Proxmox VE 5

Lecture 11
High Availability and CEPH Storage

Instructor: Hadi Alnabriss

PROXMOX

Why High Availability?

- Eliminate single point of failure
- Reduce downtime

PROXMOX

HA in Proxmox VE

- Proxmox VE make it much easier to reach high availability because they remove the “hardware” dependency.
- It also supports to setup and use redundant storage and network devices.
- So if one host fails, you can simply start services on another host within your cluster.
  - Proxmox VE provides a software stack called ha-manager, which can do that automatically for you

PROXMOX

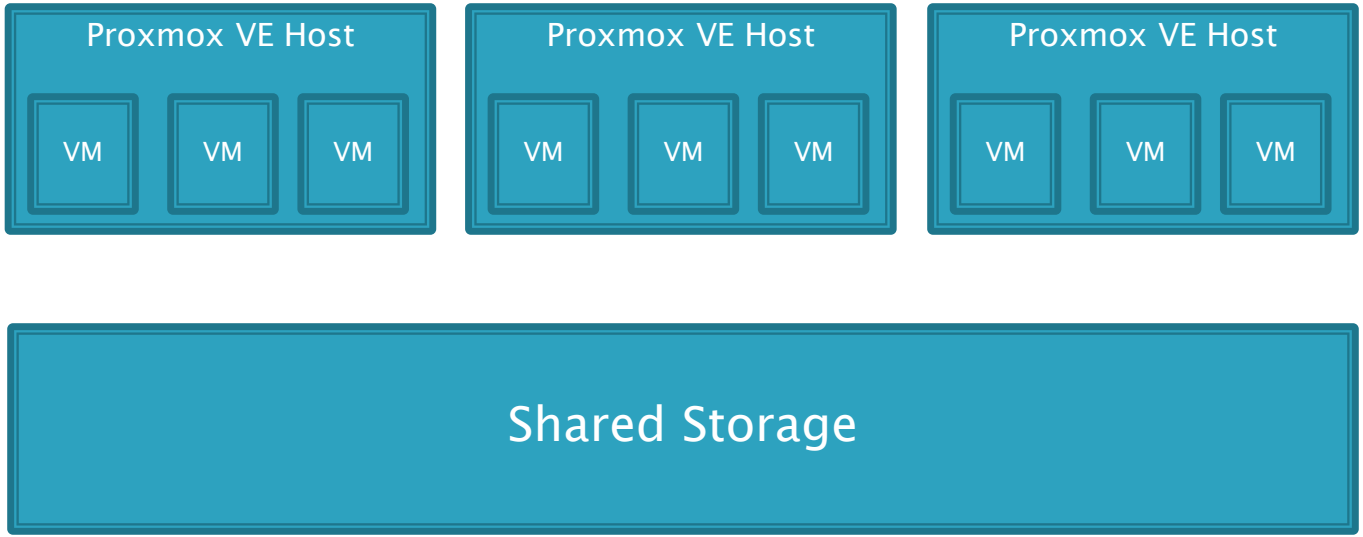
Requirements

- At least three cluster nodes (to get reliable quorum)
- Shared storage for VMs and containers
- Hardware redundancy (everywhere)
- Use reliable “server” components

PROXMOX

Why Shared Storage?

