

oVirt and Gluster hyper-converged!

HA solution for maximum resource utilization

21st of Aug 2015

Martin Sivák
Senior Software Engineer
Red Hat Czech

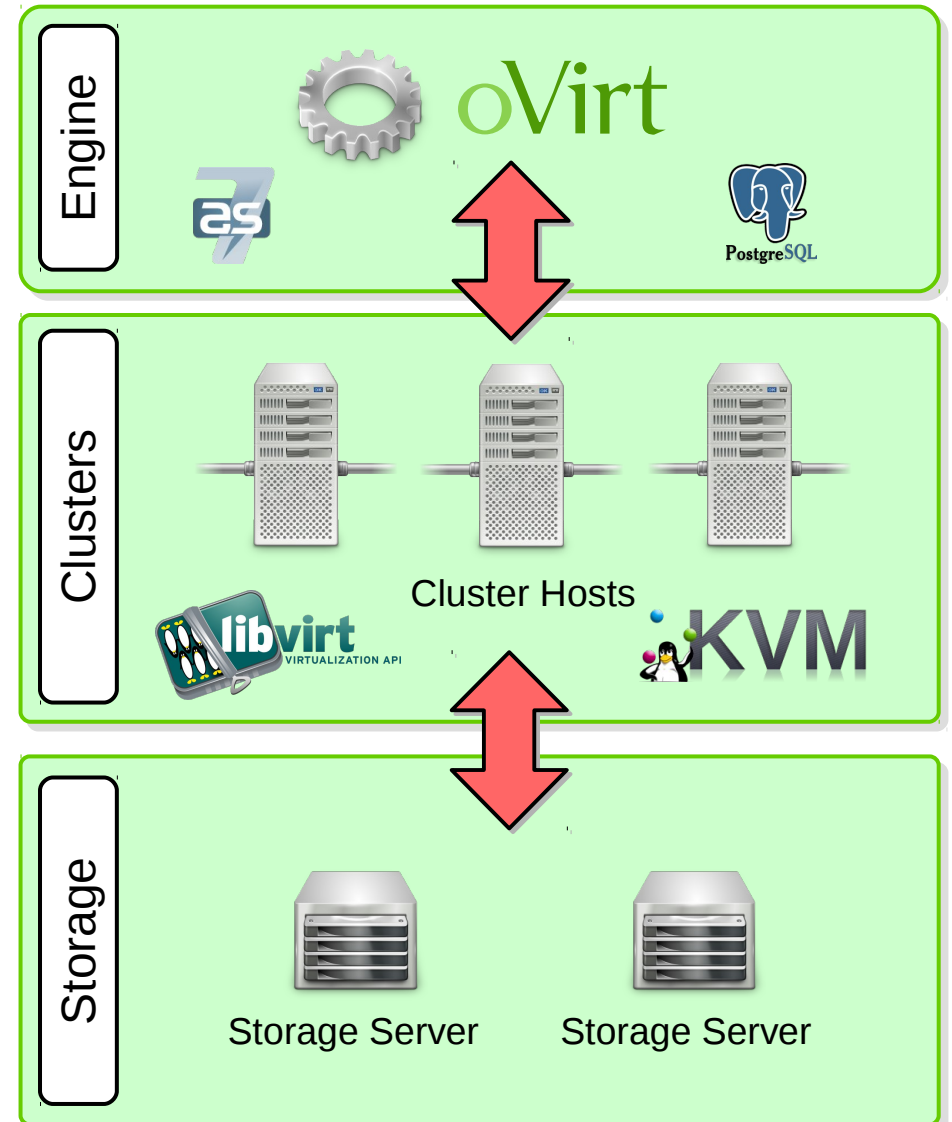
- (Storage) architecture of oVirt
- Possible failure points in standard oVirt setup
- Hosted engine refresher and improvements
- Gluster in a nutshell
- Putting it all together – hyper converged infrastructure
 - Architecture
 - Setup
 - Management

oVirt and its Architecture



oVirt is a virtualization platform to manage virtual machines, storage and networks

- **Engine (ovirt-engine)**
Manages the oVirt hosts, and allows system administrators to create and deploy new VMs
- **Host Agent (VDSM)**
oVirt engine communicates with VSDM to manage the VMs, storages and networks



