Enabling High Availability Service with oVirt Virtualization and CephFS

By Michael Poat & Dr. Jérôme Lauret

Outline

- Motivation
- oVirt Virtualization and CephFS
- How we plan to use oVirt
- Standard and Self-Hosted oVirt Deployments
- Live Migration and High Availability
- Use Case and Conclusion

Motivation

- The STAR online computing environment has been generating demand for highly available services (HAS) – STAR operation = Big Money
- Current deployment of critical services on bare metal creates a problem if the fundamental hardware fails or power outage occurs – manual intervention in the event of failure

If we were to move to a virtualized environment, what would we gain?

- Virtual Machines are mobile and can easily run on multiple hardware (as long as CPU Architecture is the same)
- Easily take VM back ups, clones, snapshots, and templates
- Ability to grow/shrink size of VM and add disk space as needed (Thin provisioned VM's)
- KVM Virtualization has a proven track record to be very stable
- VMs can be stored and accessed from a common storage (i.e CephFS)

STAR's CephFS Cluster

- Ceph is a distributed storage system based on RADOS (Reliable Autonomic Distributed Object Store)
- STAR has deployed a 30 node 220TB CephFS cluster with replication 3
- We have done many studies with multiple Ceph configurations and feature testing
  - M. Poat, J. Lauret – "Performance and Advanced Data Placement Techniques with Ceph's Distributed Storage System" ", *J. Phys.: Conf. Ser.* **762** 012025 doi:10.1088/1742-6596/762/1/012025 (2016). (ACAT 2016)
    - Tested Primary Affinity w/ SSD OSDs, Journals on SSDs, and Ceph's Cache Tiering methods.
  - M. Poat, J. Lauret – "Achieving Cost/Performance Balance Ratio Using Tiered Storage Caching Techniques: A Case Study with CephFS", (2016). (CHEP 2016)
    - Tested dm-cache and bcache disk caching techniques with SSDs
- CephFS has proven to be stable within our cluster and provides a common storage medium for our online infrastructure and users
- **Plan Forward:** Leverage CephFS as a main storage backend for a Virtualization cluster

oVirt

- oVirt – Open Virtualization Management Application
  - Based on RHEV (oVirt -> Opensource)
  - oVirt – “upstream” project for new features and development
  - Use of multiple storage technologies (CephFS, GlusterFS, NFS, Other POSIX Storage Systems, iSCSI, FCP, etc.) -> oVirt 4.1 supports CephFS Storage Domains

<u>Two Main Components:</u>

- **oVirt Engine**
  - Manages Hosts & VM’s
  - GUI Web Interface with sophisticated authentication & authorization
  - Ability to Monitor Resources & Manage Quotas (storage, compute, network)

- **oVirt Hypervisors**
  - Nodes that run virtual machines
  - Intuitively migrate live running VM’s from one hypervisor to another
  - Supports high availability solutions (with proper hardware)

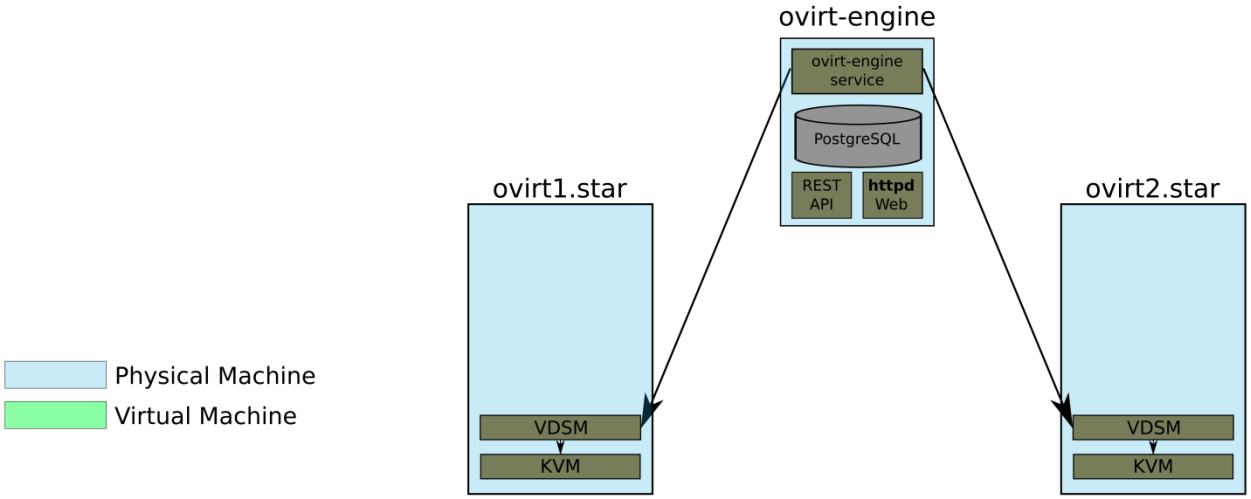
How we plan to use oVirt

- Configure a resilient oVirt cluster with no single point of failure
  - Reuse our over provisioned nodes with proper hardware requirements to deploy multiple highly available virtual machines
  - Ensure power management is implemented correctly + UPS backup power
  - Leverage our CephFS distributed storage for seamless access to the VM's
  - Self Hosted Engine Feature (see future slides)

Our first target use case:

