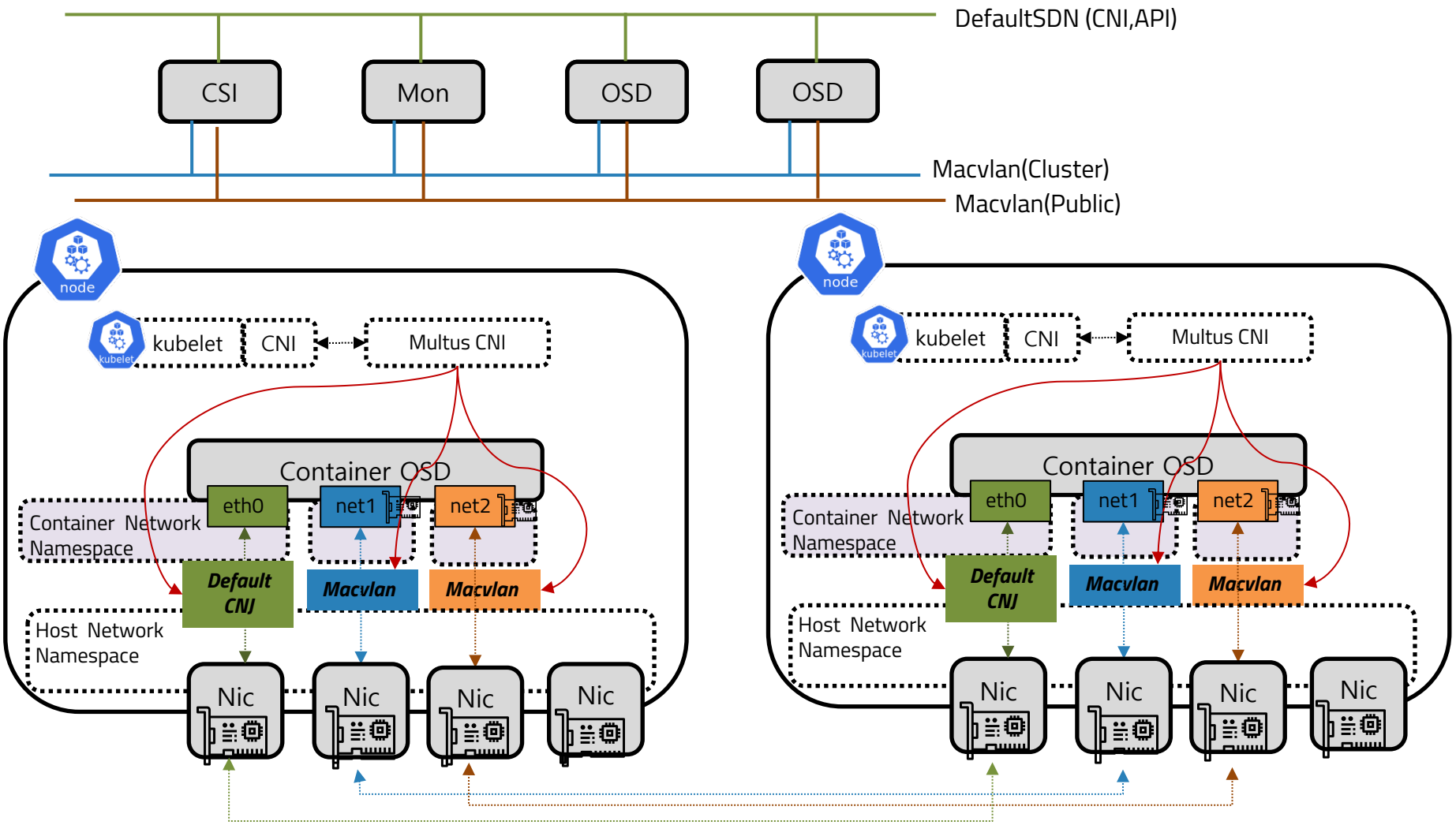
- Host networking - runs on host network namespace and uses host IP. All host's network stack is visible
  - Better Performance than Default SDN(CNI)
  - Low Security
  - Unable to define client access (Defualt monitor access ; Kubernetes node ip binding)



How to deploy rook ceph ?

Rook CephCluster Networking Methods

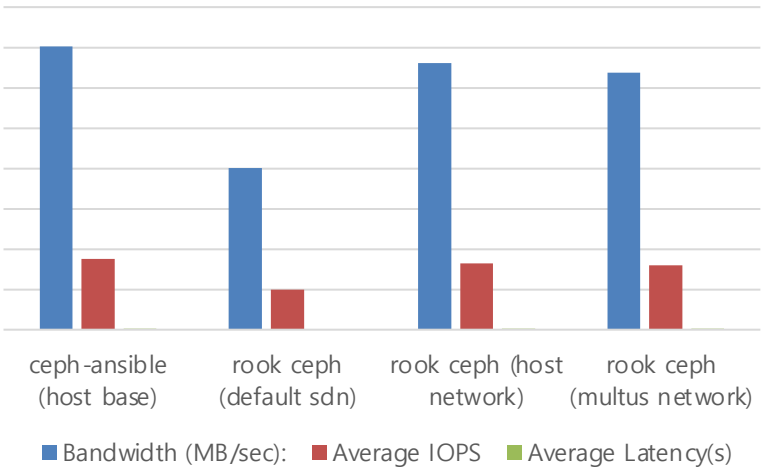
- Multus networking - Rook supports addition of public and cluster network for Ceph
  - Security
  - Better Performance (But, communication between the OSDs of the same host also goes through the top SW)



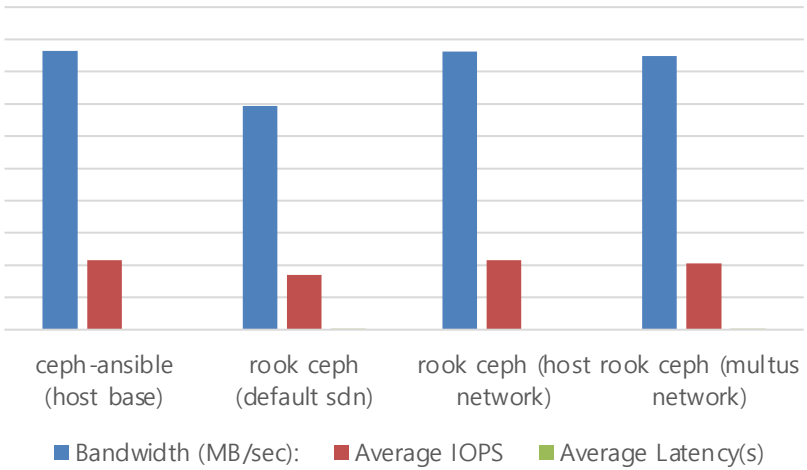
How to deploy rook ceph ?

Rook CephCluster Networking Methods

- Comparison of network performance according to network type and deployment method
  - Default SDN: calico vxlan crosssubnet
  - All the same hardware setup



radosbench write



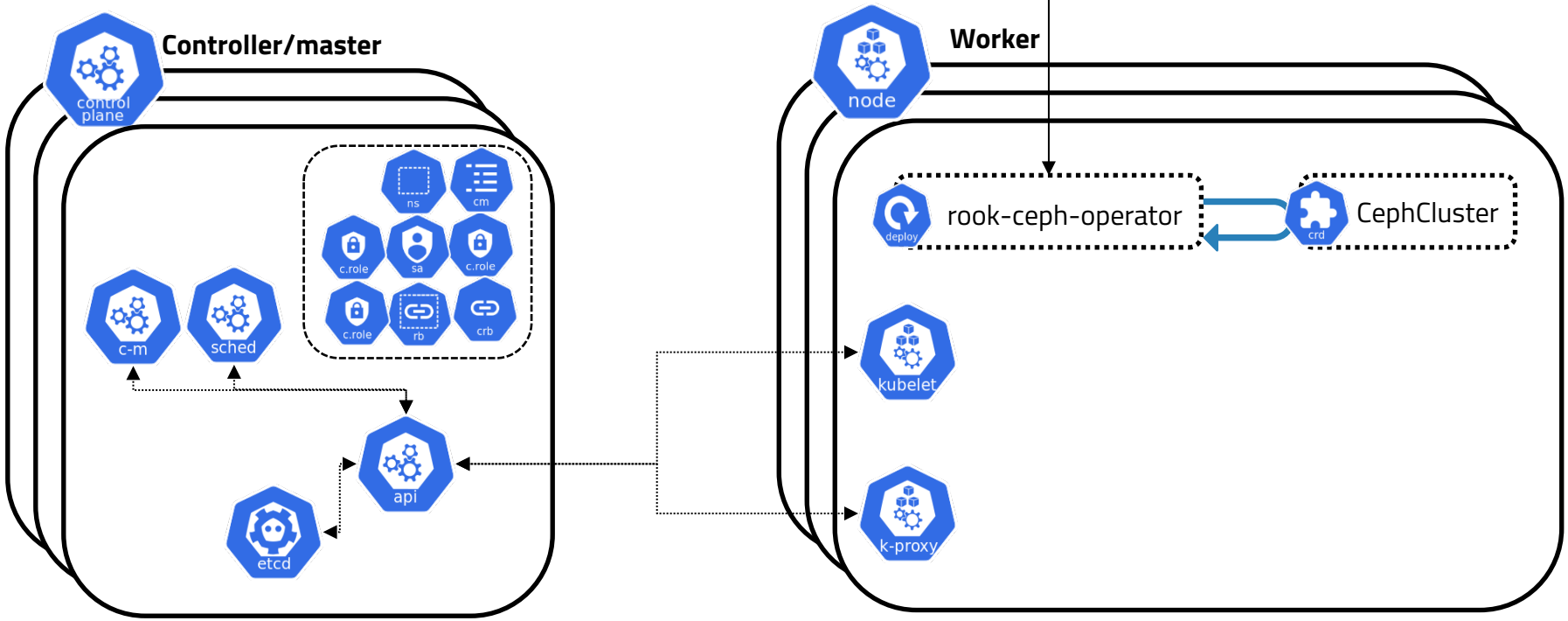
radosbench seq read

How to deploy rook ceph ?

Installation Rook(ceph)

- Deploying of look-config-override cm to be used as ceph.conf and deploying of ceph clusters by applying it as a CephClusterkind through cluster.yaml.

```
# kubectl create -f rook-config-override.yaml
# kubectl create -f cluster.yaml
```

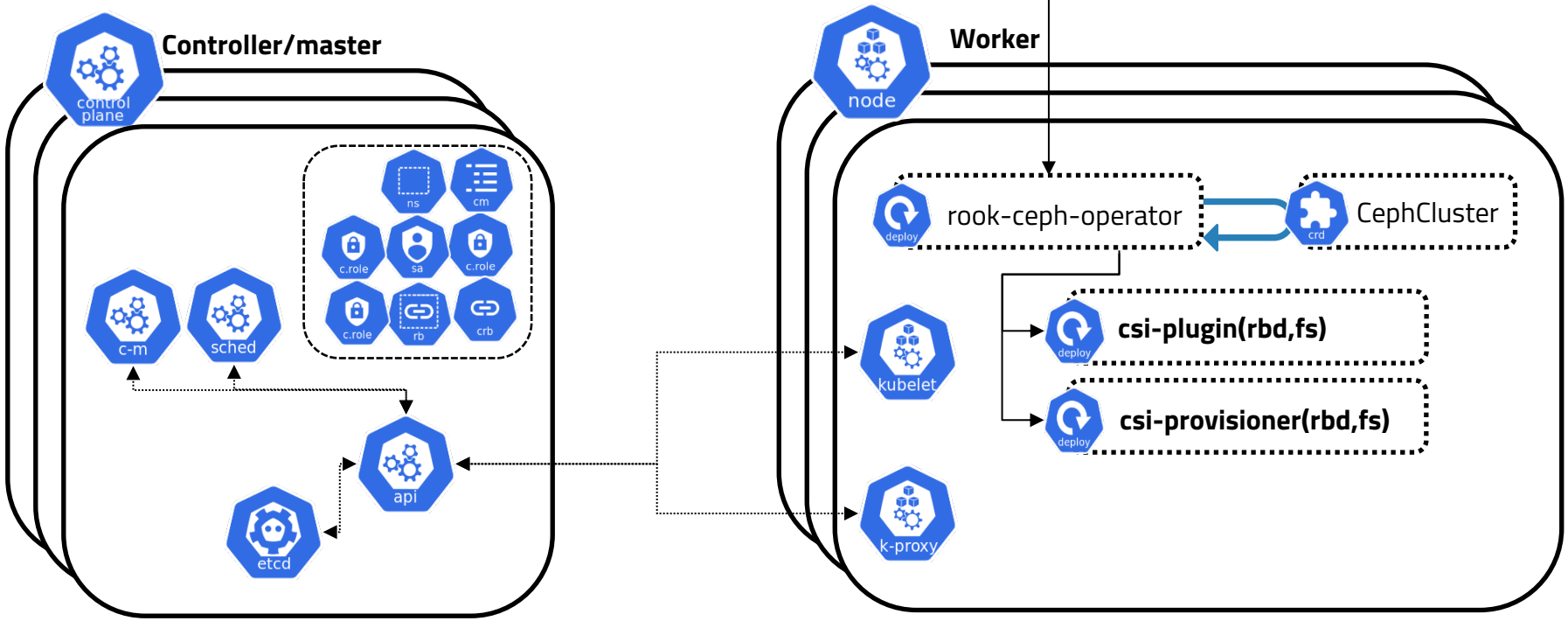


How to deploy rook ceph ?

