KubeCon | CloudNativeCon

North America 2019

Intro + Deep Dive: Kubernetes Storage SIG

November 21, 2019

Agenda

- **Kubernetes SIG-Storage Intro** by Saad Ali
- **Kubernetes-CSI Update** by Michelle Au
- **Volume Snapshots Update** by Xing Yang and Xiangqian Yu
- **General Q&A** by SIG Storage Panel

KubeCon | CloudNativeCon

North America 2019

Kubernetes SIG-Storage Intro

Saad Ali

Who is SIG Storage?

Group of Kubernetes Contributors responsible for:

- Ensuring file and block storage (whether ephemeral or persistent, local or remote) are available wherever a container is scheduled.
- Provisioning, attaching, mounting, unmounting, detaching, and deleting volumes
- Influencing scheduling of containers based on storage (data gravity, availability, etc.).
- Storage capacity management (container ephemeral storage usage, volume resizing, etc.)

Who is SIG Storage?

- Some notable examples of features owned by SIG Storage:
  - Persistent Volume Claims and Persistent Volumes
  - Storage Classes and Dynamic Provisioning
- Kubernetes volume plugins
  - Container Storage Interface (CSI)
  - Secret, ConfigMap, DownwardAPI Volumes
  - And lots more!
- Team page:
  - https://github.com/kubernetes/community/tree/master/sig-storage

Many Contributors!

- Amazon
- Dell EMC
- Diamanti
- Google
- Hitachi Data Systems
- IBM
- Kasten
- Linbit
- Mayadata
- Microsoft
- NetApp
- Nutanix
- OpenSDS
- Quantum (Rook)
- Red Hat
- Salesforce
- OpenStack
- Oracle
- IBM
- Portworx
- PURE Storage
- Robin
- StorageOS
- VMware
- Unaffiliated/Independent
- And more!

What does SIG Storage do?

- Code features, write tests, fix bugs for volume related features.
- Meet virtually every two weeks to plan and discuss.
- Meet face-to-face every now and then to close on bigger issues.
- Help each other and the community via slack and google groups.

What have we been working on?

Kubernetes 1.16

- Beta: CSI Volume cloning
- Beta: CSI Volume expansion
- Beta: CSI Ephemeral volumes

What are we working on?

Kubernetes 1.17

- GA: CSI Topology
- GA: Volume attach limits (in-tree + CSI)
- Beta: CSI Volume Snapshots
- Beta: CSI Migration with AWS EBS and GCE PD drivers

How to get involved w/SIG Storage?

- 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.
  - Zoom meeting! Attend from anywhere.
  - Agenda doc -- feel free to add items for discussion to this doc.
  - Next one December 5
- Familiarize yourself with the code. Start from main method walk through it.
  - Help fix a bug!
  - 272 open SIG storage Issues (as of 11/13/19)
  - Filter by “Help wanted” label.
- Help write tests!

How to get involved w/SIG Storage?

- Help write features!
  - There is a new Kubernetes version released every quarter (e.g. v1.9, v1.10, v1.11…)
- Release schedules:
  - github.com/kubernetes/sig-release/tree/master/releases/
- SIG Storage Planning Spreadsheet
  - Beginning of every quarter: planning and assignments
  - During quarter: help needed on assigned items & sometimes new items pop up.
- Every feature must have:
  - Enhancement issue in github.com/kubernetes/enhancements/
  - KEP in github.com/kubernetes/enhancements/tree/master/keps/sig-storage
- Need more contributors!! (Especially for SIG-owned CSI drivers).

KubeCon San Diego Presentations

- Tuesday
  - *Beyond Storage Management*
    - by Andrew Large & Yinan Li
  - *Building Blocks: How Raw Block PVs Changed the Way We Look at Storage*
    - by Jose A. Rivera & Rohan Gupta
  - *How to Backup and Restore Your Kubernetes Cluster*
    - by Annette Clewett & Dylan Murray
- Wednesday
  - *Storage on Kubernetes - Learning From Failures*
    - by Hemant Kumar & Jan Šafránek, Red Hat
  - *Kubernetes Storage Cheat Sheet for VM Administrators*
    - by Manu Batra & Jing Xu
  - *CSI Volume Snapshots: On the Way to Faster and Better Backups*
    - by Adnan Abdulhussein & Nolan Brubaker

Recordings are online!

