Kubernetes@eBay Storage Management

Chengyuan Li <chengyli@ebay.com>

tess.IO

Agenda

- Storage Overview in Kubernetes
  - Basic Concepts
    - Volumes Plugins
    - PV/PVC
    - Dynamic Volumes Provisioning
  - Kubernetes Components for Volumes
- Storage Management in Kubernetes@eBay
  - Storage Class Definition
  - Network Volume
  - Local Volume Management
    - Motivation
    - Solution in eBay
  - Health Monitoring
  - Future Work

Volumes Plugins

Network Volumes

cinder
nfs
iscsi
rbd
cephfs
gcePersistentDisk
awsElasticBlockStore
flocker
glusterfs
azureFileVolume
azureDisk
vsphereVolume
Quobyte

Local Volumes

emptyDir
hostPath

Expose Kubernetes API

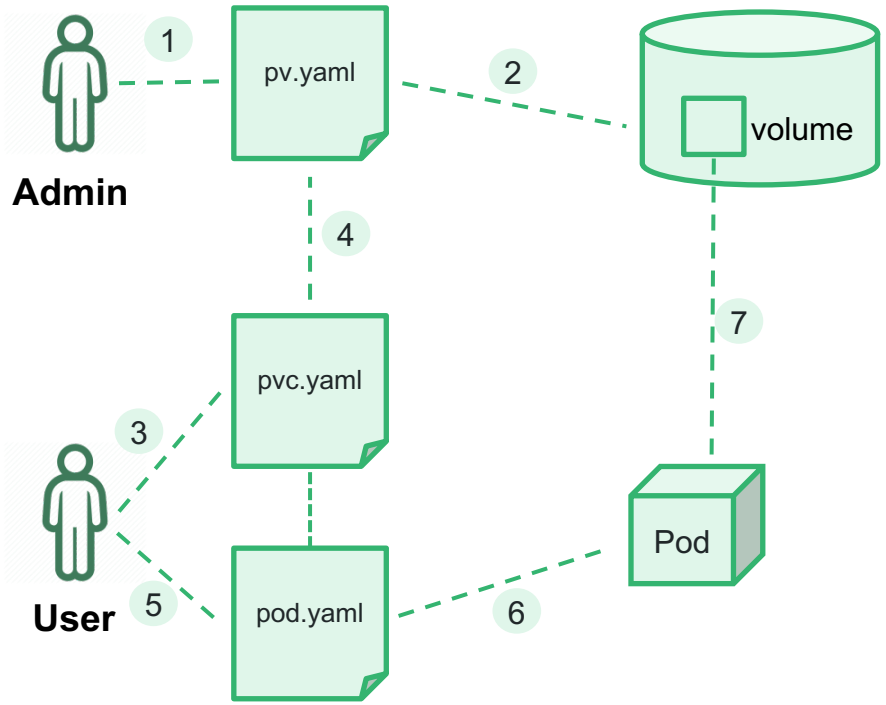
secret
downwardAPI

```
apiVersion: v1
kind: Pod
metadata:
  name: test-pd
spec:
  containers:
  - image: gcr.io/google_containers/test-webserver
    name: test-container
    volumeMounts:
    - mountPath: /cache
      name: cache-volume
  volumes:
  - name: cache-volume
    emptyDir: {}
```

PV/PVC

- PersistentVolume (PV)
  - **A piece of networked storage** in the cluster that has been **provisioned by an administrator**.
  - Have a **lifecycle independent** of any individual pod that uses the PV.

- PersistentVolumeClaim (PVC)
  - A request for storage by a **user**
  - Claims can request **specific size** and **access modes**

