



WELCOME

Artem Torubarov

14th October 2026, Zurich





NOW ALSO IN ROOK

OSD Replacement

an imperative operation in a declarative world

Artem Torubarov
October, 2026

CLY'SO

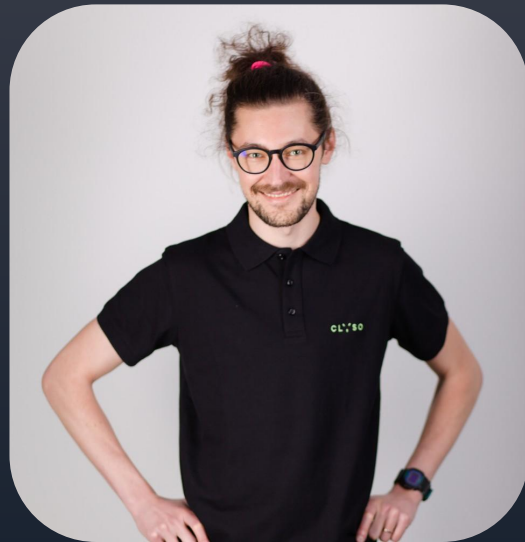
whoami

Artem Torubarov

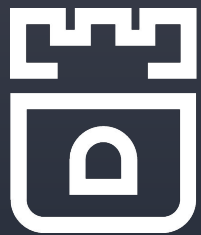
Software engineer at Clyso GmbH

Rook contributor

github.com/arttor



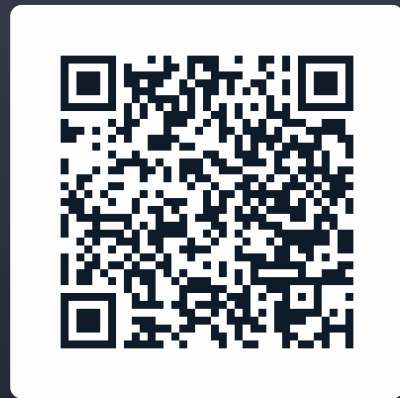
Why this talk



ROOK

v1.21 Release

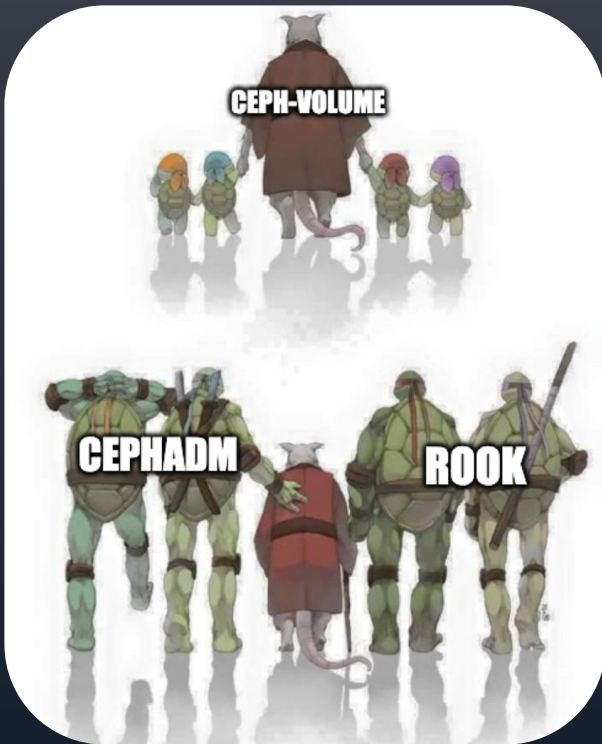
Automates OSD replacement



Release blog post

Agenda

- 01 OSD replacement: DIY
- 02 How **cephadm** automated it
- 03 What we need to know about K8s
- 04 How it is done in **Rook**
- 05 How to automate it even more



