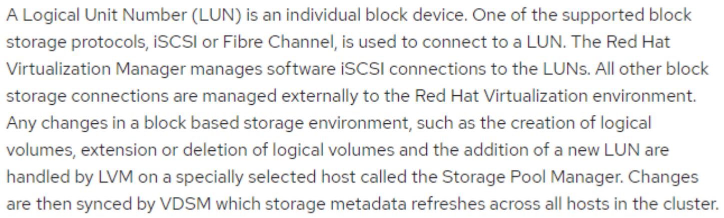
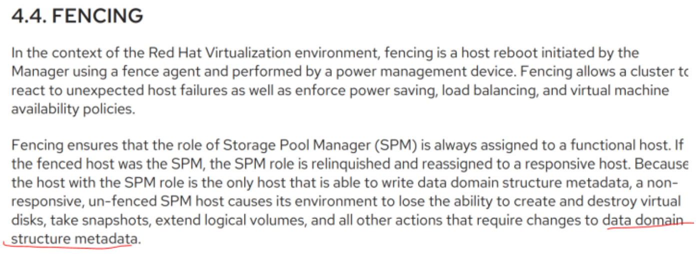
Operations LUN is managed and handled by SPM



A Logical Unit Number (LUN) is an individual block device. One of the supported block storage protocols, iSCSI or Fibre Channel, is used to connect to a LUN. The Red Hat Virtualization Manager manages software iSCSI connections to the LUNs. All other block storage connections are managed externally to the Red Hat Virtualization environment. Any changes in a block based storage environment, such as the creation of logical volumes, extension or deletion of logical volumes and the addition of a new LUN are handled by LVM on a specially selected host called the Storage Pool Manager. Changes are then synced by VDSM which storage metadata refreshes across all hosts in the cluster.

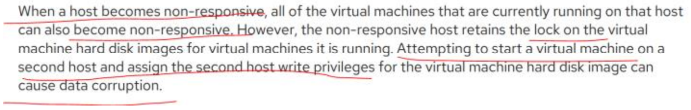
Fencing



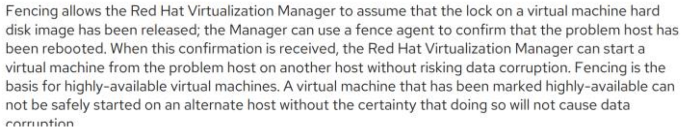
4.4. FENCING

In the context of the Red Hat Virtualization environment, fencing is a host reboot initiated by the Manager using a fence agent and performed by a power management device. Fencing allows a cluster tc react to unexpected host failures as well as enforce power saving, load balancing, and virtual machine availability policies.

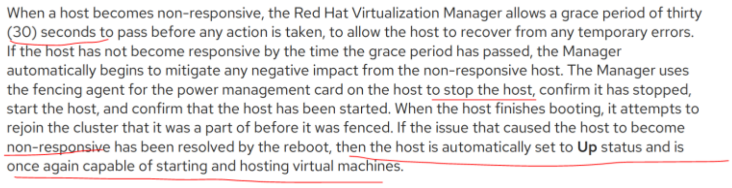
Fencing ensures that the role of Storage Pool Manager (SPM) is always assigned to a functional host. If the fenced host was the SPM, the SPM role is relinquished and reassigned to a responsive host. Because the host with the SPM role is the only host that is able to write data domain structure metadata, a non-responsive, un-fenced SPM host causes its environment to lose the ability to create and destroy virtual disks, take snapshots, extend logical volumes, and all other actions that require changes to data domain structure metadata.



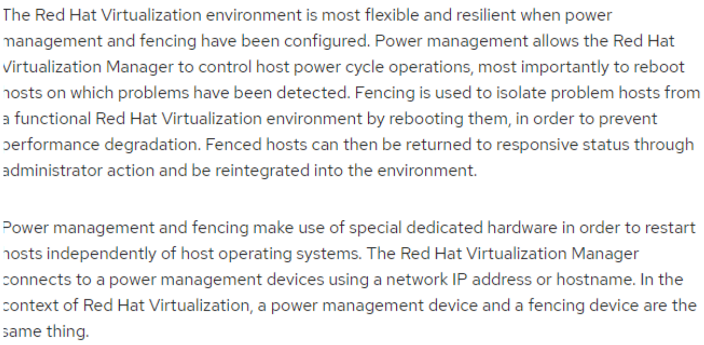
When a host becomes non-responsive, all of the virtual machines that are currently running on that host can also become non-responsive. However, the non-responsive host retains the lock on the virtual machine hard disk images for virtual machines it is running. Attempting to start a virtual machine on a second host and assign the second host write privileges for the virtual machine hard disk image can cause data corruption.