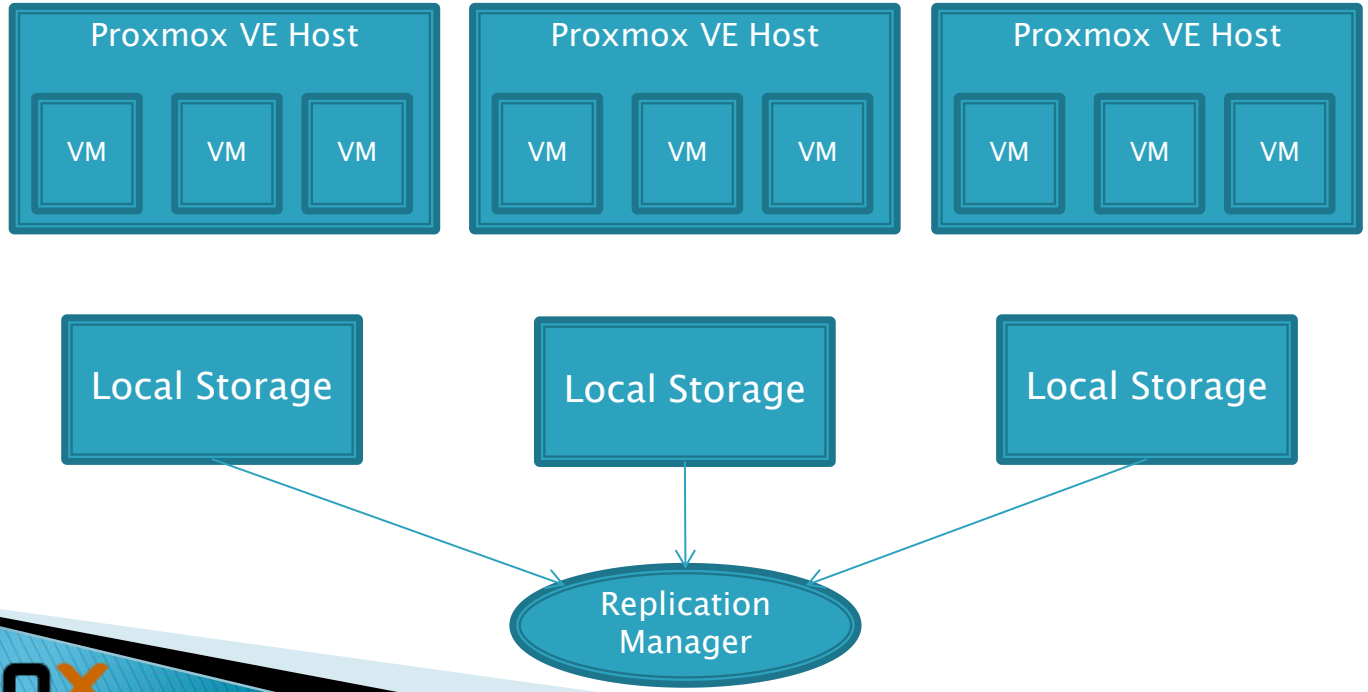
- All the Hosts have access to the same storage



PROXMOX

Why Shared Storage?

- CEPH Uses Local Storage as Shared Storage



What is CEPH?

- Ceph is a distributed replicated clustered filesystem
- Ceph is a distributed object store and file system designed to provide excellent performance, reliability and scalability.

Why Ceph?

- Easy setup and management with CLI and GUI support on Proxmox VE
- Thin provisioning
- Snapshots support
- Self healing
- No single point of failure
- Scalable to the exabyte level
- Runs on economical commodity hardware
- No need for hardware RAID controllers
- Easy management
- Open source

PROXMOX

Ceph Components

- (1) OSDs (Object Storage Device)
  - Corresponds to a physical disk. An OSD is actually a directory that Ceph uses, residing on a regular filesystem
    - (eg. /var/lib/ceph/osd-1)

Ceph Components

- (2) Placement Groups
  - Placement groups used for tracking metadata for objects
  - It represents a mostly-static mapping to one or more underlying OSDs.
  - All PGs in a pool will replicate stored objects into multiple OSDs.
  - PG calculator
    - https://ceph.com/pgcalc/

PROXMOX

Ceph Components

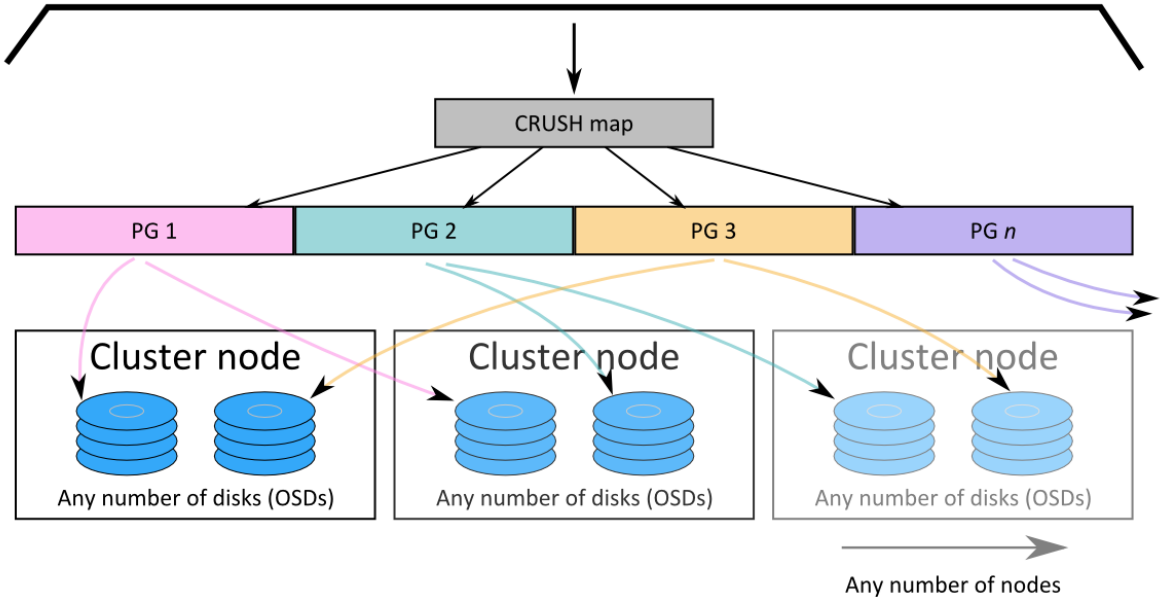
- (3) CRUSH maps
  - Ensures that replicas don’t end up on the same disk/host/rack/etc,