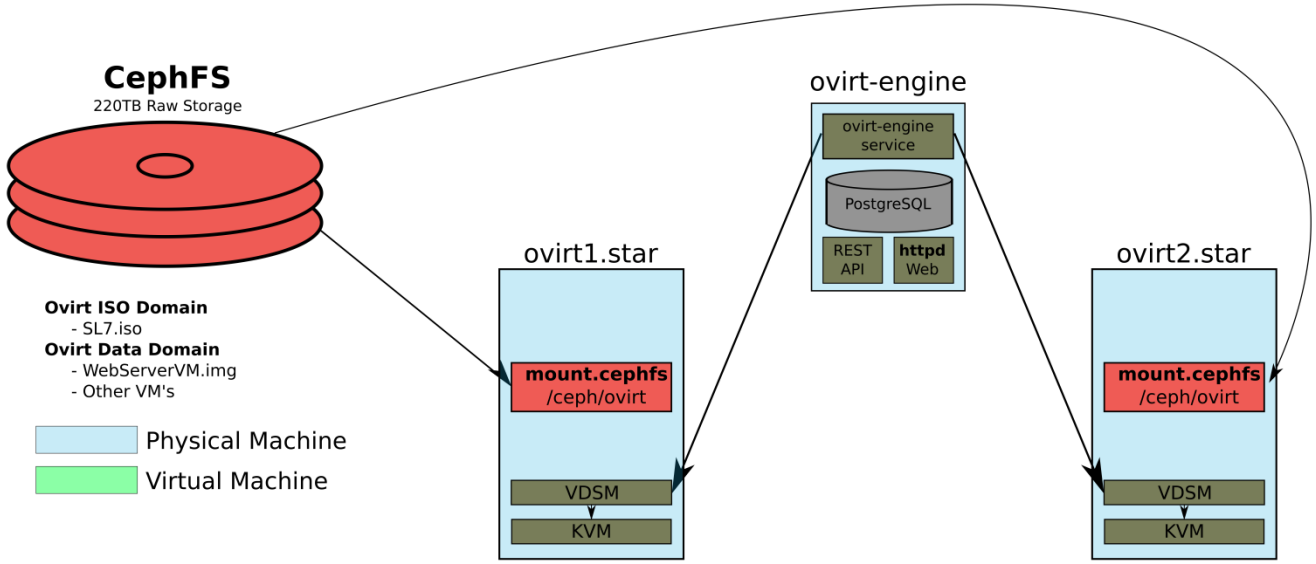
- **Webserver**
  - STAR's online webserver serves as a critical piece of infrastructure especially during RHIC runs as it serves all monitoring, status updates, and more for STAR
- **Identity provider for Single Sign On**
  - A plan for STAR has been to implement a Single Sign On service, the Identity provider is a high availability service as new logon's will be impossible if the IDP node is down
- **Database(s)**
  - Select Critical database servers may become virtualized and set to high availability

Standard oVirt Deployment



- 1 physical Node configured as the "oVirt-Engine"
  - ovirt-engine
- 2 physical nodes configured as oVirt "Hypervisors"
  - ovirt1.star
  - ovirt2.star

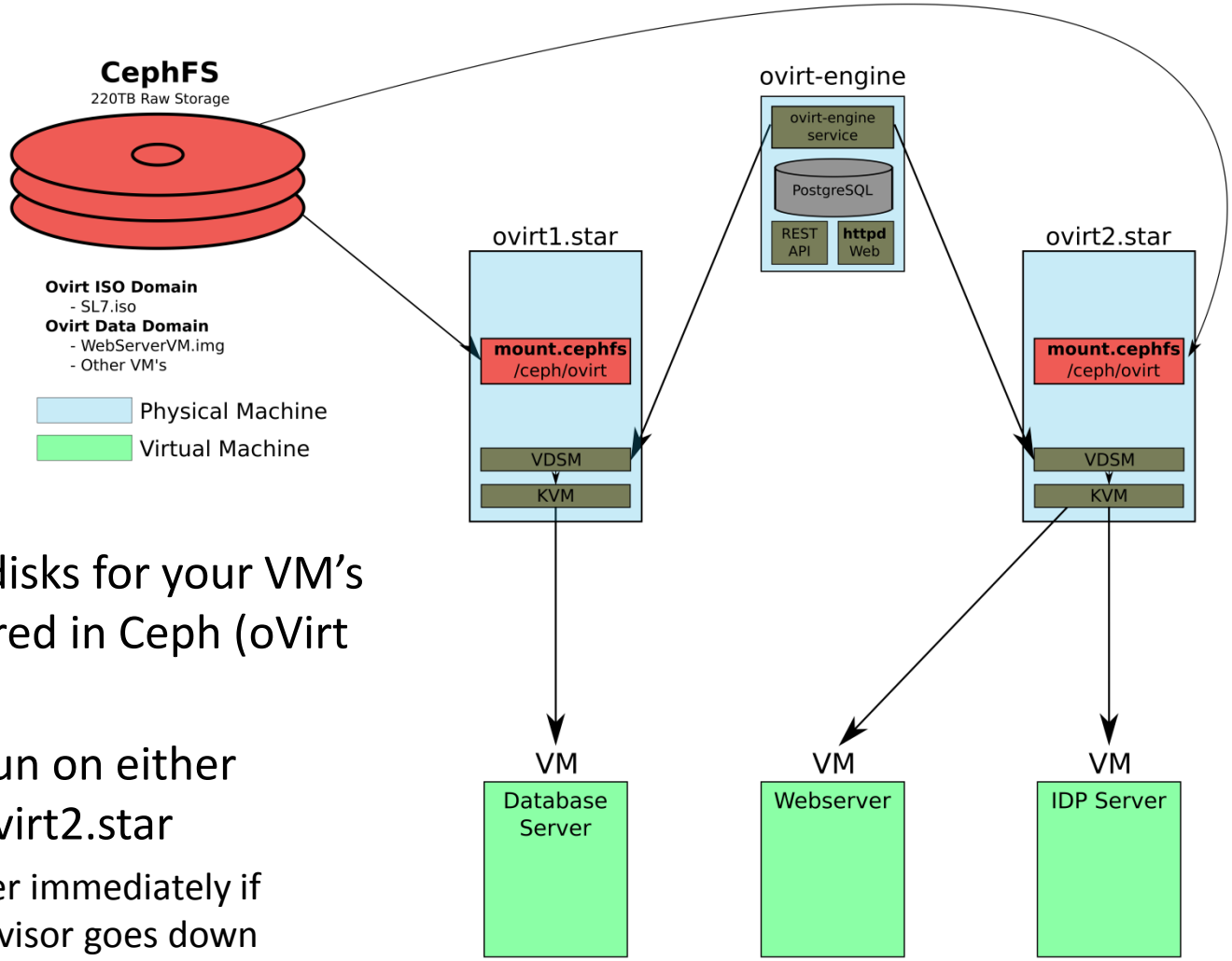
Standard oVirt Deployment



- Create directories in CephFS for oVirt Storage domains
  - ISO Domain – used for .iso image files
  - Data Domain – used to store VM's