Fencing allows the Red Hat Virtualization Manager to assume that the lock on a virtual machine hard disk image has been released; the Manager can use a fence agent to confirm that the problem host has been rebooted. When this confirmation is received, the Red Hat Virtualization Manager can start a virtual machine from the problem host on another host without risking data corruption. Fencing is the basis for highly-available virtual machines. A virtual machine that has been marked highly-available can not be safely started on an alternate host without the certainty that doing so will not cause data corruption



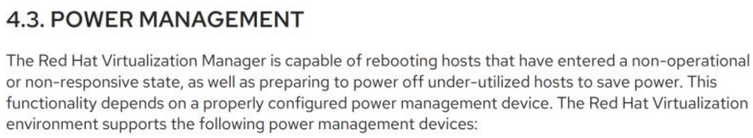
When a host becomes non-responsive, the Red Hat Virtualization Manager allows a grace period of thirty (30) seconds to pass before any action is taken, to allow the host to recover from any temporary errors. If the host has not become responsive by the time the grace period has passed, the Manager automatically begins to mitigate any negative impact from the non-responsive host. The Manager uses the fencing agent for the power management card on the host to stop the host, confirm it has stopped, start the host, and confirm that the host has been started. When the host finishes booting, it attempts to rejoin the cluster that it was a part of before it was fenced. If the issue that caused the host to become non-responsive has been resolved by the reboot, then the host is automatically set to **Up** status and is once again capable of starting and hosting virtual machines.



The Red Hat Virtualization environment is most flexible and resilient when power management and fencing have been configured. Power management allows the Red Hat Virtualization Manager to control host power cycle operations, most importantly to reboot hosts on which problems have been detected. Fencing is used to isolate problem hosts from a functional Red Hat Virtualization environment by rebooting them, in order to prevent performance degradation. Fenced hosts can then be returned to responsive status through administrator action and be reintegrated into the environment.

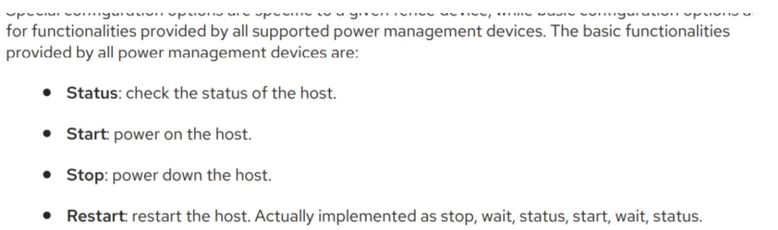
Power management and fencing make use of special dedicated hardware in order to restart hosts independently of host operating systems. The Red Hat Virtualization Manager connects to a power management devices using a network IP address or hostname. In the context of Red Hat Virtualization, a power management device and a fencing device are the same thing.

Power management



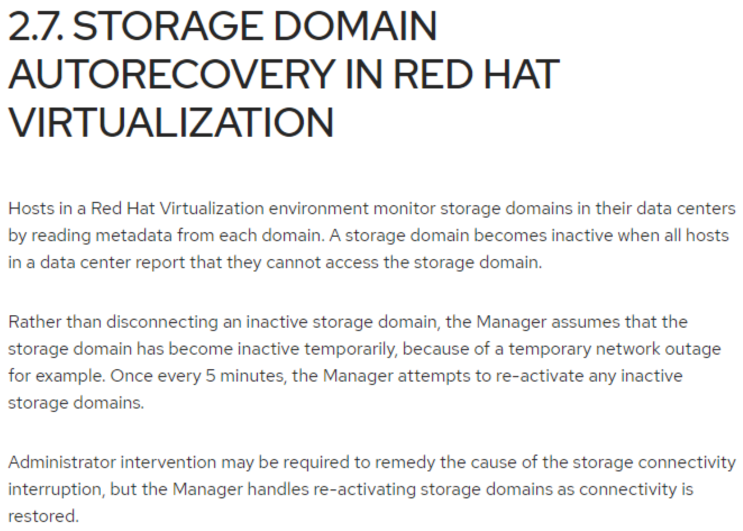
4.3. POWER MANAGEMENT

The Red Hat Virtualization Manager is capable of rebooting hosts that have entered a non-operational or non-responsive state, as well as preparing to power off under-utilized hosts to save power. This functionality depends on a properly configured power management device. The Red Hat Virtualization environment supports the following power management devices:


Special configuration options are specific to a given fence device, while basic configuration options a for functionalities provided by all supported power management devices. The basic functionalities provided by all power management devices are:

- **Status**: check the status of the host.
- **Start**: power on the host.
- **Stop**: power down the host.
- **Restart**: restart the host. Actually implemented as stop, wait, status, start, wait, status.

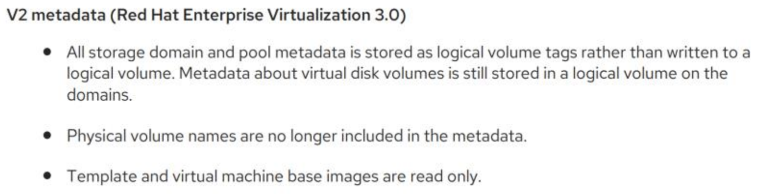
Storage Domain Metadata


2.7. STORAGE DOMAIN AUTORECOVERY IN RED HAT VIRTUALIZATION

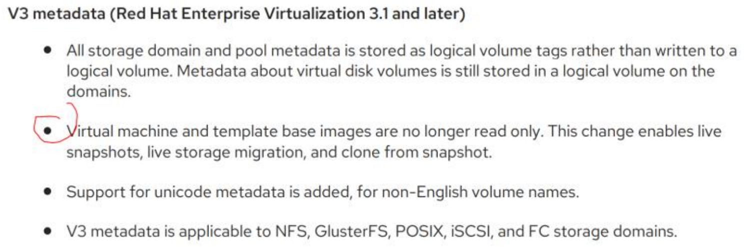
Hosts in a Red Hat Virtualization environment monitor storage domains in their data centers by reading metadata from each domain. A storage domain becomes inactive when all hosts in a data center report that they cannot access the storage domain.

Rather than disconnecting an inactive storage domain, the Manager assumes that the storage domain has become inactive temporarily, because of a temporary network outage for example. Once every 5 minutes, the Manager attempts to re-activate any inactive storage domains.

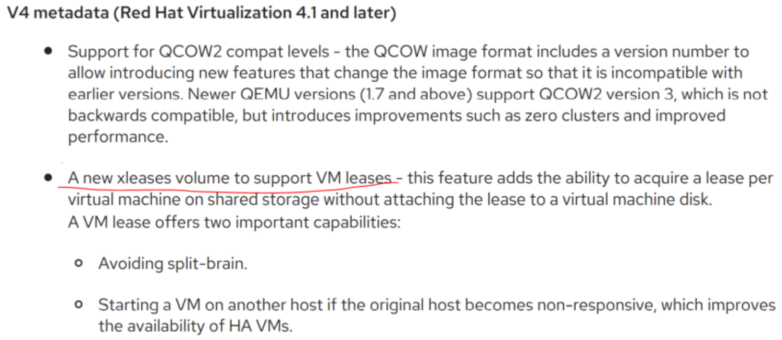
Administrator intervention may be required to remedy the cause of the storage connectivity interruption, but the Manager handles re-activating storage domains as connectivity is restored.


V2 metadata (Red Hat Enterprise Virtualization 3.0)

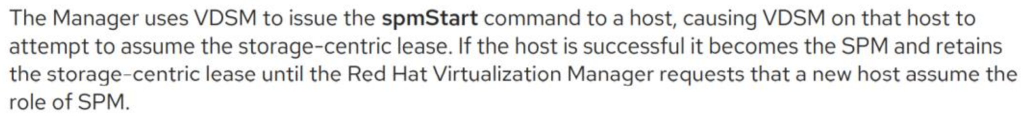
- All storage domain and pool metadata is stored as logical volume tags rather than written to a logical volume. Metadata about virtual disk volumes is still stored in a logical volume on the domains.
- Physical volume names are no longer included in the metadata.
- Template and virtual machine base images are read only.