oVirt storage

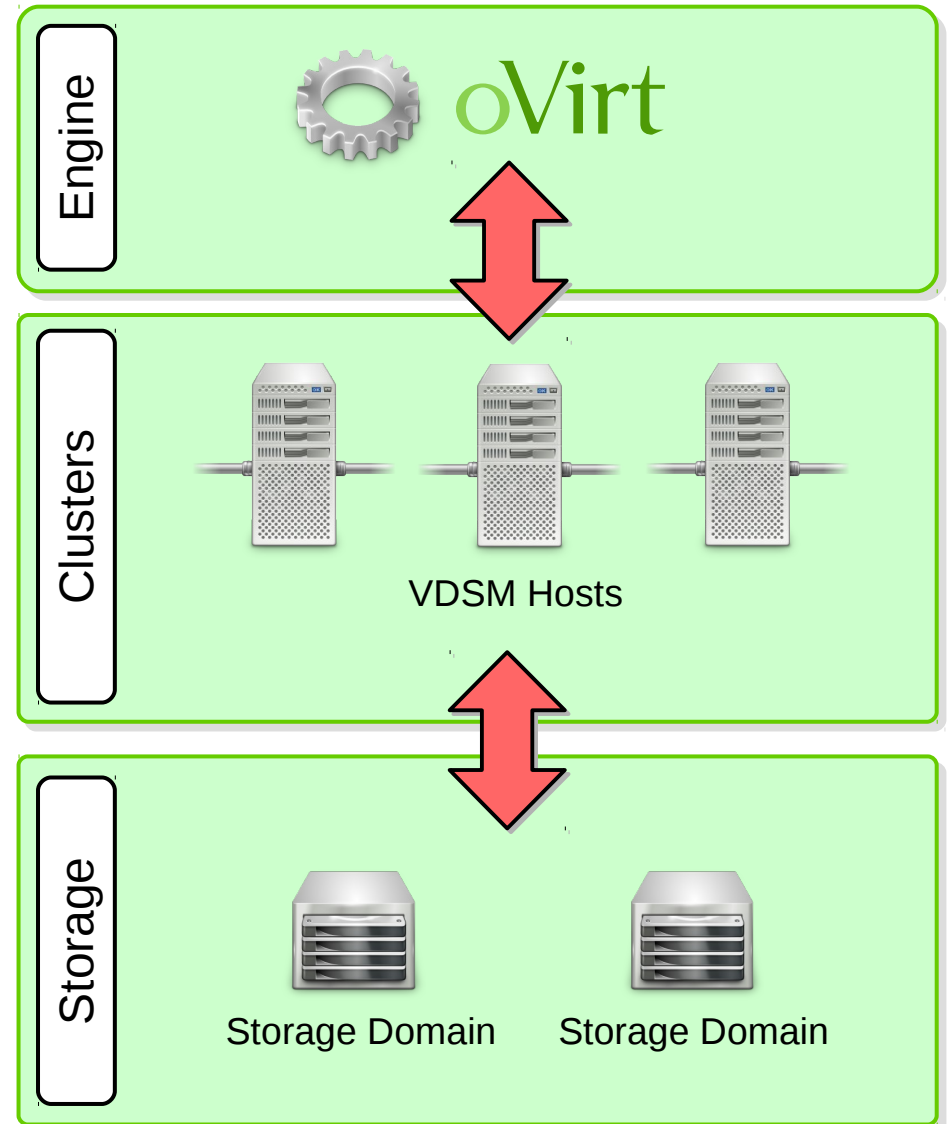


- Storage Domains

- Centralized storage system (images, templates, etc.)
- A standalone storage entity
- Stores the images and associated metadata
- Only real persistent storage for VDSM
- Used for synchronization (sanlock)

