KubeCon | CloudNativeCon

North America 2023

Agenda

- Who we are

- What we did in 1.28

- What we are working on in 1.29

- Features in design/prototyping

- How to get involved

SIG Storage

- Co-Chairs: Saad Ali (Google) and Xing Yang (VMware)
- Tech Leads: Michelle Au (Google) and Jan Šafránek (Red Hat)
- #sig-storage slack channel members: 5700+
- #csi slack channel members: 1500+
- #sig-storage-cosi slack channel members: 340+
- #csi-windows slack channel members: 240+
- SIG-Storage zoom meeting attendees: 25
- Unique approvers for SIG-owned packages: 30

What we do

- Defined in SIG Storage Charter
  - Persistent Volume Claims and Persistent Volumes
  - Storage Classes and Dynamic Provisioning
  - Kubernetes volume plugins
  - Secret, ConfigMap, DownwardAPI, Projected and EmptyDir Volumes (co-owned with SIG-Node)
  - Container Storage Interface (CSI)
    - CSI sidecars
    - Most CSI drivers are owned by SIG Cloud Provider or other community
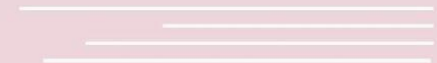 - Container Object Storage Interface (COSI)
    - Alpha since 1.25

What we did in 1.28

What we did in 1.28

- GA
  - [Reconcile Default Storage Class Assignment](#)
    - Allows existing unbound PVCs without "storageClassName" updated to use the new default StorageClass
  - (SIG-Node) [Non-graceful Node Shutdown](#)

Non-Graceful Node Shutdown

- [Graceful node shutdown](#)
  - "`ssh <node> poweroff`"
  - Shutdown command detected by Kubelet
  - Will stop containers + unmount volumes
- Non-graceful Node Shutdown
  - Can't ssh to node because of:
    - HW failure
    - Network lost
    - Kernel panic
    - …
  - Shutdown command not detected by Kubelet

Non-Graceful Node Shutdown

- Allows stateful workloads to restart on a different node if the original node is shutdown unexpectedly or ends up in a non-recoverable state perhaps due to hardware failure or unresponsive OS.
- GA in 1.28; Beta in 1.26; introduced as Alpha in 1.24
  a. GA [blog](#)
- Expected workflow
  a. Detect a node is unhealthy.
  b. Shut down the node (IPMI, cloud API, …)
  c. Apply taint to the **shutdown** node:
    - `node.kubernetes.io/out-of-service: "NoExecute"`

-> All Pods are force-deleted + all volumes are detached *immediately*.

What we did in 1.28 (cont.)

- Staying in Beta
  - [Selinux Relabeling with Mount Options](#)
    - Speeds up container startup by mounting volumes with the correct SELInux label
    - [Beta Blog](#)
  - [Robust VolumeManager Reconstruction after kubelet restart](#)
    - Allows kubelet to populate information about how existing volumes are mounted

- Alpha
  - [Recovering from Resize Failure](#)
    - Allow users to recover from volume expansion failure
    - Introduced as Alpha in 1.23
    - API changes In 1.28 - renamed `pvc.Status.ResizeStatus` to `pvc.Status.AllocatedResourceStatus` and made it a map.
  - [PV Last Phase Transition Time](#)
    - A new PersistentVolumeStatus field which holds a timestamp of when a PersistentVolume last transitioned to a different phase.

CSI Migration Schedule

Core CSI Migration is GA in 1.25

Driver	Alpha	Beta (in-tree deprecated)	Beta (on-by-default)	GA	Target "in-tree plugin" removal
OpenStack Cinder	1.14	1.18	1.21	1.24	1.26
Azure Disk	1.15	1.19	1.23	1.24	1.27
Azure File	1.15	1.21	1.24	1.26	1.30 (Target)
AWS EBS	1.14	1.17	1.23	1.25	1.27
GCE PD	1.14	1.17	1.23	1.25	1.28
vSphere *	1.18	1.19	1.25	1.26	1.30 (Target)
Portworx	1.23	1.25			

