
KubeCon
CloudNativeCon
North America 2022
BUILDING FOR THE ROAD AHEAD
DETROIT 2022
TiKV
TiKV
HELM

BUILDING FOR THE ROAD AHEAD

DETROIT 2022

Kubernetes SIG Storage Deep Dive

Xing Yang, VMware & Mauricio Poppe, Google

Agenda

- Who we are
- What we did in 1.25
- What we are working on in 1.26
- Deep Dive: CSI Windows
- How to get involved
- Q & A

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: 5100+
- #sig-storage-cosi slack channel members: 240+
- #csi slack channel members: 1300+
- #csi-windows slack channel members: 210+
- SIG-Storage zoom meeting attendees: 25
- Unique approvers for SIG-owned packages: 32

What we do

- Defined in SIG Storage Charter
  - Persistent Volume Claims and Persistent Volumes
  - Storage Classes and Dynamic Provisioning
  - Kubernetes volume plugins
  - Secret Volumes, ConfigMap Volumes, DownwardAPI Volumes, 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 in 1.25

What we did

What we did in 1.25

- GA
  - CSI ephemeral inline volumes
  - Local ephemeral storage capacity isolation
  - core CSI Migration
    - AWS EBS
    - GCE PD
- Beta
  - On-going effort: CSI migration
    - vSphere (Beta, on-by-default)
    - Portworx (Beta, off-by-default)

CSI Inline Ephemeral Volume

- GA in 1.25
- Ephemeral storage provided by “special” CSI driver
- Volume defined in the pod spec
- Example implementations
  - Secret Store CSI Driver
  - Cert-Manager CSI Driver
- Does not support features such as snapshotting, cloning, expansion, etc.
- Blog

```yaml
apiVersion: storage.k8s.io/v1
kind: CSIDriver
metadata:
  name: secrets-store.csi.k8s.io
spec:
  podInfoOnMount: true
  attachRequired: false
  volumeLifecycleModes:
  - name: Ephemeral
```

```yaml
apiVersion: v1
kind: Pod
metadata:
  name: some-pod
spec:
  containers:
    ...
  volumes:
      - name: vol
        csi:
          driver: csidriver.example.io
          volumeAttributes:
              foo: bar
```

Local Ephemeral Storage Capacity Isolation

- GA in 1.25
- Without this feature, pods can be evicted due to other pods filling the local storage.
- This feature allows users to manage local ephemeral storage.
- Each container of a Pod can specify either or both of the following. Pod evicted if exceeding limits
  - spec.containers[].resources.limits.ephemeral-storage
  - spec.containers[].resources.requests.ephemeral-storage
- Resource quota and limitRange for local ephemeral storage
- Blog

```yaml
apiVersion: v1
kind: Pod
metadata:
  name: frontend
spec:
  containers:
  - name: app
    image: images.my-company.example/app:v4
    resources:
      requests:
        ephemeral-storage: "8Gi"
      limits:
        ephemeral-storage: "12Gi"
    volumeMounts:
    - name: ephemeral
      mountPath: "/tmp"
  - name: log-aggregator
    image: images.my-company.example/log-aggregator:v6
    resources:
      requests:
        ephemeral-storage: "2Gi"
    volumeMounts:
    - name: ephemeral
      mountPath: "/tmp"
  volumes:
    - name: ephemeral
      emptyDir: {}
        sizeLimit: 5Gi
```

What we did in 1.25 (cont.)

- Alpha
  - SELinux relabeling with mount options
    - If possible, mount volumes with the correct SELinux context using -o context=XYZ mount option and avoid recursive change of all files on the volume.
  - NodeExpandSecret
    - Allows secrets to be passed in to CSI driver during volume expansion on the node
  - Reconcile Default Storage Class Assignment
    - Allows existing PVCs without “storageClassName” to be updated to use the new default StorageClass.
  - Object Storage API (COSI)

