

Proxmox VE 5

Lecture 11
High Availability and CEPH Storage

Instructor: Hadi Alnabriss

Why High Availability?

- ▶ Eliminate single point of failure
- ▶ Reduce downtime

HA in Proxmox VE

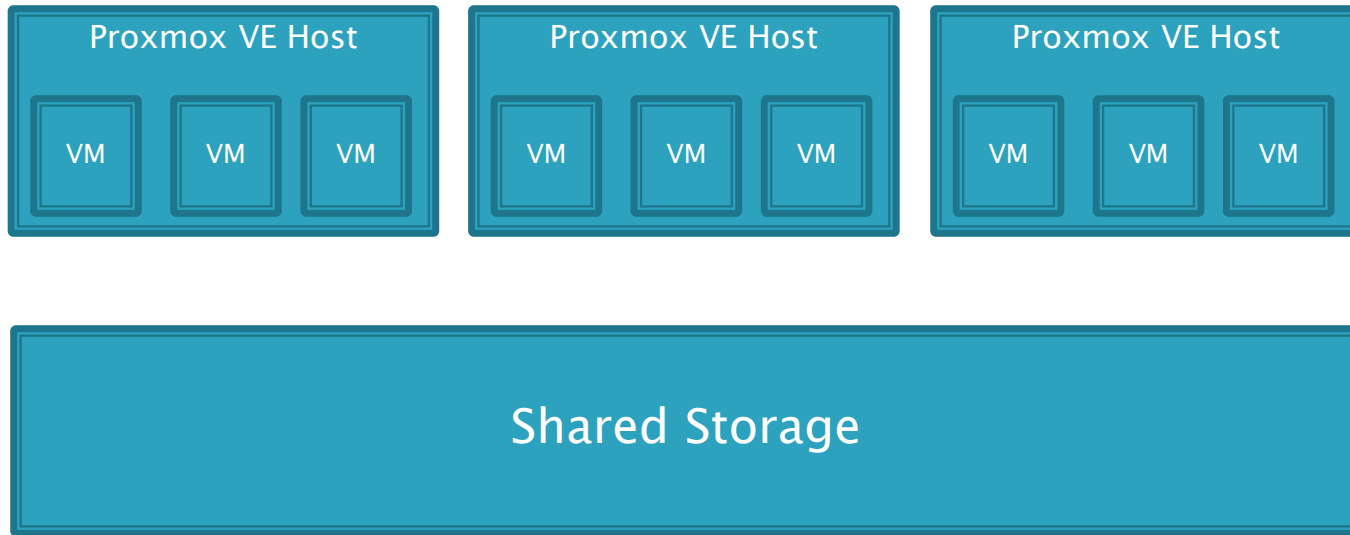
- ▶ Proxmox VE make it much easier to reach high availability because they remove the “hardware” dependency.
- ▶ It also supports to setup and use redundant storage and network devices.
- ▶ So if one host fails, you can simply start services on another host within your cluster.
 - Proxmox VE provides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