Installation Rook(ceph)

- Deploy csi(container storage interface) plugin,provisoner

```
# kubectl create  -f rook-config-override.yaml
# kubectl create  -f cluster.yaml
```

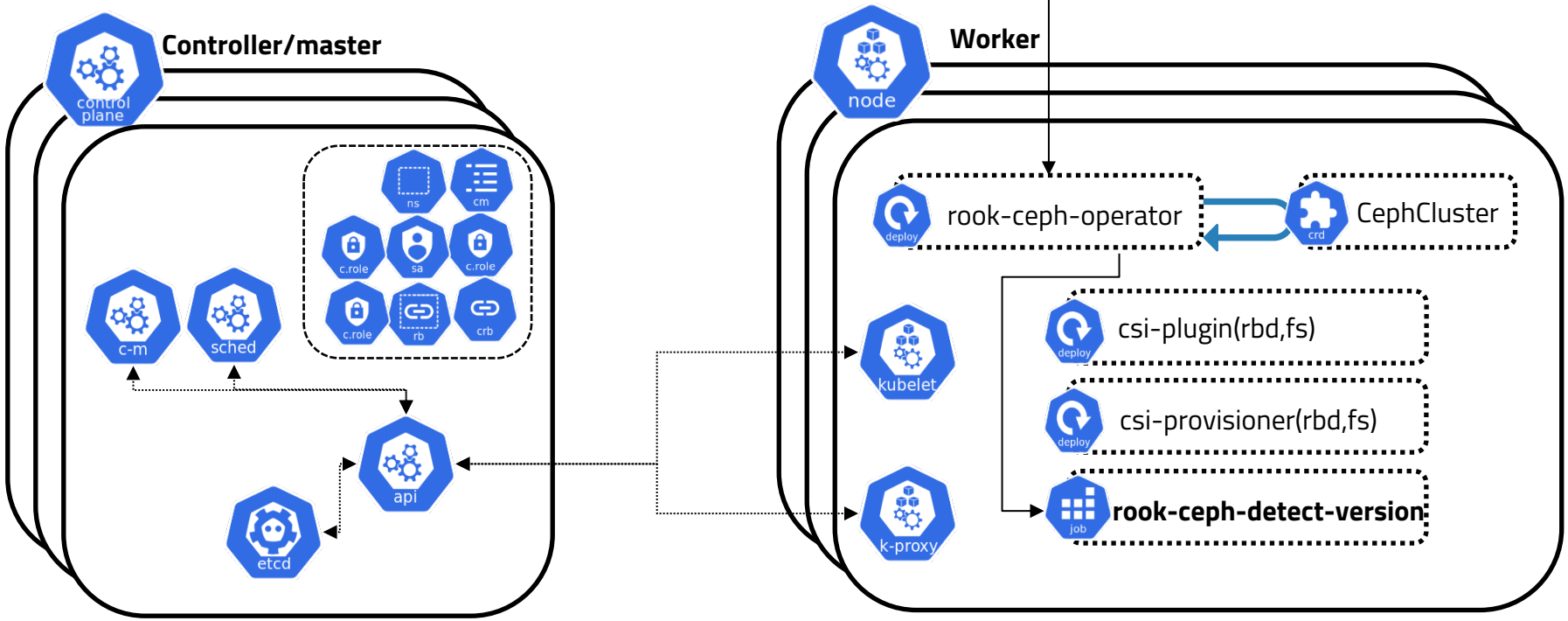


How to deploy rook ceph ?

Installation Rook(ceph)

- The Rook Ceph operator creates a Job called rook-ceph-detect-version to detect the full Ceph version used by the given cephVersion.image

```
# kubectl create  -f rook-config-override.yaml
# kubectl create  -f cluster.yaml
```

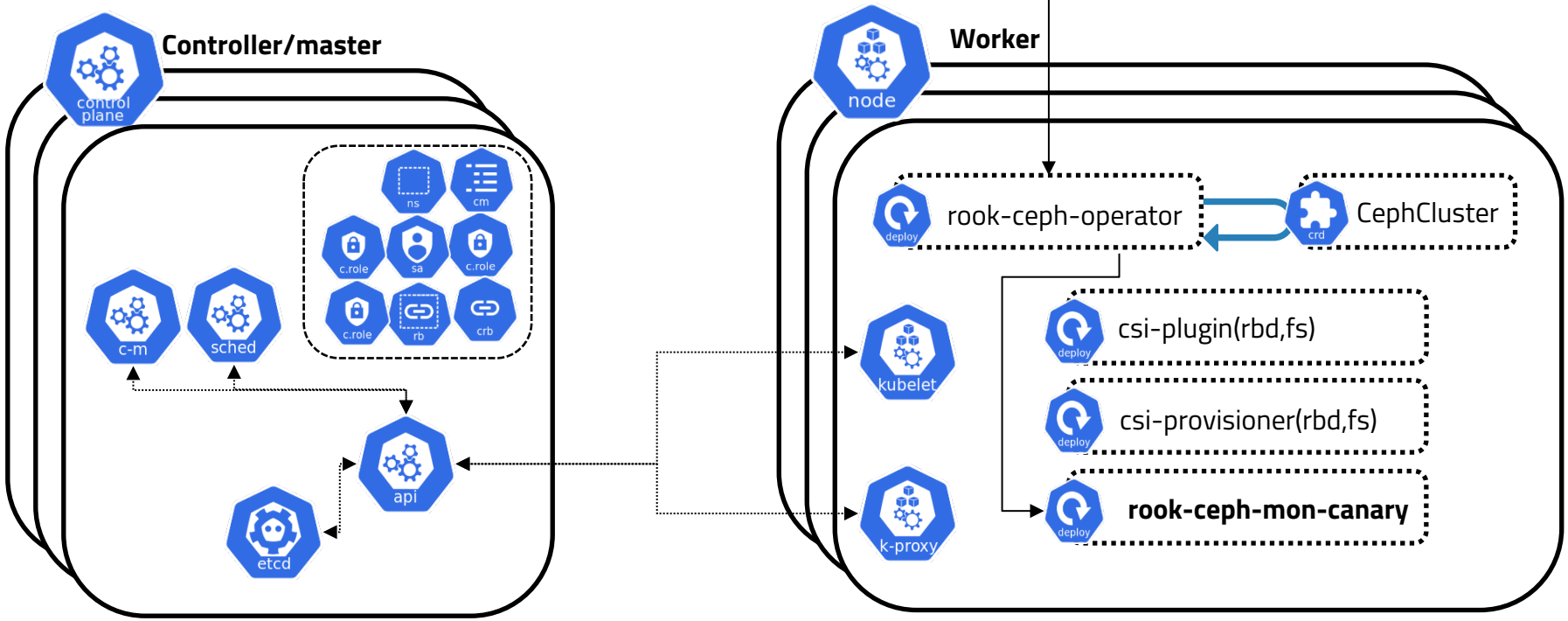


How to deploy rook ceph ?

Installation Ceph Monitor

- Deploy Ceph Monitor Canary Pod

```
# kubectl create  -f rook-config-override.yaml
# kubectl create  -f cluster.yaml
```

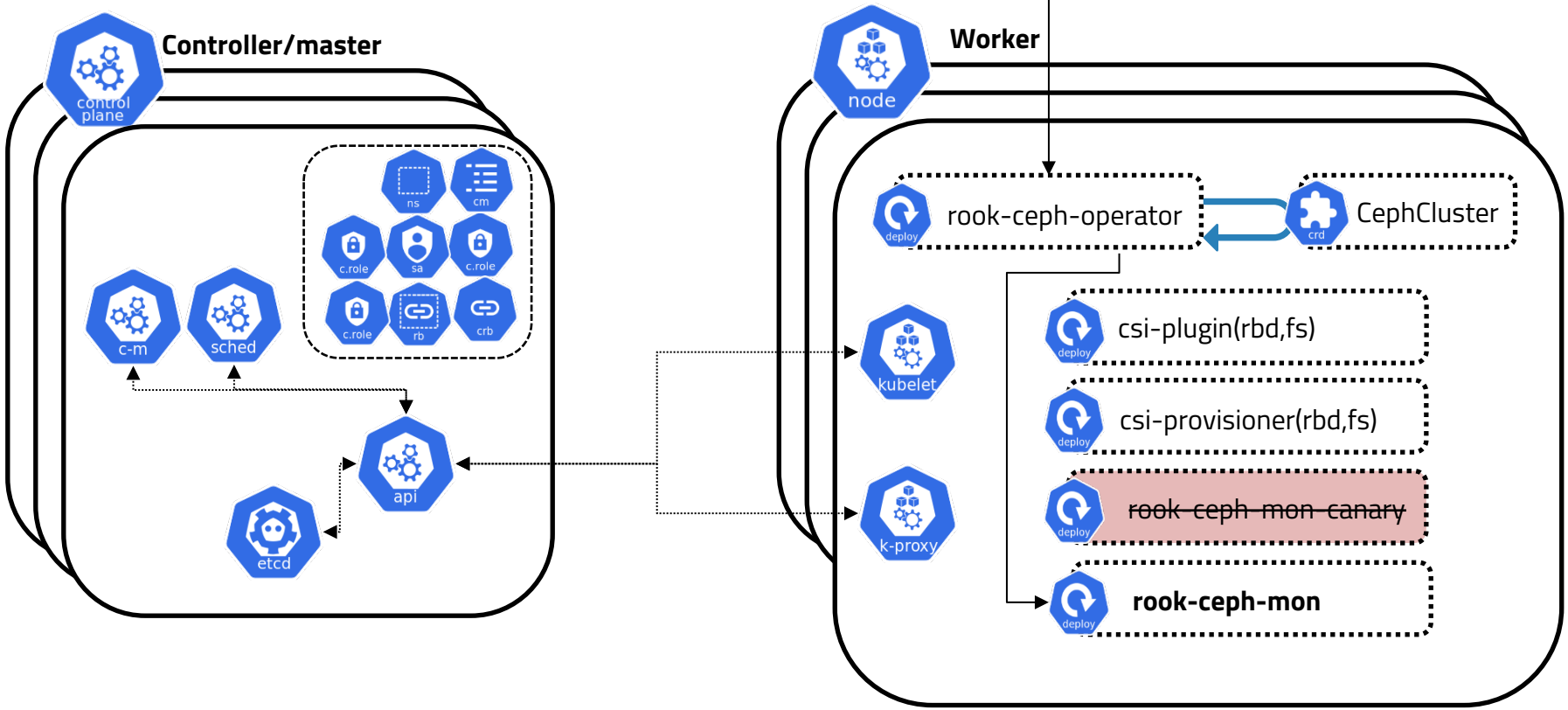


How to deploy rook ceph ?

Installation Ceph Monitor

- Deploy Ceph Monitor Pod

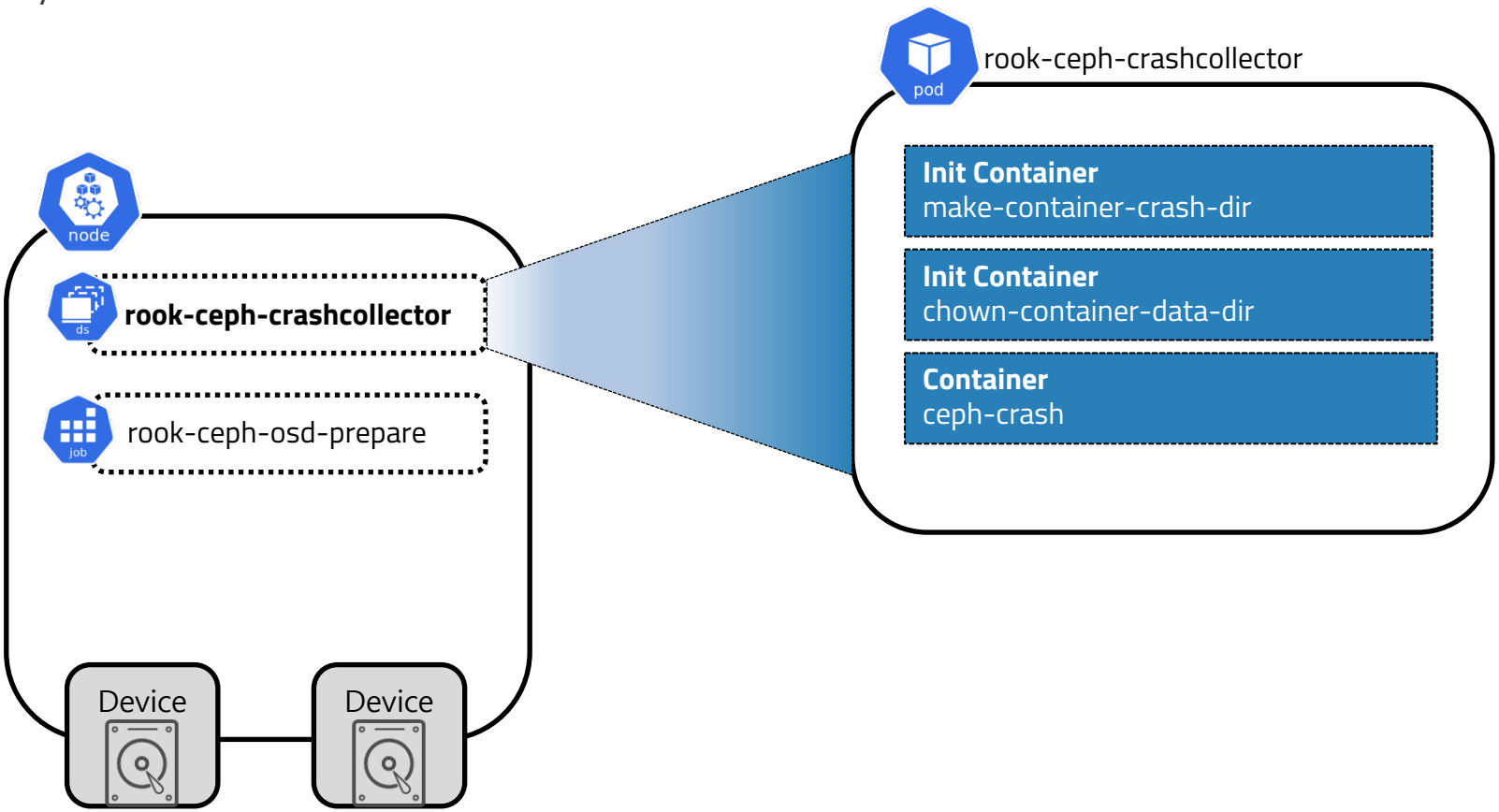
```
# kubectl create  -f rook-config-override.yaml
# kubectl create  -f cluster.yaml
```



How to deploy rook ceph ?