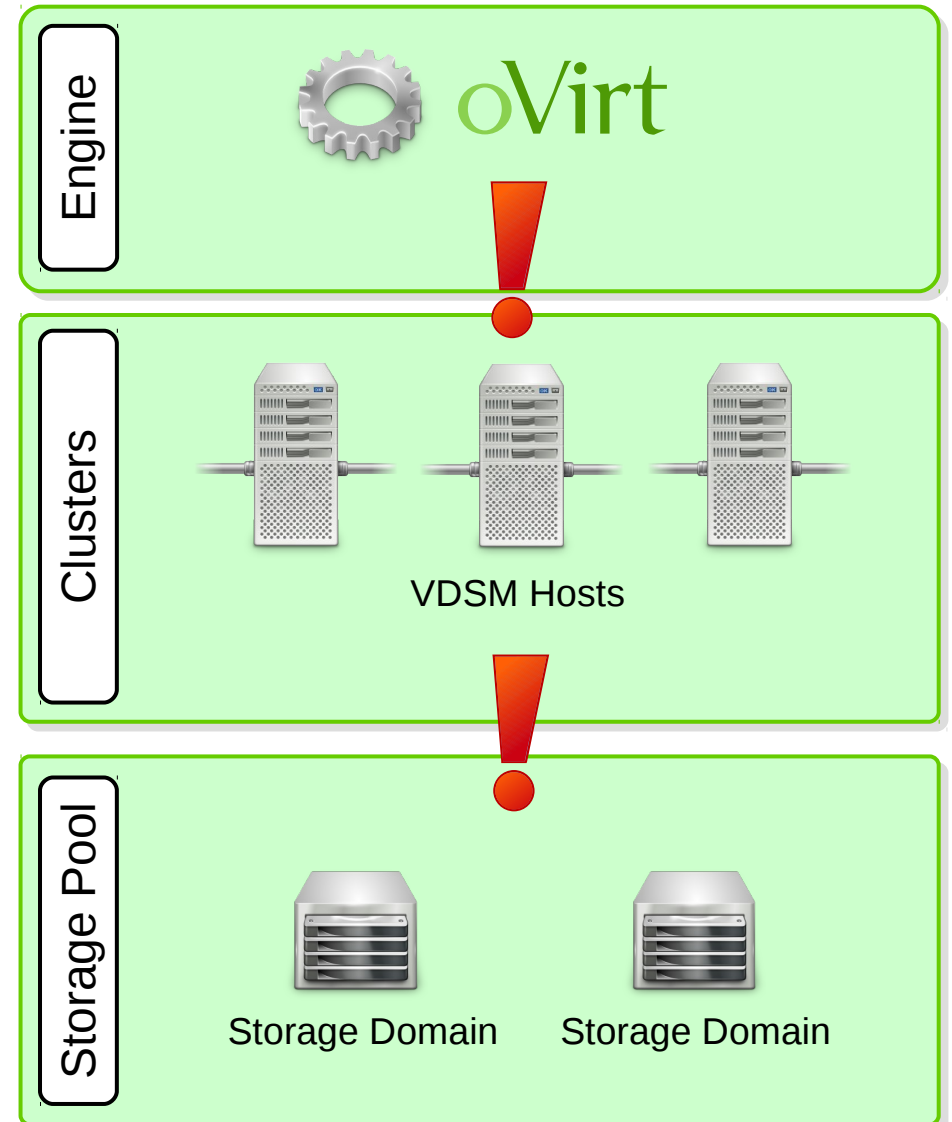
- Storage types

- NFS, FCP, iSCSI
- Gluster



Possible failure points

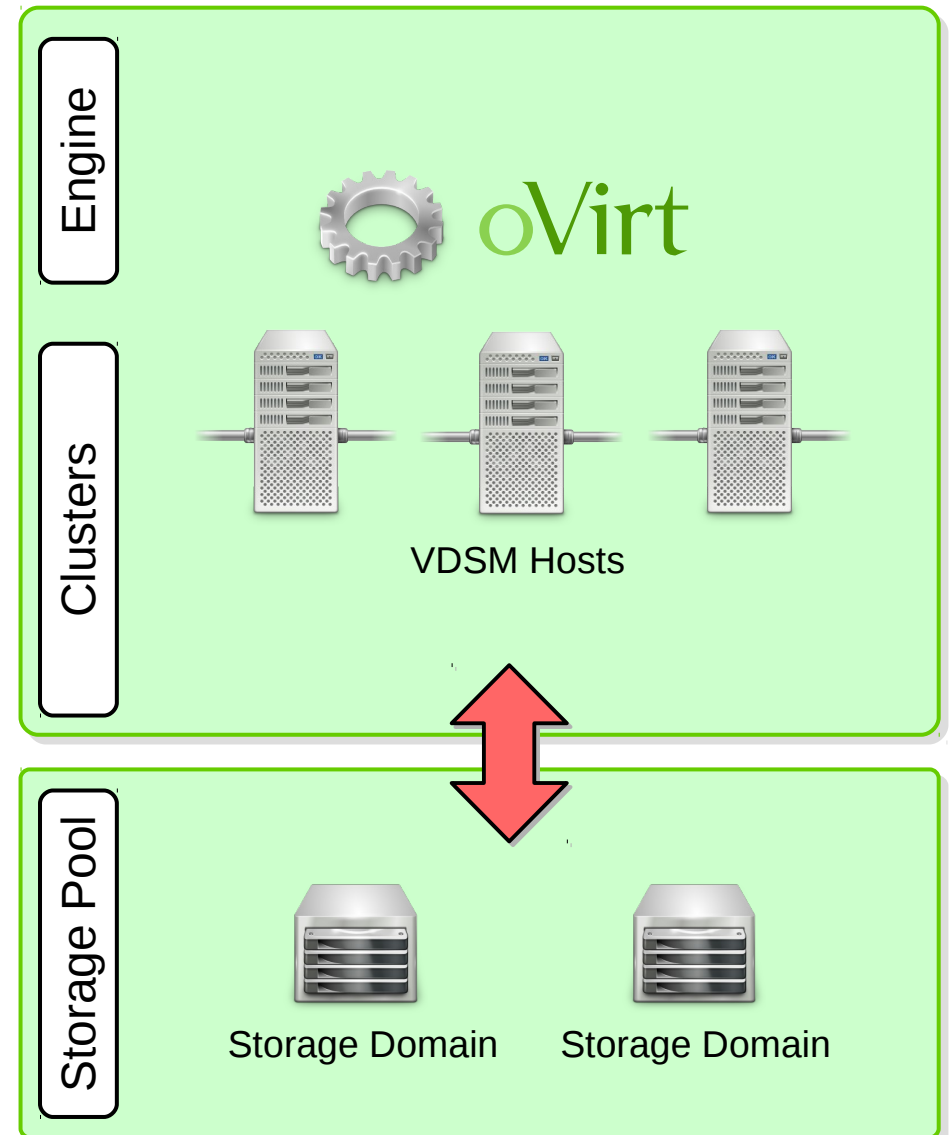
- Engine machine
 - Single point of failure
 - Cluster paralyzed without engine
- Storage connection
 - Data safe but unreachable
 - All synchronization in oVirt is storage based
 - neither NFS nor iSCSI provide redundancy



- Single ovirt-engine host manages the whole datacenter
 - Using a VM to run ovirt-engine reduces HW failure risks
 - **Hosted Engine**
- Single storage access infrastructure provides data
 - Data itself are safe – can be replicated using RAID
 - Infrastructure is not – distributed access mechanism is needed
 - **Gluster**

Hosted engine

- Management running inside a VM
- Can be migrated to a different node
- High availability
- Special agent for monitoring
- **Storage based synchronization**
- Bootstrap deployment needed