Requirements

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

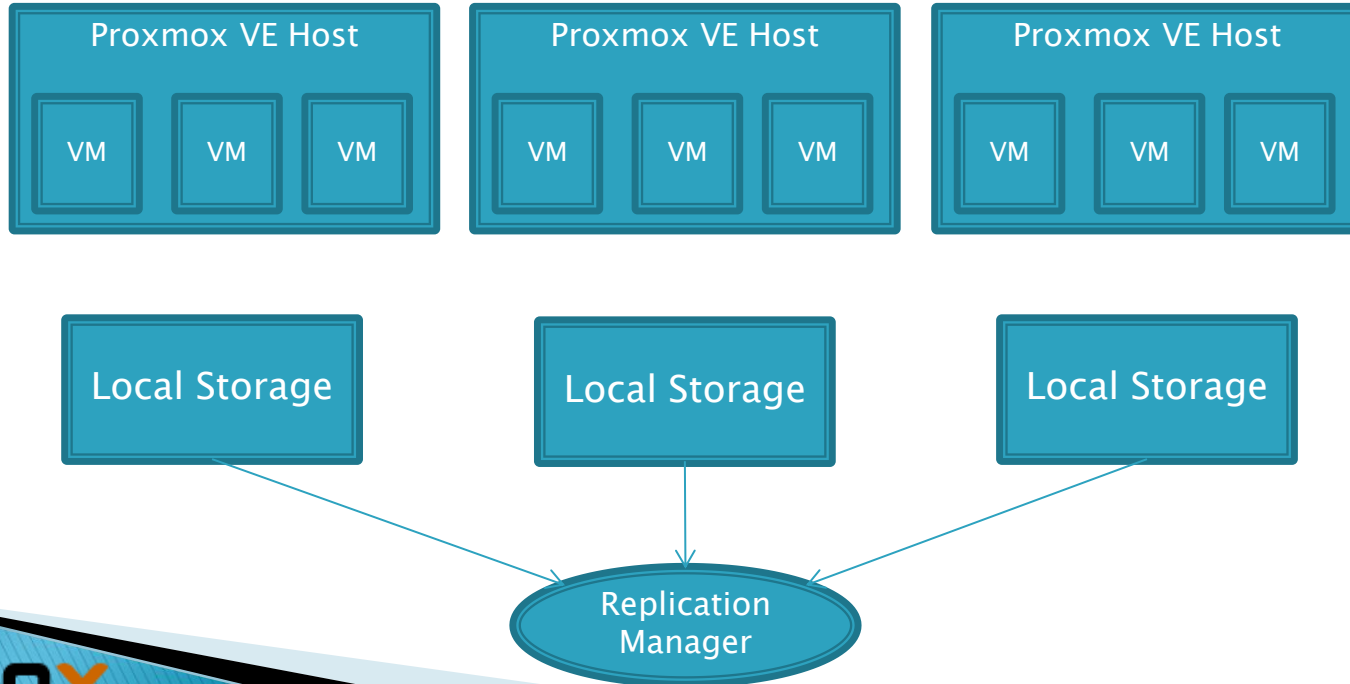
Why Shared Storage?

- ▶ All the Hosts have access to the same storage



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

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/>

Ceph Components

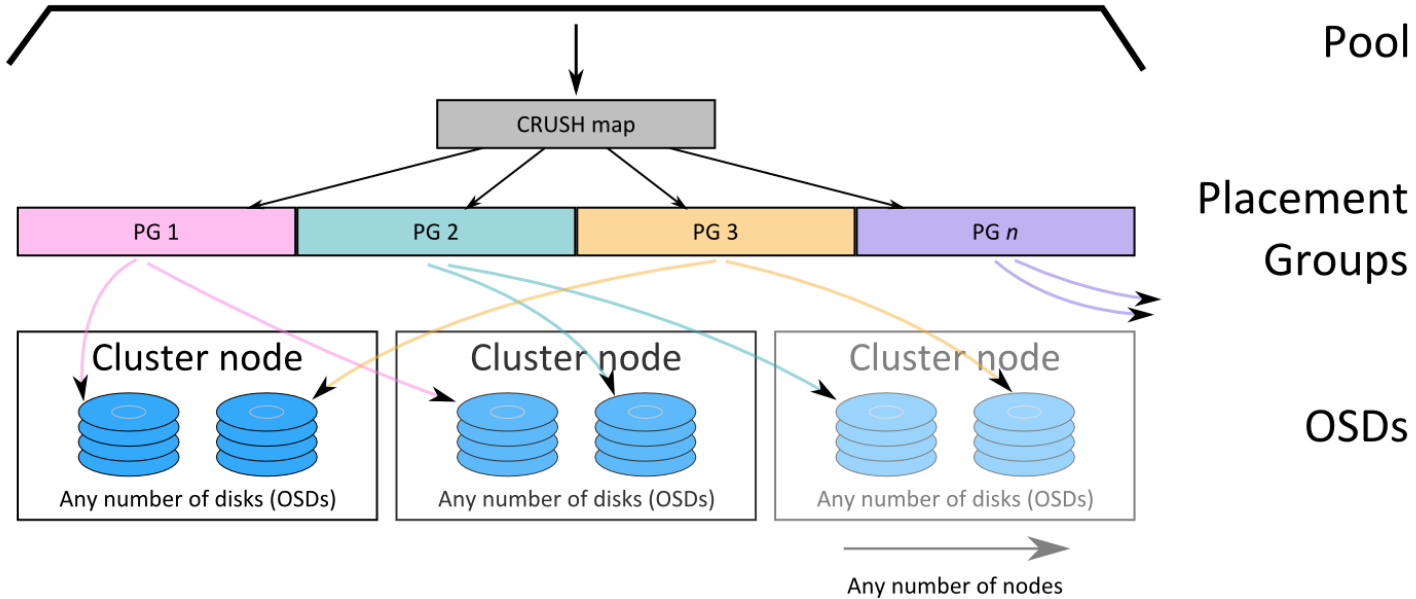
- ▶ (3) CRUSH maps
 - Ensures that replicas don't end up on the same disk/host/rack/etc,