Ceph Components

- (4) Pools
  - A pool is the layer at which most user-interaction takes place.
  - This is the important stuff like GET, PUT, DELETE actions for objects in a pool.
  - Pools contain a number of PGs, not shared with other pools (if you have multiple pools).
  - The number of PGs in a pool is defined when the pool is first created, and can't be changed later.
  - You can think of PGs as providing a hash mapping for objects into OSDs, to ensure that the OSDs are filled evenly when adding objects to the pool.

PROXMOX

Ceph Components



Pool

Placement Groups

OSDs

PROXMOX

Before Installing Ceph

- We are going to use our 3-nodes cluster that we have created in our previous lecture
- On each Proxmox VE host:
  - Add additional Disks (i.e 2 x 2GB)
  - Add additional NIC
    - And make a new subnet for Ceph (i.e 10.0.0.0/24)

PROXMOX

Install Ceph

- Install Ceph on each node
  - *pveceph install --version luminous*
  - This will download some packages from the internet
  - You might face disk space problem if you are using 8GB disk for your host

PROXMOX

Configure Ceph Network

- Configure network on one node
  - *pveceph init --network 10.0.0.0/24*
  - This creates an initial config at /etc/pve/ceph.conf

Create Monitors

- The Ceph Monitor maintains a master copy of the cluster map.
- For HA you need to have at least 3 monitors.
- You can create monitor service using the command:
  - *pveceph createmon*
- Or from GUI
  - You should run 3 monitors, one on each node

Creating Monitors From GUI

PROXMOX Virtual Environment 5.1-41 | Search | You are logged in as 'root@pam' | Documentation | Create VM

Server View

- Datacenter
  - node01
    - local (node01)
    - local-lvm (node01)
  - node02
    - 100 (test01)
    - local (node02)
    - local-lvm (node02)
  - node03
    - local (node03)
    - local-lvm (node03)

Node 'node03' | Restart | Shutdown | Shell | Bu

- System
  - Network
  - DNS
  - Time
  - Syslog
- Updates
- Firewall
- Disks
- Ceph
  - Configuration
  - Monitor
  - OSD
  - Pools

Start | Stop | Create | Remove

Name	Host	Quorum	Address
mon.node01	node01	Yes	10.0.0.1:6789/0
mon.node02	node02	Yes	10.0.0.2:6789/0
mon.node03	node03	Yes	10.0.0.3:6789/0