Object Storage API

- Container Object Storage Interface (COSI) provides a standard way for provisioning and consuming object storage in Kubernetes. Alpha in 1.25
- COSI Components
  - COSI ControllerManager: validates, authorizes and binds COSI created buckets to BucketClaims.
  - COSI Sidecar: watches COSI K8s API objects and calls COSI Driver.
  - COSI Driver: communicates with object storage providers to conduct bucket related operations.
- COSI K8s APIs
  - Bucket, BucketClaim, BucketClass
  - BucketAccess, BucketAccessClass
- COSI gRPC interfaces for object storage providers to provision buckets
- References: KEP, COSI repos, Blog

What we are working on in 1.26

- Targeting GA
  - Delegate FSGroup to CSI Driver instead of Kubelet
  - On-going effort: CSI Migration
    - Azure File
    - vSphere

Delegate FSGroup to CSI Driver instead of Kubelet

- Targeting GA in 1.26
- Allows CSI Drivers to indicate whether or not they support modifying a volume's ownership or permissions when the volume is being mounted.
- FSGroupPolicy in CSIDriver spec
  - ReadWriteOnceWithFSType: default; examined at mount time
  - File: Always apply modifications
  - None: Never apply modifications
- Blog

```
apiVersion: storage.k8s.io/v1
kind: CSIDriver
metadata:
  name: hostpath.csi.k8s.io
spec:
  # Supports persistent and ephemeral inline
volumes.
  volumeLifecycleModes:
  - Persistent
  - Ephemeral
  # To determine at runtime which mode a volume
uses,
  # pod info and its
  # "csi.storage.k8s.io/ephemeral" entry are
needed.
  podInfoOnMount: true
  fsGroupPolicy: None
```

What we are working on in 1.26 (cont.)

- Targeting Beta
  - Control Volume Mode Conversion between Source and Target
    - Prevent unauthorised volume mode conversion
  - Reconcile Default Storage Class Assignment
    - Allows existing PVCs without “storageClassName” updated to use the new default StorageClass
  - (SIG-App) Auto remove PVCs created by statefulset
    - Adds an option to allow PVCs created by StatefulSet to be removed automatically
  - On-going effort: CSI migration
    - RBD (Beta, off-by-default)
  - Non-graceful Node Shutdown

Non-graceful Node Shutdown

- Targeting Beta in 1.26; introduced as Alpha in 1.24
- Allows stateful workloads to failover to a different node after the original node is shutdown or in a non-recoverable state such as hardware failure or broken OS.
- Different from graceful node shutdown
- How does the feature work
  - Enable NodeOutOfServiceVolumeDetach feature gate for kube-controller-manager.
  - Apply taint to the shutdown node
    - node.kubernetes.io/out-of-service: "NoExecute"
    - node.kubernetes.io/out-of-service:" NoSchedule"
- Next steps
  - Move the feature to GA after it has reached Beta
  - Try to find an automatic approach
- KubeCon Session: How to Handle Node Shutdown in Kubernetes

What we are working on in 1.26 (cont.)

- Targeting Alpha
  - Provision volumes from Cross-namespace snapshots
    - Allows volumes to be provisioned from a snapshot in a different namespace
  - On-going effort: CSI Migration
    - CephFS
  - VolumeGroup and VolumeGroupSnapshot
    - Introduce a VolumeGroup API to manage multiple volumes together and a VolumeGroupSnapshot API to take a snapshot of a VolumeGroup.

VolumeGroup and VolumeGroupSnapshot

- Targeting Alpha in 1.26
- Introduce a VolumeGroup API to manage multiple volumes together and a VolumeGroupSnapshot API to take a snapshot of a VolumeGroup.
- Kubernetes APIs
  - VolumeGroup, VolumeGroupContent, VolumeGroupClass
  - VolumeGroupSnapshot, VolumeGroupSnapshotContent, VolumeGroupSnapshotClass
- New gRPC interfaces in CSI spec
  - create/delete/modify/list/get volume group
  - create/delete group snapshot

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 (Target)
Azure Disk	1.15	1.19	1.23	1.24	1.26 (Target)
Azure File	1.15	1.21	1.24	1.26 (Target)	1.28 (Target)
AWS EBS	1.14	1.17	1.23	1.25	1.27 (Target)
GCE PD	1.14	1.17	1.23	1.25	1.27 (Target)
vSphere *	1.18	1.19	1.25 (Target)	1.26 (Target)	1.28 (Target)
Ceph RBD	1.23	1.26 (Target)			
CephFS	1.26 (Target)				
Portworx	1.23	1.25	1.27 (Target)		