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.

Ceph Components



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)

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

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 5.14.1

You are logged in as 'root@pam'

Documentation Create VM

Server View

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

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

Node 'node03'

Restart Shutdown Shell Bu

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
Mav 22 11:18:11	Mav 22 11:18:36	node03	root@pam	Ceph Monitor mon.node03 - Create	OK

Create OSDs

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

The screenshot shows the Proxmox VE 5.1-41 web interface. The left sidebar shows a tree view of the cluster with 'node01' selected. The main panel shows the configuration for 'Node 'node01''. The 'Create OSD' button is highlighted with a red circle. Below it, the 'OSD' tab is also highlighted with a red circle. The 'Create OSD' button is located in the top right corner of the main panel. The 'OSD' tab is located in the bottom left corner of the main panel. The 'Create OSD' button is located in the top right corner of the main panel. The 'OSD' tab is located in the bottom left corner of the main panel.

Server View

Node 'node01'

Restart Shutdown Shell Bulk Actions Help

Create OSD Read Set noout

No OSD selected Start Stop Out In Destroy

Name	Type	C	C	Status	weight	reweight	Used	Latency (ms)		
							%	Total	A...	C...
default	root									

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

Create the Pool

- ▶ The pool == our storage

The screenshot shows the Proxmox VE 5.1-41 web interface. The left sidebar has a 'Pools' menu item circled in red. The main content area shows 'Node node01' with a 'Create' button circled in red. A 'Create: Ceph Pool' dialog box is open, showing the following configuration:

- Name: myPool
- Size: 3
- Min. Size: 2
- Crush Rule: replicated_rule
- pg_num: 64
- Add Storages: ☐

The 'Cluster log' table at the bottom shows the following entries:

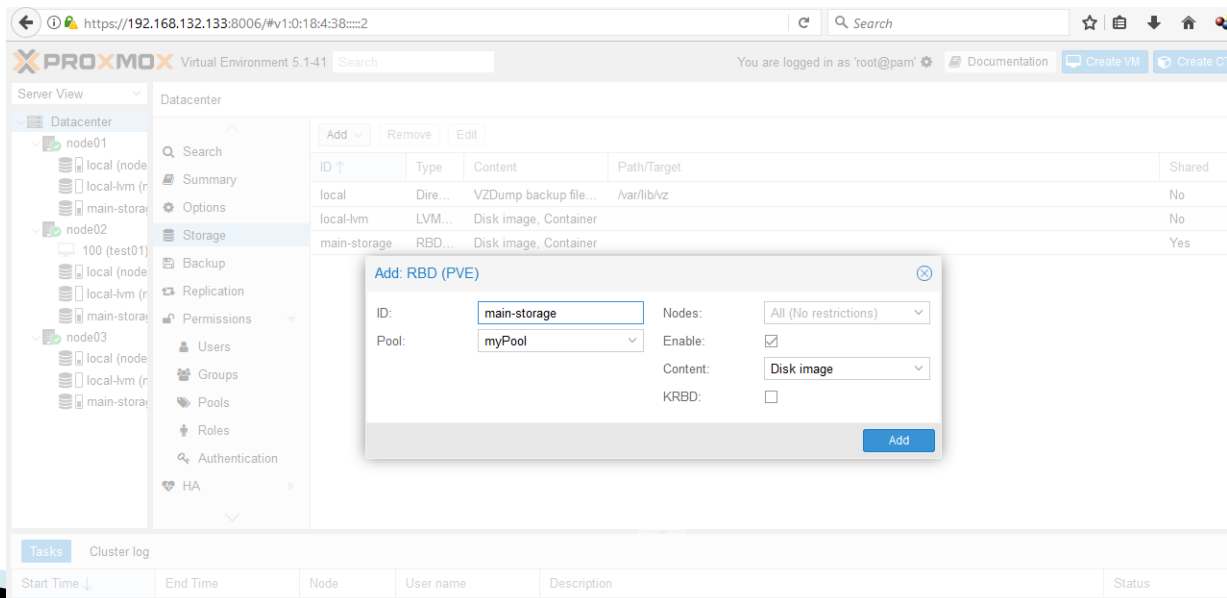
Start Time	End Time	Node	User name	Description	Status
May 22 11:29:25	May 22 11:29:45	node03	root@pam	Ceph OSD sdc - Create	OK
May 22 11:29:45	May 22 11:29:45	node03	root@pam	Ceph OSD sdb - Create	OK

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

Create Storage

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



The screenshot shows the Proxmox VE web interface. The left sidebar displays the 'Datacenter' tree with nodes 'node01', 'node02', and 'node03'. The main panel shows the 'Storage' tab for 'node01'. A dialog box titled 'Add: RBD (PVE)' is open, allowing the creation of a new storage. The dialog contains the following fields:

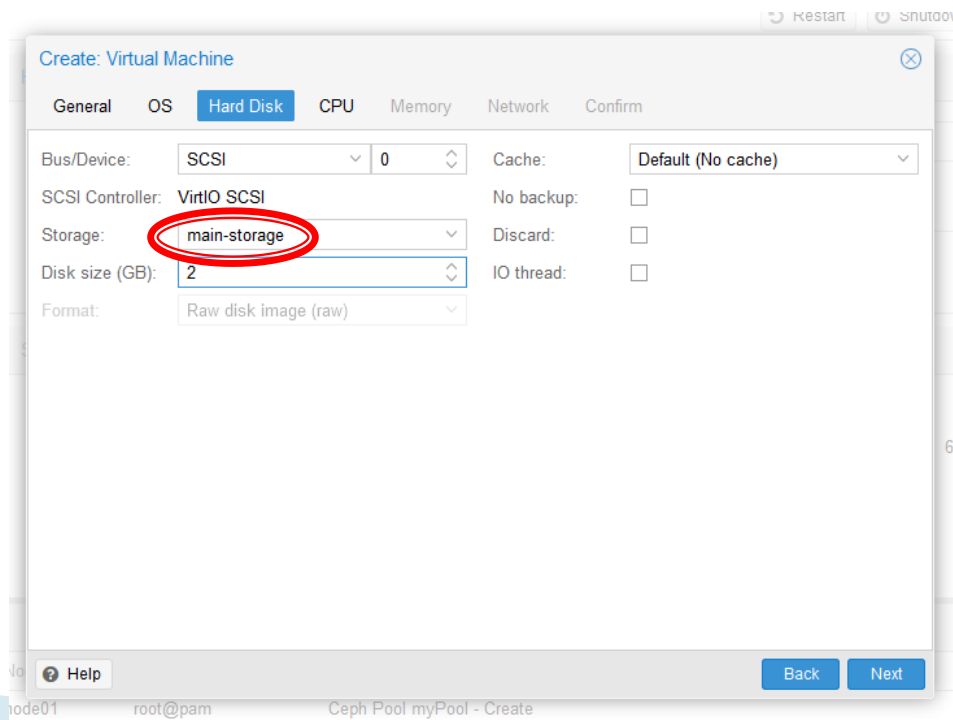
- ID:
- Pool:
- Nodes:
- Enable: ☒
- Content:
- KRBD: ☐