Replacing an OSD by hand

01

REMOVE

```
ceph osd out 5
```

PGs start moving to other OSDs

```
ceph osd safe-to-destroy 5
```

wait until no PGs are left on osd.5

```
# stop the osd.5 daemon
```

osd.5 **down**

```
ceph osd destroy 5 \  
  --yes-i-really-mean-it
```

osd.5 **destroyed**: keys deleted, ID, CRUSH position and weight kept

Replacing an OSD by hand

01

REMOVE

```
ceph osd out 5
```

PGs start moving to other OSDs

```
ceph osd safe-to-destroy 5
```

wait until no PGs are left on osd.5

```
# stop the osd.5 daemon
```

osd.5 **down**

```
ceph osd destroy 5 \  
  --yes-i-really-mean-it
```

osd.5 **destroyed**: keys deleted, ID, CRUSH position and weight kept

03

SWAP

swap the disk

04

PROVISION

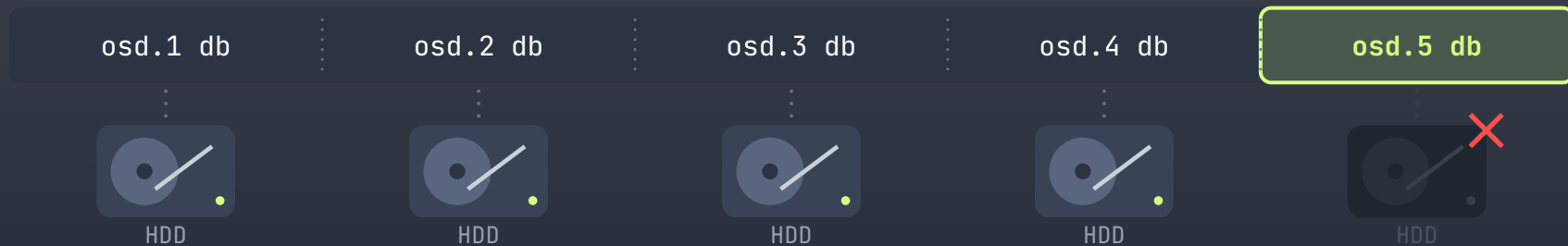
```
lsblk
```

```
ceph-volume lvm create \  
  --osd-id 5 --data /dev/sdX
```

osd.5 **up, in**: the same PGs move back to it

The real case: shared metadata, encryption

NVMe



```
# ceph-volume lvm list 5
===== osd.5 =====
[db]          /dev/ceph-1f0e.../osd-db-7c6b...
encrypted      1
osd id         5
type           db
devices        /dev/nvme0n1

# ceph-volume lvm create --osd-id 5 --data /dev/sdX \
  --block.db ceph-1f0e.../osd-db-7c6b... --dmccrypt
```

Replacing an OSD with cephadm

01

REMOVE

```
ceph orch set-unmanaged osd.hdd-nvme
```

cephadm **stops creating** OSDs from this spec

```
ceph orch osd rm 5 --replace
```

out, drain, destroy, old disk **zapped**

02

WAIT FOR OK

```
ceph orch osd rm status
```

wait until osd.5 is **destroyed**

Replacing an OSD with cephadm

01 REMOVE	<pre>ceph orch set-unmanaged osd.hdd-nvme</pre>	cephadm stops creating OSDs from this spec
	<pre>ceph orch osd rm 5 --replace</pre>	out, drain, destroy , old disk zapped
02 WAIT FOR OK	<pre>ceph orch osd rm status</pre>	wait until osd.5 is destroyed
03 SWAP	swap the disk	
04 PROVISION	<pre>ceph orch set-managed osd.hdd-nvme</pre>	the new disk becomes osd.5 , DB on the NVMe

So why did Rook take until 1.21?

Kubernetes API

spec | desired state | changed by user

```
mon:  
  count: 3
```

status | reality | changed by env

mon a	node-1	running
mon b	node-3	running
mon c	node-2	down

Controller

spec | desired state | changed by user

```
mon:  
  count: 3
```

status | reality | changed by env

```
mon a  node-1  running  
mon b  node-3  running  
mon c  node-2  down
```

Compare spec VS status

```
# Act if not match:  
create mon d on node-4  
remove mon c
```



Controller

spec | desired state | changed by user

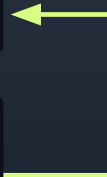
```
mon:  
  count: 3
```

status | reality | changed by env

```
mon a  node-1  running  
mon b  node-3  running  
mon d  node-4  running
```

Compare spec VS status

Act if not match:



GitOps

clusters/prod/ceph-cluster.yaml		
↑	@@ -5,5 +5,5 @@	metadata:
5	5	namespace: rook-ceph
6	6	spec:
7	7	cephVersion:
8	-	image: quay.io/ceph/ceph:v19.2.3
8	+	image: quay.io/ceph/ceph:v20.2.4
9	9	allowUnsupported: false

GitOps

```
clusters/prod/ceph-cluster.yaml
```

		@@ -5,5 +5,5 @@ metadata:
5	5	namespace: rook-ceph
6	6	spec:
7	7	cephVersion:
8	-	image: quay.io/ceph/ceph:v19.2.3
8	+	image: quay.io/ceph/ceph:v20.2.4
9	9	allowUnsupported: false

- No Ceph Access
- No Ceph client
- Only Git creds - SSO
- Protected branch rules
- Reproducible

Aaaaand git history...

	1 <code>kind: CephBlockPool</code> 2 <code>metadata:</code> 3 <code>name: three-replica-pool</code> 4 <code>spec:</code> 5 <code>replicated:</code>
2 years ago  emergency, only temporarily	6 <code>size: 2</code>

The crux

spec

```
storage:  
  deviceFilter: "^sd[b-f]"
```

status

```
osd.5  down
```

Which line do I change?

Imperative operation

```
$ kubectl rollout restart deployment/web
```

Imperative operation

```
$ kubectl rollout restart deployment/web
```

```
spec:
```

```
  template:
```

```
    metadata:
```

```
      annotations:
```

```
+       kubectl.kubernetes.io/restartedAt: "2026-10-08T14:03:12+02:00"
```

Annotations

```
apiVersion: apps/v1
kind: Deployment
metadata:
  name: mon
  annotations:
    example.com/anything: "any value"
spec:
  replicas: 3
  template: ...
status:
  availableReplicas: 2
```



argocd.argoproj.io/refresh



cert-manager.io/cluster-issuer

Replacing an OSD with Rook

01

REMOVE

```
$ kubectl -n rook-ceph annotate deployment rook-ceph-osd-5 \
  osd.rook.io/replace=yes-really-replace-osd-5
```

Rook answers

01

REMOVE

```
$ kubectl -n rook-ceph annotate deployment rook-ceph-osd-5 \
  osd.rook.io/replace=yes-really-replace-osd-5
```

02

WAIT FOR OK

```
$ kubectl -n rook-ceph get deploy rook-ceph-osd-5 \
  -o jsonpath='{.metadata.annotations}'
{...
"osd.rook.io/replace-ready-for-swap":"true", ...}
```

Rook answers

01

REMOVE

```
$ kubectl -n rook-ceph annotate deployment rook-ceph-osd-5 \
  osd.rook.io/replace=yes-really-replace-osd-5
```

02

WAIT FOR OK

```
$ kubectl -n rook-ceph get deploy rook-ceph-osd-5 \
  -o jsonpath='{.metadata.annotations}'
{...
  "osd.rook.io/replace-ready-for-swap":"true", ...}
```

03

SWAP

swap the disk

04

PROVISION

automatic: the new disk becomes osd.5 again

It is the Kubernetes API

It is the Kubernetes API

- No Ceph Access, No Ceph client
- K8s RBAC
- Other k8s apps can also use it

A SMART move



your failure detector

disk failing → osd.5

Slack: "replace osd.5?"

Approve

\$ `kubectl` annotate



drain, destroy

\$ `ceph` device light on



ready-for-swap

Rook

Self-driving drives



your failure detector

disk failing → osd.5

Slack: "replace osd.5?"

Approve

\$ `kubectl annotate`

\$ `ceph device light on`

LED off, close ticket

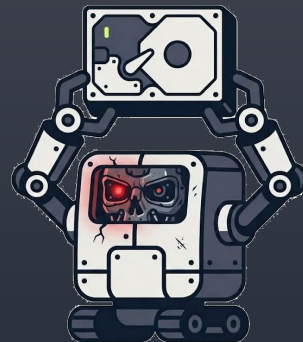
Rook

drain, destroy

ready-for-swap

osd.5 back

your DC automation



TARGET ACQUIRED: osd.5

swap!

Thank you!