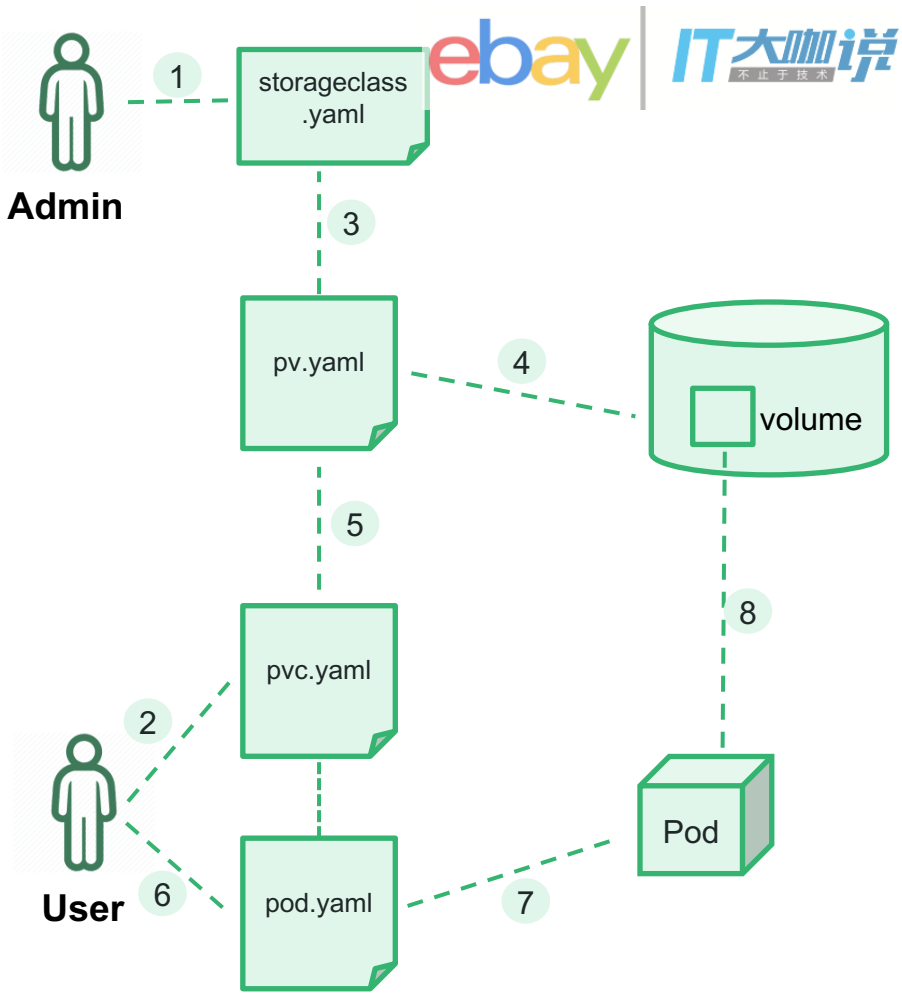


Dynamic Volumes Provisioning

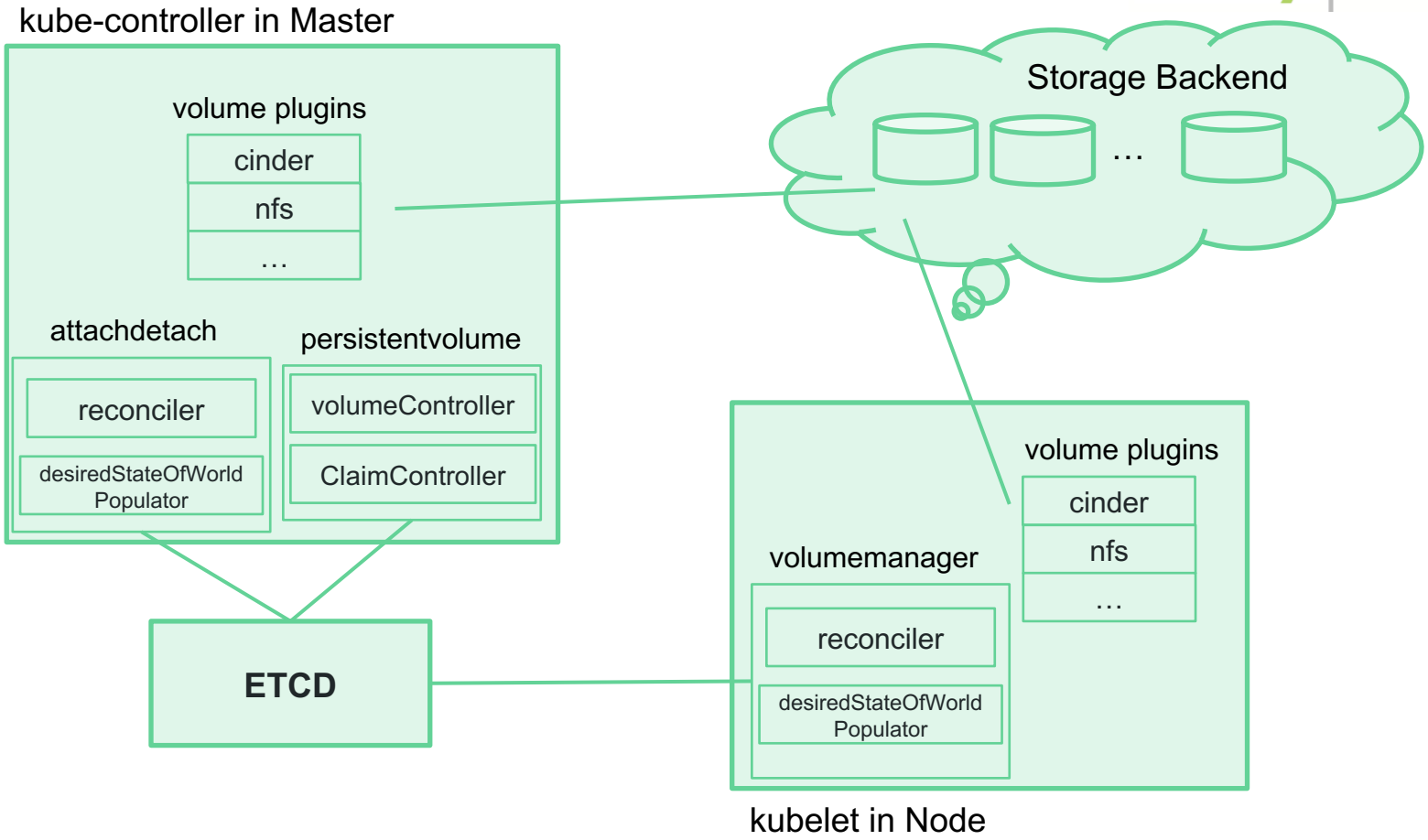
- Automatically provisions storage when it is requested by **users**
- Eliminates the need for **cluster administrators** to pre-provision storage
- The dynamic provisioning is based on **StorageClasses**

```
kind: StorageClass
apiVersion: storage.k8s.io/v1beta1
metadata:
  name: cinder-standard
provisioner: kubernetes.io/cinder
parameters:
  type: standard
```

```
kind: PersistentVolumeClaim
...
volume.alpha.kubernetes.io/storage-class:
"cinder-standard"
```



Kubernetes Components for Volumes



Storage Management in Tess