Kubernetes-CSI Update

Michelle Au

CSI Driver Development

Driver development documentation

https://kubernetes-csi.github.io/docs/

Sample driver

https://github.com/kubernetes-csi/csi-driver-host-path

CSI Driver Testing

CSI Sanity

- Conformance to CSI spec
- https://github.com/kubernetes-csi/csi-test/blob/master/pkg/sanity/README.md

Kubernetes Storage E2E Test Suite

- Basic functional tests in Kubernetes
- Run against any installed CSI driver in any Kubernetes cluster
- Download e2e test binary, define driver config and storageclass, run
- Future: Conformance/validation testing, scale and stress testing
- https://kubernetes-csi.github.io/docs/functional-testing.html

```
ginkgo -p -focus='External.Storage' -skip='\[Feature:|\[Disruptive\]' e2e.test -- \
  -storage.testdriver=/tmp/hostpath-testdriver.yaml
```

CSI Migration Deep Dive

What?

- Service in-tree volume APIs with CSI backend
- Part of broader cloud provider extraction project

Why?

- All cloud provider code is being removed from core Kubernetes
- Lower security risk from vendoring unnecessary providers
- Accelerate features and bug fixes
    - CSI driver development is decoupled from Kubernetes release cycle

CSI Migration Deep Dive

Feature Status

Driver	Alpha	Beta (in-tree deprecated)	GA	Target in-tree removal
AWS EBS	1.14	1.17	1.19 (target)	1.21
GCE PD	1.14	1.17	1.19 (target)	1.21
Openstack Cinder	1.14	1.18 (target)	1.19 (target)	1.21
Azure Disk + File	1.15	1.18 (target)	1.19 (target)	1.21
Vsphere	1.18 (target)	1.19 (target)	1.20 (target)	1.21

ALL CLOUD PROVIDER CODE WILL BE REMOVED IN 1.21

CSI Migration Deep Dive

How do I try this out?

- Using managed service? No action required
- Self-deployed? Deployer must also deploy equivalent CSI driver, turn on CSIMigration (default on in 1.17) and CSIMigration<provider> feature gates.
  - Ideally deployed as part of external cloud provider controllers (kubernetes/cloud-provider-<provider>)

Get Involved!

- Slack: #csi-migration

CSI Windows Deep Dive

Problem

- Windows containers can’t be privileged
- CSI drivers need to perform privileged operations like mount

Solution

- CSI Proxy binary runs directly on the host, performs all privileged operations
- CSI drivers communicate to proxy via gRPC API
  - APIs for common protocols: block, SMB, iSCSI
- Alpha under development

Get Involved!

- Slack: #csi-windows

CSI Ephemeral Volumes Deep Dive

What?

- Volume lifecycle follows pod
- Volume specified directly in Pod spec
- Beta in 1.16

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

Examples

- image-populator: https://github.com/kubernetes-csi/csi-driver-image-populator
- cert-manager: https://github.com/jetstack/cert-manager-csi
- secrets-store: https://github.com/deislabs/secrets-store-csi-driver

Roadmap

Feature Graduation ~ first half 2020

- GA: Skip attach
- GA: Pod info on mount
- GA: Raw block
- GA: Cloning
- GA: Resizing
- GA: Snapshots
- Alpha: Windows

Feature Graduation ~ second half 2020

- GA: Ephemeral volumes
- GA: CSI Migration for all in-tree cloud plugins
- Beta: Windows

Roadmap

In Design/Prototyping

- Volume health
- Operational metrics
  - User-centric: how long does it take to attach/mount a volume?
  - Plugin-centric: how long did plugin take to attach/mount?
  - What’s the Kubernetes overhead?
  - Error ratios by error code
- Storage pool capacity reporting
  - To support local PV dynamic provisioning and ephemeral volumes
- Application snapshots and backups
- Group snapshots and consistency groups

We Need Your Help!

Community-maintained CSI drivers

- nfs
- iscsi
- fc
- flex-adapter

Testing and release infrastructure

- Staging and publishing images following Kubernetes processes
- Improving release notes generation
- Improving modularity of test scripts
- Adding new K8s releases to test jobs
- Adding more test cases to csi-test
- Scalability testing
- K8s conformance testing

How To Get Involved?