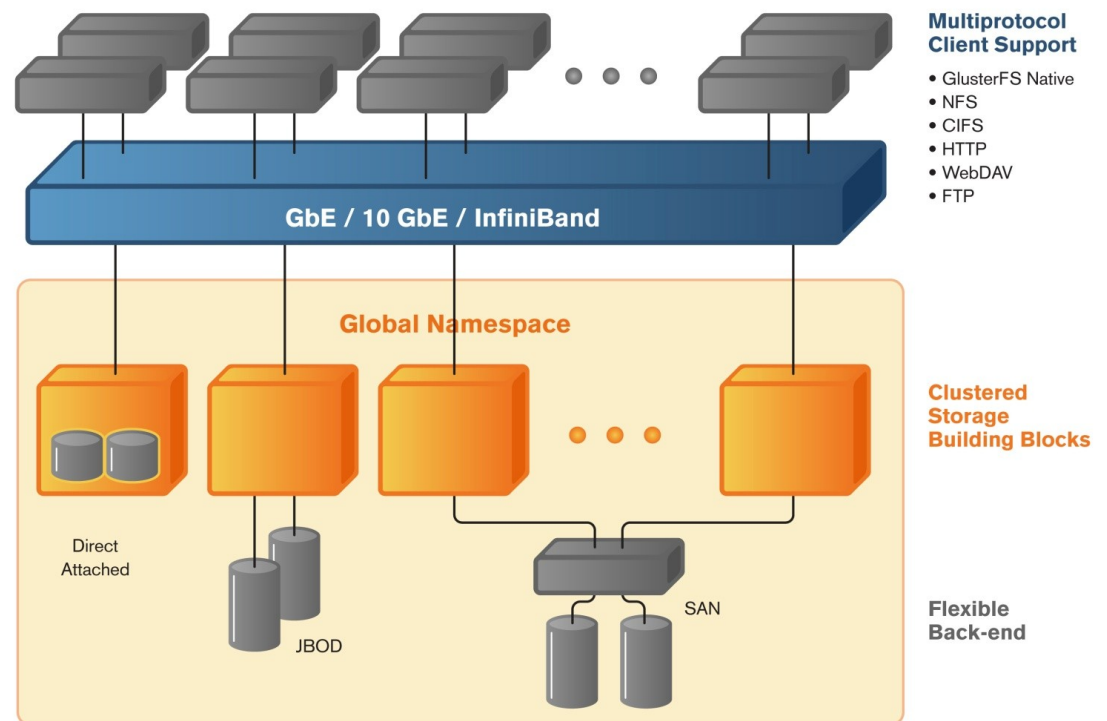


Improvements needed for HC

- GlusterFS support re-added to setup
 - With gfapi support!
- oVirt-engine appliance
 - preconfigured management VM
 - cloud-init based customization
- Shared configuration
 - all nodes see the same configuration data
 - upgrade path from oVirt 3.5
- Management GUI for the oVirt-engine VM and HE

GlusterFS and its Architecture

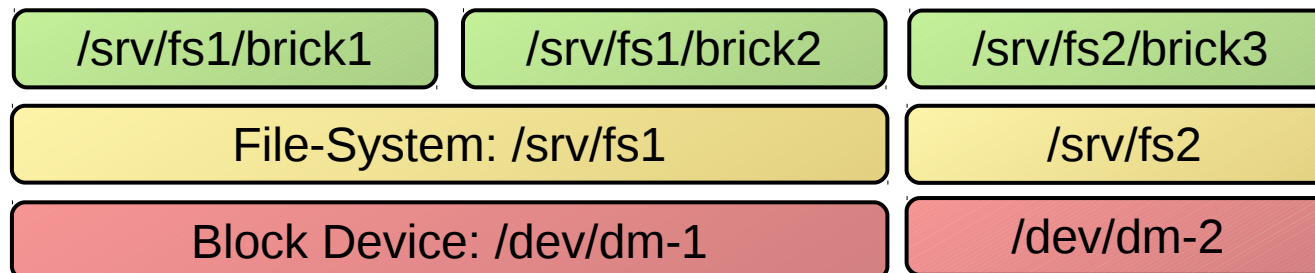
- GlusterFS is a general purpose scale-out distributed file-system supporting thousands of clients
- Aggregates storage exports over network interconnect to provide a single unified namespace
- File-system completely in userspace, runs on commodity hardware
- Layered on disk file systems that support extended attributes



