Lessons Learned Containerizing GlusterFS and Ceph with Docker and Kubernetes

Huamin Chen
@root_fs
github: rootfs
Emerging Technologies
Red Hat

Outline

- Background
- Containerizing Ceph and Gluster
- Working with Docker Containers
- Deploying Glusterfs and Ceph using Kubernetes and Ansible
- Working with Kubernetes
- Q&A

Background

Emerging technologies for software packaging, deployment, and orchestration

- Packaging: rpm/deb vs. Docker
- Deployment: Ansible/Puppet/Chef for large cluster software deployment
- Orchestration: Kubernetes/Mesos/Swarm to orchestrate containers/applications

Packaging

- Then
  - To install Ceph and Glusterfs: yum install glusterfs ceph
  - Issues:
    - platform dependent: yum or apt
    - Package dependent
    - Poor upgrade experience

Multiple distributions:

..		
debian_ceph_repository.yml	rollback previous change for ceph-common change	2 months ago
install_on_debian.yml	Deduplicate RBD client directory creation	8 days ago
install_on_redhat.yml	Merge pull request #696 from stpierre/dedup-rbd-client-dirs	4 days ago
install_rgw_on_debian.yml	rollback previous change for ceph-common change	2 months ago
install_rgw_on_redhat.yml	rollback previous change for ceph-common change	2 months ago
install_rh_storage_on_debian.yml	adds the rh storage apt-key for jewel on ubuntu	a month ago
redhat_ceph_repository.yml	rollback previous change for ceph-common change	2 months ago

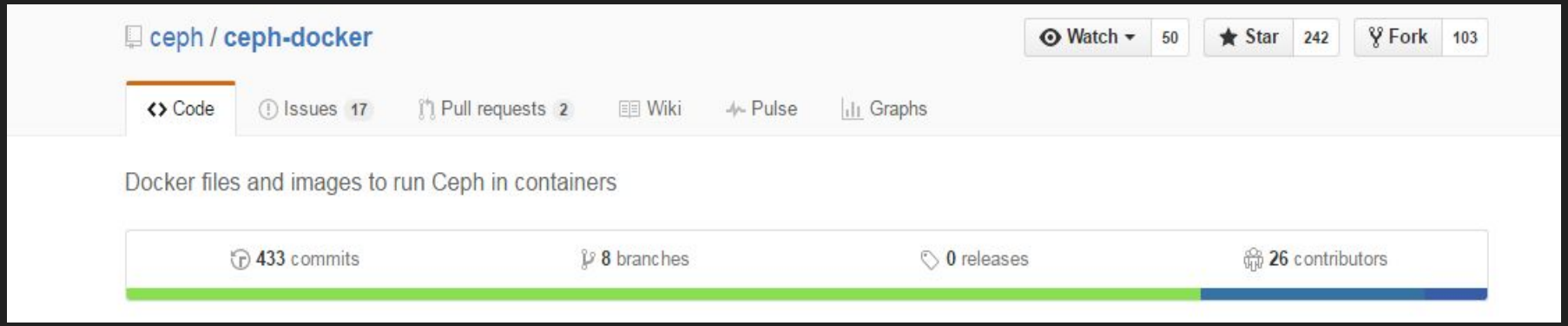
https://github.com/ceph/ceph-ansible/tree/master/roles/ceph-common/tasks/installs

Single Distribution, Multiple Releases:

```
---
- name: add ceph extra
  apt_repository:
    repo: "deb http://ceph.com/packages/ceph-extras/debian {{ ansible_lsb.codename }} main"
    state: present
  when: ansible_lsb.codename in ['natty', 'oneiric', 'precise', 'quantal', 'raring', 'sid', 'squeeze', 'wheezy']

# NOTE (leseb): needed for Ubuntu 12.04 to have access to libapache2-mod-fastcgi if 100-continue isn't being used
- name: enable multiverse repo for precise
  apt_repository:
    repo: "{{ item }}"
    state: present
  with_items:
    - deb http://archive.ubuntu.com/ubuntu {{ ansible_lsb.codename }} multiverse
    - deb http://archive.ubuntu.com/ubuntu {{ ansible_lsb.codename }}-updates multiverse
    - deb http://security.ubuntu.com/ubuntu {{ ansible_lsb.codename }}-security multiverse
  when:
    ansible_lsb.codename in ['precise'] and not
    http_100_continue
```

Get Containerized!



- Containerize Ceph releases (Hammer, Infernalis and upcoming Jewel)
- All daemons in one container: MON, OSD, RGW
- Bootstrap from scratch or from KV store

Run Containers

- Install and run container images
  - docker run -d ceph/daemon ...
- Platform independent
  - Containers have all the necessary bits, no more package dependency.
  - Same command on RHEL (including Atomic host), CoreOS, Ubuntu ...
- Easy to switch and upgrade
  - upgrade: docker pull ceph/daemon:latest
  - switch: docker run -d registry.access.redhat.com/rhceph/rhceph-1.3-rhel7 ...