Crashcollector?

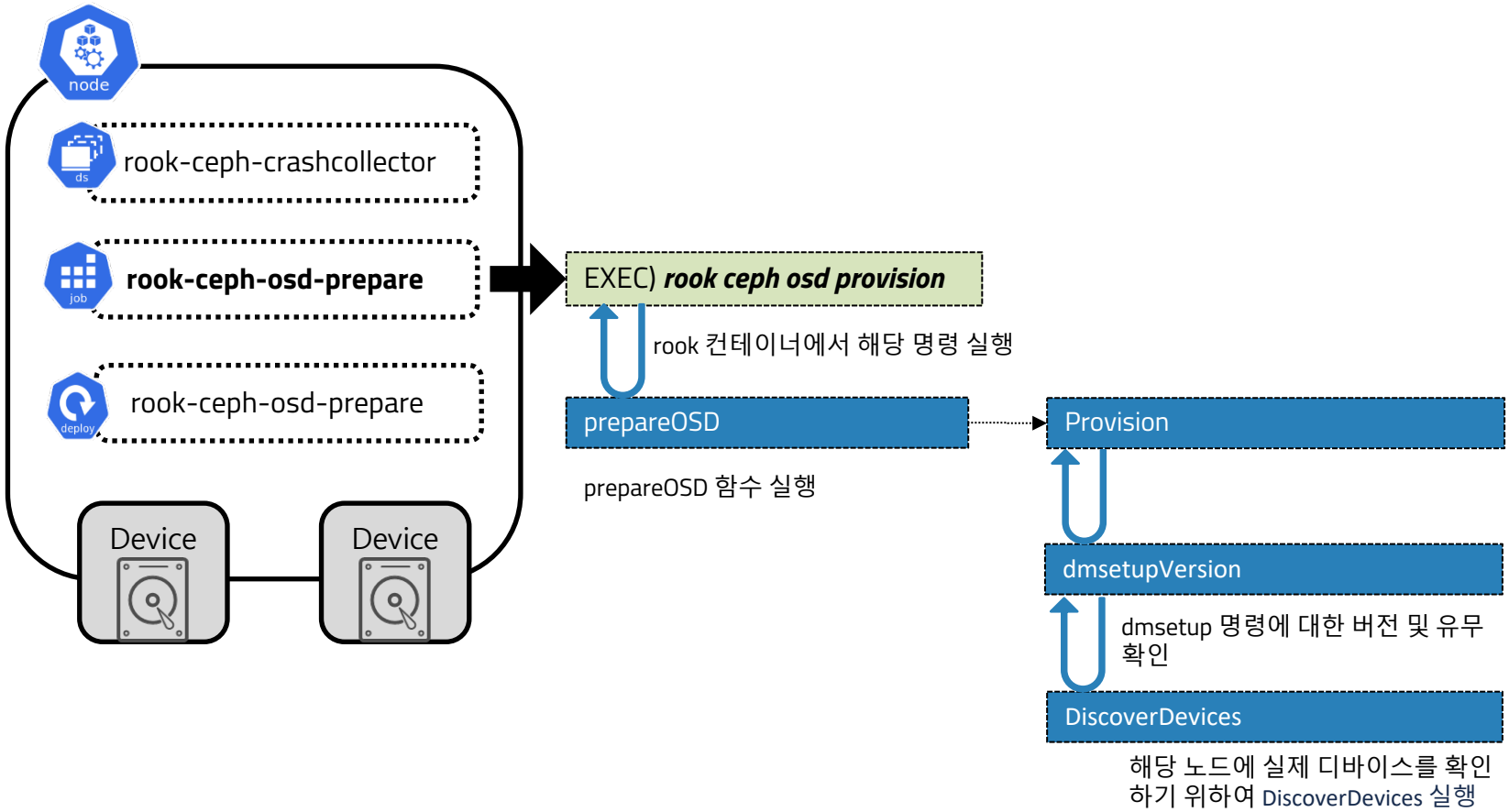
- Various system crash logs occurring in osd or mon can be checked by checking the crash log, but in the case of a container, it cannot be collected because it is terminated before the crash occurs.
- System crash logs of containers generated from all nodes with daemonset can be checked in pod and managed integrally.



How to deploy rook ceph ?

Prepare OSD

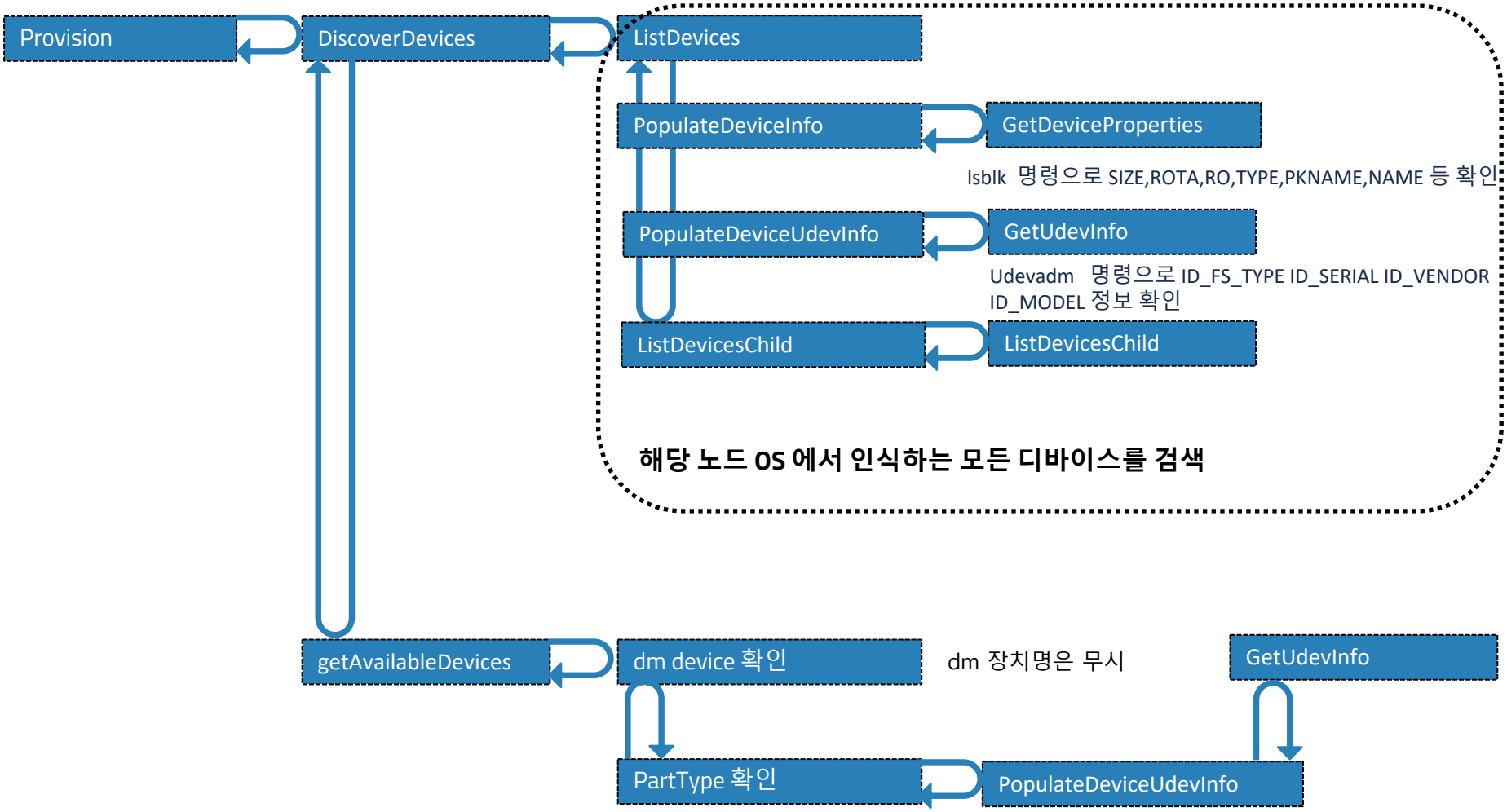
- Prepare process for executing osd by performing a look-ceph-osd-prepare job.
- Perform the rook cephosd provision command in pod.
- Execute the prepareOSD function.



How to deploy rook ceph ?

Prepare OSD

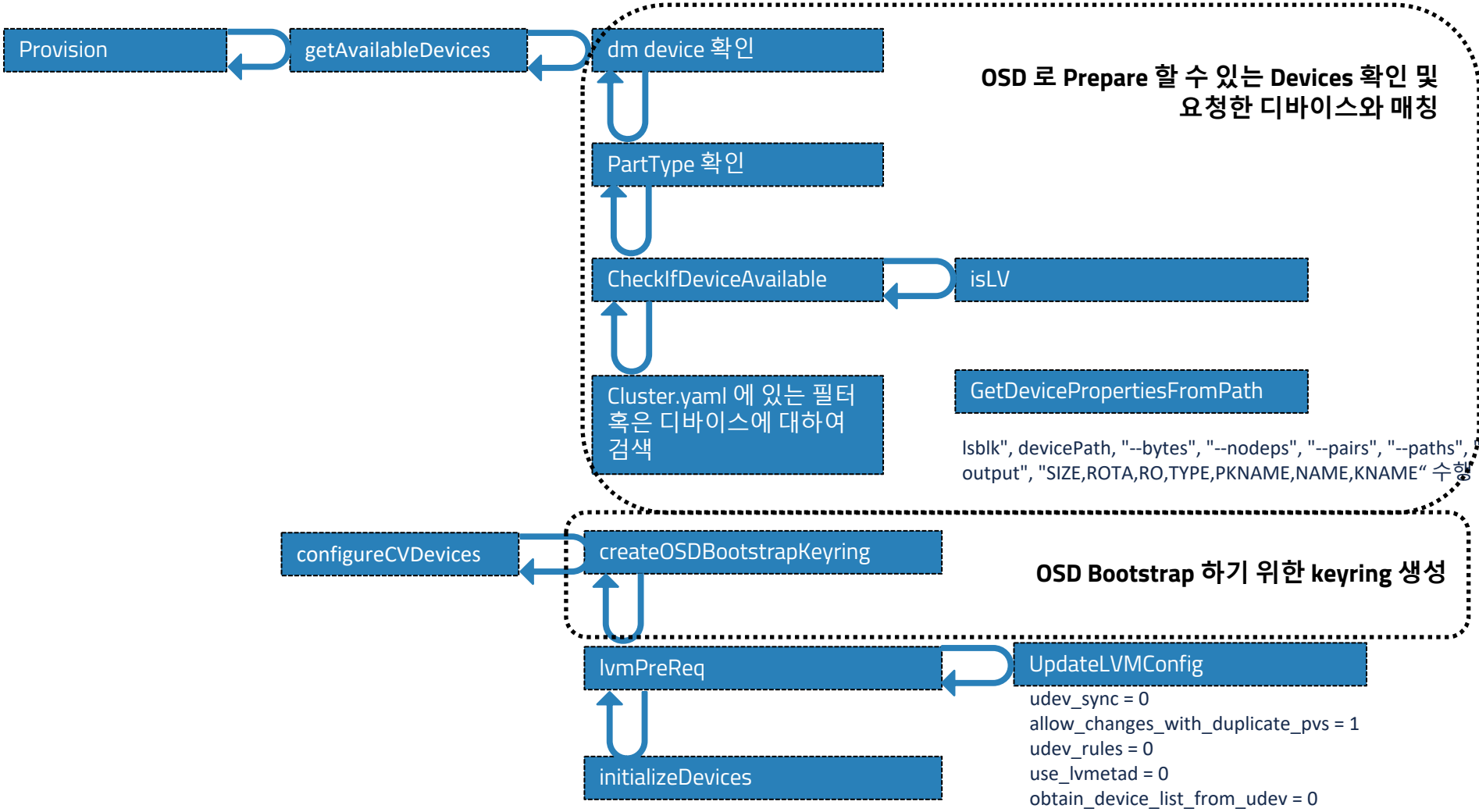
- Get all the actual physical device information of the node to be distributed and obtain information except for unnecessary devices.