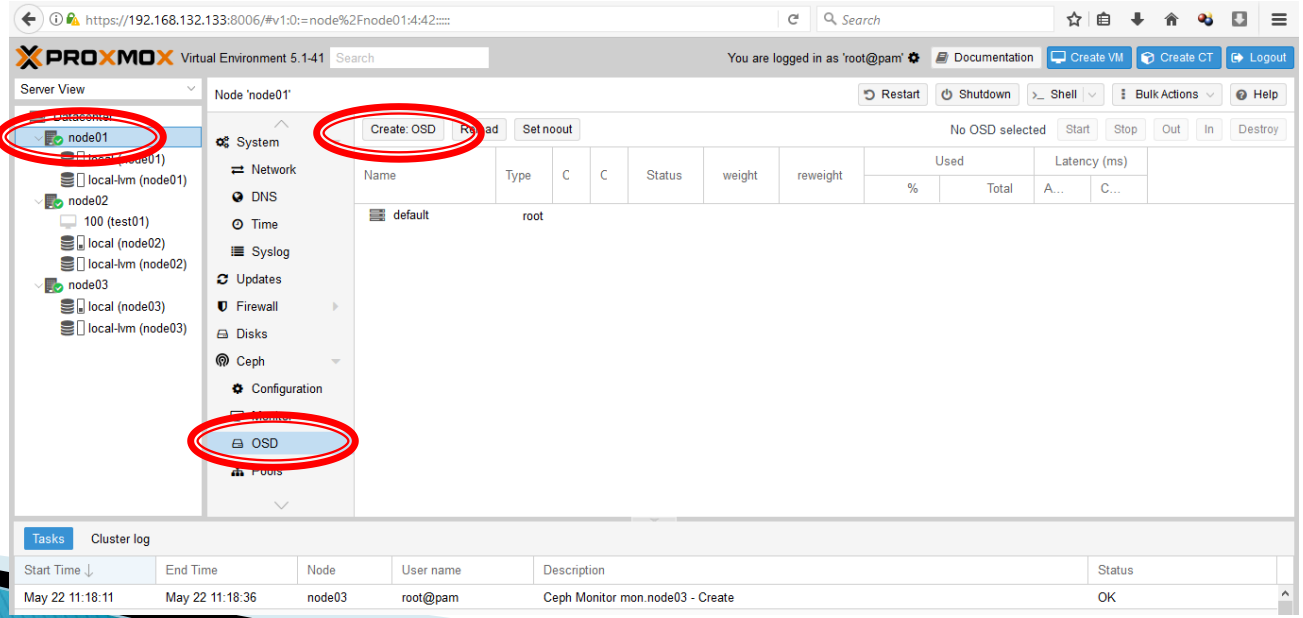
Tasks | Cluster log

Start Time	End Time	Node	User name	Description	Status
May 22 11:18:11	May 22 11:18:36	node03	root@pam	Ceph Monitor mon.node03 - Create	OK