Working with Systemd in Containers

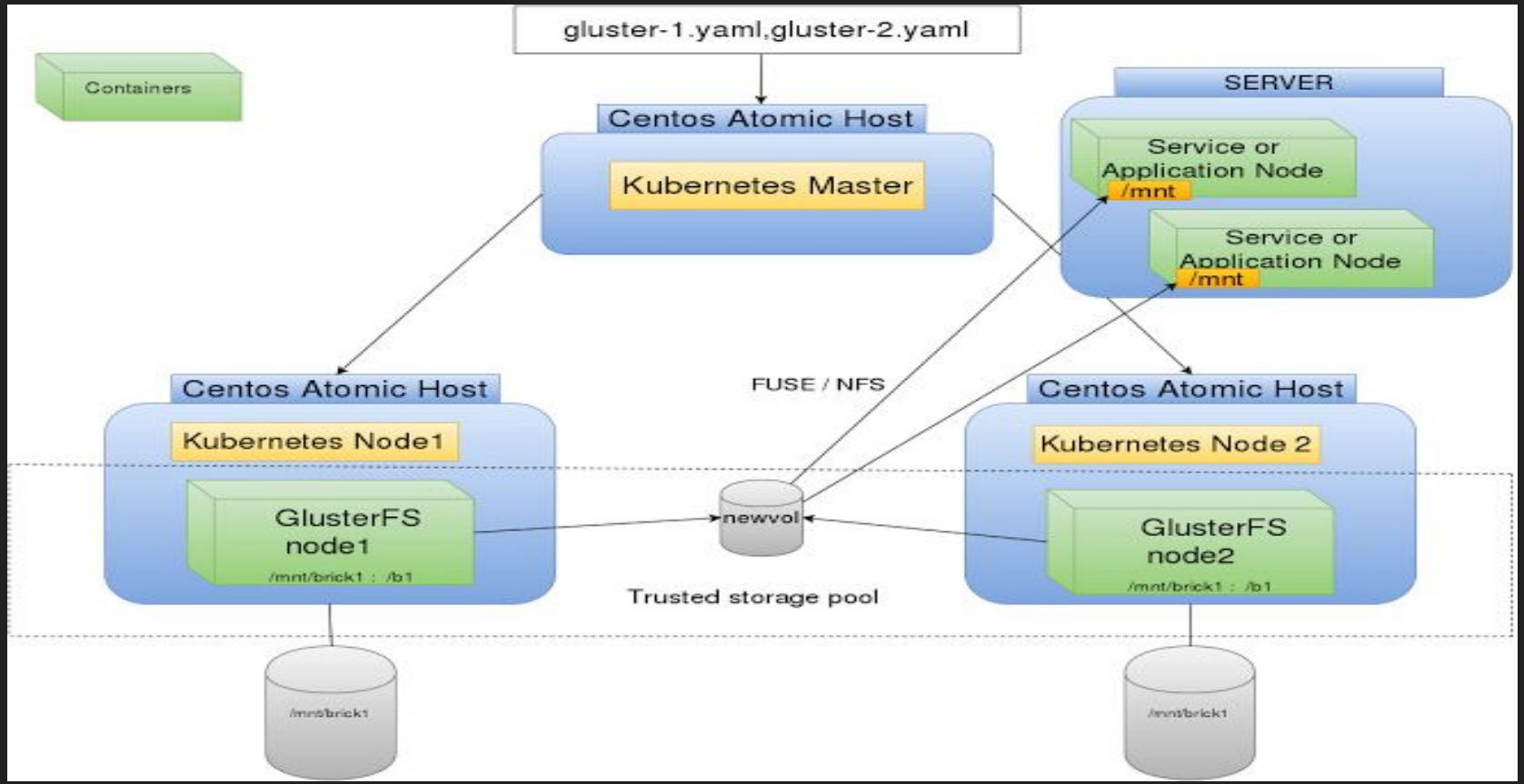
Systemd and daemon containers both want to manage host resources and trigger handler processes. But they do not always work well with each other.

- udev
  - Glusterfs: with lvmetad in place, host and container have different views of logic volumes
  - Ceph: udev rules triggers ceph-disk, which in turn starts ceph-osd daemon containers (work in progress)
- Managing daemon process
  - Containerized Glusterfs: in-container systemd manages gluster daemon.
  - Containerized Ceph: on-host systemd manages Ceph daemons, so OSD container can respond to udev trigger.