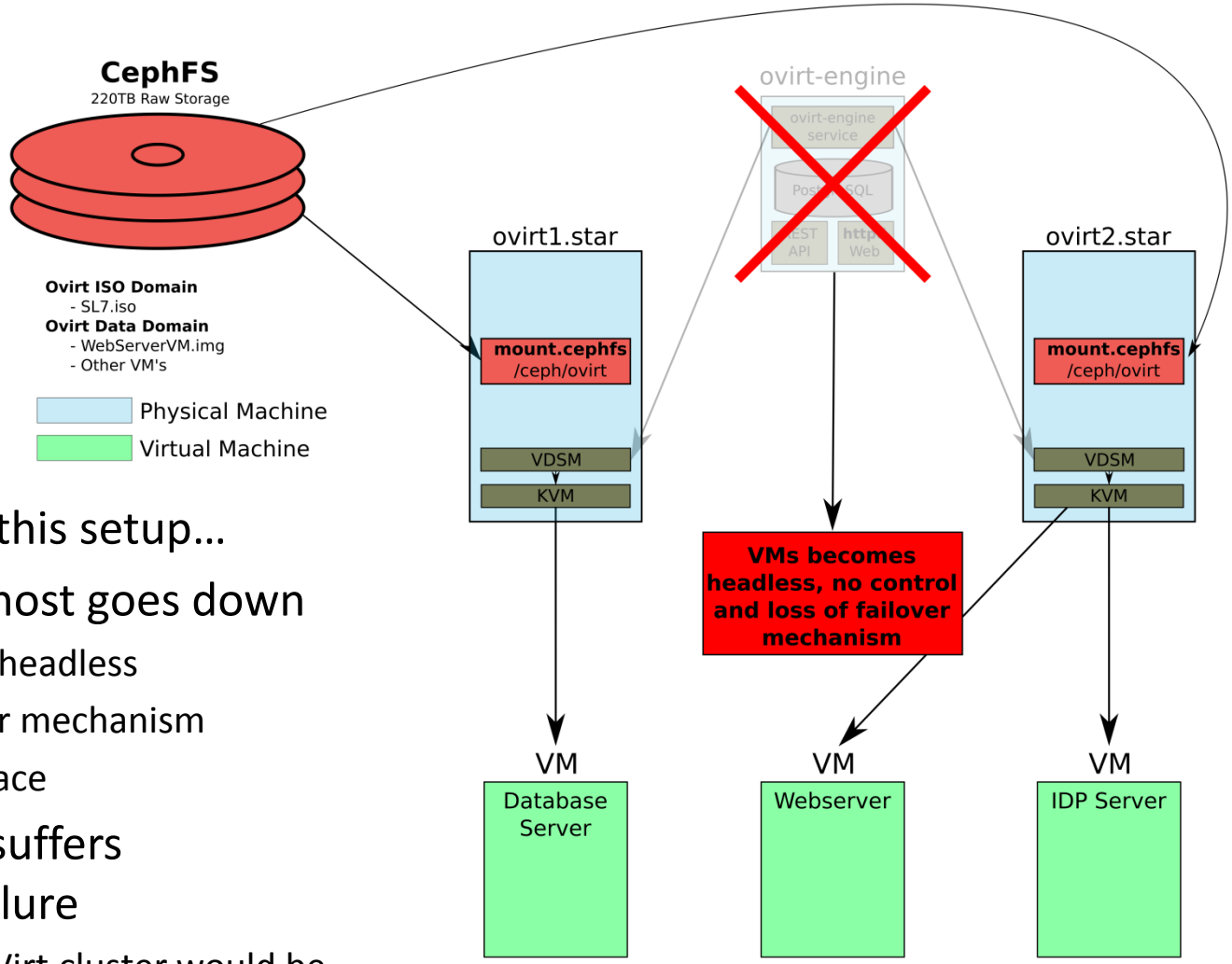
ACAT 2017
BROOKHAVEN NATIONAL LABORATORY
U.S. DEPARTMENT OF ENERGY Office of Science
STAR

Standard oVirt Deployment



- Create Virtual disks for your VM's that will be stored in Ceph (oVirt Data Domain)
- The VM's can run on either ovirt1.star or ovirt2.star
  - VM will failover immediately if running hypervisor goes down
  - VM's can be migrated live between hypervisors

Standard oVirt Deployment



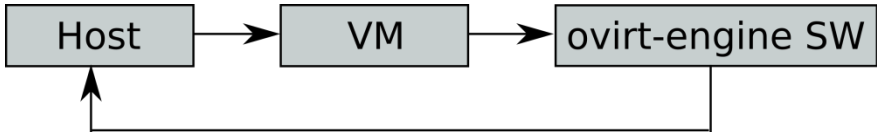
- Problems with this setup…
- If ovirt-engine host goes down
  - VM's become headless
  - Loss of failover mechanism
  - No web interface
- If ovirt-engine suffers catastrophic failure
  - Recovery of oVirt cluster would be difficult
  - Downtime of VM -> not true HA!

There is another solution…

Self-Hosted Engine

- "A self-hosted engine is a virtualized environment in which the oVirt Engine software runs on a virtual machine on the hosts managed by that engine" Say Wha?

- What is it?
  - Standard oVirt Installation
  - Running on a highly available VM
  - The VM is managed… by the engine it's hosting



What's required?

