Photo by NASA on Unsplash

Kevin Hrpcek

ceph-csi-cephfs/rbd to ceph-csi-operator

This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977.

ceph-csi

- https://github.com/ceph/ceph-csi
- Container storage interface driver for kubernetes
- Allows creation/deletion/modification of RBD and cephfs volumes
- Mounts volumes as assigned across the cluster

What it is missing

- CSI Addons
  - https://github.com/csi-addons/kubernetes-csi-addons
  - Enables additional features like reclaiming space and volume replication via kubernetes objects
- Not a kubernetes operator
  - Operators focus on extending functionality via custom resource definitions
  - Improves automation

Migrating Restrictions

- 40+ volumes per cluster already using the old ceph-csi (mostly rbd)
- The driver definitions are already using the default names, don’t want to come up with non default names if not necessary
- Minimize downtime
- New driver creates RBDs with different feature set
- Don’t want to have to clone volumes to new rbd
- Don’t want to change storage class name
- Changing storage class details will break the existing volumes if they need to be remounted
  - Can keep running as long as a kubernetes pod doesn’t restart
- Not many docs exist

Prerequisites

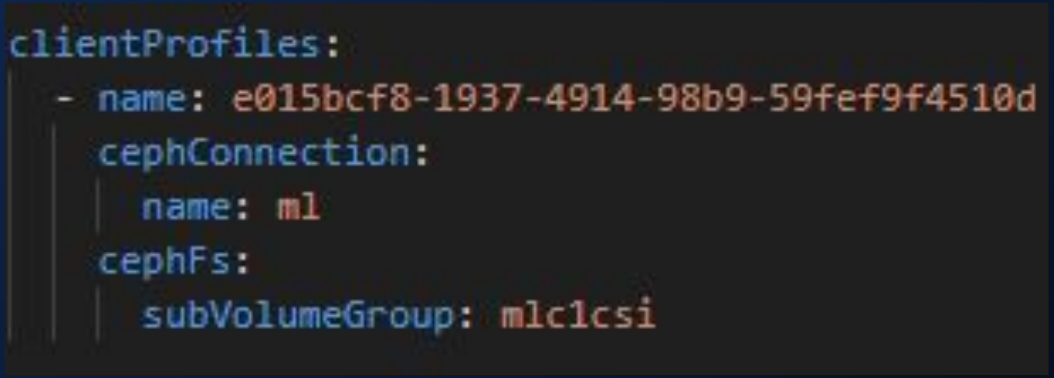
- Kubernetes
- Ceph cluster
- Rbd pools set up
- To enable rbd mirroring
  - 2nd ceph cluster with rbd mirroring enabled & daemons running

Migrating

- Storage class used to use the cluster ID, now it has to be the name of the client profile
  - This was the important thing that made a nearly seamless transition possible
- Delete old ceph-csi-* deployments & existing storageclasses
  - no effect on existing pvc/pv but they can't be remounted now
- Delete rbd & cephfs csi drivers
- Deploy ceph-csi-operator portion
  - Sets up CRDs
  - Only update the kubernetes cluster domain
- Deploy ceph-csi-drivers
  - Contains a lot more modifications and a bug that requires a lot of extra defaults to be defined
  - Contains ceph cluster connection information (similar to previous method)

Migrating

- This important bit to a stress free transition

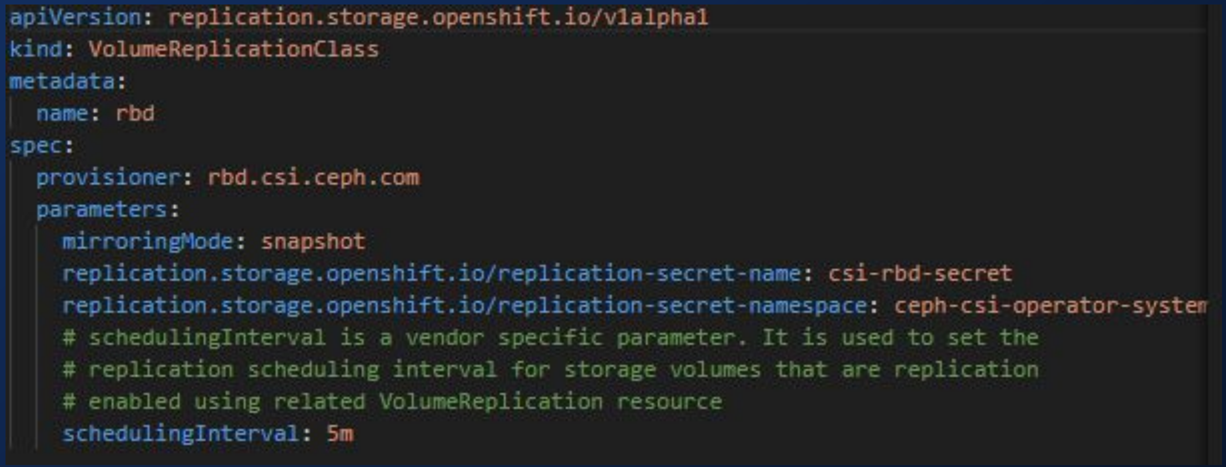


```
clientProfiles:
  - name: e015bcf8-1937-4914-98b9-59fef9f4510d
    cephConnection:
      name: ml
    cephFs:
      subVolumeGroup: mlc1csi
```