Deployment

- Traditionally storage systems are deployed and managed by storage admins
  - Mostly script based deployment
    - ceph-deploy written in python, thousands lines of code
    - Similar Glusterfs installer written in bash also claims thousands lines
- But increasingly DevOps are playing the “admin” roles.
  - New goals:
    - Repeatable: can be executed by anybody anywhere
    - Reusable: integrated with other frameworks (e.g. Kubernetes and Ansible)
    - Readable: declarative as in Kubernetes and Ansible

Deploy Glusterfs on Kubernetes

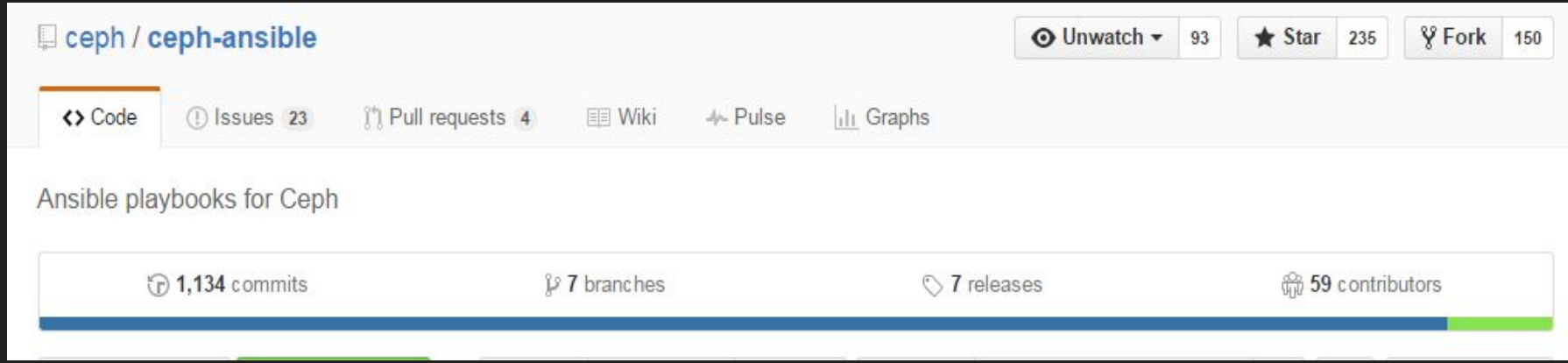


http://website-humblec.rhcloud.com/gluster_containers_in_kubernetes_cluster/

Glusterfs Pod

```yaml
apiVersion: v1
kind: Pod
metadata:
   name: gluster-1
labels:
   name: gluster-1
spec:
  # use host network and host IP address
  hostNetwork: true
  # node affinity - only run container on selected nodes
  nodeSelector:
     name: worker-1
  containers:
   - name: glusterfs
     image: gluster/gluster-centos
     volumeMounts:
      - name: brickpath
        mountPath: "/mnt/brick1"
     securityContext:
        # run as privileged container
        privileged: true
volumes:
   - name: brickpath
      # use local directory
      hostPath:
      path: "/mnt/brick1"
```

Ceph-ansible



- Deploy multiple Ceph releases (Hammer, Infernalis, and upcoming Jewel)
- Deploy on CentOS/RHEL 6 and 7 and multiple Ubuntu releases
- Deploy on Atomic Host and CoreOS
- Deploy both Ceph packages as well as ceph containers
- Deploy on bare metal, VMs (libvirtd and VirtualBox), and OpenStack