Slack: #csi

Issues

- search for help-wanted label in https://github.com/kubernetes-csi

KubeCon | CloudNativeCon

North America 2019

Volume Snapshots Update

Jing Xu, Xing Yang, and Xiangqian Yu

What’s New in 1.17

- Snapshot API is alpha since 1.12. It goes to beta in 1.17
- API revamp
- Controller splitting

Dynamic v.s. Pre-Provisioned

- Dynamic creation of volume snapshots
  - User creates namespaced VolumeSnapshot (with PVC as source) to trigger creation of a new snapshot which will be represented by a newly created VolumeSnapshotContent.
- Manually bind to pre-provisioned volume snapshots
  - Admin manually creates VolumeSnapshotContent to represent a pre-existing snapshot.
  - User creates VolumeSnapshot to point to the desired VolumeSnapshotContent.
  - Controller binds them if VolumeSnapshotContent also points back to the VolumeSnapshot.

API Design Principles

- Spec
  - Represents desired state: configuration settings provided by the user, properties initialized or otherwise changed after creation by other ecosystem components.
- Status
  - Represents actual state: information updated by controller.
  - Recoverable from spec by controller.
  - User cannot specify status during object creation.

VolumeSnapshot

```
type VolumeSnapshotSpec struct {
        Source VolumeSnapshotSource
        ~~Source core_v1.ObjectReference~~
        VolumeSnapshotClassName *string
        ~~SnapshotContentName string~~
}
// Exactly one of its members MUST be specified
type VolumeSnapshotSource struct {
        // +optional
        PersistentVolumeClaimName *string
        // +optional
        VolumeSnapshotContentName *string
}
```

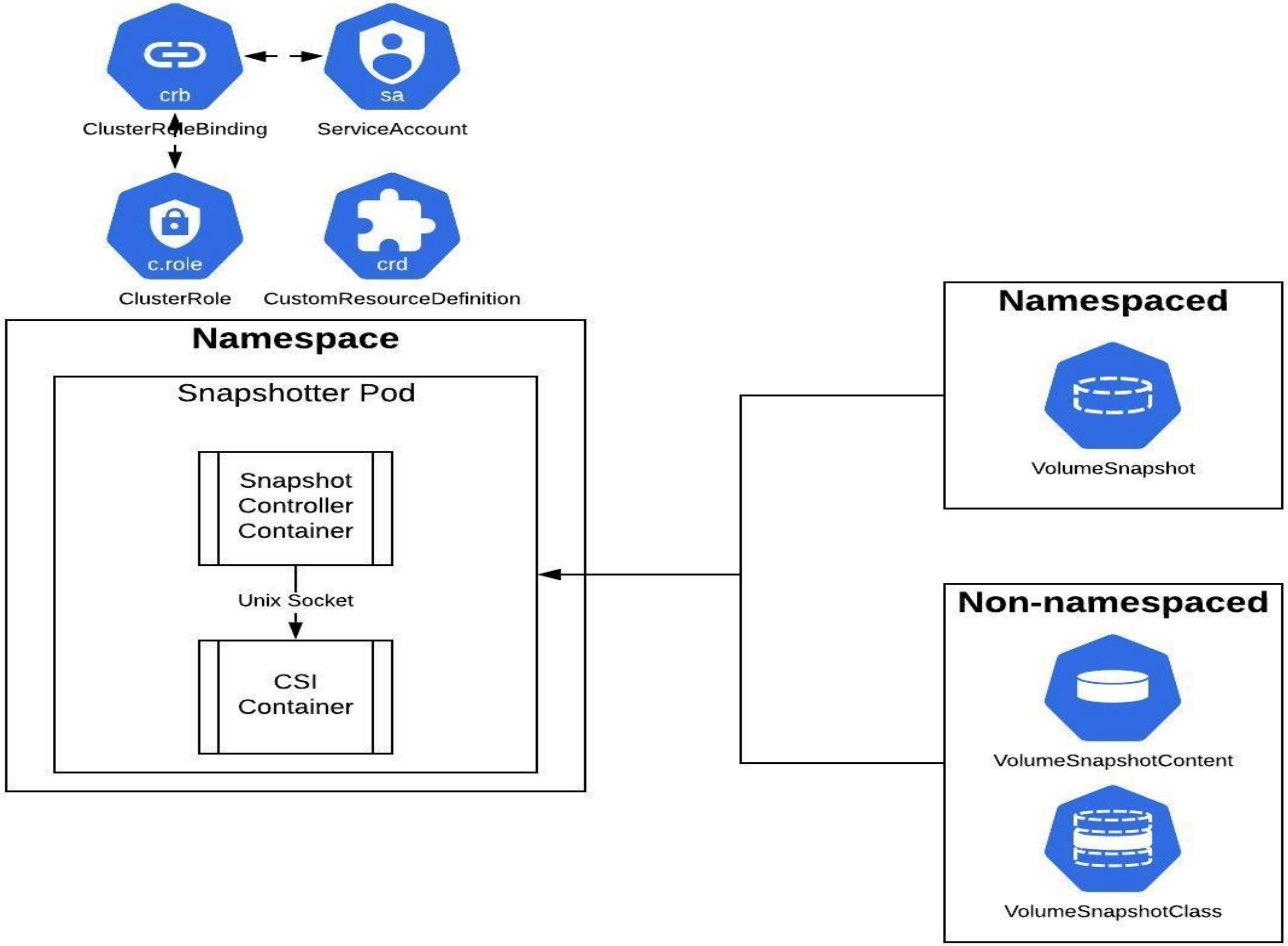
```
type VolumeSnapshotStatus struct {
        BoundVolumeSnapshotContentName *string
        CreationTime *metav1.Time
        ReadyToUse *bool
        RestoreSize *resource.Quantity
        Error *VolumeSnapshotError
}
```

