Ovirt Storage Overview

Nov 2nd, 2011

Ayal Baron

Agenda

- Defining the problem
- Storage Domain
- Storage Pool
- Example
- Domain Classes
- Domain Types
- File Domains
- Block Domains
- SPM
- Thin Provisioning
- Roadmap
- How to contribute
- Q&A

Defining the problem

- The requirements
  - Manage tens of thousands of virtual disk images
  - Each image potentially accessible from hundreds (thousands?) of nodes
  - Support both file based storage as well as raw block devices
- The assumptions
  - Exclusive access when used
  - Big (X GBs)
  - Centralized manager (can be HA in and by itself)

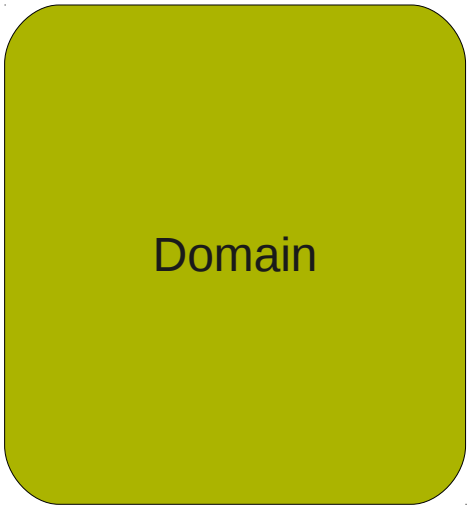
Defining the problem - guidelines

- High level, image centered API
- Cluster safe
- High performance (not in data path)
- Highly available
  - no single point of failure
  - Continues working in the absence of the manager
- Backing storage agnostic

Storage Domain

- A standalone storage entity
- Stores the images and associated metadata (but not VMs)

  Only true persistent storage for VDSM



Domain Classes

- Data
  - Master
- ISO (NFS only)
- Backup (NFS only)
- Domain classes are planned for deprecation

Domain Types

oVirt

- File Domains (NFS, local dir)
  - Use file system features for segmentation
  - Use file system for synchronizing access
- Block Domains (iSCSI, FCP, FCoE, SAS, ...)
  - Use LVM for segmentation
  - Use reserved space to ensure writes do not overlap
  - Very specialized use of LVM

File Domains

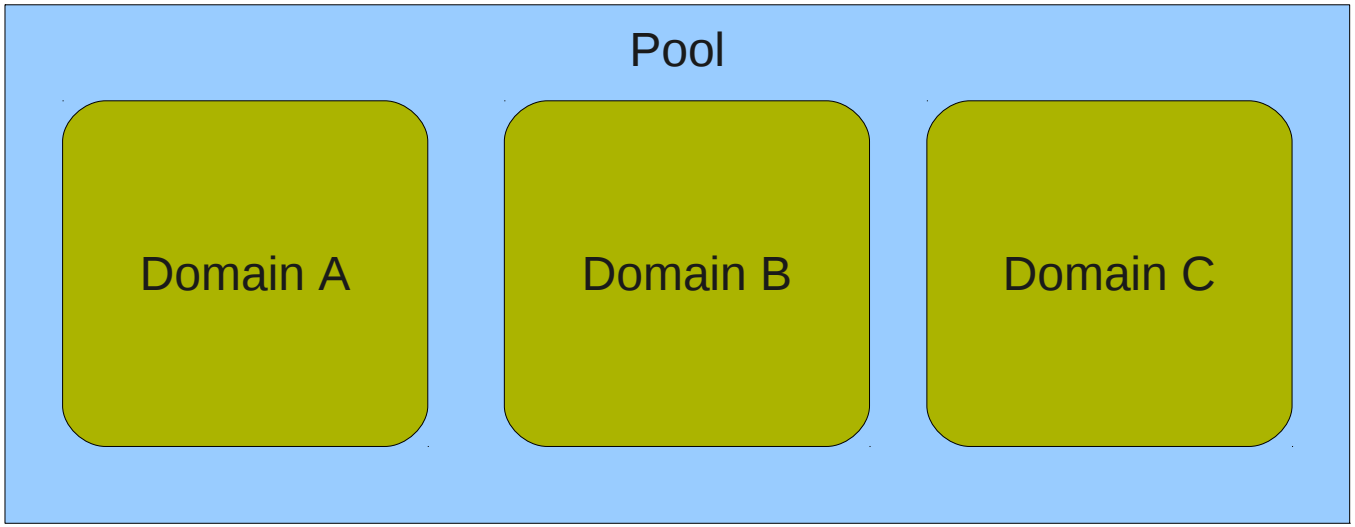
- Sparse files
- Better image manipulation capabilities
- Volumes and metadata are files
- 1:1 Mapping between domain and dir / NFS export
- NFS - Different error handling logic for data path and control path