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

In-Tree Storage Driver Removal

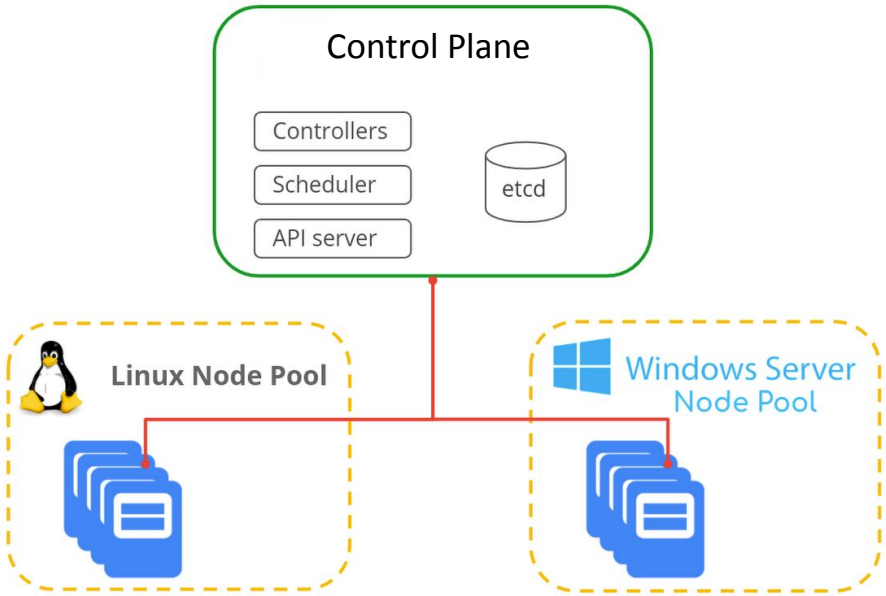
Driver	Deprecated	Code Removal
Flocker	1.22	1.25
GlusterFS	1.25	1.26 (Target)
Quobyte	1.22	1.25
ScaleIO	1.16	1.22
StorageOS	1.22	1.25

KubeCon CloudNativeCon North America 2022
BUILDING FOR THE ROAD AHEAD
DETROIT 2022

CSI Windows Deep Dive

Kubernetes, Windows and CSI

- Windows workloads run in a Windows nodepool.
  - kubernetes 1.25 supports these Windows Server OS: LTSC2019, LTSC2022, SAC.
- Control plane components run in Linux.
- In CSI, the node component of CSI Driver, livenessprobe and node-driver-registrar run in Windows.
- DaemonSet images built through a multiarch build pipeline.



Cluster with Linux and Windows nodepools
Source: Windows Server applications

CSI Node Operations in Windows

Set up

NodeStageVolume - Create a Windows volume, format it to NTFS, create a partition access path in the node (global mount).

NodePublishVolume - Create a symlink from the kubelet Pod-PVC path to the global path (pod mount).

Tear down

NodeUnpublishVolume - Remove the symlink created above.

NodeUnstageVolume - Remove the partition access path.