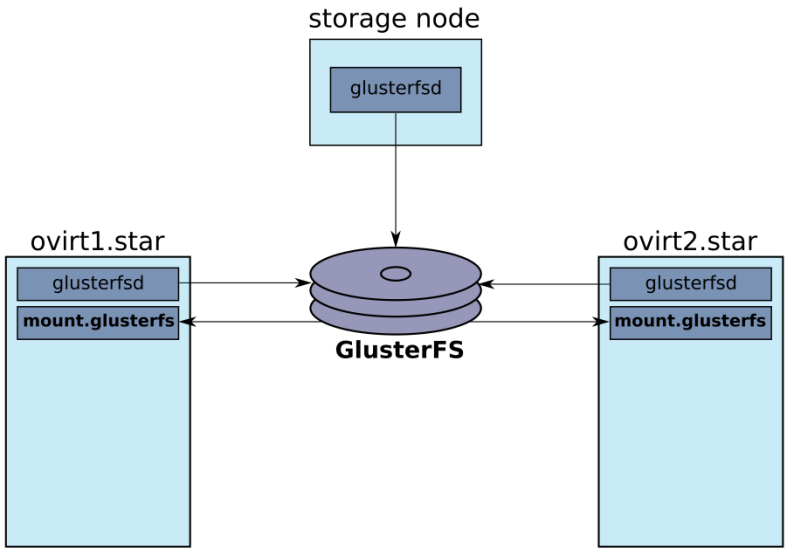
- Two (or more) physical nodes that both run software called the 'ovirt-ha-agent' & 'ovirt-ha-broker'
- The two (or more) hosts will negotiate using 'ovirt-ha-agent' determining which host should run the VM based on an internal scoring system
  - Scores based on: Memory usage, CPU load, available memory, network status, etc.
- One host will run the ovirt-engine VM as a highly available VM

Self-Hosted Engine: No Ceph Support

- Currently oVirt does not have support to store the Self-Hosted Engine VM in CephFS
  - Available options: GlusterFS, iscsi, fc, nfs3, or nfs4
- Next best thing? -> Replication 3 GlusterFS filesystem
  - Think of it as a NFS storage replicated across multiple nodes
- GlusterFS
  - Scalable network filesystem
  - Open source
  - Redundancy and POSIX Complaint
- Setup:
  - 2 oVirt hypervisor nodes + 1 spare node
  - 3 disks per node = a small 9 disk replication 3 GlusterFS setup

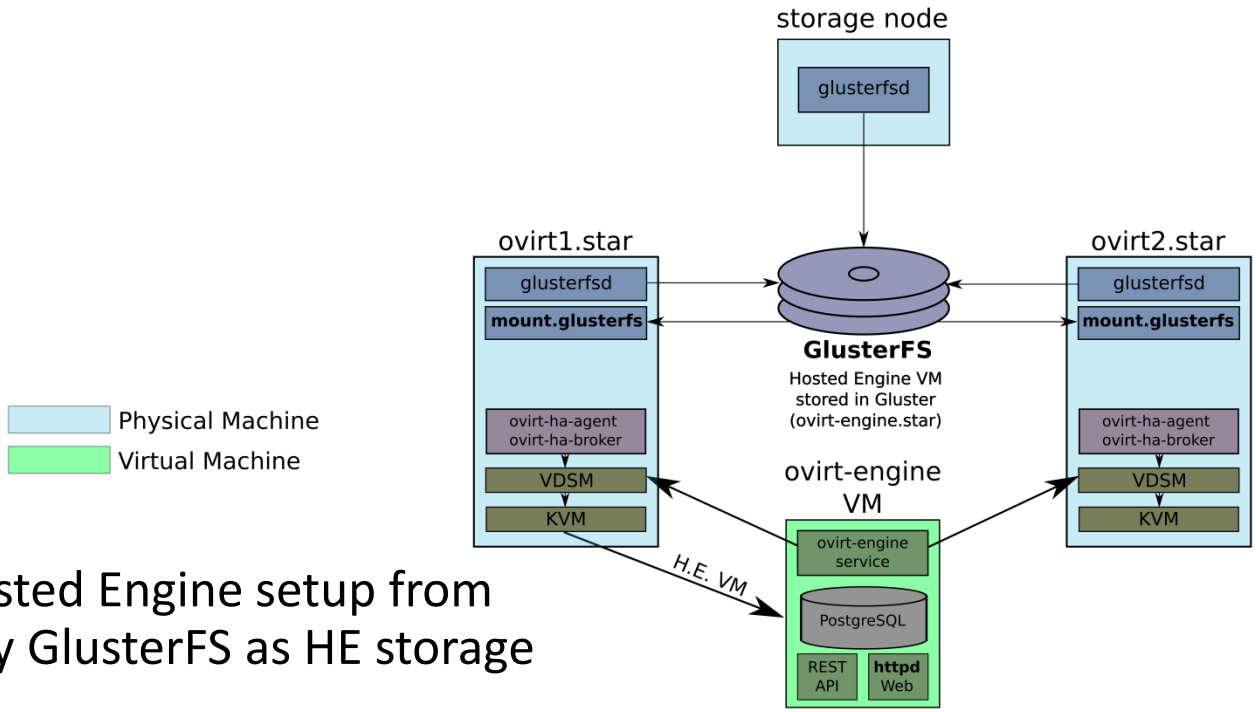
- Until oVirt supports CephFS for the Hosted Engine storage domain, GlusterFS will provide the redundancy and high availability we seek for the time being