Block Domains

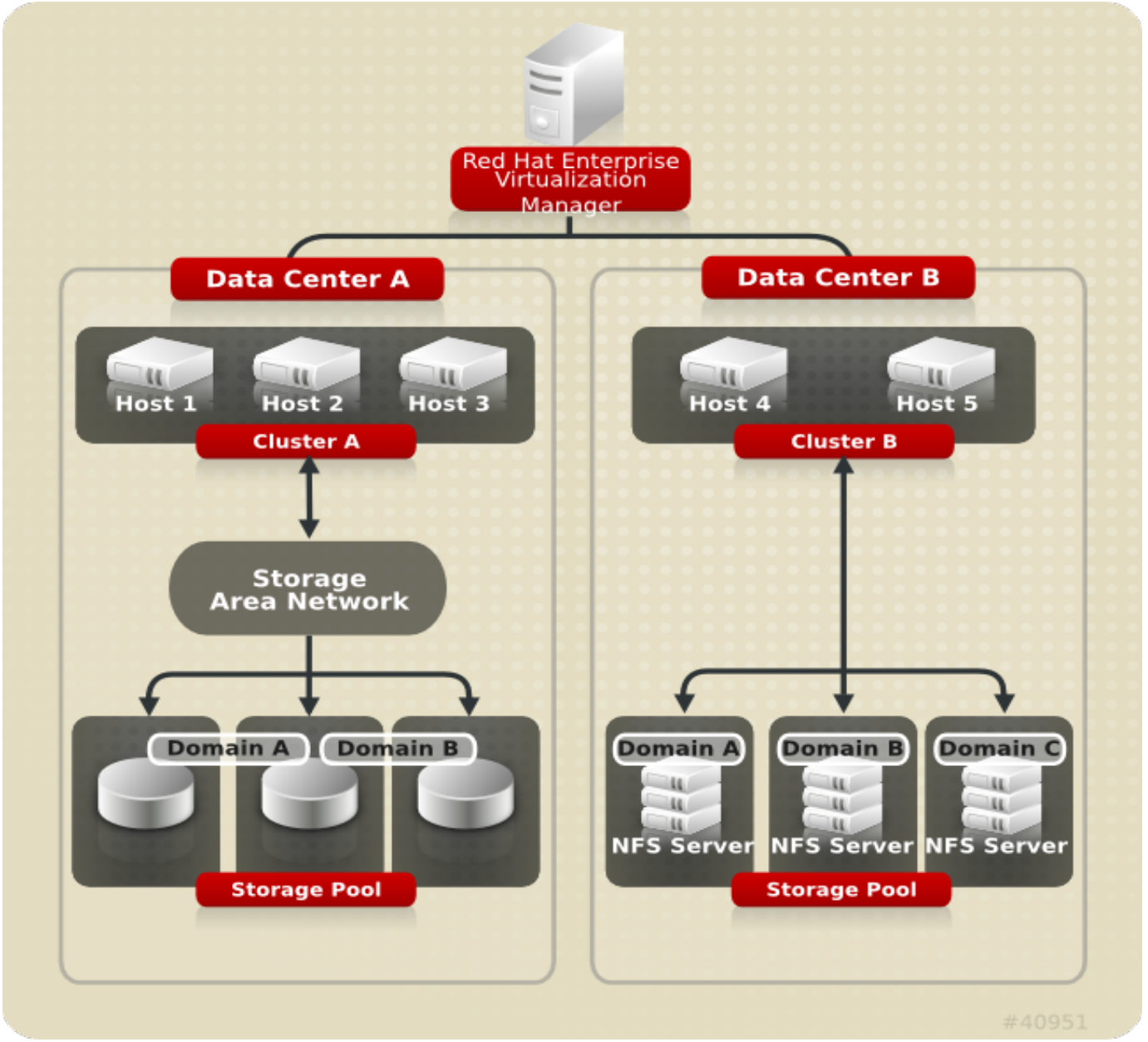
- Mail box
- Thin provisioning
- Slower image manipulation
- Devices managed by device-mapper and multipath
- Domain is a VG
- Metadata is stored in a single LV and in lvm tags
- Volumes are LVs