VolumeSnapshotContent

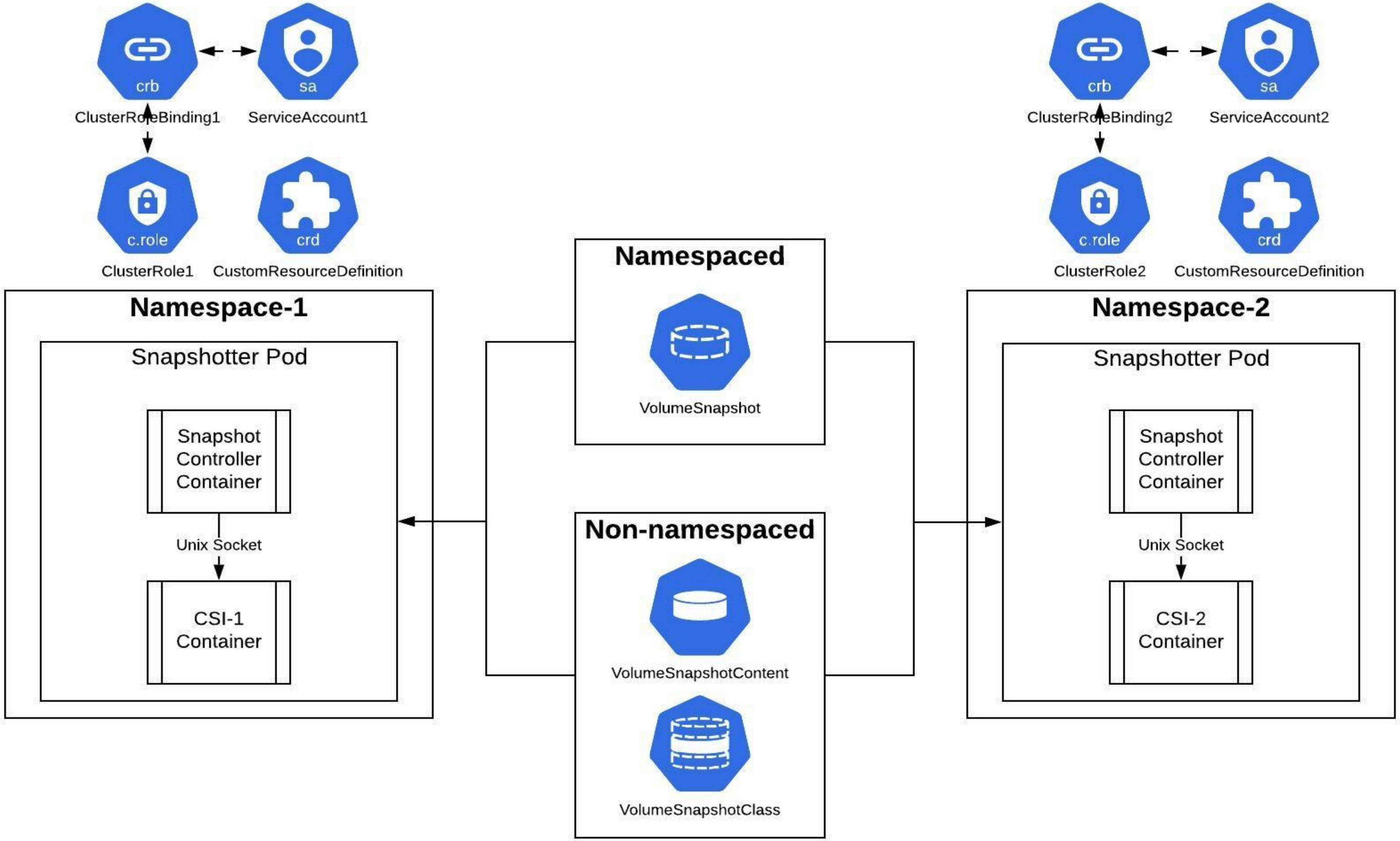
```
type VolumeSnapshotContentSpec struct {
        VolumeSnapshotRef core_v1.ObjectReference
        ~~PersistentVolumeRef core_v1.ObjectReference~~
        Source VolumeSnapshotContentSource
        DeletionPolicy DeletionPolicy
        Driver string
        SnapshotClassName *string
}

type VolumeSnapshotContentSource struct {
        // +optional
        VolumeHandle *string
        // +optional
        SnapshotHandle *string
}
```

```
type VolumeSnapshotContentStatus struct
{
        CreationTime *int64
        ReadyToUse *bool
        RestoreSize *int64
        Error *VolumeSnapshotError
        SnapshotHandle *string
}
```