- Storage Management in Kubernetes@eBay
  - Storage Class Definition
  - Network Volume
  - Local Volume Management
    - Motivation
    - Solution in eBay
  - Health Monitoring
  - Future Work

Storage Class Definition

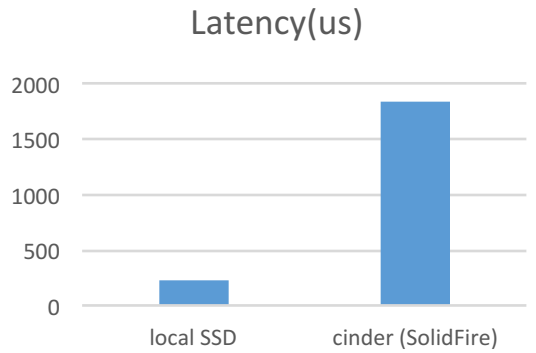
Storage Class	IOPS/Throughput	Use Cases	Storage Solutions
Hot Tier	>50K IOPS 500 MB/s	NoSQL, eBay In-house App	Local SSD
Warm Tier	10000 IOPS 300 MB/s	Hadoop, Kafka	SolidFire, Local HDD
Standard Tier	300 IOPS 150 MB/s	Archive	Ceph HDD based

Network Volume

- Kubernetes Cinder Volume
  - Why?
  - Two Classes
    - cinder-standard
    - cinder-performance
  - Stability
    - Openstack
    - Qemu
    - Kubernetes