Storage Pool

- A group of storage domains
- Supposed to simplify cross domain operations
- Being deprecated

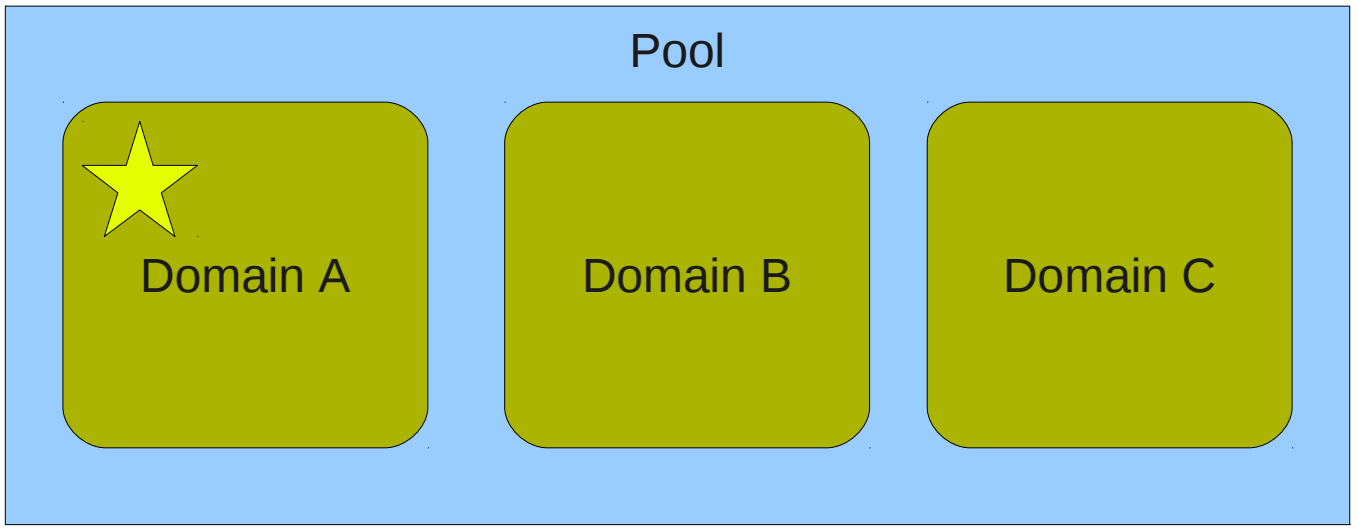


Example


Red Hat Enterprise Virtualization Manager
Data Center A
Data Center B
Host 1
Host 2
Host 3
Host 4
Host 5
Cluster A
Cluster B
Storage Area Network
Domain A
Domain B
Domain A
Domain B
Domain C
NFS Server
NFS Server
NFS Server
Storage Pool
Storage Pool
#40951

Master Domain

- Used to store:
  - Pool metadata
  - Backup of OVFs (treated as blobs)
  - Async tasks persistent data
- Contains the clustered lock for the pool