How to deploy rook ceph ?

Prepare OSD

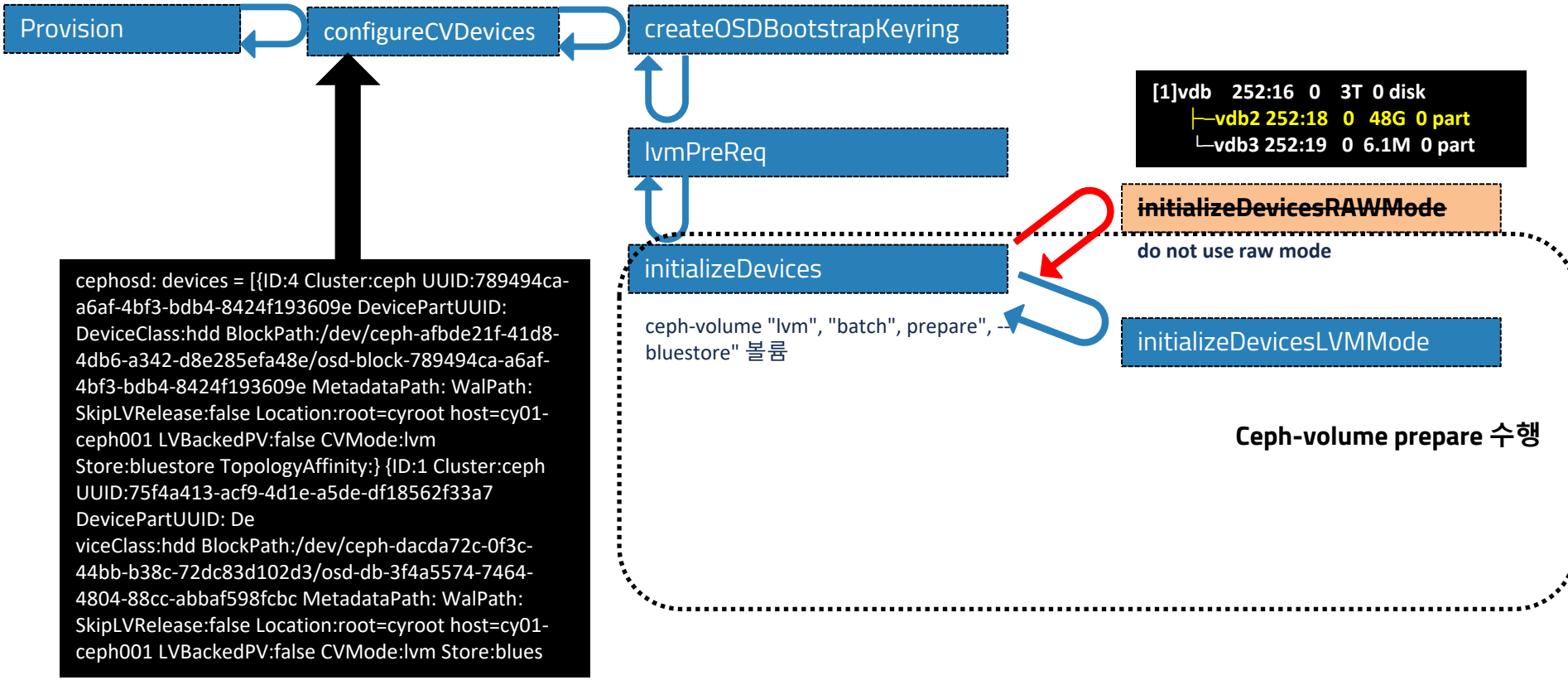
- Check information between the desired device and the actual physical device and proceed with deployment validation. Check the settings and binary file for Lvm.



How to deploy rook ceph ?

Prepare OSD

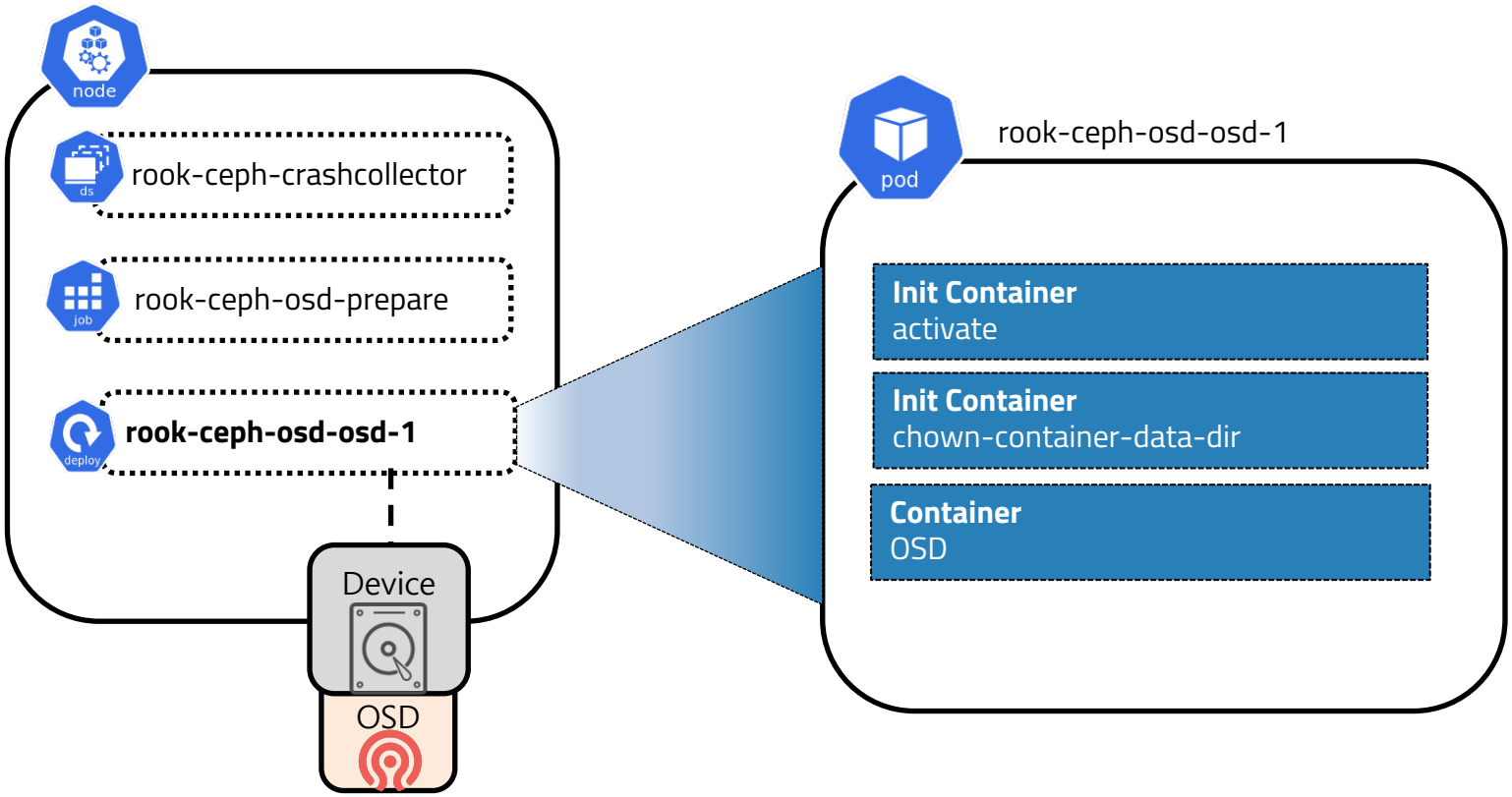
- Raw mode is currently unavailable due to issue [1] and is prepared to lvm mode by default.
- Proceed prepare by executing the ceph-volume command



How to deploy rook ceph ?

Run OSD Pod

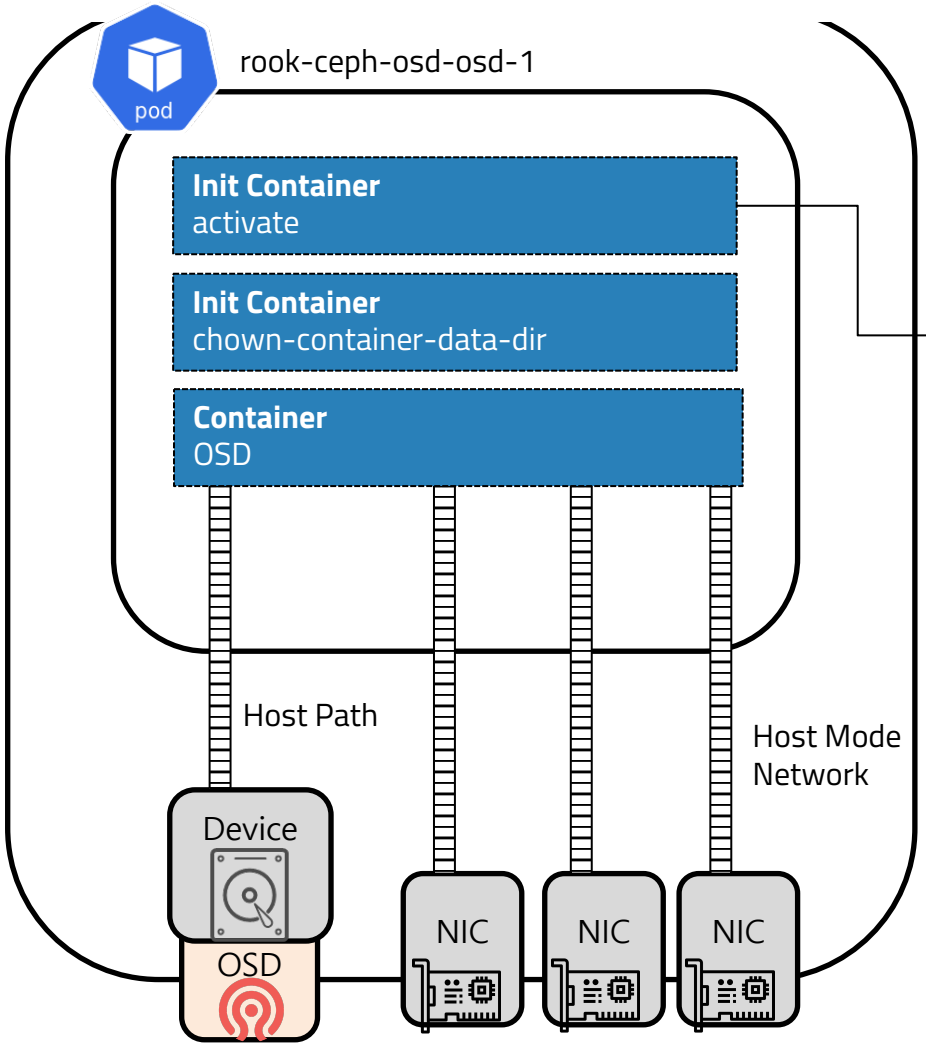
- OSD Pod is separated into init container and cord container, and OSD operates normally only when the init process is performed normally.



How to deploy rook ceph ?

Run OSD Pod

- In active container, it plays a role in activating OSDs that have been prepare completed or previously terminated.