PROXMOX

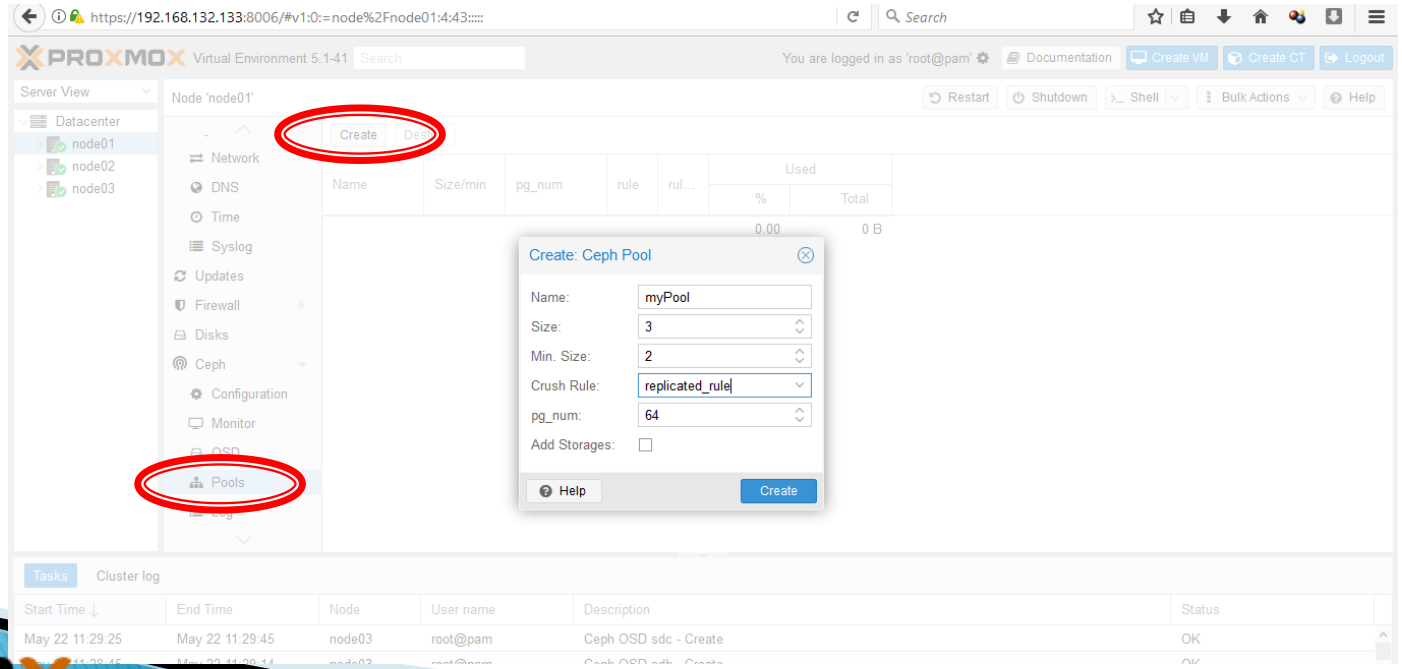
Create OSDs

- Add the required disks (6 disks)
  - Again: Don’t use RAID



Create the Pool

- The pool == our storgae



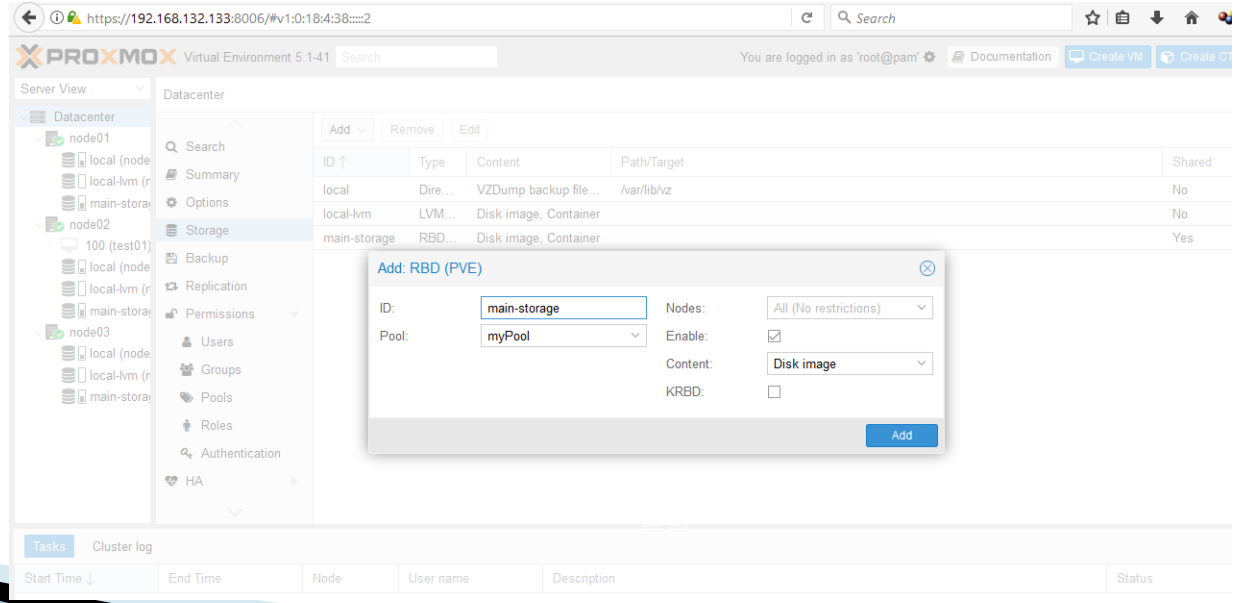
Keyring

- To build a trust relationship between ceph and Proxmox VE (Requires for external Ceph)
  - mkdir /etc/pve/priv/ceph
  - cp /etc/pve/priv/ceph.client.admin.keyring /etc/pve/priv/ceph/my-ceph-storage.keyring
- As we use local Ceph, this step is not required