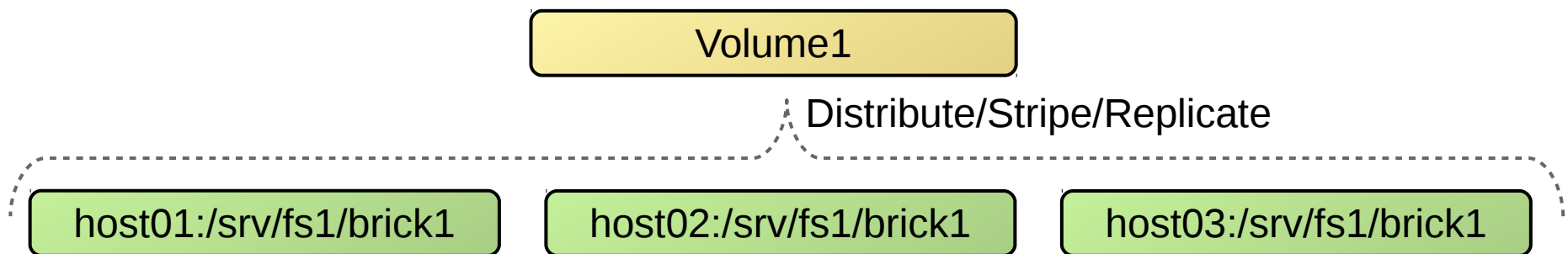
GlusterFS Bricks



- A brick is an export directory located on a specific node (e.g. host-01:/srv/fs1/brick1)
- Each brick inherits limits of the underlying file-system
- No limit on the number bricks per node (as best-practice each brick in a cluster should be of the same size)

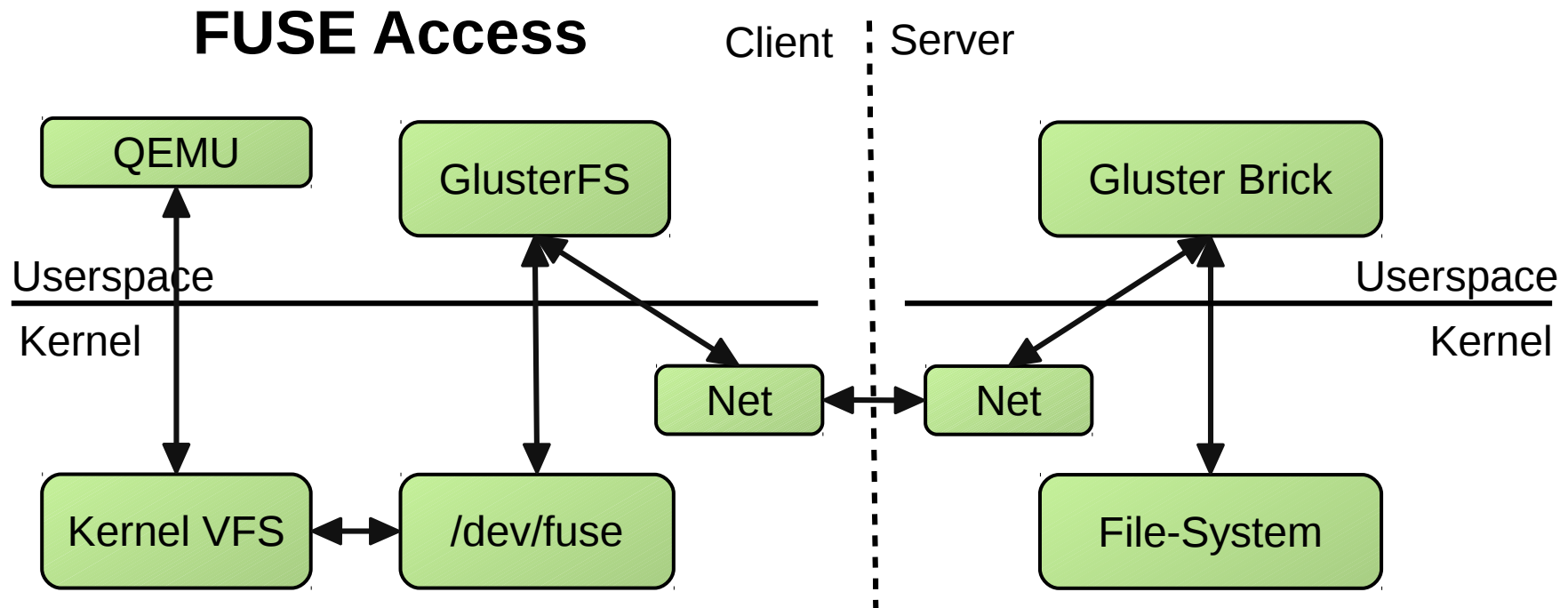


- A volume (the mountable entity) is a logical collection of bricks
- Bricks from the same node can be part of different volumes
- Different types of Volumes
 - Distribute, Stripe, Replicate (+ combinations), **Quorum**
- Type of a volume is specified at the time of volume creation and determines how and where data is placed