Alpha Controller Architecture



Alpha Controller Architecture



Challenges

1. Deployment of multiple CSI drivers (CRD, RBAC etc).

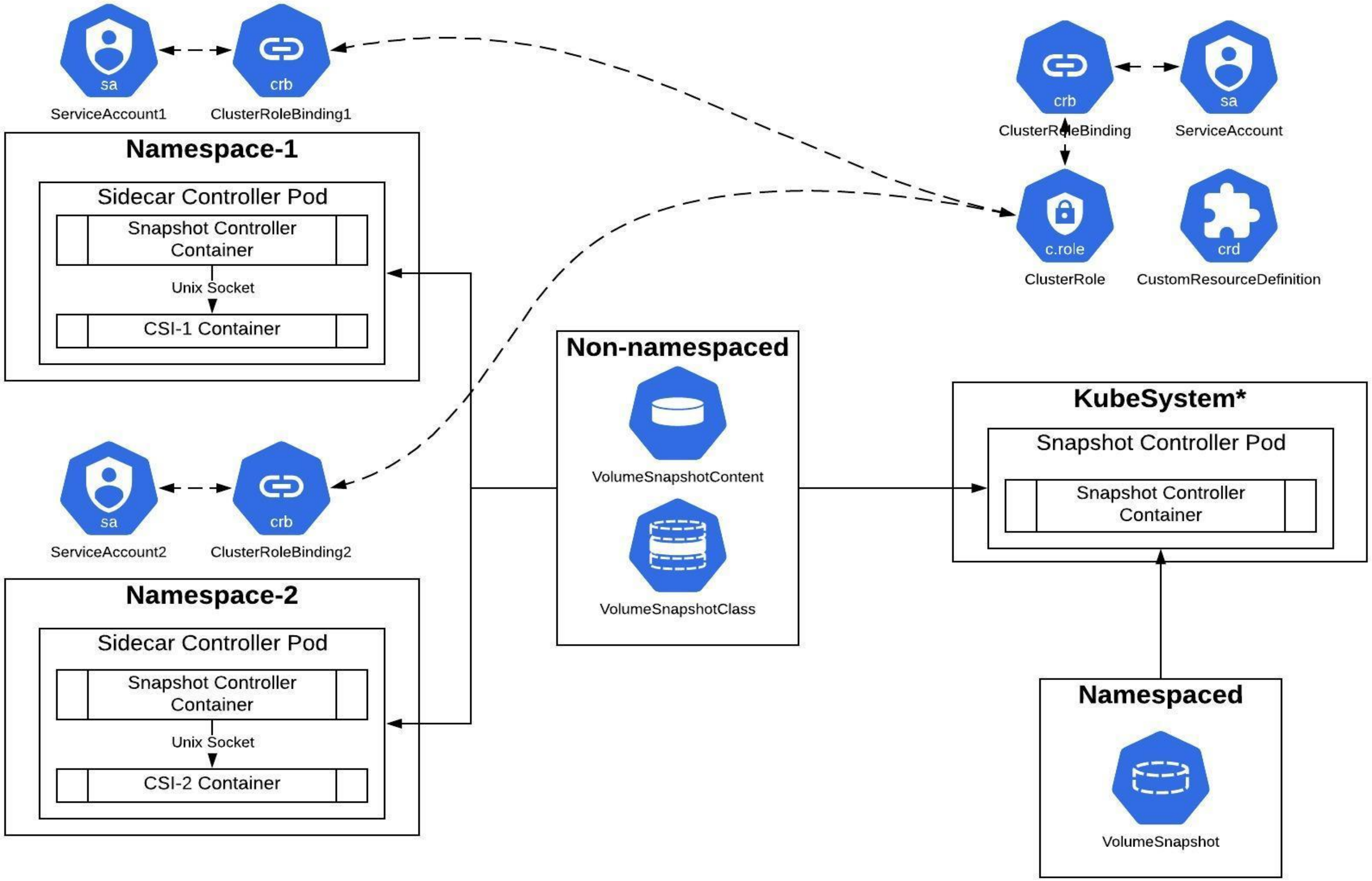
2. Observability signals collection.