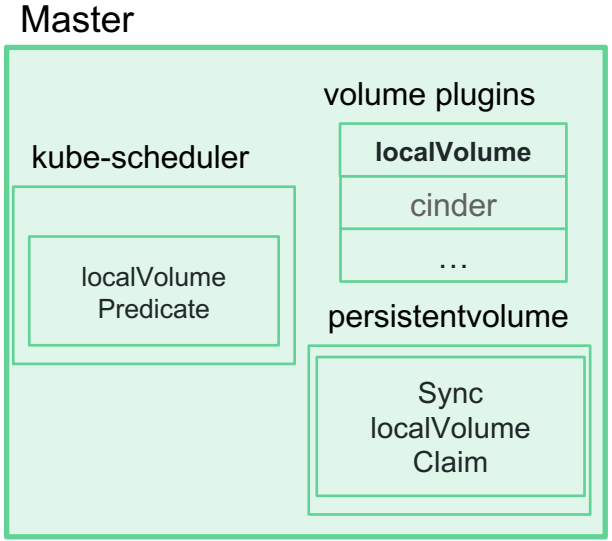
Local Volume Management

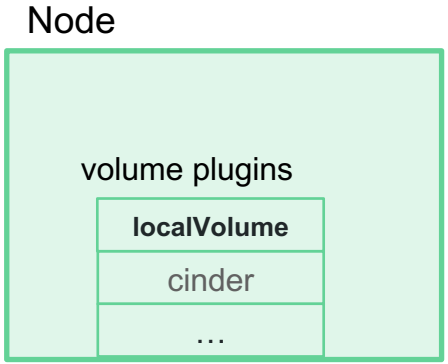
- Motivation
  - Compared with network volume
    - Performance
    - Cost
    - Availability
  - Limitation of EmptyDir and HostPath Volume in Kubernetes
    - No PV/PVC support
    - No “Guaranteed”, size/IOPS



Local Volume Management

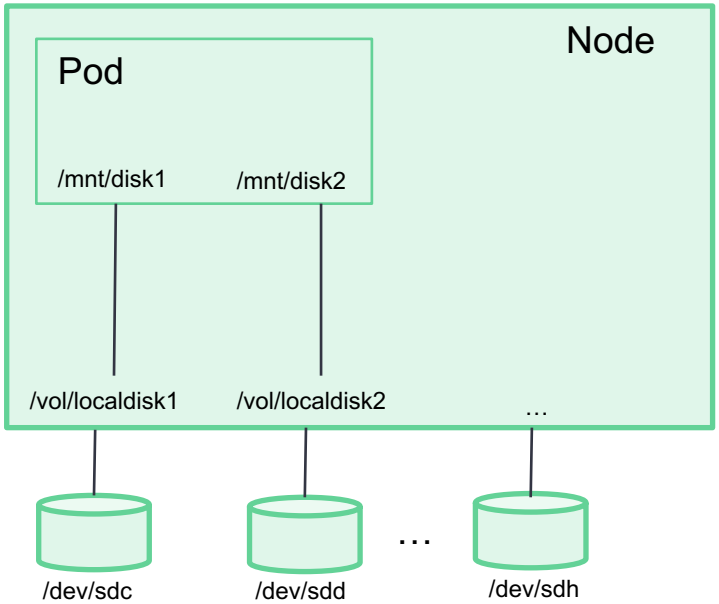
- Local Volume @eBay
  - LocalVolume plugin, supports both
    - Disk/Partition
    - LVM
  - PV/PVC
  - Dynamic provisioning and two storage-classes
  - Scheduler
    - Pod scheduler
    - PV allocation





Local Volume Management - Example

```
apiVersion: v1
kind: Pod
metadata:
  name: ld-test
spec:
  containers:
  - image: busybox
    command:
    - sleep
    - "3600"
    imagePullPolicy: IfNotPresent
    name: ld-test
    volumeMounts:
    - mountPath: /mnt/disk1
      name: test-vol1
    - mountPath: /mnt/disk2
      name: test-vol2
  volumes:
    - localVolume:
        type: disk
    name: test-vol1
    - localVolume:
        type: disk
    name: test-vol2
  restartPolicy: Always
```



Local Volume Management - Example

```
kind: StorageClass
apiVersion: storage.k8s.io/v1beta1metadata:
  name: local-ssd
provisioner: kubernetes.io/local-volume
parameters:
  type: disk
```

```
apiVersion: v1
kind: PersistentVolumeClaim
metadata:
  name: myclaim
  annotations:
    volume.beta.kubernetes.io/storage-class:"local-ssd"
spec:
accessModes:
- ReadWriteOnce
resources:
  requests:
    storage: 100Gi
```

```
apiVersion: v1
kind: Pod
metadata:
  name: ld-test
spec:
  containers:
  - image: busybox command:
  - sleep - "3600"
  imagePullPolicy: IfNotPresent
  name: ld-test
  volumeMounts:
  - mountPath: /mnt/vol
    name: test-vol
  volumes:
  - PersistentVolumeClaim:
      claimName: myclaim
    name: test-vol
  restartPolicy: Always
```