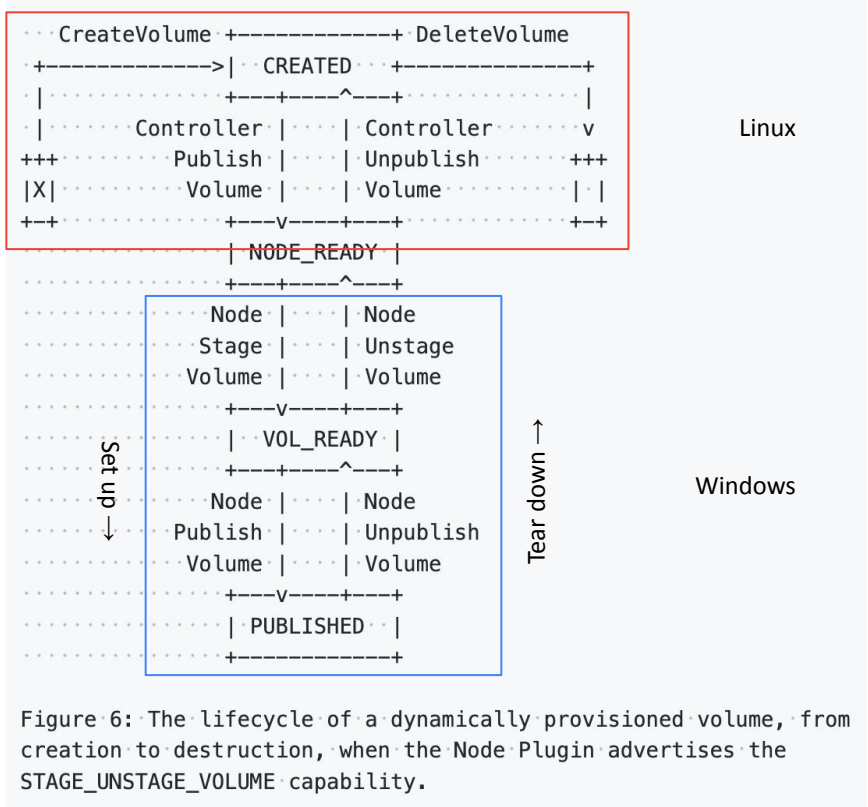
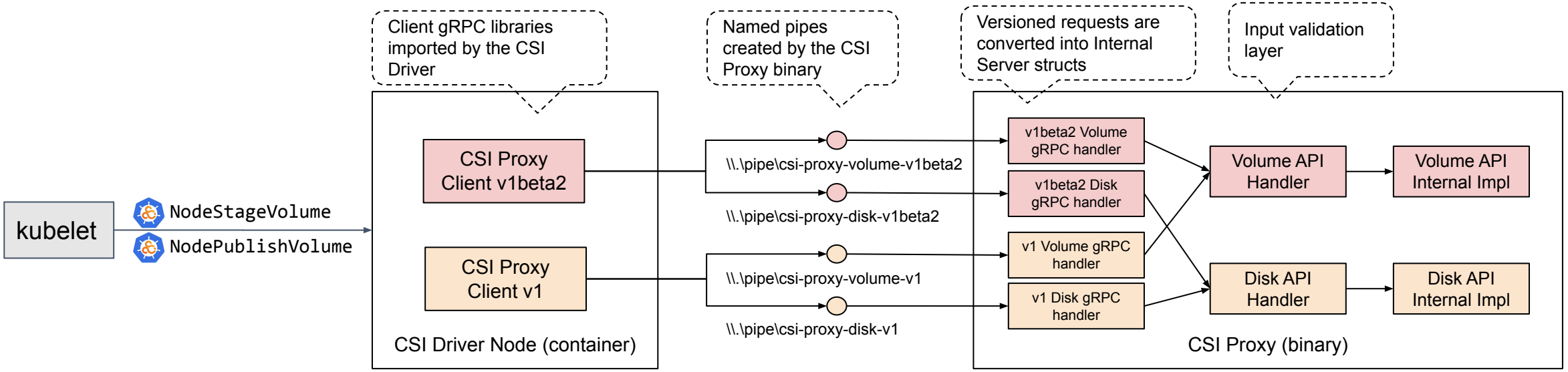


Figure 6: The lifecycle of a dynamically provisioned volume, from creation to destruction, when the Node Plugin advertises the STAGE_UNSTAGE_VOLUME capability.

CSI Operations in a cluster with Windows nodepools

Storage Operations in Windows

- The volume format/mount are privileged operations and can't be done through a Windows container*.
- Windows nodes run a binary as a service on the host (CSI Proxy) that performs privileged storage operations on behalf of a CSI Driver. GA in 1.22.
- Drawbacks: additional host component, difficult maintenance (multiple API versions, different lifecycle to the CSI Driver), tiny increase in latency.