3. Any controller release requires storage vendors’ involvement.

Split Controllers

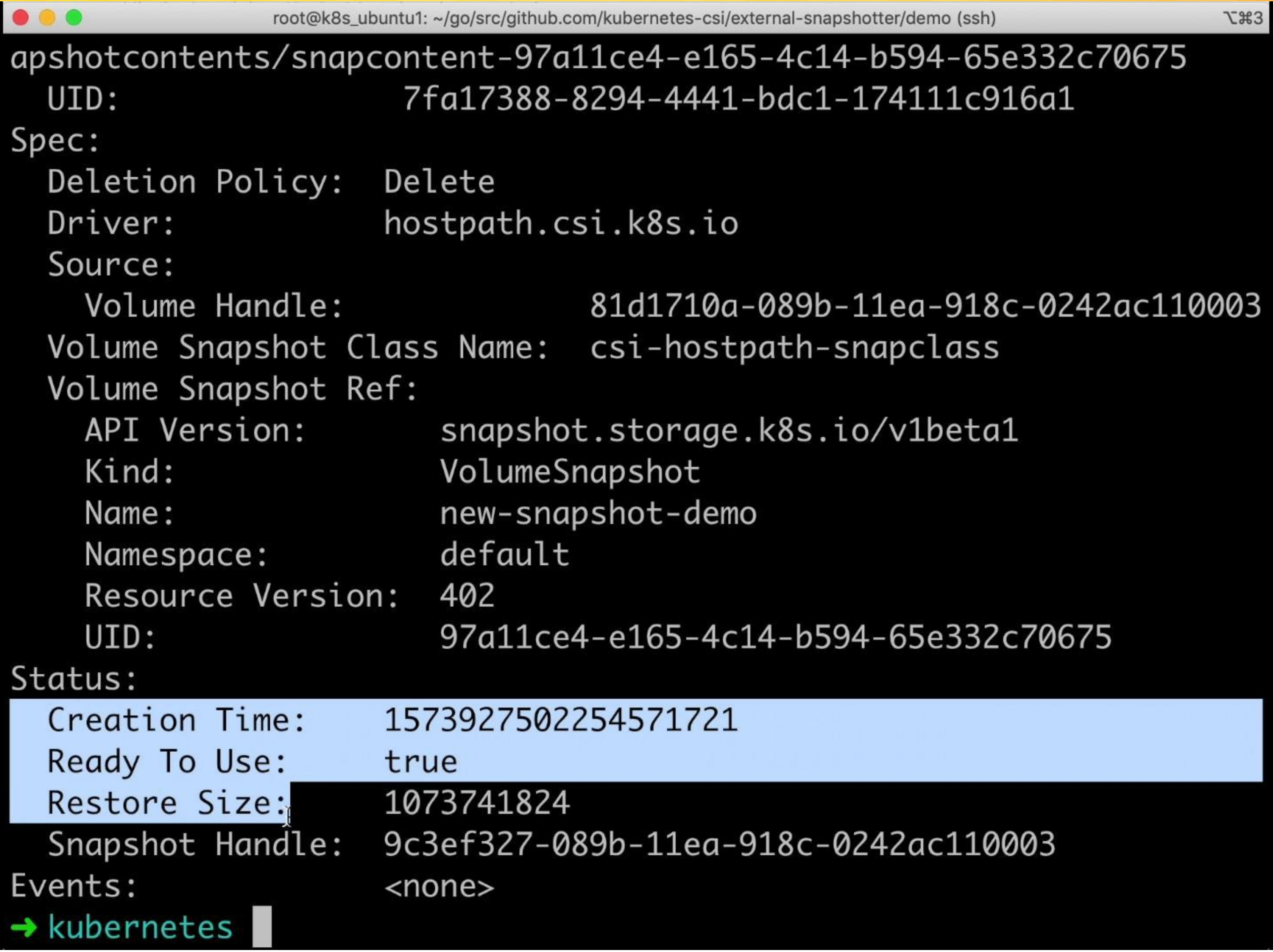
- Snapshot Controller (deployed by cluster deployer)
    - Deployed along with CRD
    - Works on both VolumeSnapshot and VolumeSnapshotContent
    - Not aware of CSI
- Sidecar Controller (deployed with CSI driver)
    - Conduct CSI calls
    - Works only on VolumeSnapshotContent
    - Keep it simple!

