A Full Stack journey to reach Efficient Container Storage Cloning

Niels de Vos
Gluster Maintainer
Engineer at Red Hat
ndevos@redhat.com

@nixpanic
Agenda

• Brief Introduction into Kubernetes Storage
• KubeVirt storage requirements
• Containerized-Data-Importer and Cloning
• Linux filesystem support for copy_file_range()
• Passing copy_file_range() on to Gluster
Start of the trip with Kubernetes and KubeVirt
Brief Introduction into Kubernetes Storage

- Dynamic provisioned persistent storage
- PersistentVolumeClaim and StorageClass
- Storage provisioners
  - Internal
  - External Storage project
  - Container Storage Interface
kind: StorageClass
apiVersion: storage.k8s.io/v1
metadata:
  name: glusterfile
provisioner: gluster.org/glusterfile
parameters:
  resturl: "http://127.0.0.1:8081"
  restuser: "admin"
  restsecretnamespace: "default"
  restsecretname: "heketi-secret"
  #clusterid: "454811fcedbec6316bc10e591a57b472"
  volumetype: "replicate:3"
  volumeoptions: "features.shard enable"
  volumenameprefix: "dept-dev"
kind: PersistentVolumeClaim
apiVersion: v1
metadata:
  name: my-persistent-data
spec:
  storageClassName: "glusterfile"
  accessModes:
    - ReadWriteMany
  resources:
    requests:
      storage: 4Gi
KubeVirt storage requirements

- Virtual Machines in containers managed by Kubernetes
- Disk-image on a PersistentVolume
- Cloning new VMs from a Template
- Scalable and resilient
Brief stop at the Containerized-Data-Importer
Containerized-Data-Importer

- DataVolume (CRD) as an extension to PVCs
- Pre-populating new disk-images
- Download image from the network
- Delay ‘readiness’ until the PVC is provisioned
apiVersion: cdi.kubevirt.io/v1alpha1
kind: DataVolume
metadata:
  name: "example-import-dv"
spec:
  source:
    http:
      url: "https://download.cirros-cloud.net/0.4.0/cirros-0.4.0-x86_64-disk.img" # Or S3
      secretRef: "" # Optional
  pvc:
    accessModes:
      - ReadWriteOnce
    resources:
      requests:
        storage: "64Mi"
CDI and Cloning

- Supports cloning of PVCs by Annotations
- `CloneRequest` is passed in the PVC
- `CloneOf` is returned by the provisioner
- Host-Assisted cloning as fallback
CDI and Host-Assisted Cloning
apiVersion: cdi.kubevirt.io/v1alpha1
kind: DataVolume
metadata:
  name: "example-clone-dv"
spec:
  source:
    pvc:
      name: source-pvc
      namespace: example-ns
    pvc:
      accessModes:
      - ReadWriteOnce
      resources:
        requests:
          storage: "128Mi"
apiVersion: v1
kind: PersistentVolumeClaim
metadata:
  name: "target-pvc"
  namespace: "target-ns"
  labels:
    app: Host-Assisted-Cloning
  annotations:
    k8s.io/CloneRequest: "source-ns/golden-pvc"
spec:
  accessModes:
  - ReadWriteOnce
  resources:
    requests:
      storage: 10Gi
Short detour passing Linux filesystems
Linux filesystem support for `copy_file_range()`

- Linux VFS uses `clone_file_range()`
- XFS implements `clone_file_range()` with ‘reflinks’
- FUSE supports `clone_file_range()` and passes it on to the userspace process handling the filesystem
- Fallback in glibc when `copy_file_range()` is unsupported
Linux filesystem support for `copy_file_range()`

- Fallback in glibc when `copy_file_range()` is unsupported
- Very similar to Host-Assisted cloning
Reaching the destination on a solid brick road
Passing `copy_file_range()` on to Gluster

- Restricted to single mountpoint
- PV to be placed in a subdirectory
- Requires support in FUSE and the client-side mount process
- Network protocol additions from client to server
Handling `copy_file_range()` in Gluster

- Receive `copy_file_range()` calls from clients
- Bricks need to be on a filesystem with support for `copy_file_range()`
- Limitations with distributed Gluster Volumes
Efficient Cloning with subdirectories
References

- kubenvirt.io
  - Containerized-Data-Importer
  - Talk by Fabian Deutsch, Wed. 14:15
- Kubernetes External Storage
  - Gluster Subvol Provisioner (PR#1013)
- gluster.org
- FUSE clone_file_range() (Linux, libfuse)