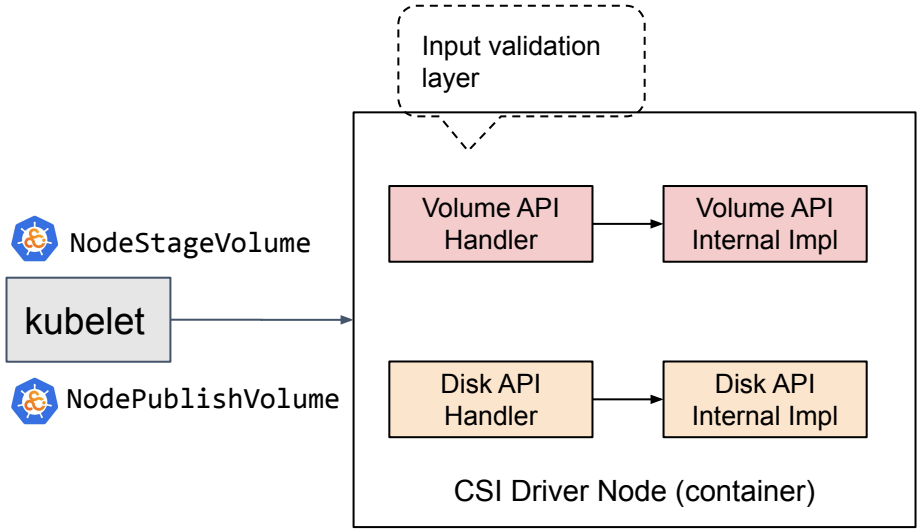


CSI Proxy architecture in a Windows node

* doable through a HostProcess Container

CSI Proxy as a Go library

- Windows HostProcess Pod is a beta feature in 1.23 and stable in 1.26.
  - Allows running containers as a process in the host.
  - No filesystem isolation.
- If a CSI Driver becomes a HostProcess Pod it can run the privileged storage operations.
- CSI Drivers still need the format/mount utility functions in CSI Proxy.
- Plan is to turn CSI Proxy to a Go library used by CSI Drivers.
- CSI Drivers become easy to develop/deploy in Windows! (very similar to Linux)
- kubernetes/enhancements #3636



CSI Proxy as a Go library in a CSI Driver

CSI Drivers as HostProcess Pods

- Updates in the CSI Driver node format/mount.

```
// Initialize the client (versioned).
client, err = v1client.NewClient()

// Make a call to CSI Proxy.
_, err = client.FormatVolume(
     ctx,
     &v1.FormatVolumeRequest{
          VolumeId: volumeID,
     },
)
```

```
// Initialize the client (unversioned).
// volume, volumeapi are imports from the library.
client, err = volume.New(volumeapi.New())

// Format the volume directly.
_, err = client.FormatVolume(
     ctx,
     &volume.FormatVolumeRequest{
          VolumeID: volumeID,
     }
)
```

- Updates in the CSI Driver deployment: add HostProcess Pod spec fields, remove hostpath mounts to the CSI Proxy named pipes, remove CSI Proxy binary.
- Thanks to Alexander Ding from Brown University for his work in CSI Proxy.
- Great time to add Windows support to your CSI Driver. Thank you SIG Windows!

How to get involved

How to Get Involved

- Start at the SIG Storage page:
  - https://github.com/kubernetes/community/tree/master/sig-storage
- Attend the bi-weekly meetings: 9 AM PT every second Thursday.
- 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

KubeCon | CloudNativeCon
North America 2022

BUILDING FOR THE ROAD AHEAD

DETROIT 2022

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

BUILDING FOR THE ROAD AHEAD

DETROIT 2022

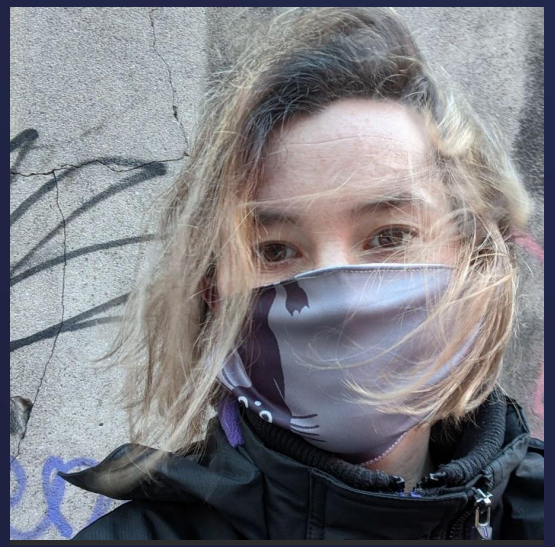
Thank You

Secure your cluster to cluster traffic, the agnostic way

KubeCon | CloudNativeCon
North America 2022
BUILDING FOR THE ROAD AHEAD
DETROIT 2022