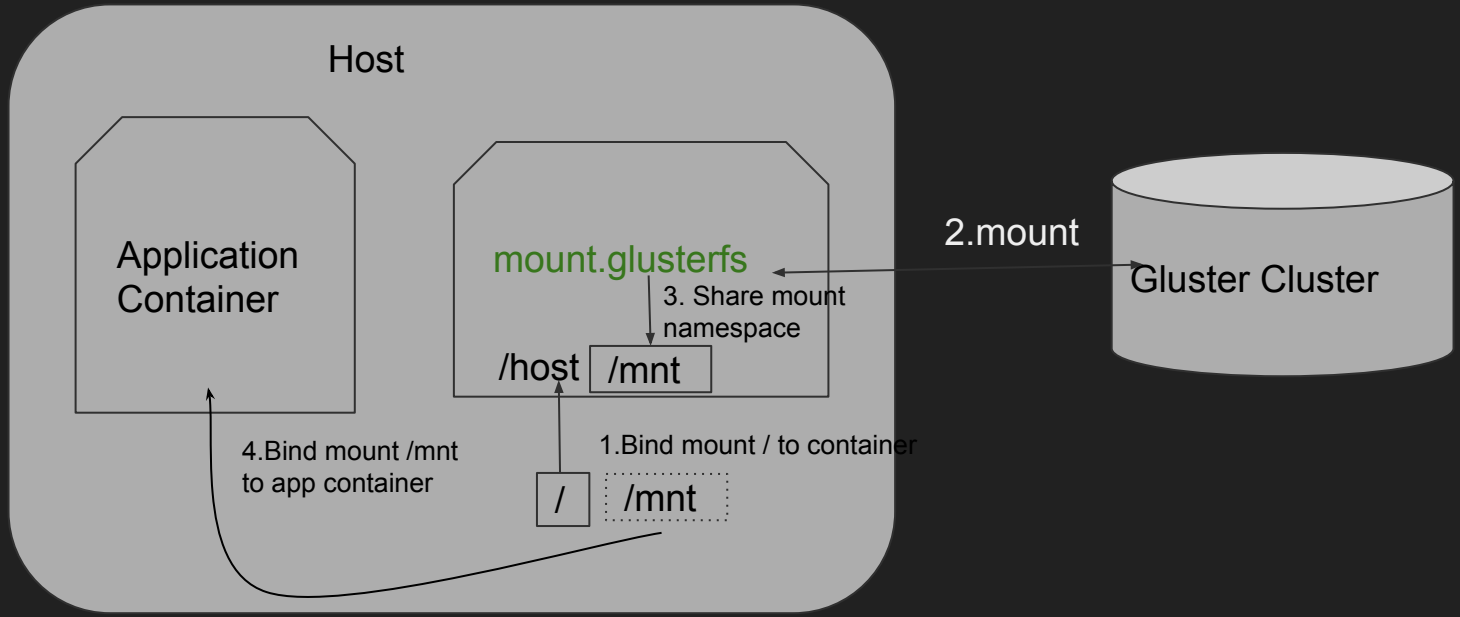
Using Glusterfs and Ceph in Containers

Problem of installing client packages

- Installing and upgrading client packages on a large cluster is not fun!
- Sometimes client packages cannot be installed
  - Atomic host and CoreOS require these packages are built into OS images

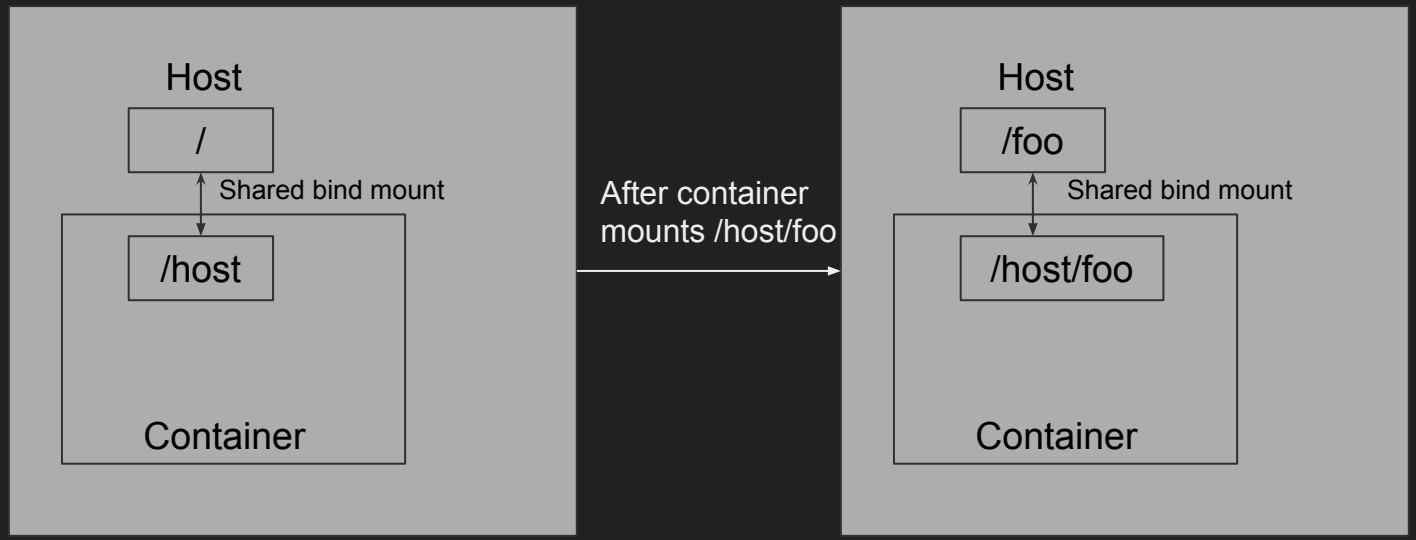
Containerized Client

- Run mount.glusterfs in a container!



Mount Namespace Propagation

docker run -v /:/host:rw,shared, available in Docker 1.10



Security

Docker leverages existing security features (SELinux/AppArmor/etc) to isolate containers. Unprivileged containers are not able to access paths that don’t have proper SELinux labels. SELinux support is critical in multi-tenant environment.

SELinux uses security.selinux namespace in inode’s extended attributes.

SELinux is supported by local filesystems (xfs, ext), Glusterfs, and NFS v4.2

Questions?