V3 metadata (Red Hat Enterprise Virtualization 3.1 and later)

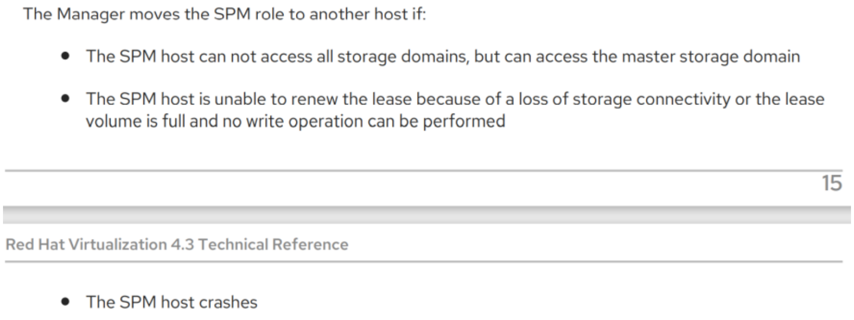
- All storage domain and pool metadata is stored as logical volume tags rather than written to a logical volume. Metadata about virtual disk volumes is still stored in a logical volume on the domains.
- Virtual machine and template base images are no longer read only. This change enables live snapshots, live storage migration, and clone from snapshot.
- Support for unicode metadata is added, for non-English volume names.
- V3 metadata is applicable to NFS, GlusterFS, POSIX, iSCSI, and FC storage domains.


V4 metadata (Red Hat Virtualization 4.1 and later)

- Support for QCOW2 compat levels - the QCOW image format includes a version number to allow introducing new features that change the image format so that it is incompatible with earlier versions. Newer QEMU versions (1.7 and above) support QCOW2 version 3, which is not backwards compatible, but introduces improvements such as zero clusters and improved performance.
- A new xleases volume to support VM leases - this feature adds the ability to acquire a lease per virtual machine on shared storage without attaching the lease to a virtual machine disk.
  A VM lease offers two important capabilities:
  - Avoiding split-brain.
  - Starting a VM on another host if the original host becomes non-responsive, which improves the availability of HA VMs.



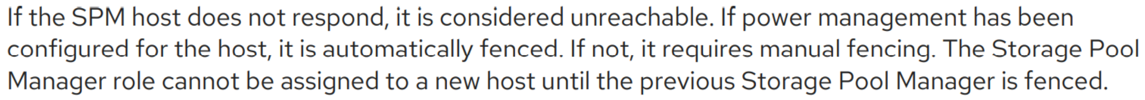
The Manager uses VDSM to issue the **spmStart** command to a host, causing VDSM on that host to attempt to assume the storage-centric lease. If the host is successful it becomes the SPM and retains the storage-centric lease until the Red Hat Virtualization Manager requests that a new host assume the role of SPM.



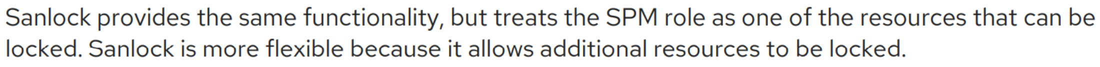
The Manager moves the SPM role to another host if:

- The SPM host can not access all storage domains, but can access the master storage domain
- The SPM host is unable to renew the lease because of a loss of storage connectivity or the lease volume is full and no write operation can be performed

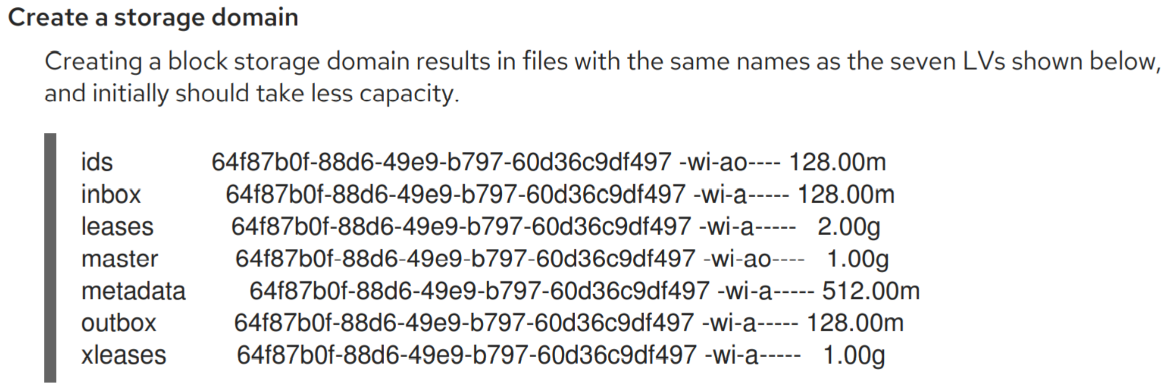
- The SPM host crashes



If the SPM host does not respond, it is considered unreachable. If power management has been configured for the host, it is automatically fenced. If not, it requires manual fencing. The Storage Pool Manager role cannot be assigned to a new host until the previous Storage Pool Manager is fenced.