Self-Hosted Engine oVirt Deployment



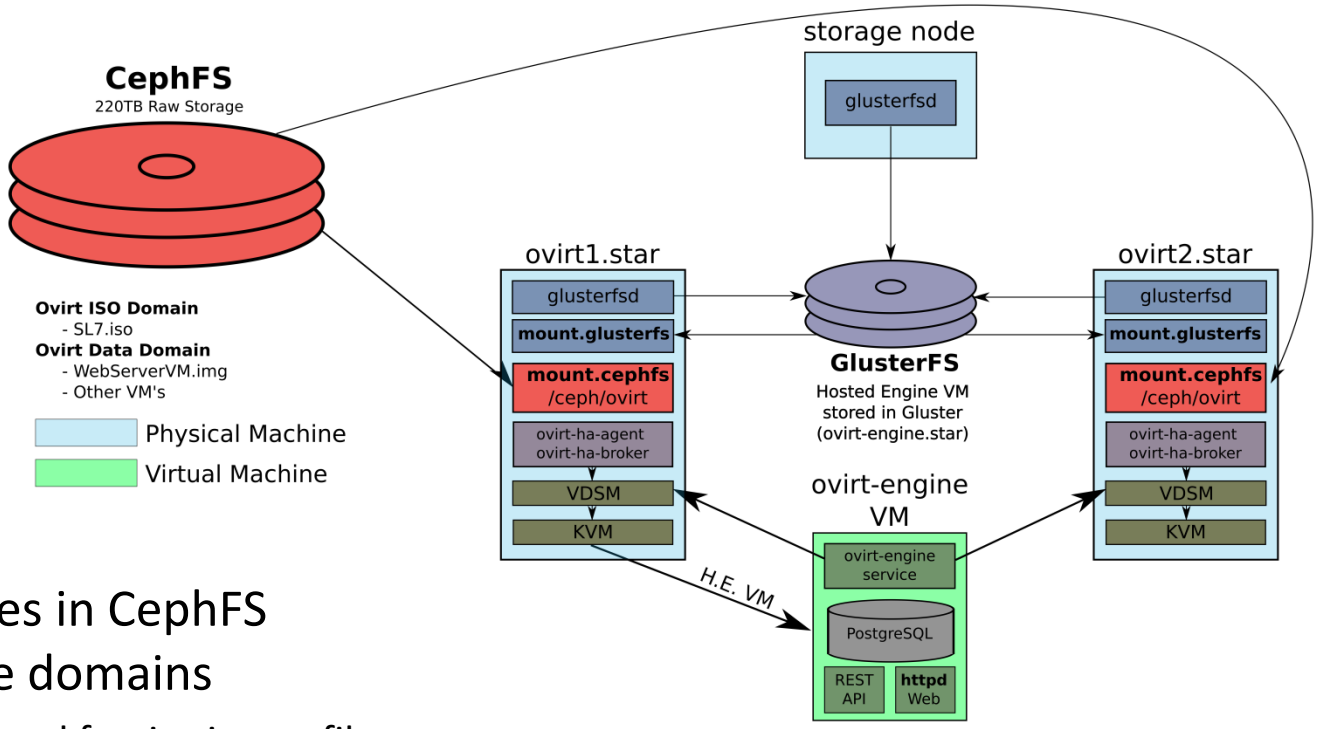
- Initial setup of GlusterFS
- Hosts: ovirt1, ovirt2 & an extra storage node compose a replication **3** GlusterFS filesystem
- ovirt1 & ovirt2 are also clients to the GlusterFS storage **mount.glusterfs**

Self-Hosted Engine oVirt Deployment



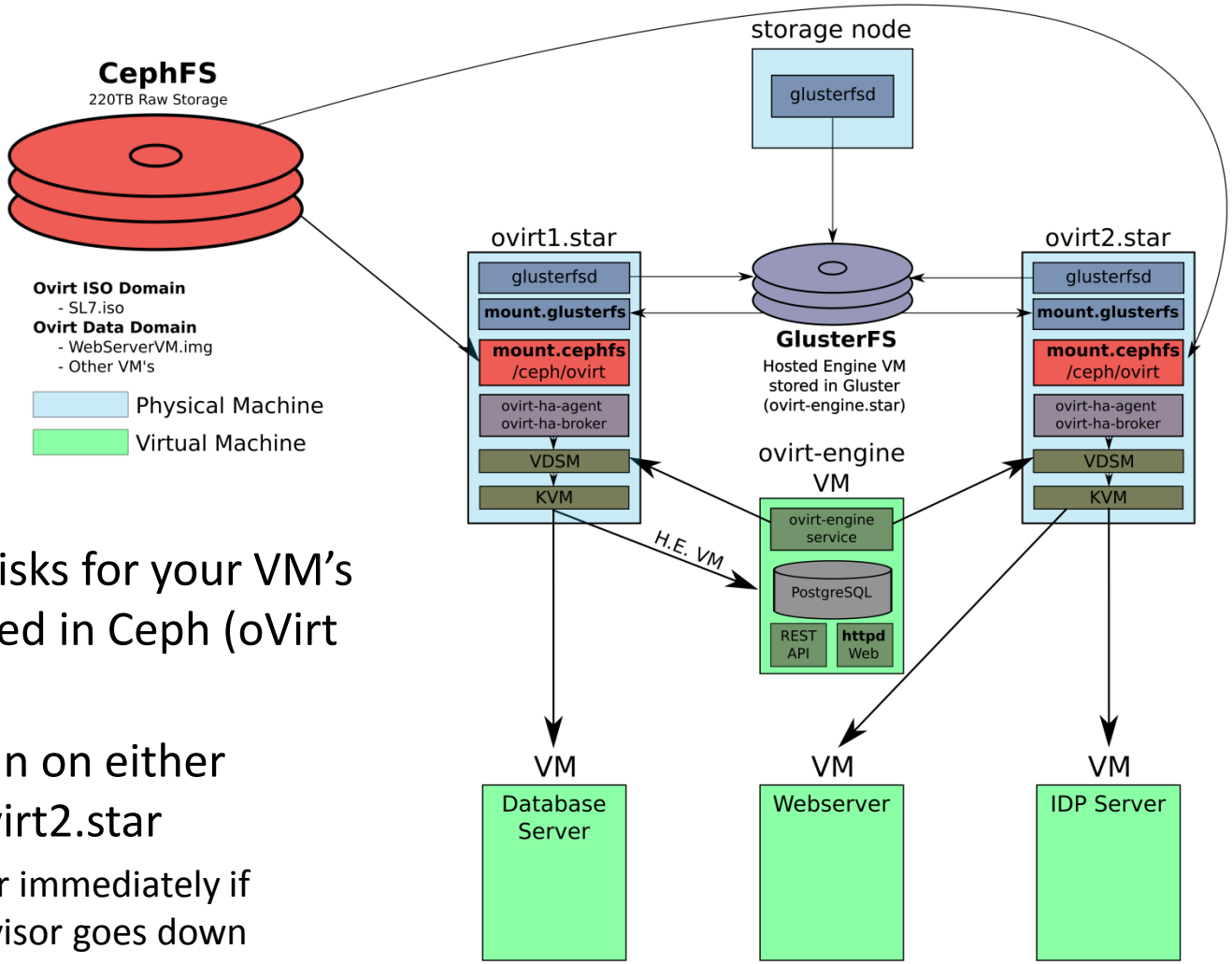
- Run the oVirt Hosted Engine setup from ovirt1 and specify GlusterFS as HE storage domain
- Create the self-hosted ovirt-engine VM
- Deploy oVirt host installation via ovirt-engine web interface to remaining hosts (ovirt2)
- The ovirt-engine service installed on the VM will then manage VDSM on ovirt1 & ovirt2