Beta Controller Architecture



Deployment

- Cluster deployer
  - Install Snapshot Beta CRDs
    - kubectl create -f config/crd
    - https://github.com/kubernetes-csi/external-snapshotter/tree/master/config/crd
  - Install Snapshot Controller
    - kubectl create -f deploy/kubernetes/snapshot-controller
    - https://github.com/kubernetes-csi/external-snapshotter/tree/master/deploy/kubernetes/snapshot-controller
- CSI Vendor
  - Install sidecar csi-snapshotter and CSI driver
    - kubectl create -f deploy/kubernetes/csi-snapshotter
    - https://github.com/kubernetes-csi/external-snapshotter/tree/master/deploy/kubernetes/csi-snapshotter



root@k8s_ubuntu1: ~/go/src/github.com/kubernetes-csi/external-snapshotter/demo (ssh)
⌥⌘3

```
apshotcontents/snapcontent-97a11ce4-e165-4c14-b594-65e332c70675
  UID:                 7fa17388-8294-4441-bdc1-174111c916a1
Spec:
  Deletion Policy:  Delete
  Driver:           hostpath.csi.k8s.io
  Source:
    Volume Handle:             81d1710a-089b-11ea-918c-0242ac110003
  Volume Snapshot Class Name:  csi-hostpath-snapclass
  Volume Snapshot Ref:
    API Version:       snapshot.storage.k8s.io/v1beta1
    Kind:              VolumeSnapshot
    Name:              new-snapshot-demo
    Namespace:         default
    Resource Version:  402
    UID:               97a11ce4-e165-4c14-b594-65e332c70675
Status:
  Creation Time:    1573927502254571721
  Ready To Use:     true
  Restore Size:     1073741824
  Snapshot Handle:  9c3ef327-089b-11ea-918c-0242ac110003
Events:             <none>
➜ kubernetes
```

Dynamic Provisioning

```yaml
apiVersion: snapshot.storage.k8s.io/v1beta1
kind: VolumeSnapshotClass
metadata:
  name: test-snapclass
driver: testdriver.csi.k8s.io
deletionPolicy: Delete
```

```yaml
apiVersion: snapshot.storage.k8s.io/v1beta1
kind: VolumeSnapshot
metadata:
  name: test-snapshot
spec:
  volumeSnapshotClassName: test-snapclass
  source:
    persistentVolumeClaimName: test-pvc
```