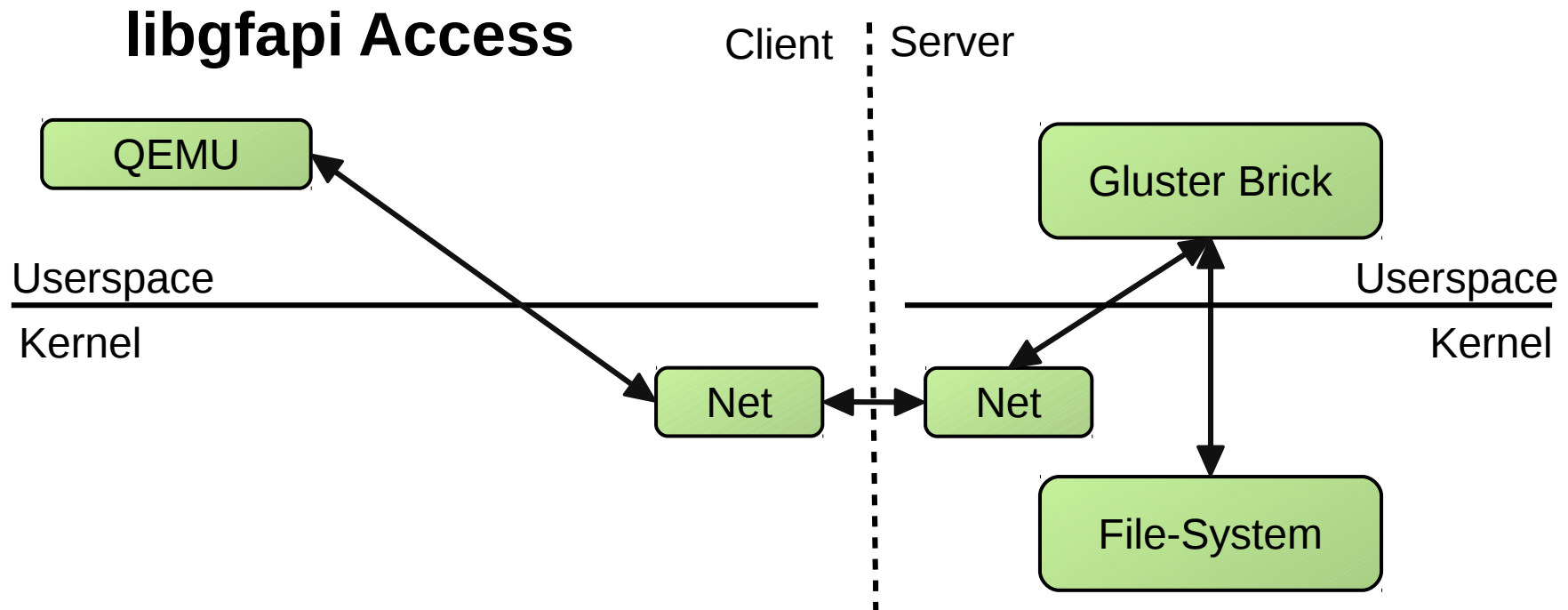
QEMU libgfapi Support

- GlusterFS exposes APIs for accessing Gluster volumes
- Reduces context switches



QEMU libgfapi Support

- GlusterFS exposes APIs for accessing Gluster volumes
- Reduces context switches



But see: https://bugzilla.redhat.com/show_bug.cgi?id=1247933

Putting it all together



- oVirt cluster
- Glusterfs backed storage domain
- Hosted engine to maintain HA of the management
- Pre-configured management using an OVF image

Are you feeling lucky?

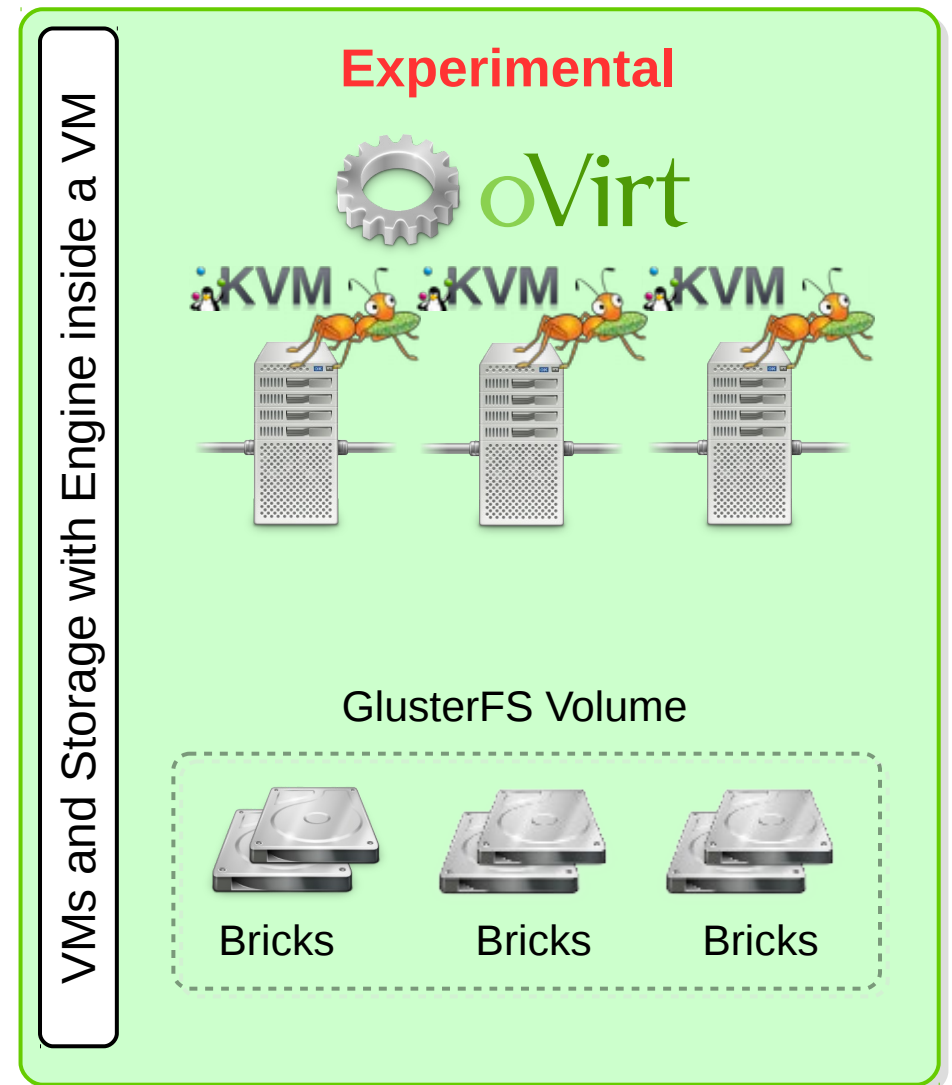
Due to unexpected issues the automatic HC deployment was **removed from 3.6**. It is still possible to configure most of the HC setup manually.