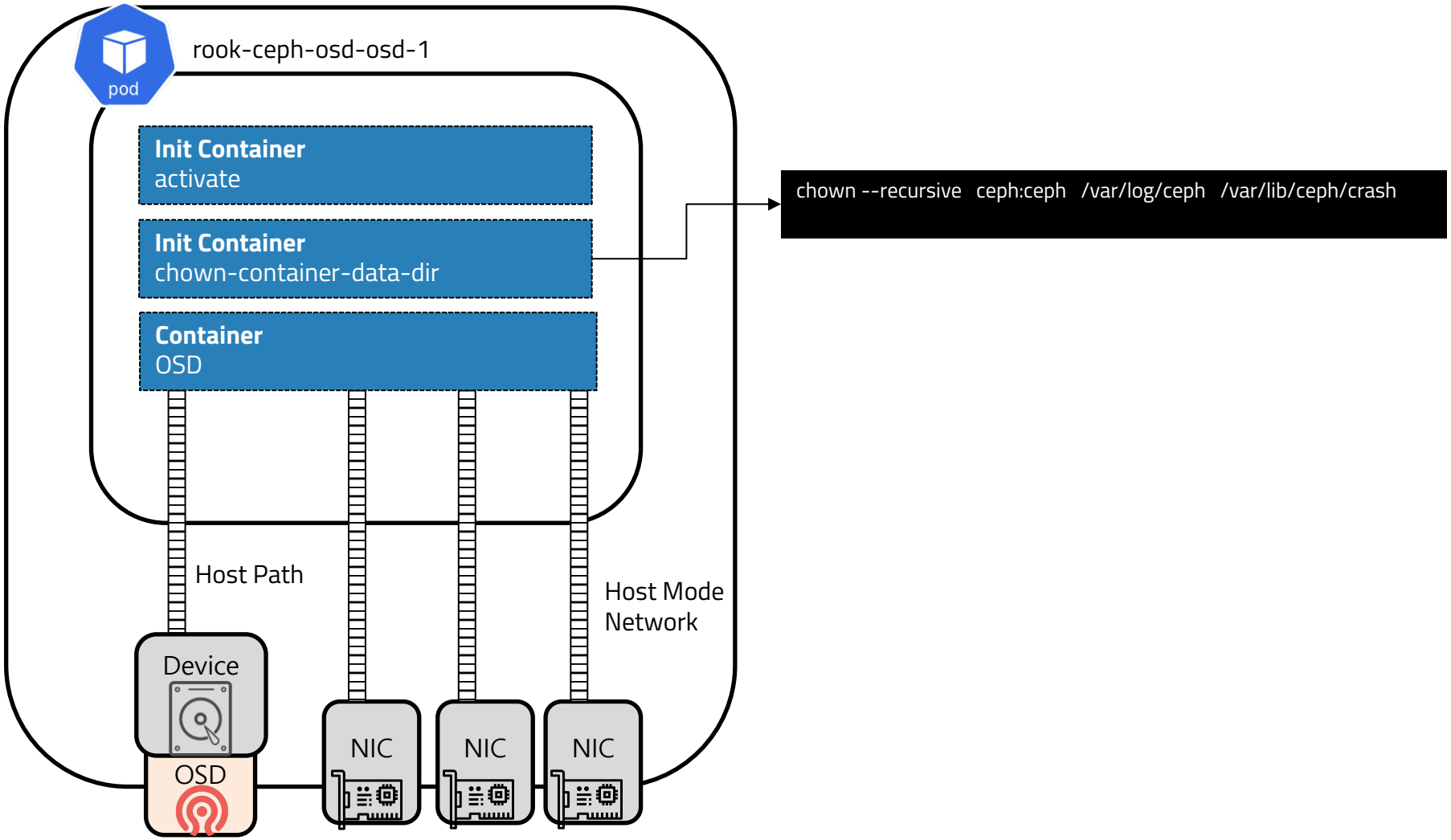



```
ceph -n client.admin auth get-or-create osd."$OSD_ID" mon 'allow
profile osd' mgr 'allow profile osd' osd 'allow *' -k /etc/ceph/admin-
keyring-store/keyring

ceph-volume lvm activate --no-systemd "$OSD_STORE_FLAG"
"$OSD_ID" "$OSD_UUID"

...
생략
...
```

How to deploy rook ceph ?

Run OSD Pod

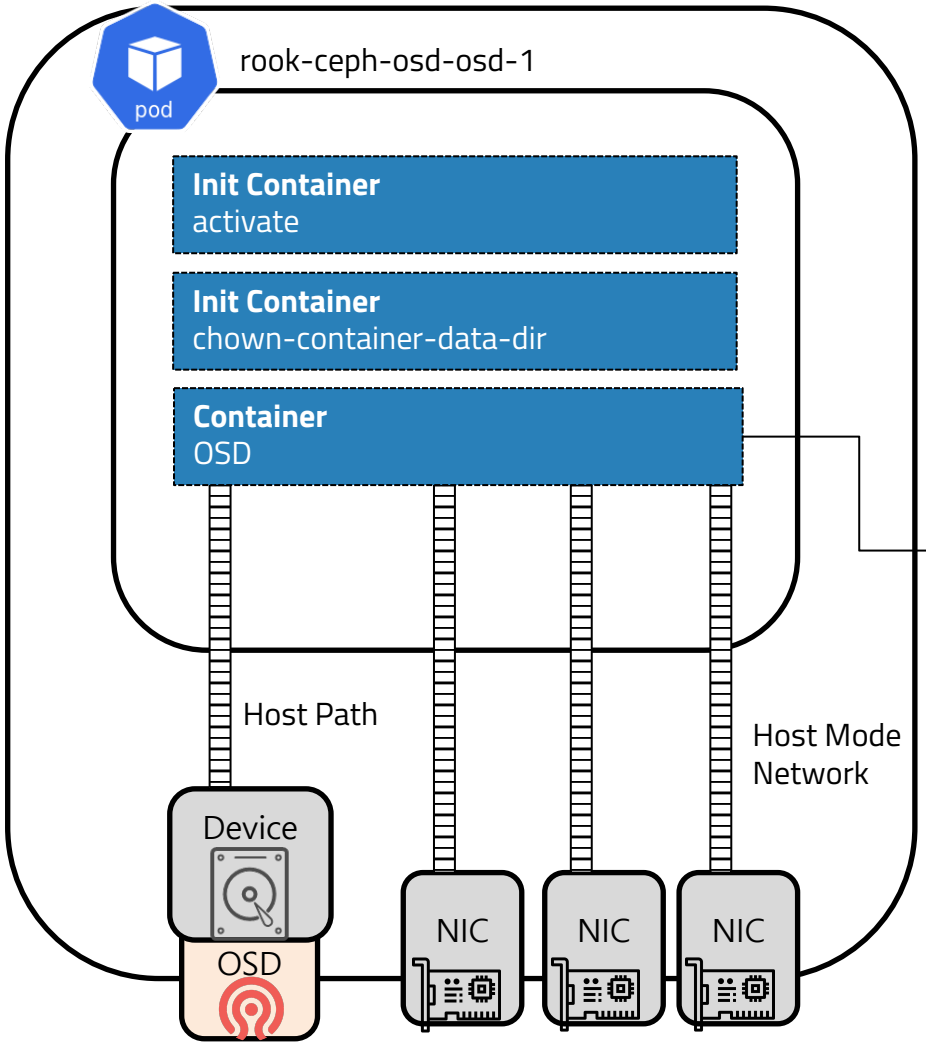
- It sets permissions to the log path that Crashcollector will collect.



How to deploy rook ceph ?

Run OSD Pod

- Execute the **ceph-osd** command by the osd container.

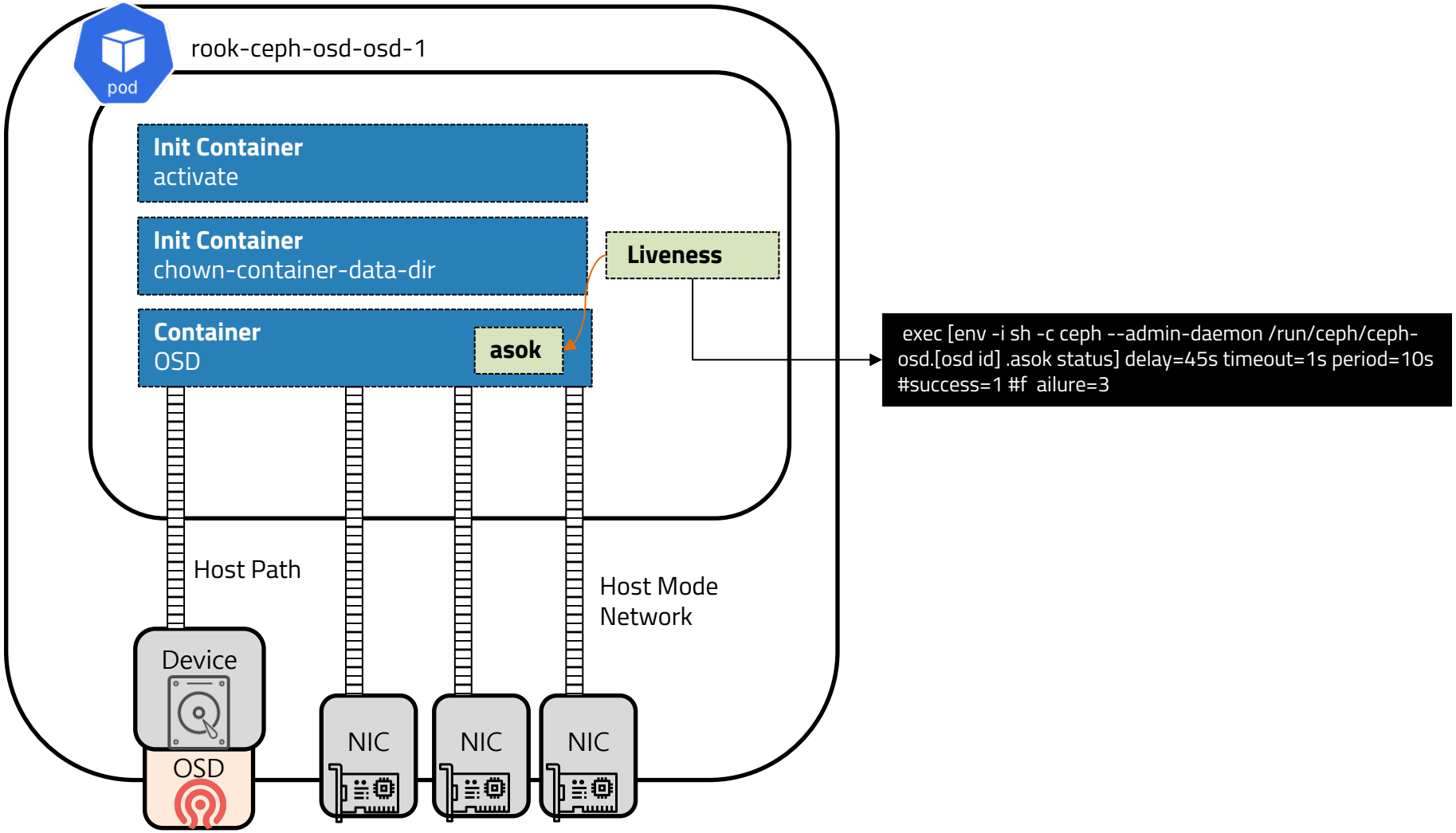



```
ceph-osd --foreground  --id  [id] --fsid [fsid]  --setuser  ceph  --
setgroup  ceph
 --crush-location=root=default host=stg1-obs-cache-001   \
 --log-to-stderr=true      \
 --err-to-stderr=true      \
 --mon-cluster-log-to-stderr=true            \
 --log-stderr-prefix=debug  \
 --default-log-to-file=false \
 --default-mon-cluster-log-to-file=false
```

How to deploy rook ceph ?

Run OSD Pod

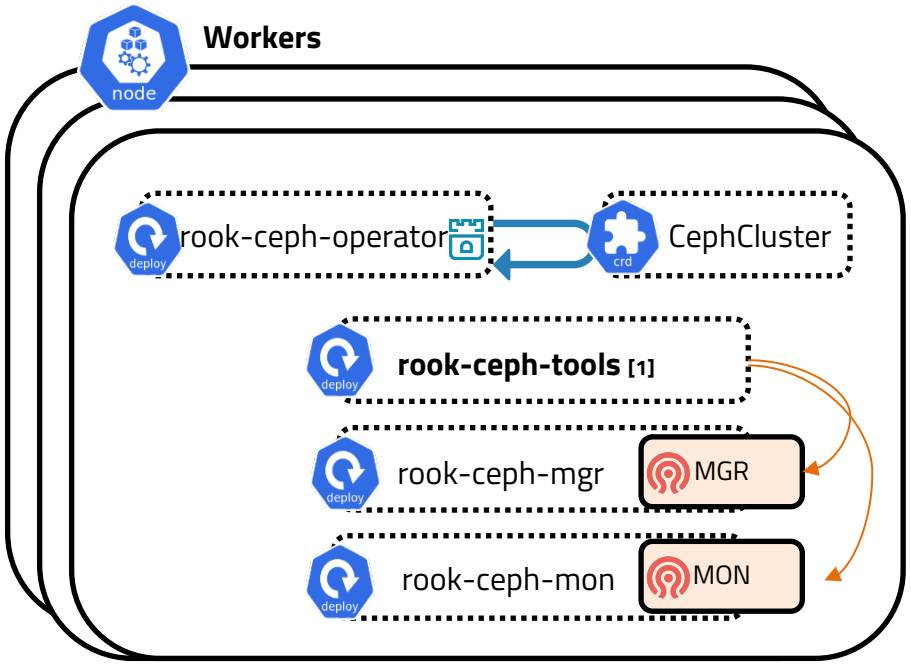
- Check Osd socket file and check liveness

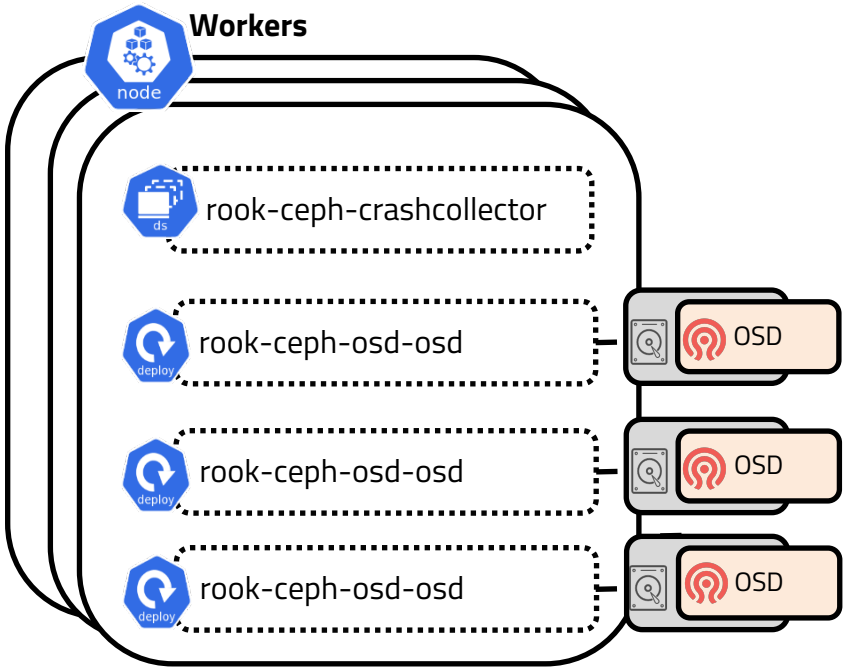


How to deploy rook ceph ?