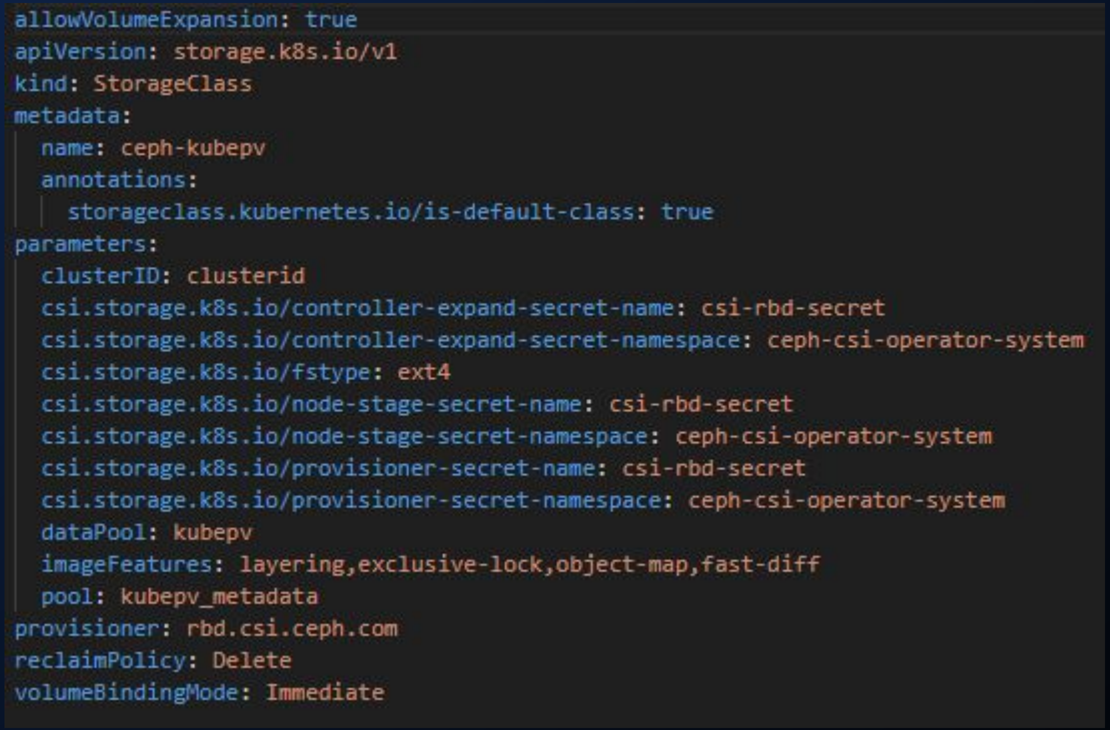
- Allows the new storage class to be identical to the previous one so legacy volume mounts don’t fail
  - Just need to keep an extra secret around with cluster credentials
- Seems to not enable rbd mirroring for old volumes. Needs more testing
- New volumes support new features
- No need to recreate all PVC

Migrating

- Operator method no longer creates StorageClasses
- Doesn't give you a method to define Volume replication classes
- Create a customization to handle these



```yaml
apiVersion: replication.storage.openshift.io/v1alpha1
kind: VolumeReplicationClass
metadata:
  name: rbd
spec:
  provisioner: rbd.csi.ceph.com
  parameters:
    mirroringMode: snapshot
    replication.storage.openshift.io/replication-secret-name: csi-rbd-secret
    replication.storage.openshift.io/replication-secret-namespace: ceph-csi-operator-system
    # schedulingInterval is a vendor specific parameter. It is used to set the
    # replication scheduling interval for storage volumes that are replication
    # enabled using related VolumeReplication resource
    schedulingInterval: 5m
```



```yaml
allowVolumeExpansion: true
apiVersion: storage.k8s.io/v1
kind: StorageClass
metadata:
  name: ceph-kubepv
  annotations:
    storageclass.kubernetes.io/is-default-class: true
parameters:
  clusterID: clusterid
  csi.storage.k8s.io/controller-expand-secret-name: csi-rbd-secret
  csi.storage.k8s.io/controller-expand-secret-namespace: ceph-csi-operator-system
  csi.storage.k8s.io/fstype: ext4
  csi.storage.k8s.io/node-stage-secret-name: csi-rbd-secret
  csi.storage.k8s.io/node-stage-secret-namespace: ceph-csi-operator-system
  csi.storage.k8s.io/provisioner-secret-name: csi-rbd-secret
  csi.storage.k8s.io/provisioner-secret-namespace: ceph-csi-operator-system
  dataPool: kubepv
  imageFeatures: layering,exclusive-lock,object-map,fast-diff
  pool: kubepv_metadata
provisioner: rbd.csi.ceph.com
reclaimPolicy: Delete
volumeBindingMode: Immediate
```

pv-migrate

- https://github.com/utkuozdemir/pv-migrate
- Useful utility to migrate contents of a PV to another
- Slightly annoying to have to use a 2 step process for every copy
  - Original pvc -> temp pvc
  - delete original pvc
  - temp pvc -> new pvc with original time

Eventually

- Migrate all pvc to new storage class to enable new features
- Contribute some improvements back to the github repo
  - Docs aren’t great
  - Typos in the default configuration
  - Helm chart problems
    - Permissions problems