Hyperconverged oVirt – GlusterFS



- The Data Center nodes are used both for virtualization and serving replicated images from the GlusterFS Bricks
- Engine runs inside a VM (Hosted Engine)
- The boxes can be standardized (hardware and deployment) for easy addition and replacement
- Support for both scaling up, adding more disks, and scaling out, adding more hosts



Hyper converged setup – ingredients



- at least 3 virtualization capable hosts (CentOS 7.1+)
- 10 GB of temporary space on the primary host
- two separate partitions for data (20GB+) on all hosts
- DHCP configured to map a MAC address to a fixed IP
- DNS configured with A and PTR names for the IP
- oVirt release package installed on all hosts

<http://resources.ovirt.org/pub/ovirt-master-snapshot-static/rpm/el7/noarch/ovirt-release-master-001-0.9.master.noarch.rpm>

- Physical console on the primary host or network access and screen package installed

- **Replica 3 volume required**

```
# execute on all hosts
yum install glusterfs-server
systemctl enable glusterfs-server
systemctl start glusterfs-server
mkdir -p /srv/gluster/hosted-engine/brick
```

```
# Execute on the first host you are going to deploy
gluster peer probe <address another host> # for each host in the HC cluster
gluster volume create hosted-engine replica 3 \
  <host1>:/srv/gluster/hosted-engine/brick \
  <host2>:/srv/gluster/hosted-engine/brick \
  <host3>:/srv/gluster/hosted-engine/brick \
  ...
gluster volume start hosted-engine
```

- **This step will be automated by the setup tool once remaining bugs are solved**

Gluster volume setup – cont.



```
# Execute on the first host you are going to deploy
gluster volume set hosted-engine cluster.quorum-type auto
gluster volume set hosted-engine network.ping-timeout 10
gluster volume set hosted-engine auth.allow \*
gluster volume set hosted-engine group virt
gluster volume set hosted-engine storage.owner-uid 36
gluster volume set hosted-engine storage.owner-gid 36
gluster volume set hosted-engine server.allow-insecure on
```

Hosted engine - recipe



```
yum install -y ovirt-engine-appliance ovirt-hosted-engine-setup
yum install -y vdsm-gluster glusterfs-server
ovirt-hosted-engine-setup
```

...

Please specify the storage you would like to use: glusterfs

Please specify the full shared storage connection path to use: <ip1>:/hosted-engine

[INFO] GlusterFS replica 3 Volume detected

...

Please specify the device to boot the VM from [disk]: disk

The following appliance have been found on your system:

[1] – The oVirt Engine Appliance image (OVA) – 20150802.0-1.el7.centos

[2] – Directly select an OVA file

Please select an appliance (1, 2): 1

...

Please specify the memory size of the appliance in MB: 16384

Would you like to use cloud-init to customize the appliance on the first boot?: Yes

Please provide the FQDN you would like to use for the engine appliance: <engine fqdn>

...

Hosted engine - recipe



```
...  
You may specify a unicast MAC address for the VM: <MAC assoc. with the FQDN>  
...
```

```
--== Configuration Preview ==--  
...  
Please confirm installation settings: Yes  
...
```

- Quite lot of questions and lines were omitted for brevity, but the answers to those are not “too important” for successful installation of hosted engine.
- You can watch a full appliance installation (using NFS storage) video on YouTube: https://www.youtube.com/watch?v=ODJ_UO7U1WQ