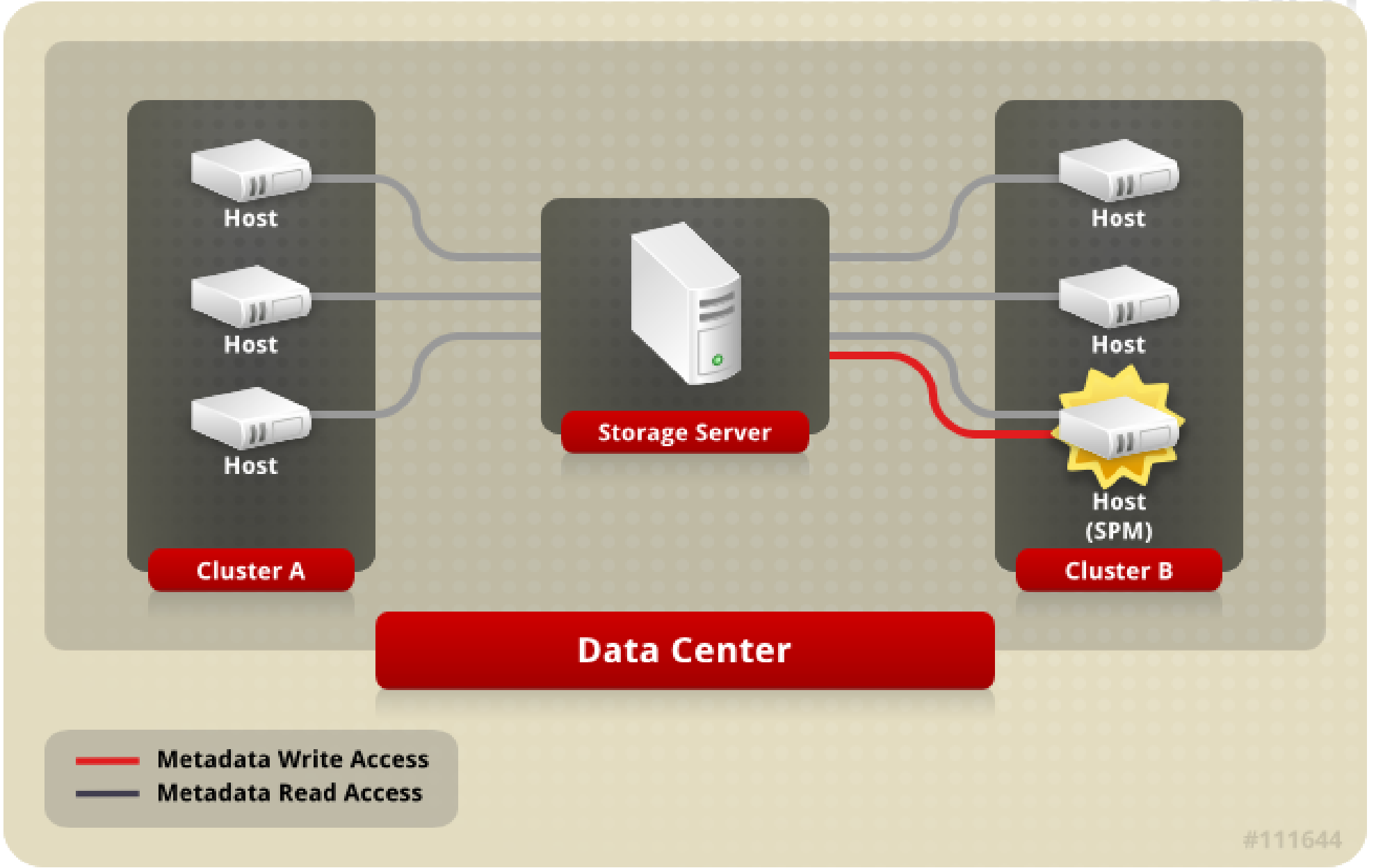
Storage Pool Manager (SPM)

A role assigned to one host in a data center granting it sole authority over:

- Creation, deletion and manipulation of virtual disk images, snapshots and templates
- Allocation of storage for sparse block devices (on SAN)
- Single meta data writer:
  - SPM lease mechanism (Chockler and Malkhi 2004, Light-Weight Leases for Storage-Centric Coordination)
  - Storage-centric mailbox.

This role can be migrated to any host in a data center.