PROXMOX

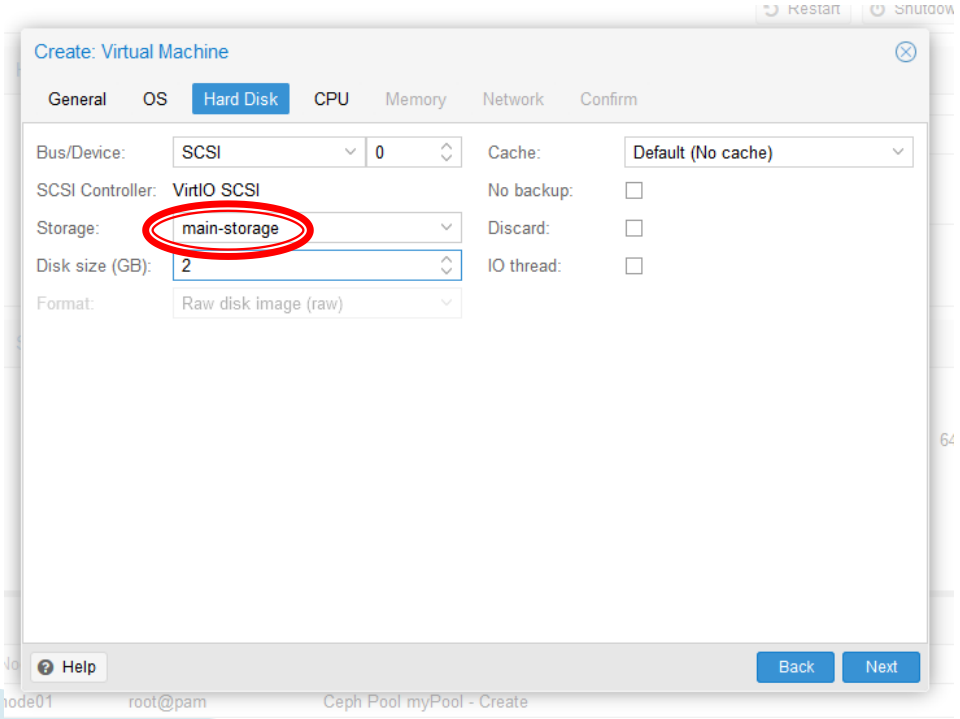
Create Storage

- Create a storage that maps to the created Pool (choos RBD PVE storage)



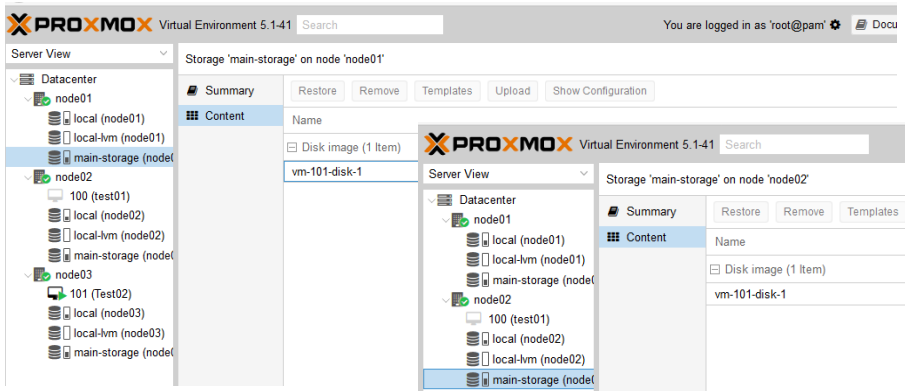
High Availability

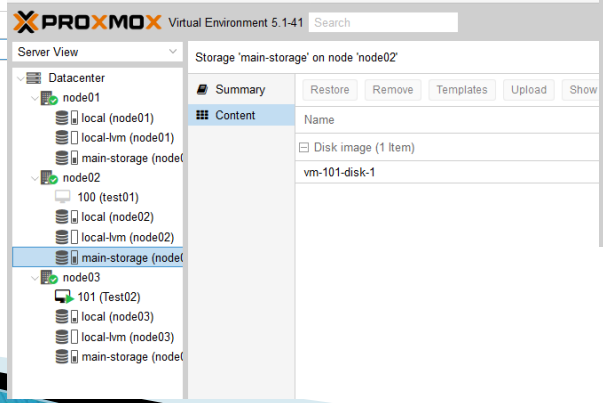
- Create a VM and choose the Ceph storage for its Hard Disk

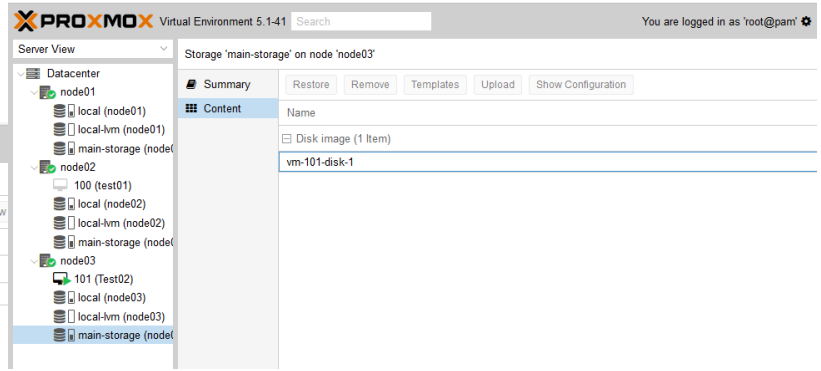


Migration

- Now you can migrate VMs and containers on the shared storage ONLINE, with no down time