Finishing setup of the oVirt cluster

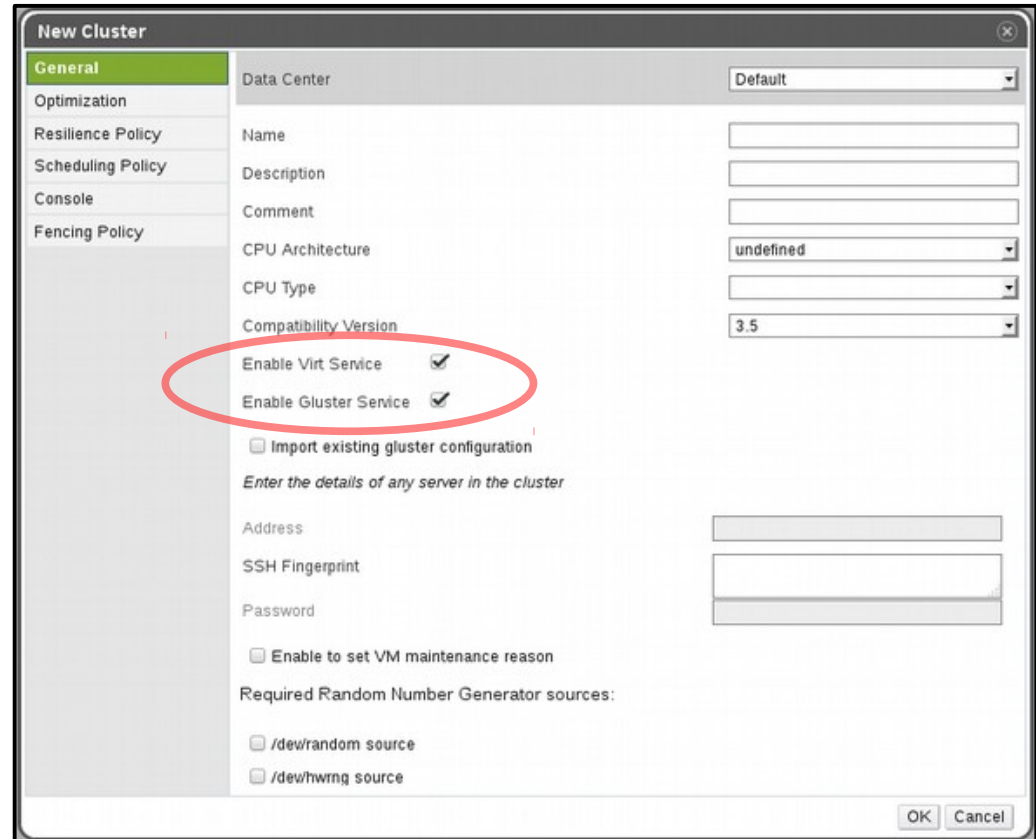


- You should now have a running single node oVirt
- Log in to the management
- Make sure Gluster support is enabled
- Add remaining nodes
- Create and add the main storage domain

Enabling GlusterFS



- Gluster Service support is located in the Cluster properties
- Deploy Hosts with GlusterFS Server support
- Enable Bricks and Volume Management from oVirt WebAdmin and REST-API
- Engine is not taking in consideration GlusterFS on Virtualization Power-Saving policies and Fencing yet



Adding additional nodes

- Simple checkbox during in the Add host dialog
- Host deploy script does everything else auto-magically

Install Host

General

Network Provider

Authentication

User Name: root

Password

SSH PublicKey

Automatically configure host firewall

Activate host after install

Deploy Hosted Engine Agent

Hosted Engine Agent Gateway

OK Cancel

Adding Gluster storage

- It is possible to create and manage Gluster Volumes from WebAdmin and using the REST-API

- Volume Profiling
- Volume Capacity Monitoring

The 'New Volume' dialog box shows configuration options for a Gluster volume. It includes dropdowns for Data Center (MixedDataCenter1) and Volume Cluster (MixedCluster1), a text field for Name (Volume1), and a dropdown for Type (Distribute). Under Transport Type, TCP is selected and RDMA is unselected. There is an 'Add Bricks' button and a status '(0 bricks selected)'. Under Access Protocols, Gluster, NFS, and CIFS are all checked. There is a text field for 'Allow Access From' with a note '(Comma separated list of IP addresses/hostnames)'. At the bottom, there is an 'Optimize for Virt Store' checkbox and 'OK' and 'Cancel' buttons.

The screenshot shows the 'Storage' tab in the WebAdmin interface. It displays a table of Gluster volumes and a detailed view of the 'Bricks' for a selected volume.

| Name | Cluster | Volume Type | Bricks | Space Use | Activities |
|-------------|---------------|-------------|---------|-----------|------------|
| ovirt-data1 | MixedCluster1 | Replicate | ▲ 2 ▼ 0 | 18% | |
| ovirt-data2 | MixedCluster1 | Replicate | ▲ 2 ▼ 0 | 18% | |

| Server | Brick Directory | Space Used | Activities |
|----------------------------|----------------------------|------------|------------|
| vm-ovirt01.vn1.bytenix.com | /srv/glusterfs/ovirt-data1 | 12% | |
| vm-ovirt02.vn1.bytenix.com | /srv/glusterfs/ovirt-data1 | 18% | |

Done!



Now just add the volume as a new storage domain, wait for data center to initialize and enjoy your new HA setup.

The next important topic is management ...

Engine's VM management

- Support for editing the Hosted Engine VM
- Memory and CPU allocation, network configuration
- Work in progress..

- Distributed to all hosted engine nodes using OVF file on the storage domain
- Hosted engine daemons pick up the configuration when the management VM is restarted

- Reporting configuration
 - State transitions
 - SMTP details
- Timeout configuration
 - Allowed downtime before forced recovery
- Host scoring constants and rules

What is missing from oVirt 3.6?



- Automated hyper-converged setup
 - Removed at the last moment because of unresolved issues
- Full support for managing the oVirt engine VM
 - Missed the feature deadline, will be available soon
- Hosted engine configuration UI
- Support for multiple Gluster brick servers not ideal
 - issue with VM startup – see qemu bug [#1247933](#)
 - but HA properly maintained during operation

THANK YOU !

<http://wiki.ovirt.org/wiki/Category:SLA>
users@ovirt.org
devel@ovirt.org

#ovirt irc.oftc.net