Storage Pool Manager (SPM) 2

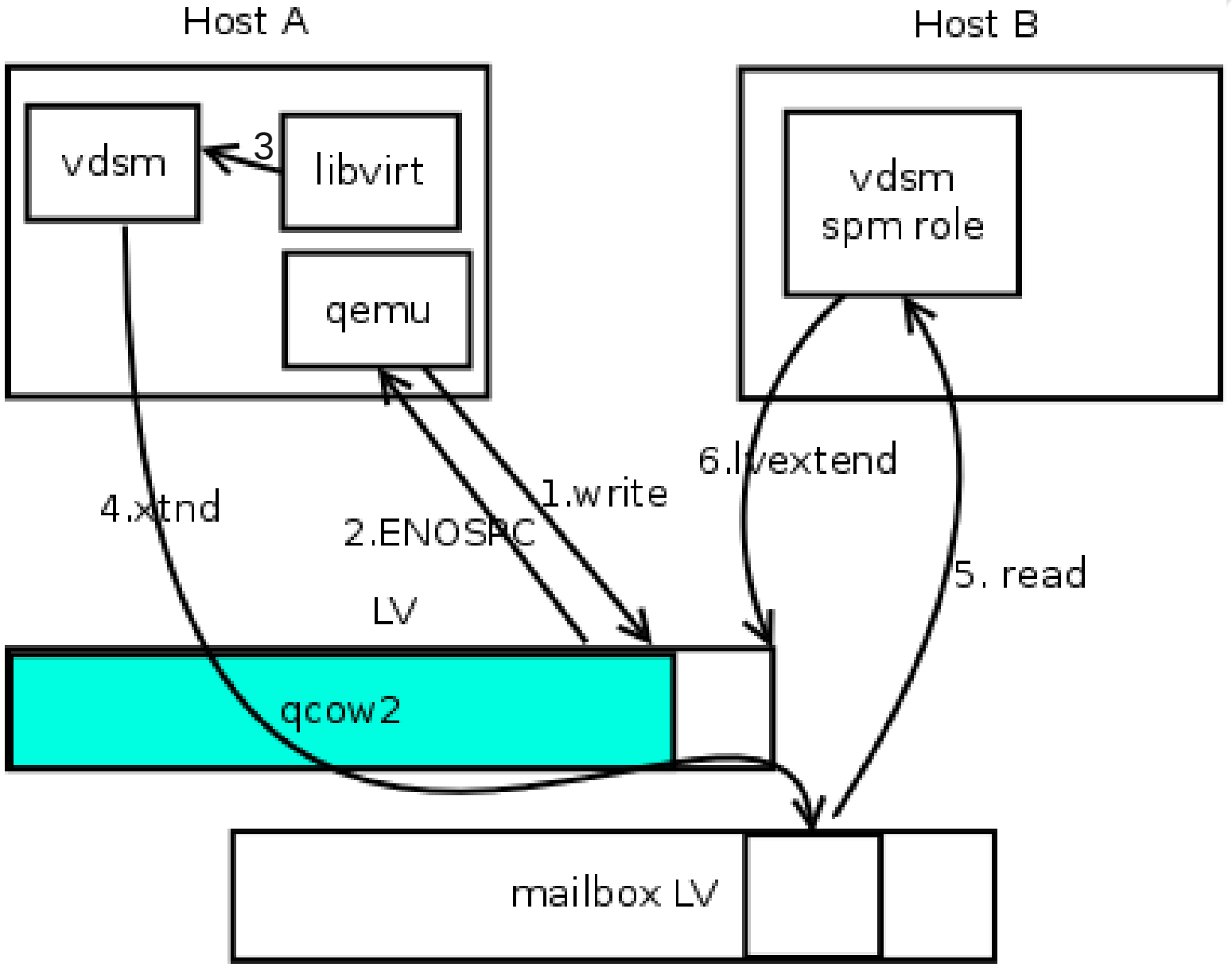


Thin Provisioning

Over-commitment is a storage function which allows oVirt to logically allocate more storage than is physically available.