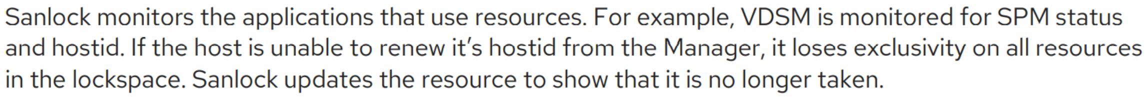
Sanlock provides the same functionality, but treats the SPM role as one of the resources that can be locked. Sanlock is more flexible because it allows additional resources to be locked.



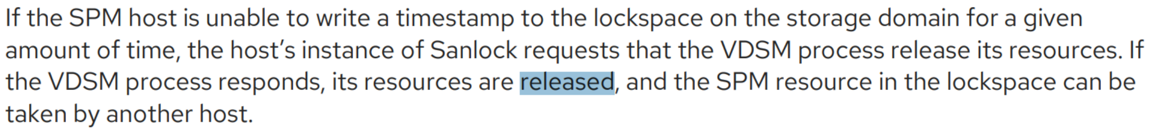
Create a storage domain

Creating a block storage domain results in files with the same names as the seven LVs shown below, and initially should take less capacity.

```
ids          64f87b0f-88d6-49e9-b797-60d36c9df497 -wi-ao---- 128.00m
inbox        64f87b0f-88d6-49e9-b797-60d36c9df497 -wi-a----- 128.00m
leases       64f87b0f-88d6-49e9-b797-60d36c9df497 -wi-a-----   2.00g
master       64f87b0f-88d6-49e9-b797-60d36c9df497 -wi-ao----   1.00g
metadata     64f87b0f-88d6-49e9-b797-60d36c9df497 -wi-a----- 512.00m
outbox       64f87b0f-88d6-49e9-b797-60d36c9df497 -wi-a----- 128.00m
xleases      64f87b0f-88d6-49e9-b797-60d36c9df497 -wi-a-----   1.00g
```



Sanlock monitors the applications that use resources. For example, VDSM is monitored for SPM status and hostid. If the host is unable to renew it's hostid from the Manager, it loses exclusivity on all resources in the lockspace. Sanlock updates the resource to show that it is no longer taken.



If the SPM host is unable to write a timestamp to the lockspace on the storage domain for a given amount of time, the host's instance of Sanlock requests that the VDSM process release its resources. If the VDSM process responds, its resources are released, and the SPM resource in the lockspace can be taken by another host.

`xlease - manage external leases`

vdsm will create a new special volume for external leases, named "xleases". This volume will be used for external leases such as VM leases. Vdsm uses this volume to create sanlock resources, and maintain the mapping from lease id to lease offset on the volume.

When creating a new VM, a user will be able to add a lease on one of the storage domains. During VM creation, engine will ask the SPM to create a lease for the VM. Vdsm will create a new sanlock resource on the selected storage domain. Vdsm will also update the mapping from the lease id to the lease offset in the xleases volume.

Xlease : lease-id → lease offset

- “FREE”: lease is not acquired by anyone
- “EXCLUSIVE”: lease is acquired by one owner
- “SHARED”: lease is acquired my multiple owners (not supported yet)

Lease.create(lease)

Starts a SPM task creating a lease on the xleases volume in the lease storage domain. Can be used only on the SPM.

Creates a sanlock resource on the domain xleases volume, and mapping from lease_id to the resource offset in the volume.

Arguments:

- lease (Lease): the lease to create

How it works

When creating or upgrading a storage domain to version 4, vdsm will create a new special volume for external leases, named “xleases”. This volume will be used for external leases such as VM leases. Vdsm uses this volume to create sanlock resources, and maintain the mapping from lease id to lease offset on the volume.