October 24-28, 2021

Dave Kerr
Software Engineer
Workday



Pauline Lallinec
Software Engineer
Workday

Title

Content

Title

KubeCon | CloudNativeCon
North America 2022
BUILDING FOR THE ROAD AHEAD
DETROIT 2022

KubeCon | CloudNativeCon
North America 2022
BUILDING FOR THE ROAD AHEAD
DETROIT 2022

October 24-28, 2021

Speaker Name
Job Title, *Company*

Title

Speaker Name
Job Title
Company

Speaker Name
Job Title
Company

Speaker Name
Job Title
Company

KubeCon | CloudNativeCon

North America 2022

BUILDING FOR THE ROAD AHEAD

DETROIT 2022

October 24-28, 2021

Speaker Name
Job Title
Company

Speaker Name
Job Title
Company

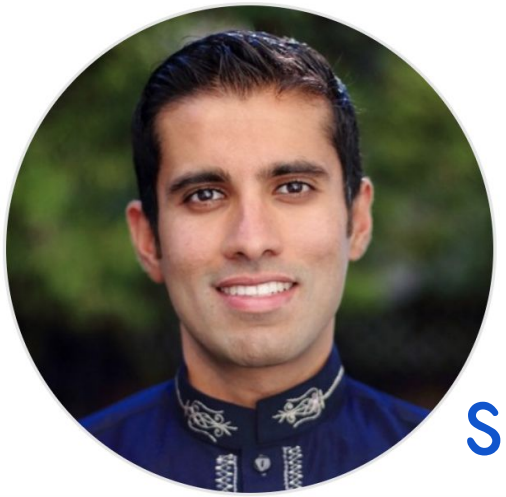
CSI Migration Testing

- What do I do?
  - Using a managed Kubernetes distribution? Check your distro's documentation. In many cases, distro will take care of everything.
  - Managing your own Kubernetes?
    - Install the replacement CSI driver for your cloud.
    - Enable CSIMigration and CSIMigrationX feature gates (X is the specific driver): detailed ordering + drain nodes.
- Caveats
  - CSI-only features do not work (e.g., snapshots, cloning, etc).
    - Manually re-import PV as CSI type.
  - Some in-tree functionality has been deprecated and won't work with migration (check Kubernetes release notes).

Deep Dive on CSI Migration

- Why? Built-in cloud providers are deprecated and target for removal in 1.26.
- What? CSI Migration allows your existing PVs and StorageClasses using these in-tree volume plugins to continue working even when built-in cloud providers are removed
  - kubernetes.io/aws-ebs
  - kubernetes.io/azure-disk
  - kubernetes.io/azure-file
  - kubernetes.io/cinder
  - kubernetes.io/gce-pd
  - kubernetes.io/vsphere-volume

SIG-Storage Leads



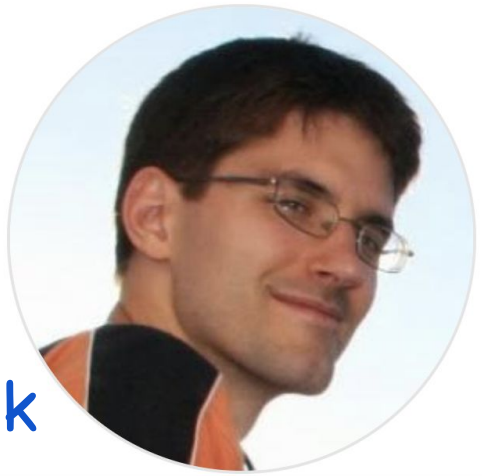
SIG-Storage
Co-Chair

Saad Ali

Xing Yang

SIG-Storage
Tech Lead

Michelle Au



Jan Šafránek