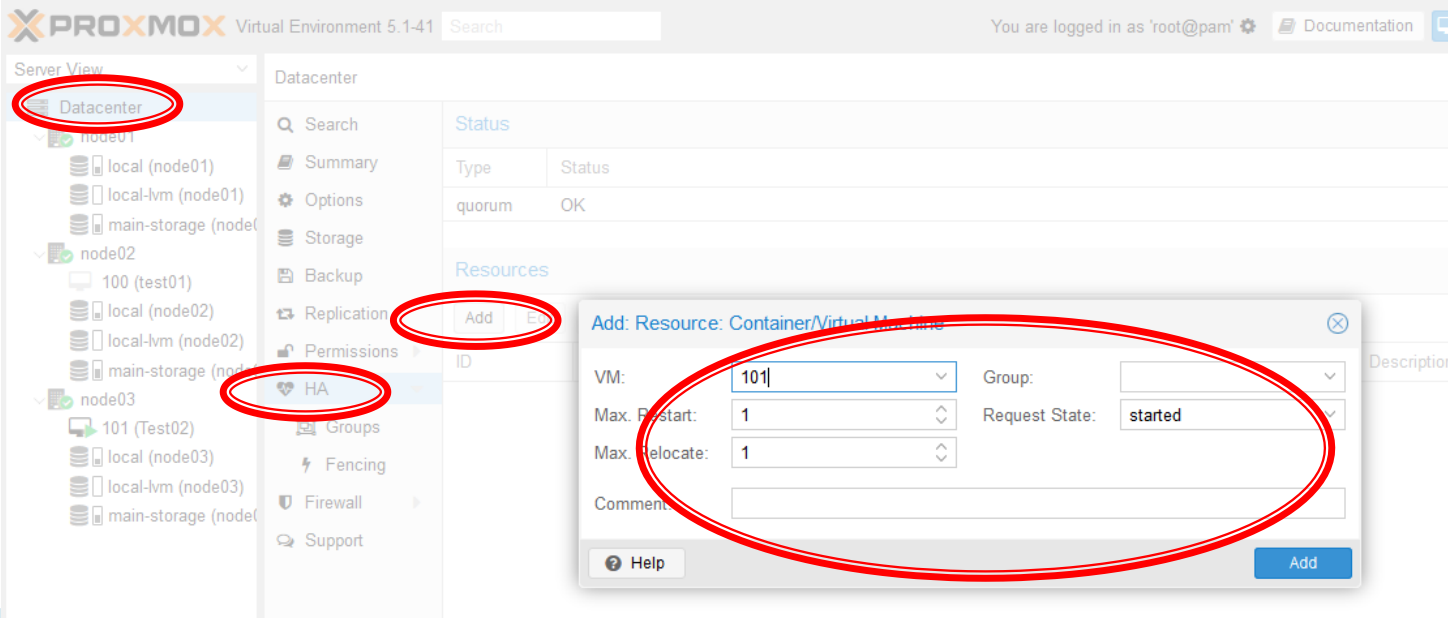


What is High Availability?

- When one of your Hosts fail, the HA manager will migrate all the HA-enabled VMs to another Host

Enable HA for VMs

- Add VM as resource to HA



Active Resources

PROXMOX Virtual Environment 5.1-41 Search You are logged in as 'root@pam'

Server View

- Datacenter
  - node01
    - local (node01)
    - local-lvm (node01)
    - main-storage (node(
  - node02
    - 100 (test01)
    - local (node02)
    - local-lvm (node02)
    - main-storage (node(
  - node03
    - 101 (Test02)
    - local (node03)
    - local-lvm (node03)
    - main-storage (node(

Datacenter

- Search
- Summary
- Options
- Storage
- Backup
- Replication
- Permissions
- HA
  - Groups
  - Fencing
- Firewall
- Support

Status

Type	Status
quorum	OK
master	node02 (active, Tue May 22 13:13:10 2018)
lrm	node01 (idle, Tue May 22 13:13:14 2018)
lrm	node02 (idle, Tue May 22 13:13:14 2018)
lrm	node03 (active, Tue May 22 13:13:14 2018)