Self-Hosted Engine oVirt Deployment



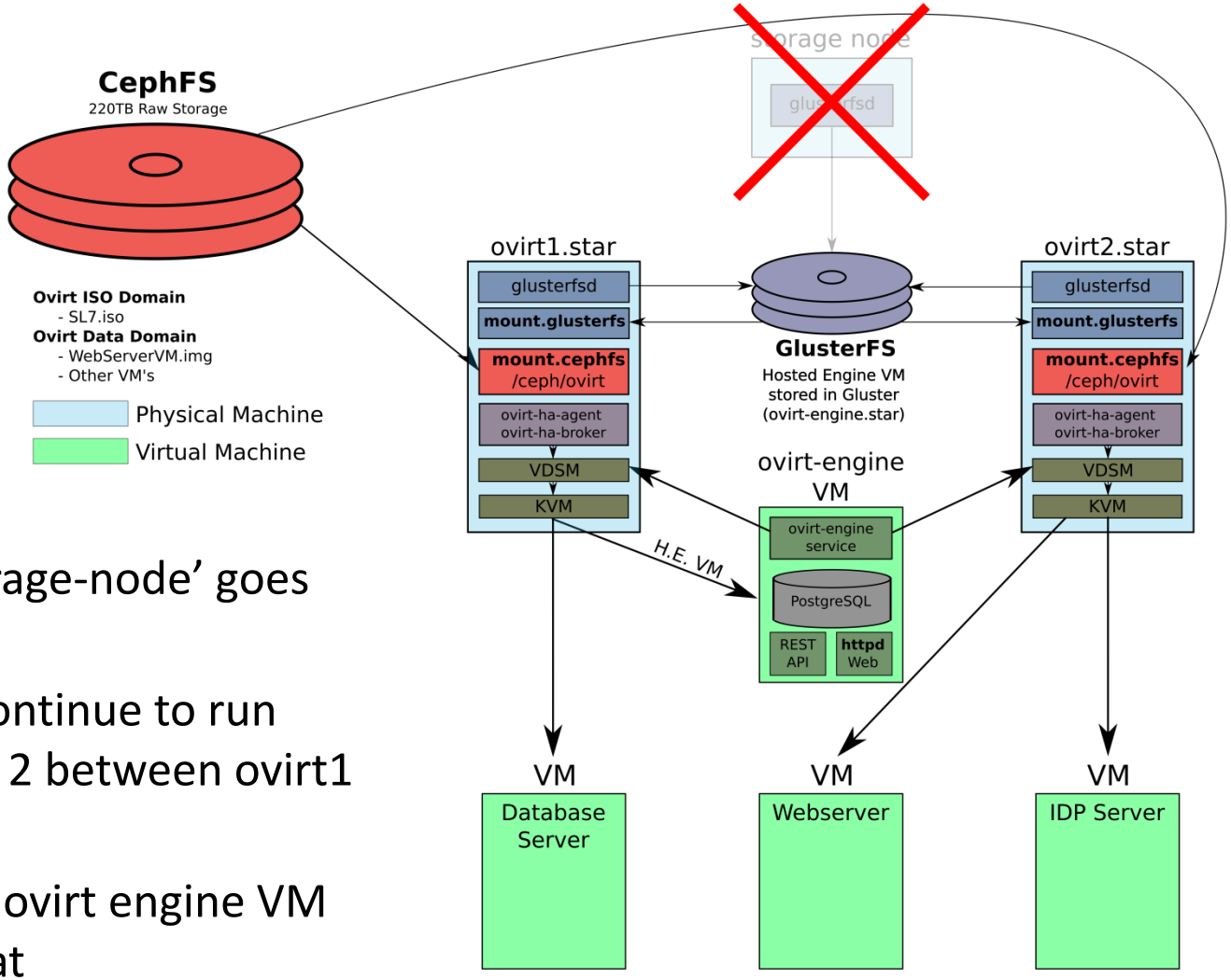
- Create directories in CephFS for oVirt Storage domains
  - ISO Domain – used for .iso image files
  - Data Domain – used to store VM's

Self-Hosted Engine oVirt Deployment



- Create Virtual disks for your VM's that will be stored in Ceph (oVirt Data Domain)
- The VM's can run on either ovirt1.star or ovirt2.star
  - VM will failover immediately if running hypervisor goes down
  - VM's can be migrated live between hypervisors