Local Volume Management - Problem

- Local Volume Recycle
- Pod with multiple localVolume PVCs

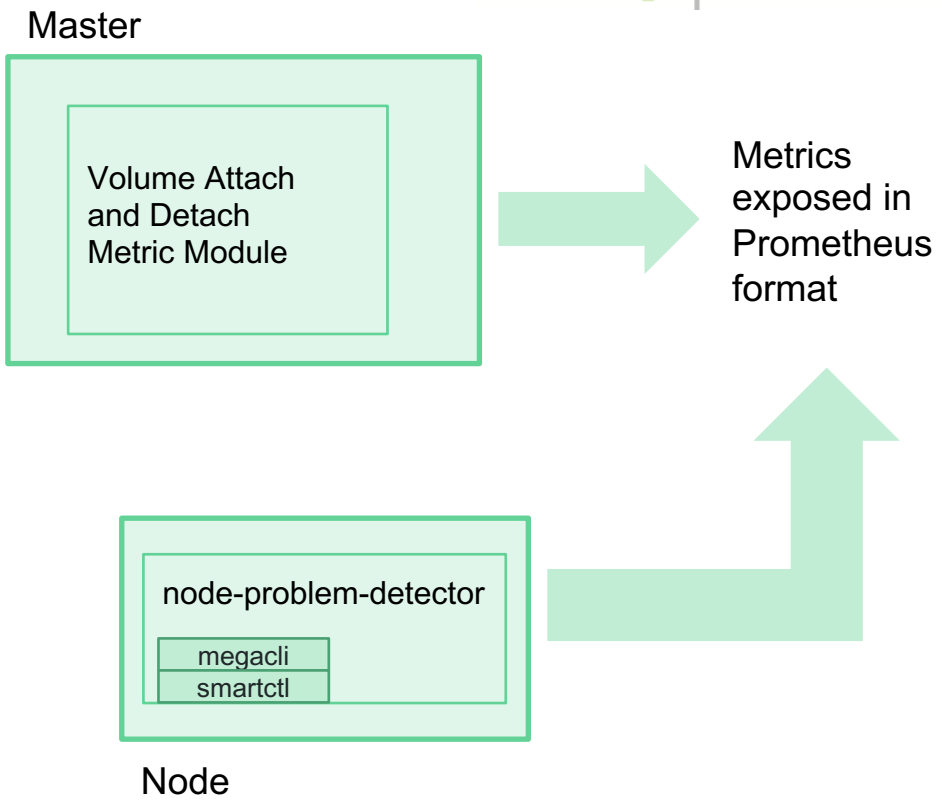
```
annotations:
    volume.beta.kubernetes.io/storage-class:"local-ssd"
    volume.beta.kubernetes.io/group: '{"name":"group1", "count":"2"}
```

- Anti-Affinity

```
annotations:
    volume.beta.kubernetes.io/storage-class:"local-ssd"
    volume.alpha.kubernetes.io/antiaffinity: '{"labelSelector":{
"matchLabels":{ "key":"ssdtest" } }, "topologyKey":"kubernetes.io/hostname"}'
```

Health Monitoring

- Network Volume
  - Collect Metrics of volume attach and detach
- Local Volume
  - Metrics collected by Raid controller tool and disk smartctl



Local Volume Management

- Future Work
  - Network Volume
    - Cinder volume interface for BM Node
    - Shared Filesystem Volume, CephFS
  - Local Volume
    - PV bindings happening in the Pod scheduler vs in PV controller?
    - DaemonSet for localVolume discovery and recycle.
    - Raw block devices management

Reference

- Overview of Kubernetes Storage

https://docs.google.com/presentation/d/1YbD1o5-NroRCOJcDlqIp1pCykeGModfdKIcaXkynf_w/pub?start=false&loop=false&delayms=3000&slide=id.g1a9b32dff3_0_376

- LocalVolume Proposal

https://github.com/chengyli/kubernetes/commit/5480f257477d802e83a946540071a1ecea0fb50e

- [Proposal] Improve Local Storage Management

https://github.com/kubernetes/community/pull/306