Resources

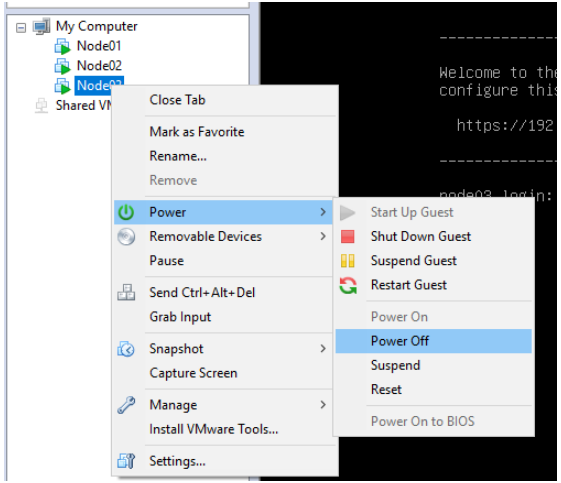
Add Edit Remove

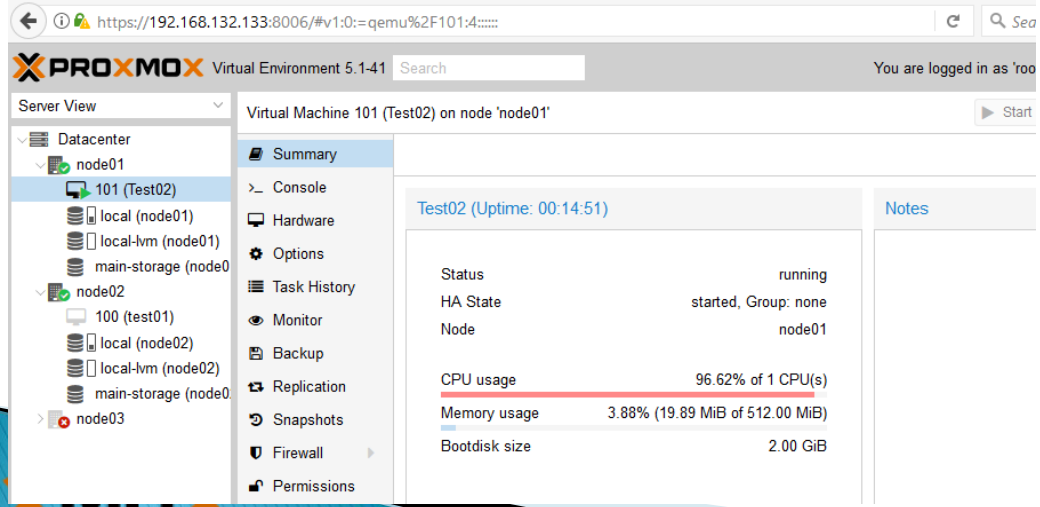
ID	State	Node	Max. Restart	Max. Reloc...	Group
vm:101	started	node03	1	1	

PROXMOX

Test HA

- Power off the Host: node03
- VM 101 must transfer to Another host





Ceph Health

https://192.168.132.133:8006/#v1:0:=node%2Fnode01:4:38:::::

PROXMOX Virtual Environment 5.1-41 | Search | You are logged in as 'root@pam' | Documentation | Create VM | Create CT | Logout

Server View

- Datacenter
  - node01
  - node02
  - node03

Node 'node01' | Restart | Shutdown | Shell | Bulk Actions | Help

- Search
- Summary
- Shell
- Shell (JS)
- System
  - Network
  - DNS
  - Time
  - Syslog
- Updates
- Firewall
- Disks
- Ceph
  - Configuration
  - Monitor
  - OSD
  - Pools

Health

Status

HEALTH_WARN

Severity	Summary
!	1/3 mons down, quorum node01,node02
!	2 osds down
!	1 host (2 osds) down
!	Degraded data redundancy: 5/15 objects degraded (33.333...
!	too few PGs per OSD (21 < min 30)

Status

Monitors

node01: ✔ node02: ✔ node03: ✖

OSDs

	In	Out
Up	4	0
Down	2	0

Total: 6

PGs

active+undersized:	60
active+undersized+degraded:	4

Performance

PROXMOX

Increase and Decrease Storage

- To add more disks:
  - Add new Disk
  - Create OSD for the new Disk

- To remove disks:
  - Stop the OSD for the required disk
  - Make the disk out

Discussion

- The disk is shared, but what about RAM?
- What if the VM has a local ISO for CD-ROM?

PROXMOX

Conclusion

- Now you must be able to:
  - Create Ceph shared storage
  - Make Online migrations
  - Enable High Availability for specific VMs

PROXMOX