The 'Add' button is at the bottom right of the dialog. The background interface shows a table of existing storage configurations:

ID	Type	Content	Path/Target	Shared
local	Dire...	VZDump backup file...	/var/lib/vz	No
local-lvm	LVM...	Disk image, Container		No
main-storage	RBD...	Disk image, Container		Yes

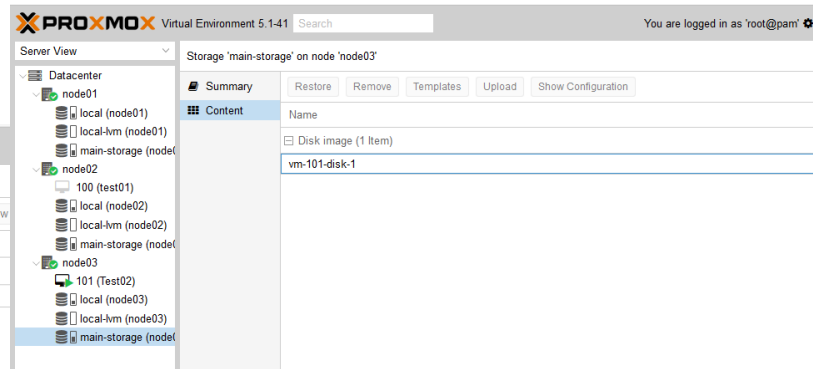
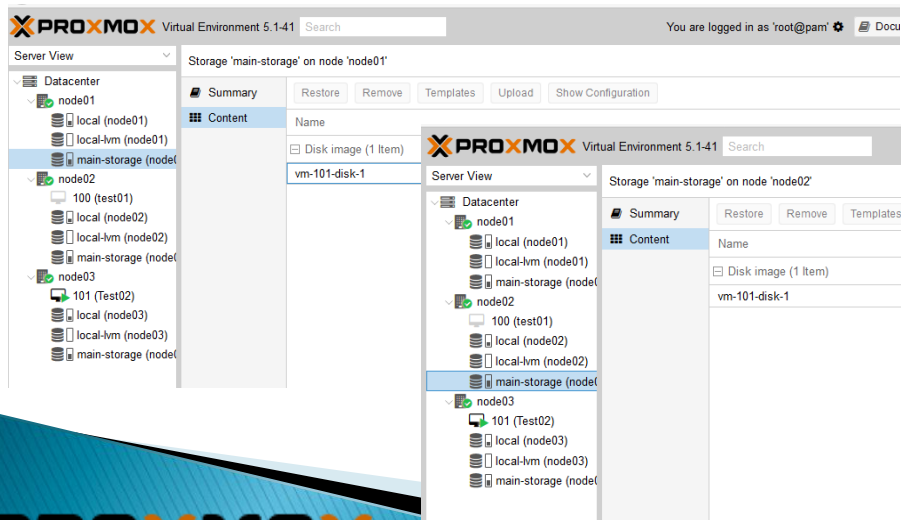
High Availability

- ▶ 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

The screenshot displays the Proxmox Virtual Environment 5.1-41 web interface. The left sidebar shows the 'Server View' with the 'Datacenter' node selected. The main panel shows the 'Datacenter' view with a list of resources. The 'HA' (High Availability) tab is selected, and the 'Add' button is highlighted. A modal window titled 'Add: Resource: Container/Virtual Machine' is open, showing the configuration for adding a new resource. The 'VM' field is set to '101', and the 'Request State' is set to 'started'. The 'Add' button in the modal is also highlighted.