How to connect a client to Ceph Cluster

- Connect directly from host by installing Ceph client binary
  - A network connected to the monitor on the client node is required.
- Use rook-ceph-tools pod [1]
  - A network connected to the monitor in the client node is not required.
- Use kubectl-rook-ceph (https://github.com/rook/kubectl-rook-ceph)
  - By using the Krew plugin, only the ceph command is still provided, under development



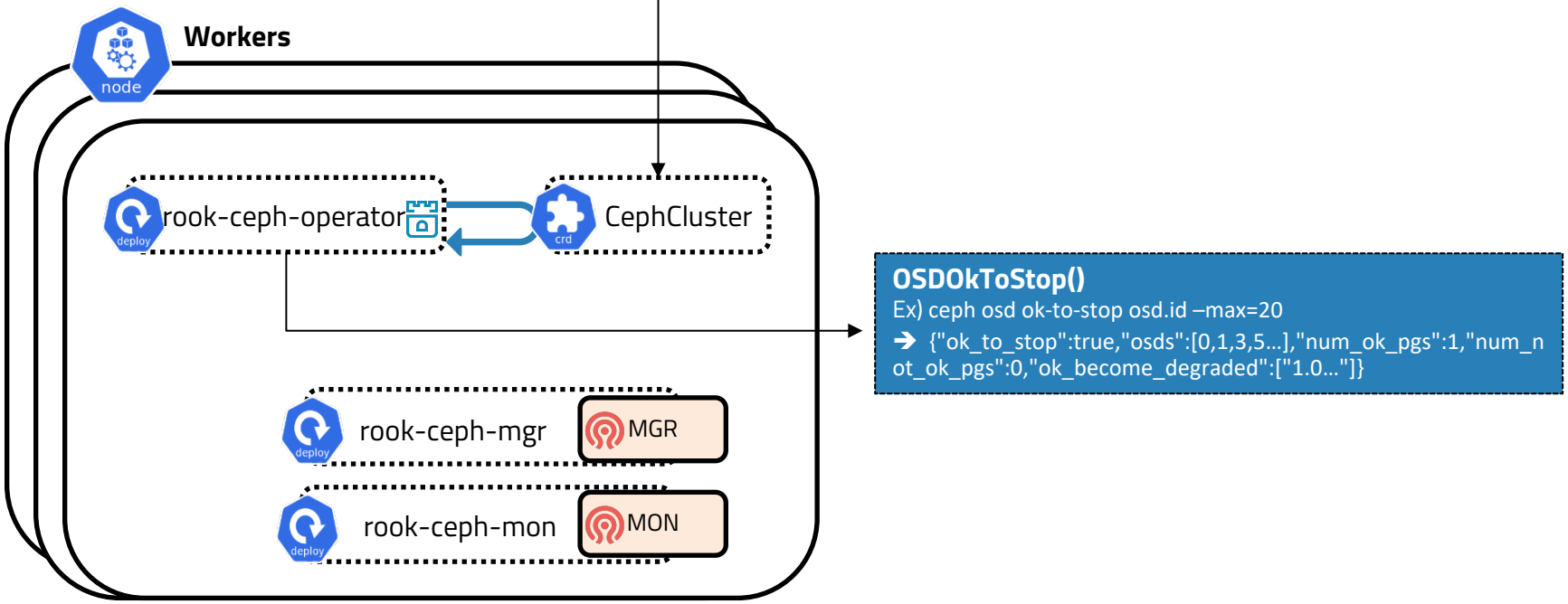


How to deploy rook ceph ?

Upgrade Ceph Version

- Easily upgrade CephCluster by pod with changing Ceph image tag of CephCluster
- Execute **ceph osd ok-to-stop** command to proceed with the upgrade while replacing the pod container images sequentially so that the impact does no

```
# kubectl  patch CephCluster $CLUSTER_NAME --type=merge -p "{\"spec\": {\"cephVersion\": {\"image\":\"$NEW_CEPH_IMAGE\"}}}"
```



How to upgrade rook ceph

Upgrade Ceph Version

- Easily upgrade CephCluster by pod with changing Ceph image tag of CephCluster
- Execute **ceph osd ok-to-stop** command to proceed with the upgrade while replacing the pod container images sequentially so that the impact does no

```
root default
    room room-1
        rack rack-1-1
            host cy01-ceph001
                osd.0
                osd.3
                osd.6
                osd.9
        rack rack-1-2
            host cy01-ceph002
                osd.1
                osd.4
                osd.7
                osd.10
        rack rack-1-3
            host cy01-ceph003
                osd.2
                osd.5
                osd.8
                osd.11
```

```
root@cy02-test050:~# ceph osd ok-to-stop 0 --max 20
{"ok_to_stop":true,"osds":[0,3,6,9],"num_ok_pgs":17,"num_not_ok_pgs":0,"ok_become_degraded"
```

Update Container Image Tag OSD 0,3,6,9.

Health OK & Upgrade Complete

```
root@cy02-test050:~# ceph osd ok-to-stop 1 --max 20
{"ok_to_stop":true,"osds":[1,4,7,10],"num_ok_pgs":17,"num_not_ok_pgs":0,"ok_become_degraded":
```

Update Container Image Tag OSD 1,4,7,10

Health OK & Upgrade Complete

```
root@cy02-test050:~# ceph osd ok-to-stop 11 --max 20
{"ok_to_stop":true,"osds":[2,5,8,11],"num_ok_pgs":17,"num_not_ok_pgs":0,"ok_become_degraded"
```

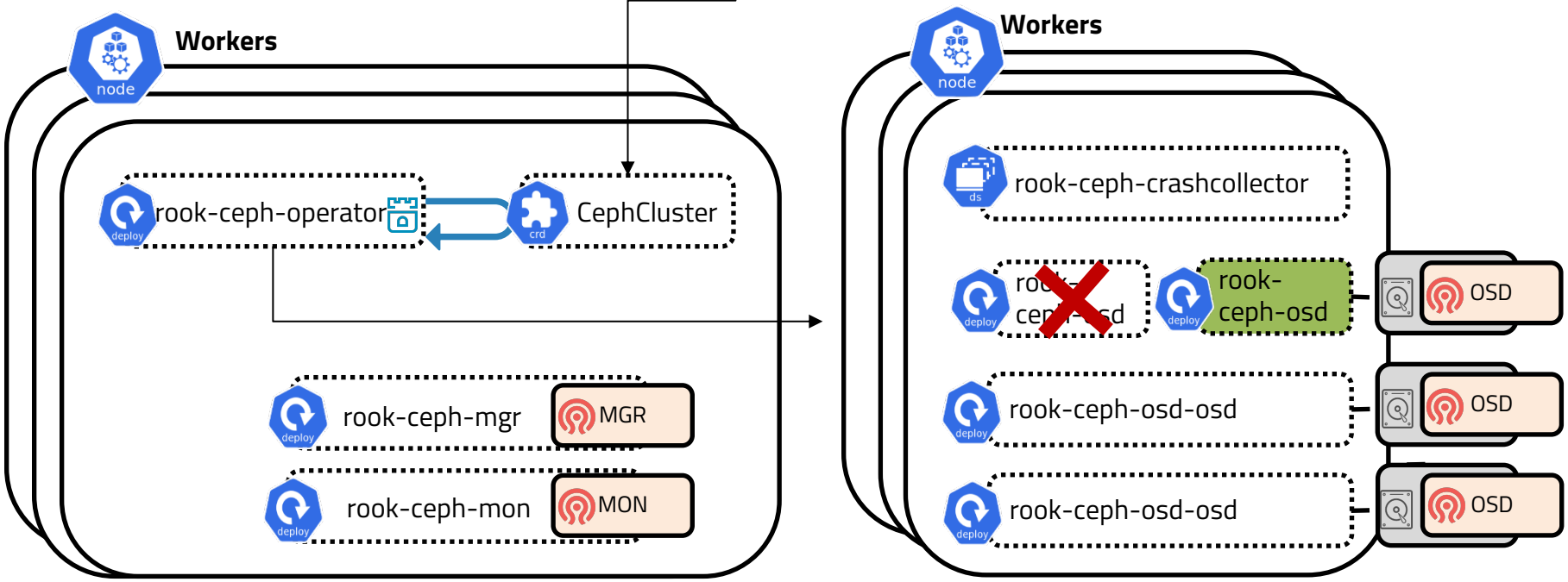
Update Container Image Tag OSD 2,5,8,11

How to upgrade rook ceph

Upgrade Ceph Version

- Easily upgrade CephCluster by pod with changing Ceph image tag of CephCluster
- Execute **ceph osd ok-to-stop** command to proceed with the upgrade while replacing the pod container images sequentially so that the impact does no

```
# kubectl  patch CephCluster $CLUSTER_NAME --type=merge -p "{\"spec\": {\"cephVersion\": {\"image\":\"$NEW_CEPH_IMAGE\"}}}"
```

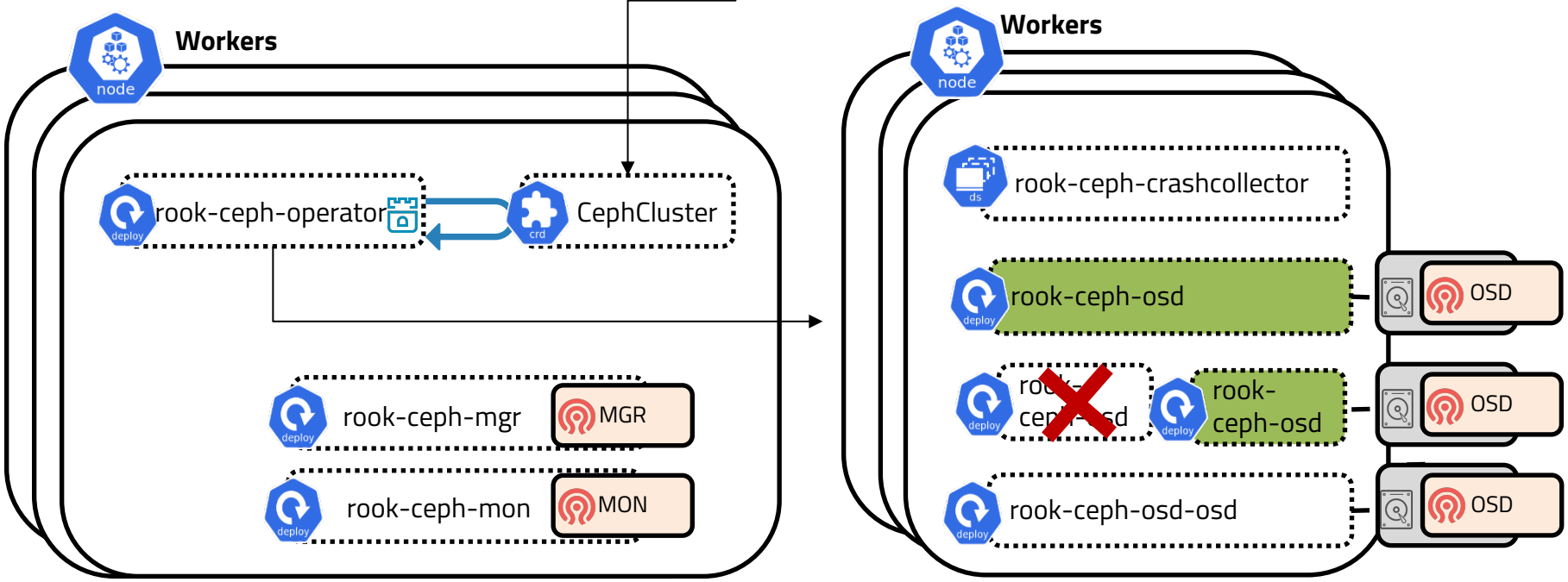


How to upgrade rook ceph

Upgrade Ceph Version

- Easily upgrade CephCluster by pod with changing Ceph image tag of CephCluster
- Execute **ceph osd ok-to-stop** command to proceed with the upgrade while replacing the pod container images sequentially so that the impact does no

```
# kubectl  patch CephCluster $CLUSTER_NAME --type=merge -p "{\"spec\": {\"cephVersion\": {\"image\":\"$NEW_CEPH_IMAGE\"}}}"
```