* vSphere version < 7.0u2 is no longer supported for in-tree vSphere volume in 1.25+

In-Tree Storage Driver Removal

Driver	Deprecated	Code Removal
Ceph RBD	1.28	1.31 (Target)
CephFS	1.28	1.31 (Target)
Flocker	1.22	1.25
GlusterFS	1.25	1.26
Quobyte	1.22	1.25
ScaleIO	1.16	1.22
StorageOS	1.22	1.25

What we are working on in 1.29 and beyond
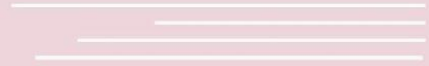
- Targeting GA
  - ReadWriteOncePod Persistent Volume Access Mode
    - A new ReadWriteOncePod access mode for PersistentVolumes that restricts access to a single pod on a single node.
  - NodeExpandSecret
    - Adds NodeExpandSecret to the CSI persistent volume source and enables the csi client to send it out as part of the nodeExpandVolume request to the csi drivers.

What we are working on in 1.29

- Targeting Beta
  - [PV Last Phase Transition Time](#)
    - Adds timestamp to PV.Status when the phase changes
  - [Prevent Unauthorized Volume Mode Conversion](#)
    - Volume Mode Conversion is only allowed if the annotation AllowVolumeModeChange is added on VolumeSnapshotContent.
    - Beta since 1.27. Feature flag will be set to true by default in external-snapshotter and external-provisioner in 1.29.
    - Action Required: For any application that relies on this workflow, code changes are required. Otherwise your application will fail.
  - (SIG-App) [Auto remove PVCs created by statefulset](#)
    - Adds an option to allow PVCs created by StatefulSet to be removed automatically
- Targeting Alpha
  - [Modifiable PVCs](#)
    - Add capabilities to the Kubernetes PersistentVolumeClaim API to allow users to dynamically control volume options through a VolumeAttributeClass.

Modifiable PVCs Example

```
apiVersion: storage.k8s.io/v1alpha1
kind: VolumeAttributesClass
metadata:
  name: silver
driver: foo.driver
parameters:
  iops: "5000"
```

```
apiVersion: storage.k8s.io/v1alpha1
kind: VolumeAttributesClass
metadata:
  name: gold
driver: foo.driver
parameters:
  iops: "10000"
```

```
apiVersion: v1
kind: PersistentVolumeClaim
metadata:
  name: my-volume
spec:
  volumeAttributesClassName: silver
  accessModes:
    - ReadWriteOnce
  resources:
    requests:
      storage: 100Gi
```

Update →

```
apiVersion: v1
kind: PersistentVolumeClaim
metadata:
  name: my-volume
spec:
  volumeAttributesClassName: gold
  accessModes:
    - ReadWriteOnce
  resources:
    requests:
      storage: 100Gi
```

Features in Design/Prototyping

- Changed block tracking
  - Provides a common method to get differential snapshots of two arbitrary CSI VolumeSnapshots for PersistentVolumes backed by Block volumes for efficient backups.
- (SIG-Apps) Volume Expansion for StatefulSets
  - Allows changes to storage requests in StatefulSets.volumeClaimTemplates to propagate to associated PVCs.
- Storage Capacity Scoring
  - Provides a way to score nodes for dynamic provisioning of PVs based on storage capacity in the VolumeBinding plugin.
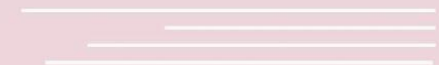- Combining CSI sidecars into a single component

How to get involved

How to get involved

- Start at the SIG Storage page:
  - https://github.com/kubernetes/community/tree/master/sig-storage
- Attend our meetings:
  - Bi-weekly meeting: 9am PT every second Thursday.
  - Issue triage: 10am PT every Wednesday.
- Mailing List:
  - kubernetes-sig-storage@googlegroups.com
- Slack channel:
  - #sig-storage
  - #csi
  - #sig-storage-cosi

Resources

- SIG Storage page
- Storage concepts
- CSI driver docs
- CSI spec
- CSI sample driver hostpath deployment example
- SIG Storage annual report

**Please scan the QR Code above
to leave feedback on this session**