- Generally, virtual machines use much less storage than what has been defined for them
- Over-commitment allows a virtual machine to operate completely unaware of the resources that are actually available
- QEMU identifies the highest offset written onto the logical volume (soon to be moved to LVM)
- VDSM monitors the highest offset marked by QEMU
- VDSM asks the SPM to extend the logical volume when needed

Thin Provisioning



Roadmap

oVirt

- SANLock
- SDM
- Live snapshots
- Live storage migration (block copy / streaming)
- Direct LUN
- Support any shared file system (not just NFS)
- NFS V.4 support
- NFS Hardmounts support
- Offload (snapshots, lun provisioning, thin provisioning, etc)

Roadmap - cont

- Image handling
  - Image Manager
  - Allocation policy (Space / Performance)
- Dynamic Connection Management
- So much more

How to contribute

- **Repository:**
  - http://git.fedorahosted.org/git/?p=vdsm.git
- **Mailing lists:**
  - vdsm-devel@lists.fedorahosted.org
  - vdsm-patches@lists.fedorahosted.org
- **IRC:**
  - #vdsm on Freenode
- **Core Team:**
  Dan Kenigsberg, Saggi Mizrahi, Igor Lvovsky, Eduardo Warszawasky, Yotam Oron, Ayal Baron

Q&A

oVirt

THANK YOU !

http://www.ovirt.org

Ovirt Storage Overview

Nov 2nd, 2011

Ayal Baron

Agenda

oVirt

- Defining the problem
- Storage Domain
- Storage Pool
- Example
- Domain Classes
- Domain Types
- File Domains
- Block Domains
- SPM
- Thin Provisioning
- Roadmap
- How to contribute
- Q&A

Defining the problem

- The requirements
  - Manage tens of thousands of virtual disk images
  - Each image potentially accessible from hundreds (thousands?) of nodes
  - Support both file based storage as well as raw block devices
- The assumptions
  - Exclusive access when used
  - Big (X GBs)
  - Centralized manager (can be HA in and by itself)

Defining the problem - guidelines

oVirt

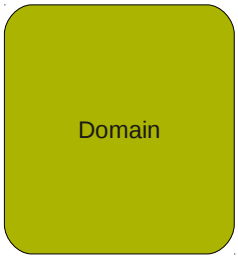
- High level, image centered API
- Cluster safe
- High performance (not in data path)
- Highly available
  - no single point of failure
  - Continues working in the absence of the manager
- Backing storage agnostic

Storage Domain

oVirt

- A standalone storage entity
- Stores the images and associated metadata (but not VMs)

  Only true persistent storage for VDSM



Domain Classes

- Data
  - Master
- ISO (NFS only)
- Backup (NFS only)
- Domain classes are planned for deprecation

Domain Types

oVirt

- File Domains (NFS, local dir)
  - Use file system features for segmentation
  - Use file system for synchronizing access
- Block Domains (iSCSI, FCP, FCoE, SAS, ...)
  - Use LVM for segmentation
  - Use reserved space to ensure writes do not overlap
  - Very specialized use of LVM

File Domains

oVirt

- Sparse files
- Better image manipulation capabilities
- Volumes and metadata are files
- 1:1 Mapping between domain and dir / NFS export
- NFS - Different error handling logic for data path and control path