VolumeSnapshot API Object

kubectl describe volumesnapshot

```
Name:         test-snapshot
Namespace:    default
Labels:       <none>
Annotations:  <none>
API Version:  snapshot.storage.k8s.io/v1beta1
Kind:         VolumeSnapshot
Metadata:
  Creation Timestamp:  2019-11-16T00:36:04Z
  Finalizers:
    snapshot.storage.kubernetes.io/volumesnapshot-as-source-protection
    snapshot.storage.kubernetes.io/volumesnapshot-bound-protection
  Generation:        1
  Resource Version:  1294
  Self Link:
/apis/snapshot.storage.k8s.io/v1beta1/namespaces/default/volumesnapshots/new-snapshot-demo
  UID:               32ceaa2a-3802-4edd-a808-58c4f1bd7869
Spec:
  Source:
    Persistent Volume Claim Name:  test-pvc
  Volume Snapshot Class Name:      test-snapclass
Status:
  Bound Volume Snapshot Content Name:
snapcontent-32ceaa2a-3802-4edd-a808-58c4f1bd7869
  Creation Time:                   2019-11-16T00:36:04Z
  Ready To Use:                    true
  Restore Size:                    1Gi
```

VolumeSnapshotContent API Object

kubectl describe volumesnapshotcontent

```
Name:         snapcontent-32ceaa2a-3802-4edd-a808-58c4f1bd7869
Namespace:
Labels:       <none>
Annotations:  <none>
API Version:  snapshot.storage.k8s.io/v1beta1
Kind:         VolumeSnapshotContent
Metadata:
  Creation Timestamp:  2019-11-16T00:36:04Z
  Finalizers:
    snapshot.storage.kubernetes.io/volumesnapshotcontent-bound-protection
  Generation:        1
  Resource Version:  1292
  Self Link:         /apis/snapshot.storage.k8s.io/v1beta1/volumesnapshotcontents/snapcontent-32ceaa2a-3802-4edd-a808-58c4f1bd7869
  UID:               7dfdf22e-0b0c-4b71-9ddf-2f1612ca2aed
Spec:
  Deletion Policy:  Delete
  Driver:           testdriver.csi.k8s.io
  Source:
    Volume Handle:            d1b34a5f-0808-11ea-808a-0242ac110003
  Volume Snapshot Class Name:  test-snapclass
  Volume Snapshot Ref:
    API Version:       snapshot.storage.k8s.io/v1beta1
    Kind:              VolumeSnapshot
    Name:              test-snapshot
    Namespace:         default
    Resource Version:  1286
    UID:               32ceaa2a-3802-4edd-a808-58c4f1bd7869
Status:
  Creation Time:    1573864564608810101
  Ready To Use:     true
  Restore Size:     1073741824
  Snapshot Handle:  127c5798-0809-11ea-808a-0242ac110003
Events:             <none>
```

Pre-Provisioned Snapshots

```
apiVersion: snapshot.storage.k8s.io/v1beta1
kind: VolumeSnapshotContent
metadata:
  name: test-content
spec:
  deletionPolicy: Delete
  driver: testdriver.csi.k8s.io
  source:
    snapshotHandle: 7bdd0de3-aaeb-11e8-9aae-0242ac110002
  volumeSnapshotRef:
    name: test-snapshot
    namespace: default
```

```
apiVersion: snapshot.storage.k8s.io/v1beta1
kind: VolumeSnapshot
metadata:
  name: test-snapshot
spec:
  source:
        volumeSnapshotContentName: test-content
```

Create Volume from Snapshot

```
apiVersion: v1
kind: PersistentVolumeClaim
metadata:
  name: pvc-restore
spec:
  storageClassName: test-sc
  dataSource:
    name: test-snapshot
    kind: VolumeSnapshot
    apiGroup: snapshot.storage.k8s.io
  accessModes:
    - ReadWriteOnce
  resources:
    requests:
      storage: 1Gi
```

Future Plan

- Web hook for validation
- Metrics for snapshot controller
- More e2e tests
- Volume group snapshots
- Volume backups
  - Change block tracking
- ……

General Q&A

SIG Storage Panel

Thank you!

- Bi-weekly meetings
  - 9 AM Thursdays every two weeks
  - See https://github.com/kubernetes/community/tree/master/sig-storage for invite
- Slack channel
  - #sig-storage on kubernetes.slack.com
- Mailing list
  - https://groups.google.com/forum/#!forum/kubernetes-sig-storage

KubeCon | CloudNativeCon

North America 2019