How to deploy Object Store

Rook Ceph Object Store (Ceph RGW)

- Deploy Ceph Object Store through CRD
  - CephObjectStore : Ceph Object Store CRD
  - CephObjectRealm: Ceph Object Realm CRD
  - CephObjectZoneGroup: Ceph Object Zone Group CRD
  - CephObjectZone: Ceph Object Zone CRD
  - ObjectBucketClaim: Ceph Object Bucket Claim
  - CephObjectStoreUser: Ceph Object Store User CRD

```
# kubectl create  -f object.yaml
```

```yaml
apiVersion: ceph.rook.io/v1
kind: CephObjectStore
metadata:
  name: my-store
  namespace: rook-ceph
spec:
  metadataPool:
    failureDomain: host
    replicated:
      size: 3
dataPool:
    failureDomain: host
    replicated:
      size: 3
gateway:
    sslCertificateRef:
    port: 80
    instances: 1
    placement:
    annotations:
    labels:
    resources:
  healthCheck:
    bucket:
      disabled: false
      interval: 60s
    livenessProbe:
      disabled: false
```

RGW Pool Setting

RadosGW Setting

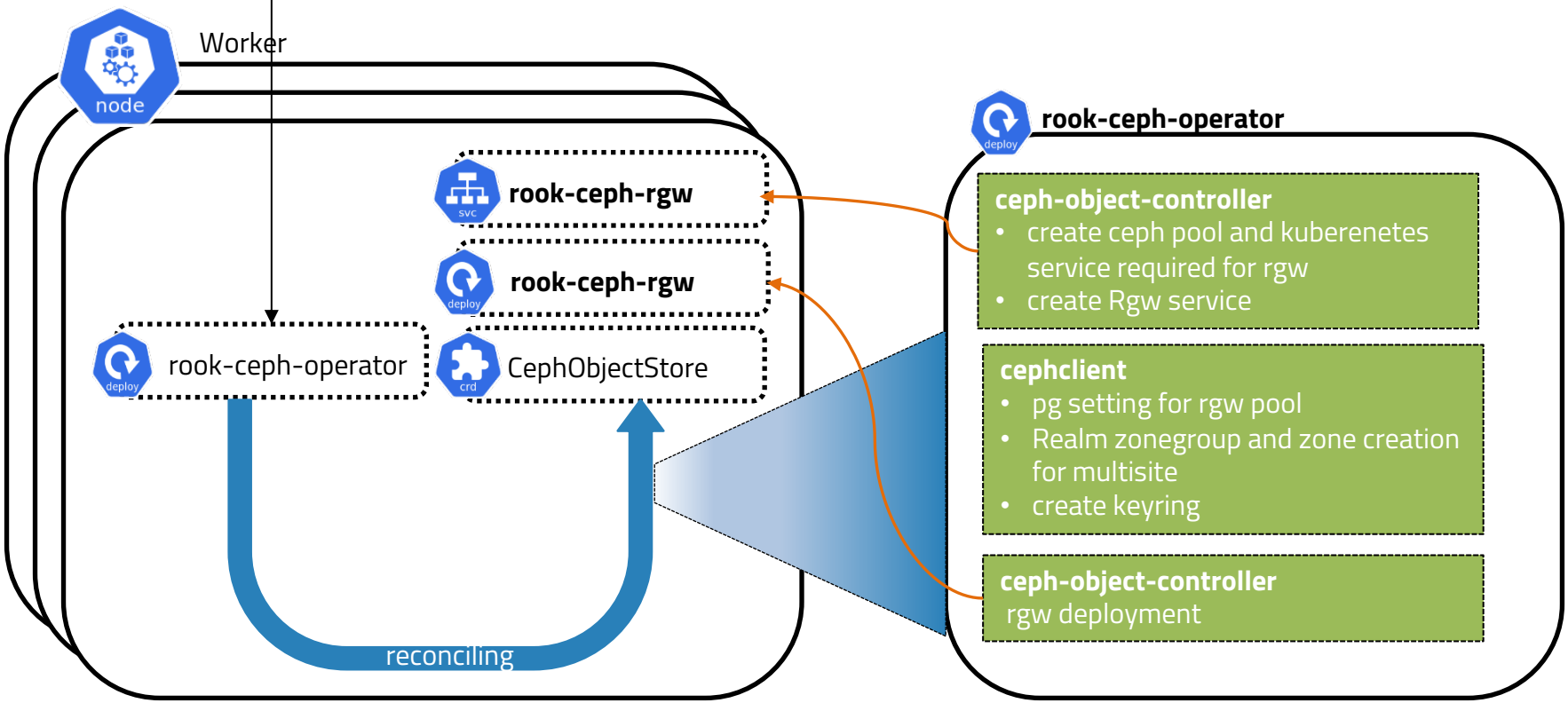
Health Check Setting

How to deploy Object Store

Rook Ceph Object Store (Ceph RGW)

- To create a radosgw pod in the operator, use the ceph client to create pool, realm, zone, zonegroup
- Deploy rgw pod when the initialization process is completed

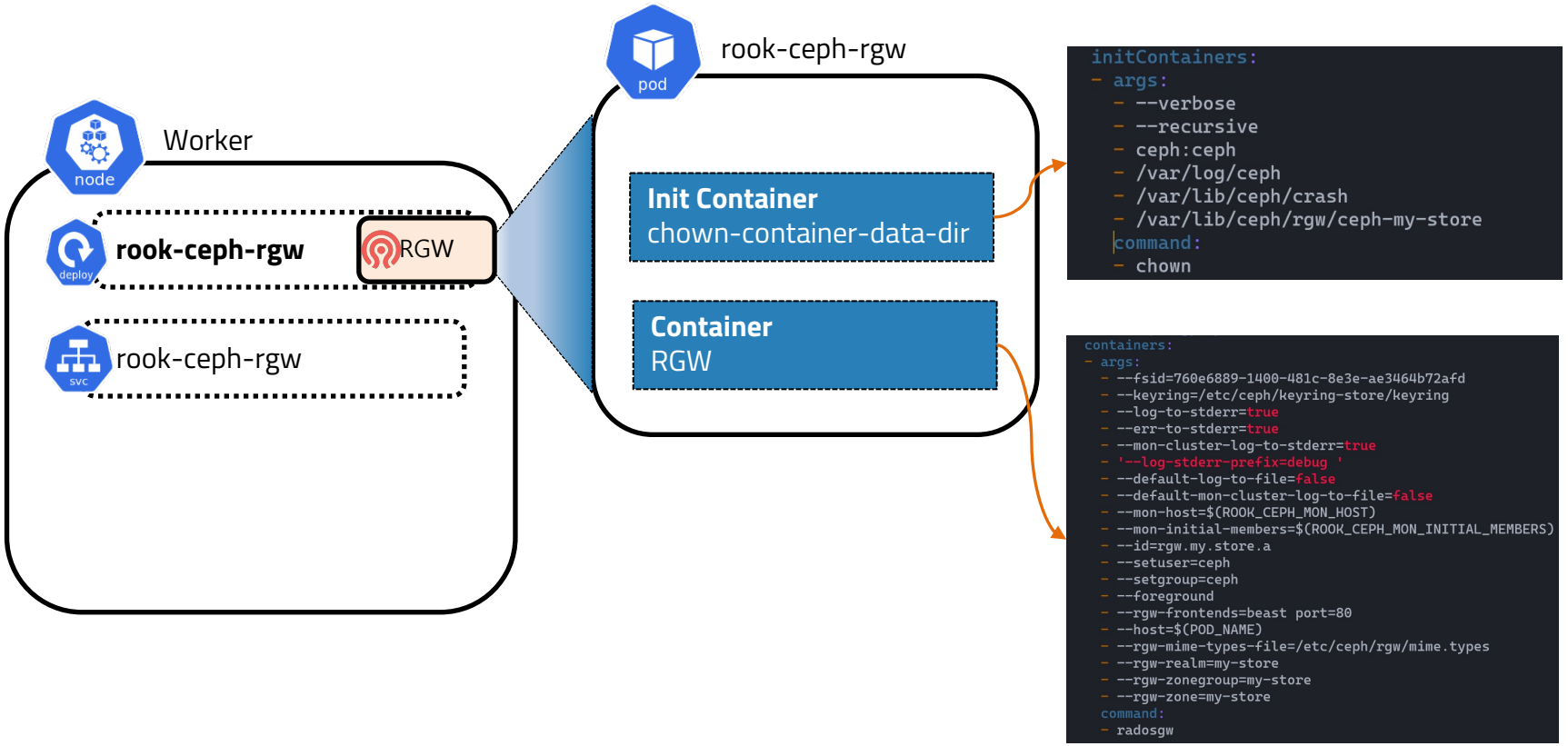
```
# kubectl create  -f object.yaml
```



How to deploy Object Store

Rook Ceph Object Store (Ceph RGW)

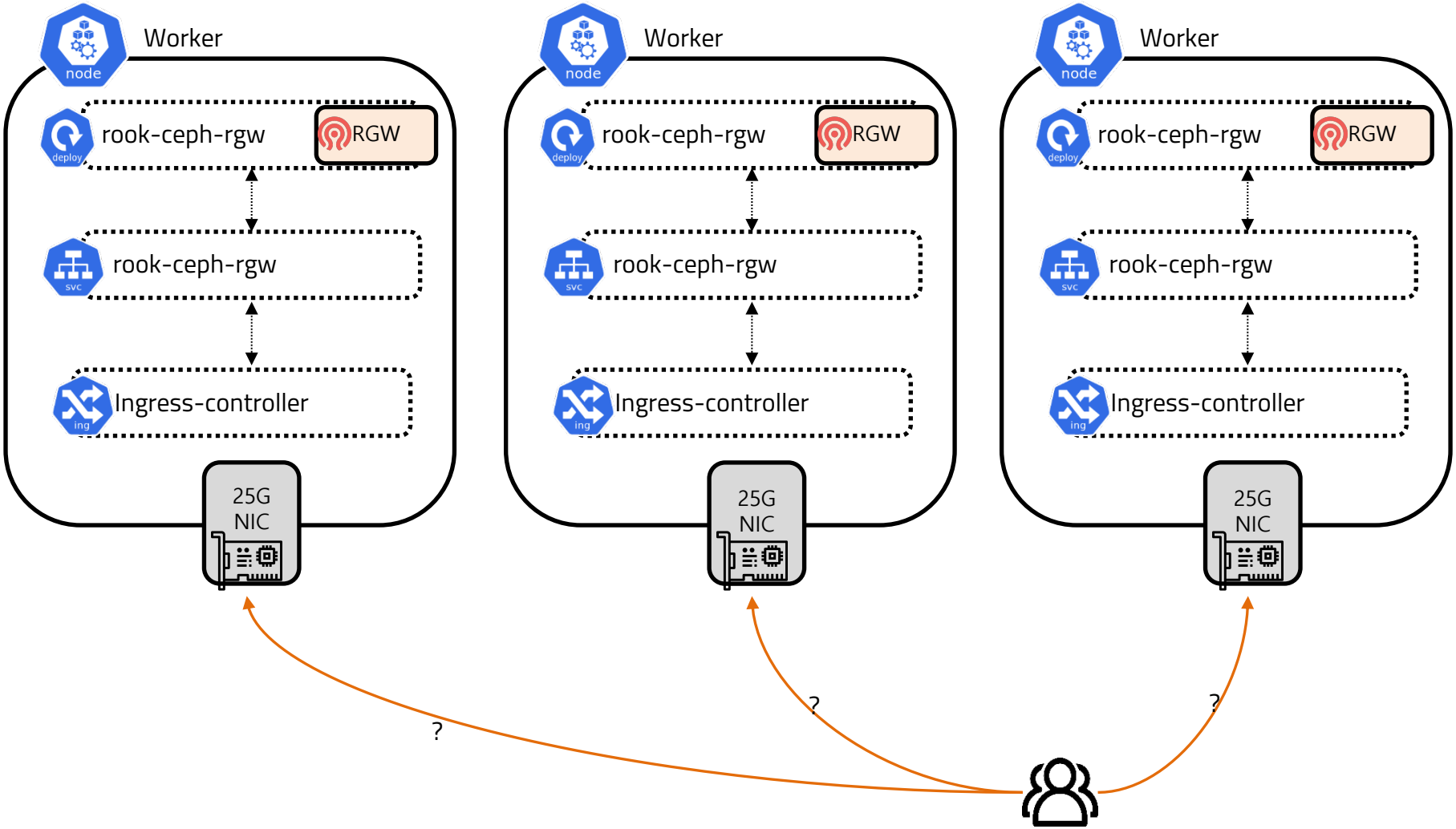
- After completion of Init Container Create an rgw pod using the radosgw command



How to access radosgw

Rook Ceph Object Store (Ceph RGW)