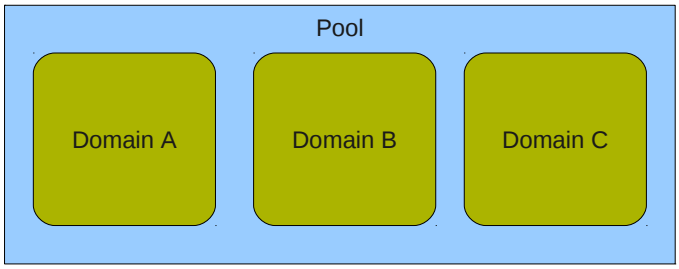
Block Domains

oVirt

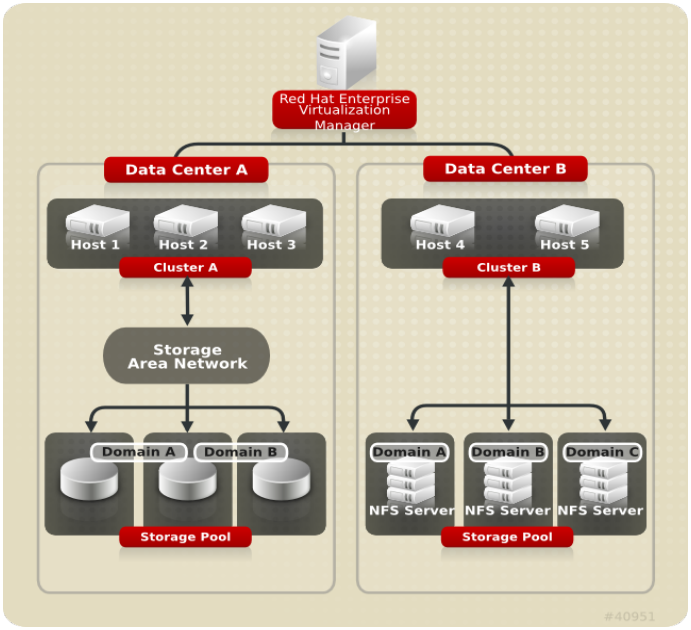
- Mail box
- Thin provisioning
- Slower image manipulation
- Devices managed by device-mapper and multipath
- Domain is a VG
- Metadata is stored in a single LV and in lvm tags
- Volumes are LVs

Storage Pool

- A group of storage domains
- Supposed to simplify cross domain operations
- Being deprecated

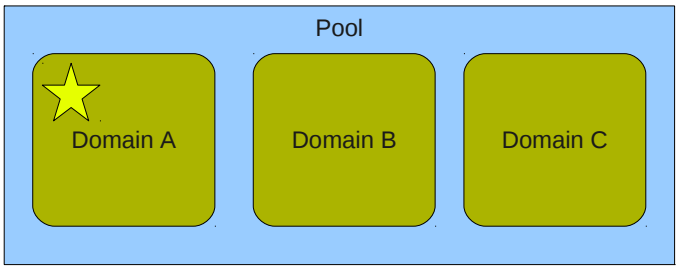


Example


Red Hat Enterprise Virtualization Manager
Data Center A
Data Center B
Host 1
Host 2
Host 3
Cluster A
Host 4
Host 5
Cluster B
Storage Area Network
Domain A
Domain B
Storage Pool
Domain A
Domain B
Domain C
NFS Server
NFS Server
NFS Server
Storage Pool
#40951

Master Domain

- Used to store:
  - Pool metadata
  - Backup of OVFs (treated as blobs)
  - Async tasks persistent data
- Contains the clustered lock for the pool



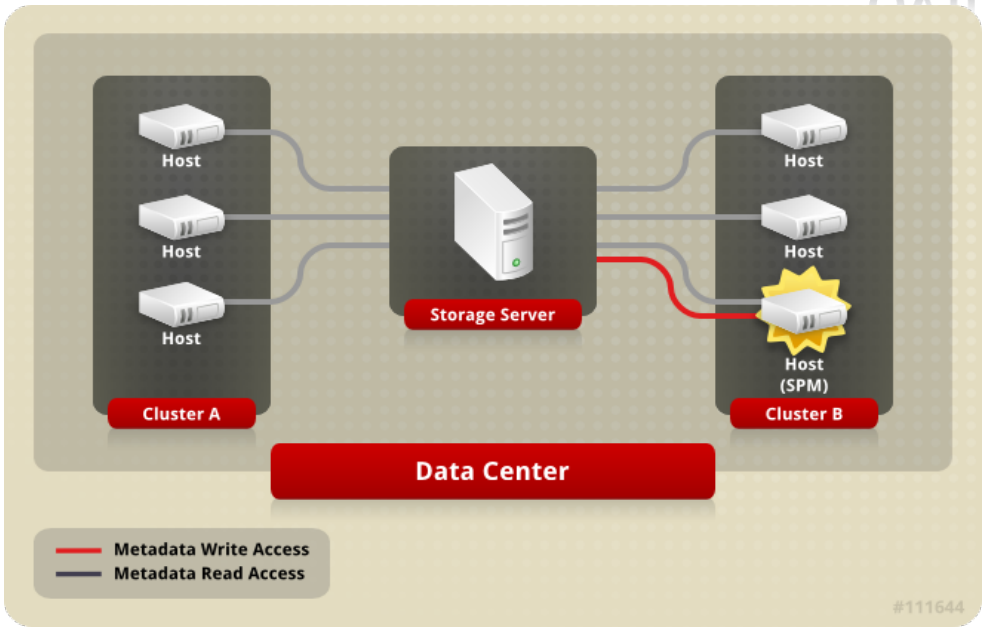
Storage Pool Manager (SPM)