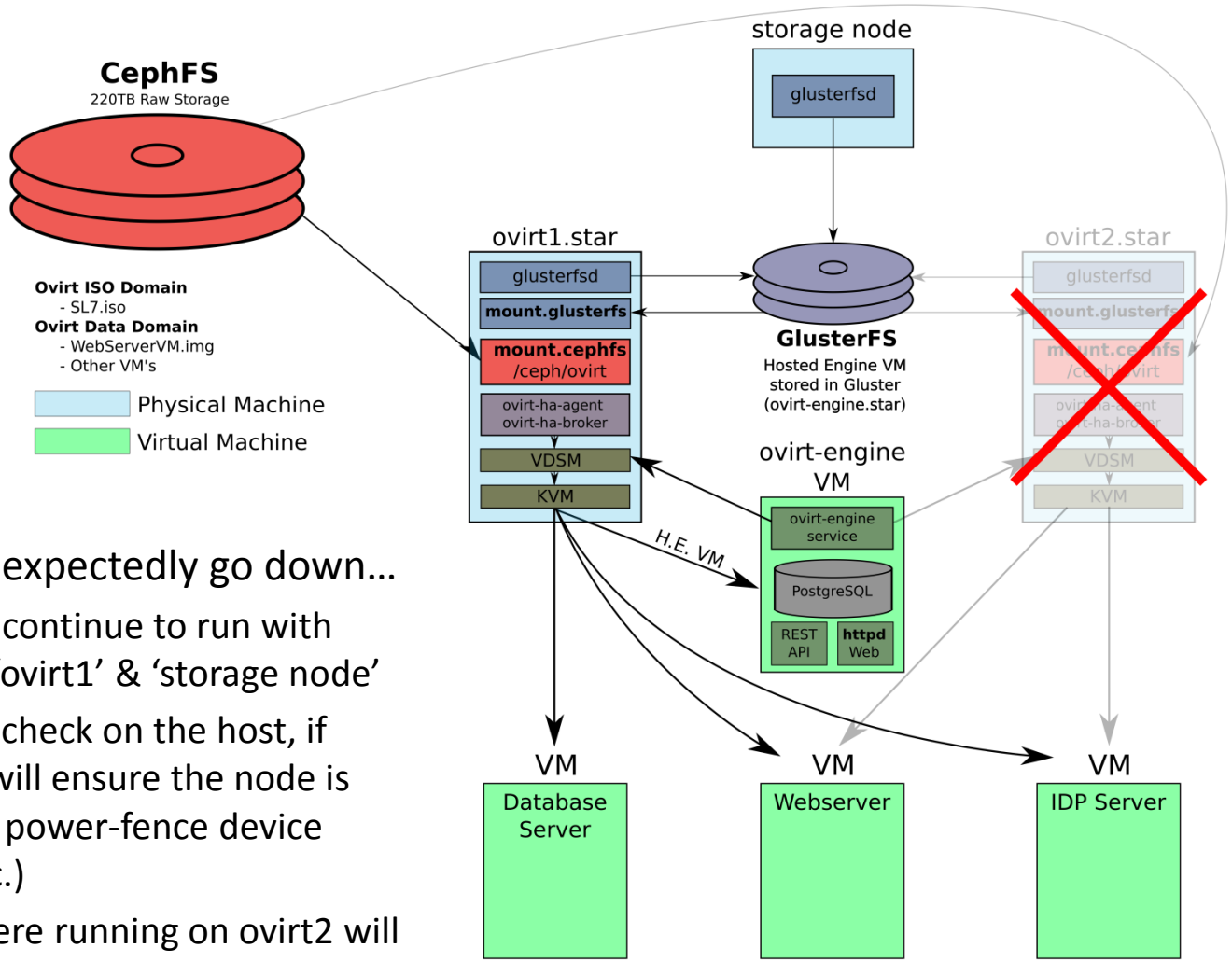
Self-Hosted Engine oVirt Deployment



- If the extra 'storage-node' goes down…
- GlusterFS will continue to run with replication 2 between ovirt1 & ovirt2
- The self-hosted ovirt engine VM won't skip a beat

ACAT 2017
BROOKHAVEN NATIONAL LABORATORY
U.S. DEPARTMENT OF ENERGY
Office of Science
STAR

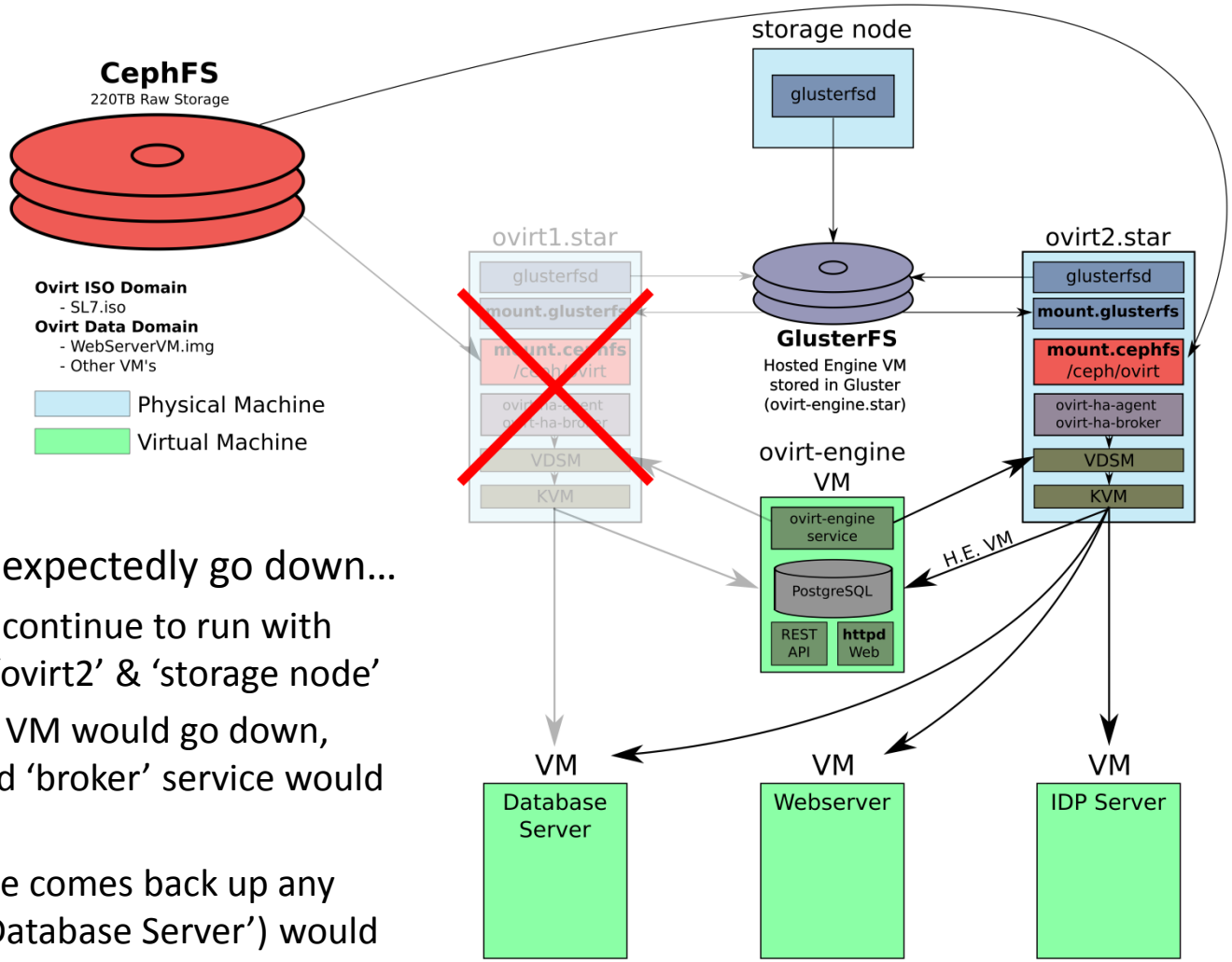
Self-Hosted Engine oVirt Deployment



- If ovirt2 were to unexpectedly go down…
  - GlusterFS would continue to run with replication 2 on 'ovirt1' & 'storage node'
  - ovirt-engine will check on the host, if unresponsive it will ensure the node is off/rebooted via power-fence device (IPMI, iDRAC, etc.)
  - The VM's that were running on ovirt2 will failover (reboot) on ovirt1
  - Downtime 3-5 minutes