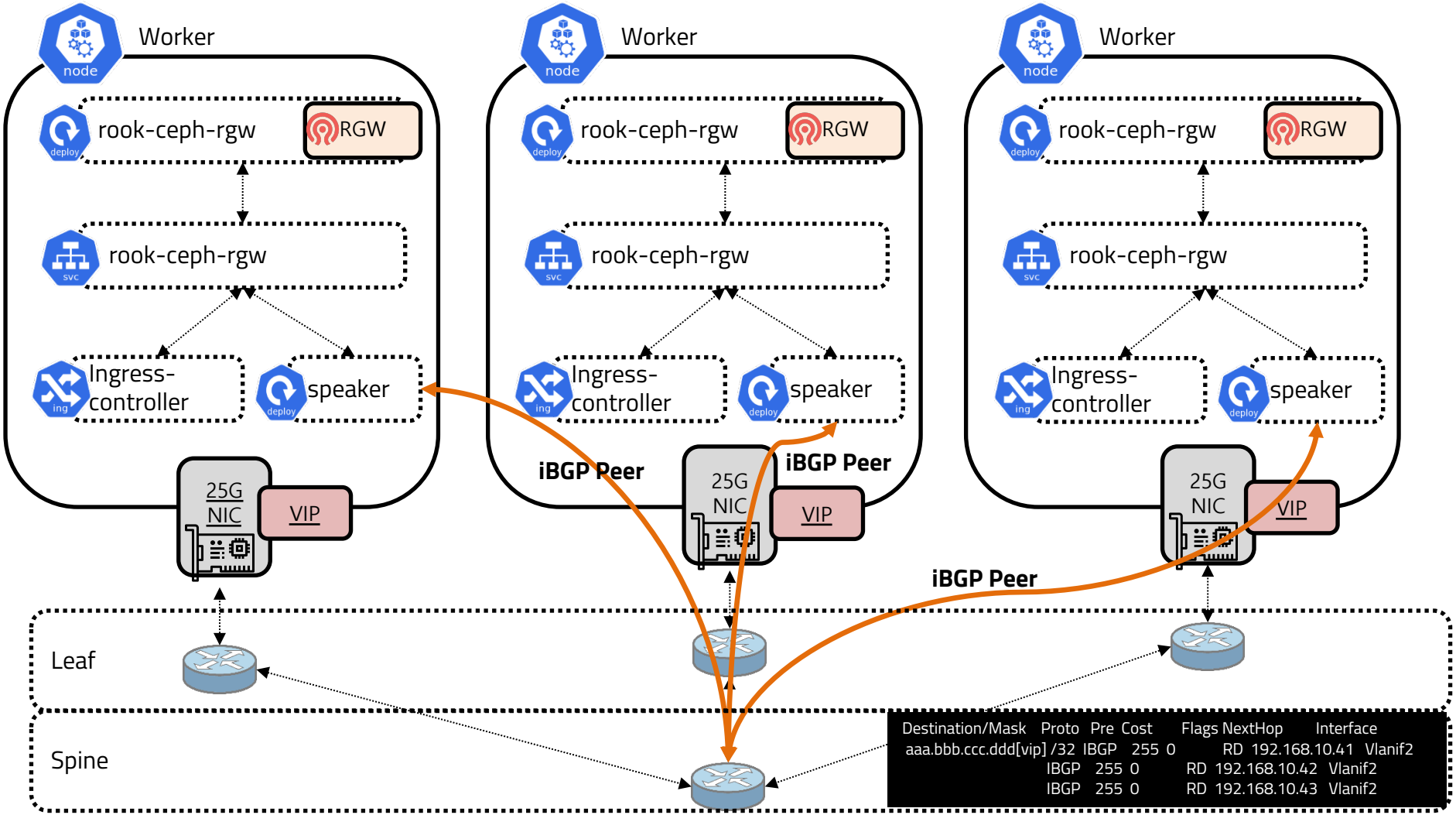
- Place rgw pods on 3 nodes to do a lot of processing
- For external exposure, nginx ingress controller was used.
- However, an lb , ecmp switch that distributes traffic to three nodes is very expensive. (using 25G nic...)



How to access radosgw

Rook Ceph Object Store (Ceph RGW)

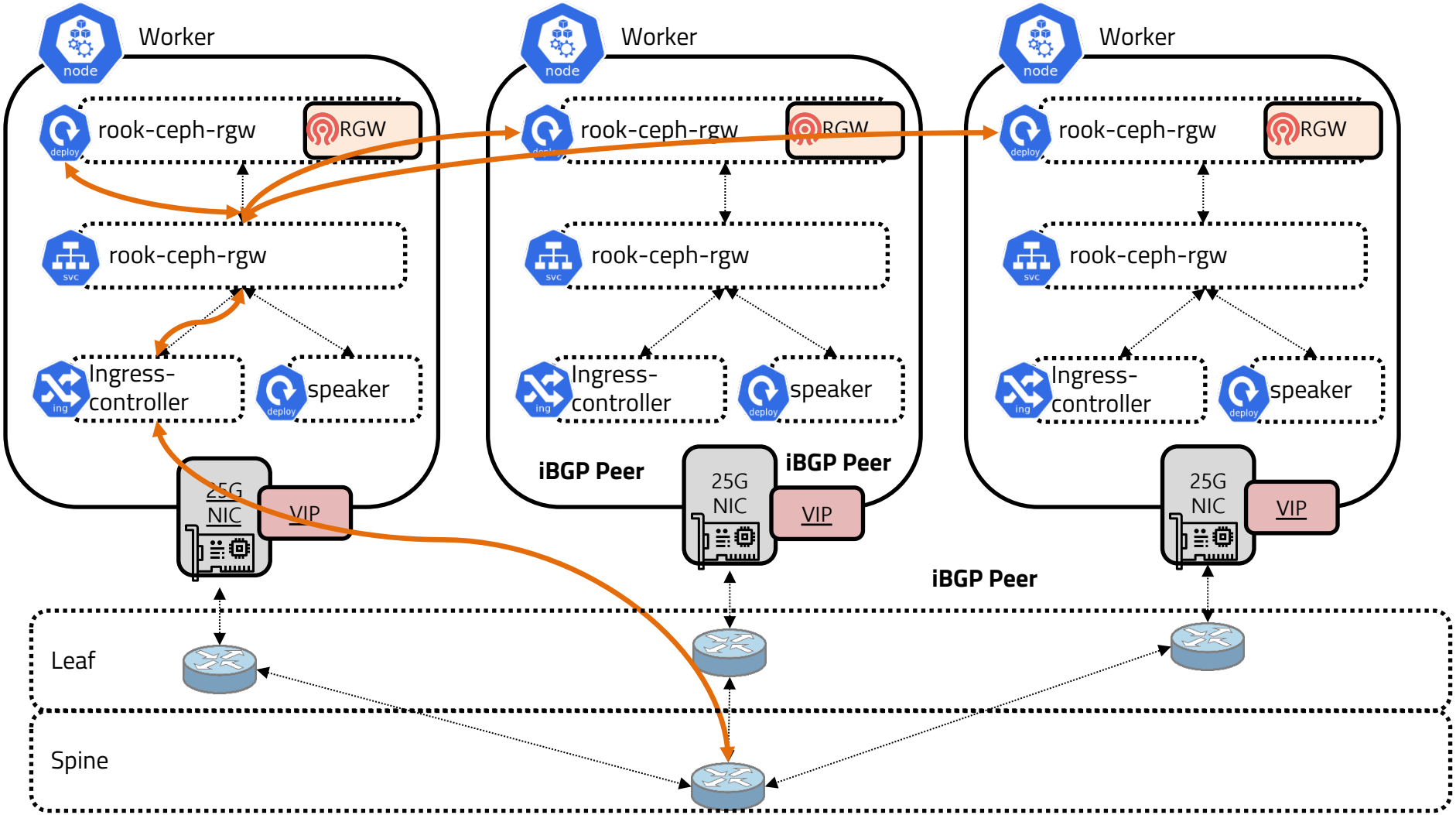
- Traffic is distributed from spine to multipath using metallb bgp mode
- A metallb speaker is deployed on the rgw node to become an ibpg peer with the spine.
- Configure vip assigned by speaker as service of ingress controller



How to access radosgw

Rook Ceph Object Store (Ceph RGW)

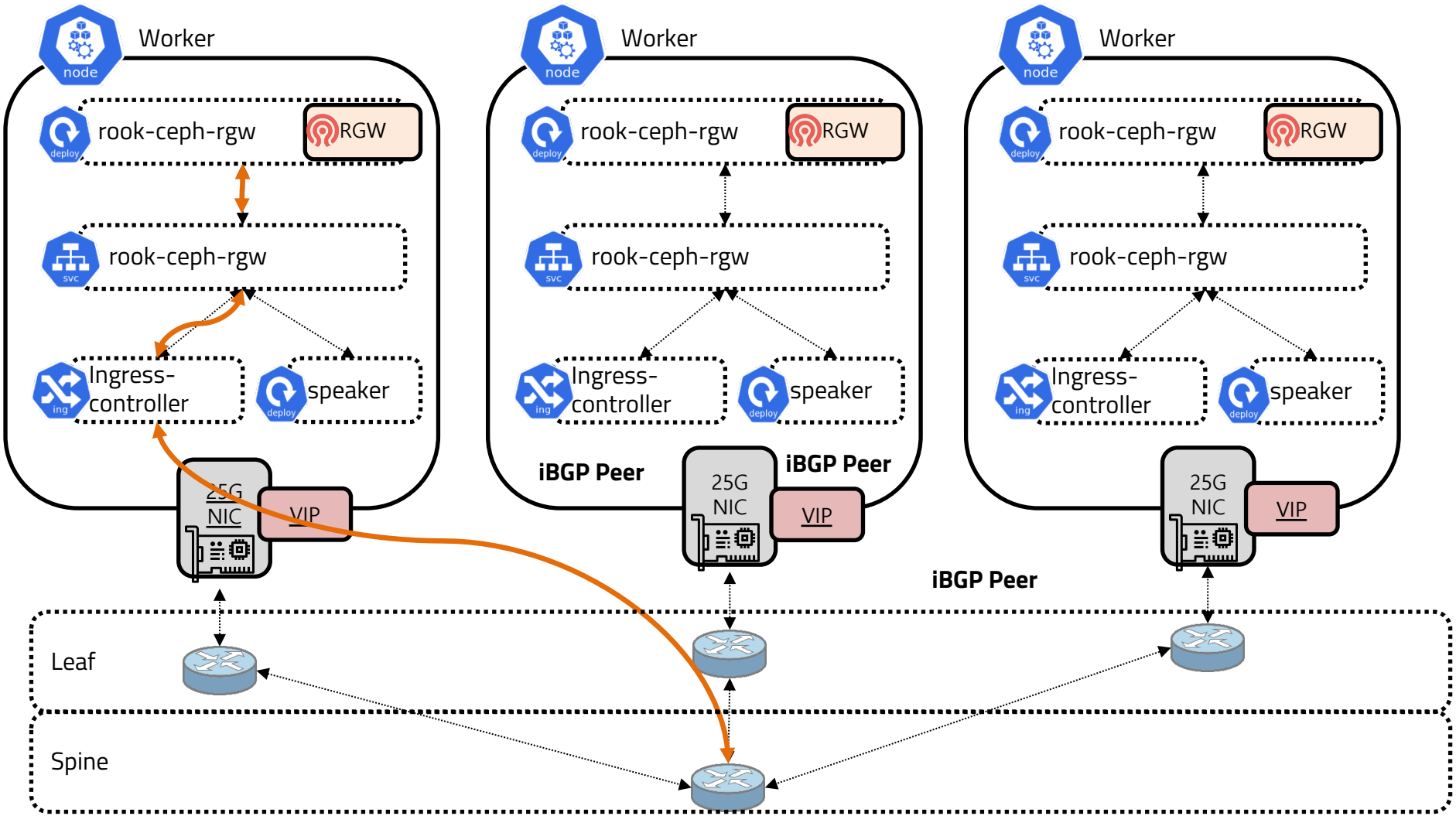
- However, the traffic coming in from the outside is distributed, but once again in the rgw service.
  - Unnecessary traffic between nodes occurs
  - Unable to preserve client source ip address



How to access radosgw

Rook Ceph Object Store (Ceph RGW)

- because the service's **externalTrafficPolicy** option is **cluster** by default.
  - By changing the **externalTrafficPolicy from cluster to host**, you can remove unnecessary traffic and preserve the client source ip



마지막으로

생각할점

- osd_memory_target 옵션은 고민해서 넣을것
- S3cmd to s5cmd
- 생각보다 많은 벤더의 참여로 빠른 버그 픽스 및 도움 주는 사람들 많음
- 업그레이드는 스테이징/테스트환경에서 많이 하고 진행
  - 자동화 된 argocd 와 같은 tool 을 이용한 code 화 된 배포 프로세스를 사용하는 것이 좋음
- Object 사용시 csi 를 사용을 하는 경우가 많은데 불필요한 csi 의 태생이 filesystem 을 중점을 만들어 졌기에 불필요 하거나 , 이슈가 많음
  - Ceph-cosi 가 정식적으로 릴리즈 된다면 많은 도움이 될 것 같다.