A role assigned to one host in a data center granting it sole authority over:

- Creation, deletion and manipulation of virtual disk images, snapshots and templates
- Allocation of storage for sparse block devices (on SAN)
- Single meta data writer:
  - SPM lease mechanism (Chockler and Malkhi 2004, Light-Weight Leases for Storage-Centric Coordination)
  - Storage-centric mailbox.

This role can be migrated to any host in a data center.

Storage Pool Manager (SPM) 2

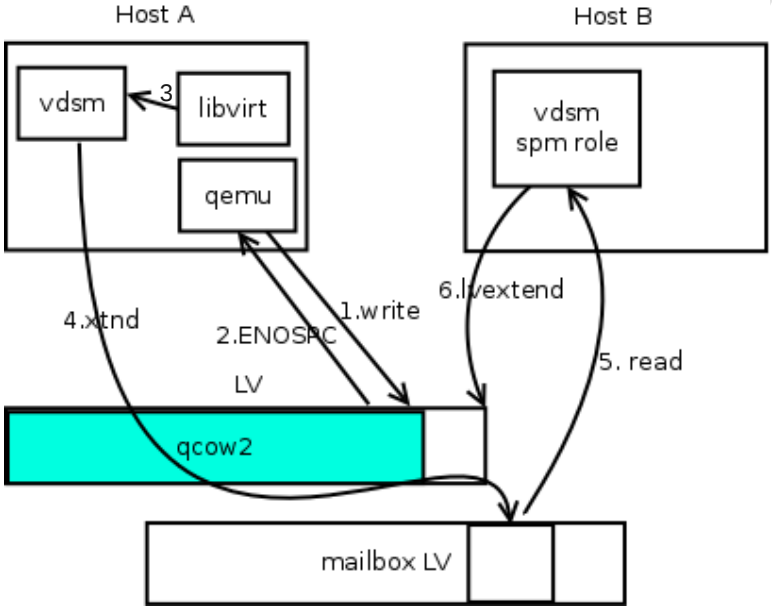


Thin Provisioning

Over-commitment is a storage function which allows oVirt to logically allocate more storage than is physically available.

- Generally, virtual machines use much less storage than what has been defined for them
- Over-commitment allows a virtual machine to operate completely unaware of the resources that are actually available
- QEMU identifies the highest offset written onto the logical volume (soon to be moved to LVM)
- VDSM monitors the highest offset marked by QEMU
- VDSM asks the SPM to extend the logical volume when needed

Thin Provisioning



Roadmap

oVirt

- SANLock
- SDM
- Live snapshots
- Live storage migration (block copy / streaming)
- Direct LUN
- Support any shared file system (not just NFS)
- NFS V.4 support
- NFS Hardmounts support
- Offload (snapshots, lun provisioning, thin provisioning, etc)

Roadmap - cont

- Image handling
  - Image Manager
  - Allocation policy (Space / Performance)
- Dynamic Connection Management
- So much more

How to contribute

- **Repository:**
  - http://git.fedorahosted.org/git/?p=vdsm.git
- **Mailing lists:**
  - vdsm-devel@lists.fedorahosted.org
  - vdsm-patches@lists.fedorahosted.org
- **IRC:**
  - #vdsm on Freenode
- **Core Team:**
  Dan Kenigsberg, Saggi Mizrahi, Igor Lvovsky, Eduardo Warszawasky, Yotam Oron, Ayal Baron

oVirt

Q&A

oVirt

THANK YOU !

http://www.ovirt.org