ACAT 2017
BROOKHAVEN NATIONAL LABORATORY
U.S. DEPARTMENT OF ENERGY
Office of Science
STAR

Self-Hosted Engine oVirt Deployment



- If ovirt1 were to unexpectedly go down…
  - GlusterFS would continue to run with replication 2 on ‘ovirt2’ & ‘storage node’
  - The ovirt-engine VM would go down, ovirt2 ‘agent’ and ‘broker’ service would reboot the VM
  - Once ovirt-engine comes back up any VM’s affected (‘Database Server’) would reboot on either ovirt1 or ovirt2

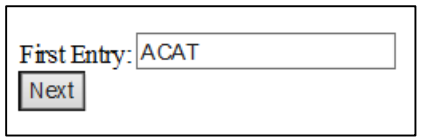
Live Migration and Test Case

Live Migration:

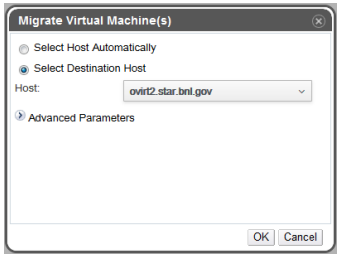
- Live Migration is a feature of the KVM hypervisor allowing you transfer a live VM from one host to another with practically 0 down time
- The VM remains powered on and user applications continue to run during relocation
- The VM's RAM is copied from the source host to destination host
- Requirements:
  - VM must be accessible from both source and destination hosts (accessed from CephFS)
  - Must be available resources on destination host (Memory, CPU, etc.)
- Nearly no guest down time
  - Would be undetected by users -> SSH connections are fine

Test Case: Create a multi step web page and test live migration while filling out page(s)

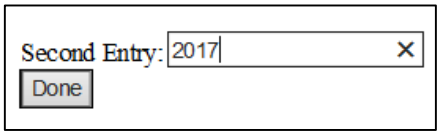
Step 1: Fill out page 1 of PHP web page then click the Next button



Step 2: Before filling out next page we migrate the VM



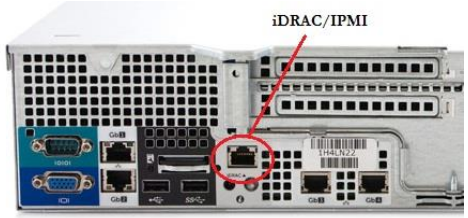
Step 3: Once the VM is migrated we fill next page and click Done



Result: VM RAM is properly transferred during live migration -

Your entry on page 1: **ACAT**
Your entry on page 2: **2017**

High Availability



High Availability:

- A feature in oVirt to ensure your VM's are always up
- Requirements:
  - Enable Power Management with use of power fencing device (iDRAC, IPMI, etc.)
  - Must be available resources on destination host (Memory, CPU, etc.)
  - More than one hypervisor host with similar CPU architecture

Test Case:

- 5 Second power loss on a host running a HA VM…

→

VM: reserve1.star.bnl.gov

Hosts: ovirt1.star.bnl.gov, ovirt2.star.bnl.gov

- oVirt-engine immediately recognizes down VM and tries to restart on another host

Aug 4, 2017 11:54:14 AM		Trying to restart VM reserve1.star.bnl.gov on Host ovirt1.star.bnl.gov
Aug 4, 2017 11:54:13 AM		Highly Available VM reserve1.star.bnl.gov failed. It will be restarted automatically.
Aug 4, 2017 11:54:13 AM		VM reserve1.star.bnl.gov is down with error. Exit message: VM has been terminated on the host.

- oVirt-engine checks status of downed host + uses IPMI interface for status check. If status returns machine is alive but 'down' -- oVirt-engine restarts node via IPMI

Aug 4, 2017 11:55:47 AM		Executing power management status on Host ovirt2.star.bnl.gov using Proxy Host ovirt1.star.bnl.gov and Fence Agent ipmilan:ovirt2-idrac.star.bnl.gov.

- Meanwhile our virtual machine comes back up in <1 minute on secondary host

reserve1.star.bnl.gov

- After ~3-5 minutes the failed host is back up and verified in oVirt

Aug 4, 2017 12:01:39 PM		Host ovirt2.star.bnl.gov power management was verified successfully.
Aug 4, 2017 12:01:39 PM		Status of host ovirt2.star.bnl.gov was set to Up.

Conclusion

- oVirt virtualization will enable STAR to create a true highly available system with no single point of failure to deploy critical services
- While HA does require resources on stand by (idle) ⇔ x2 CPU & Memory – it is a fair tradeoff to ensure our critical nodes and services are always up
- In oVirt we can easily add/remove hypervisor nodes – with live migration we can take hypervisor nodes out of service for maintenance with 0 interruption
- Good use of CephFS storage for VMs
  - Currently we massively use CephFS to store content (web content, user content, etc.)
  - Thin provisioned VM's create small storage footprint (keep VM small and access content from Ceph)
- The Self-Hosted Engine feature is a true HA setup removing single points of failure
  - Tradeoff: As of oVirt 4.1 Self-Hosted Engine cannot be stored in CephFS – Requires GlusterFS or alternative.
- Virtualization will allow STAR to better use over provisioned physical nodes to deploy a diverse set of virtual machines to meet our operational needs