PROXMOX Virtual Environment 5.1-41

You are logged in as 'root@pam'

Search

Server View

Datacenter

node01

- local (node01)
- local-lvm (node01)
- main-storage (node01)

node02

- 100 (test01)
- local (node02)
- local-lvm (node02)
- main-storage (node02)

node03

- 101 (Test02)
- local (node03)
- local-lvm (node03)
- main-storage (node03)

Datacenter

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

Status

Type	Status
quorum	OK

Resources

Add

Add: Resource: Container/Virtual Machine

VM: 101

Group:

Max. Restart: 1

Max. Relocate: 1

Request State: started

Comment:

Help

Add

Active Resources

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

Server View ▾

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

Datacenter

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

Status

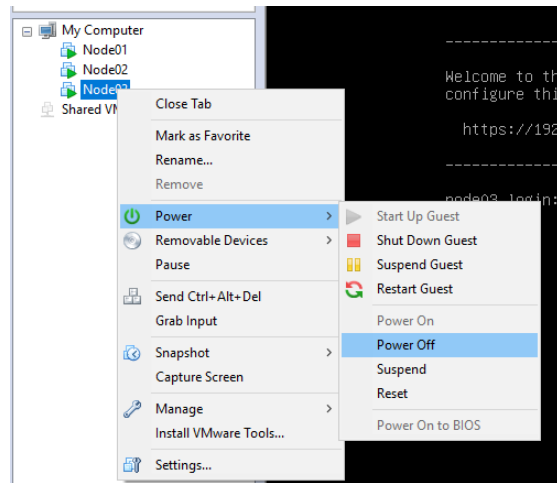
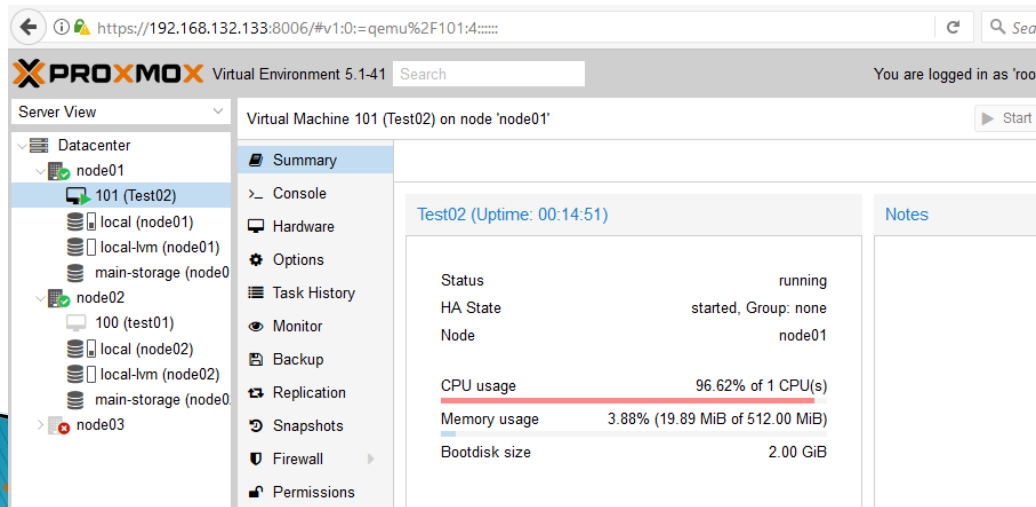
Type	Status
quorum	OK
master	node02 (active, Tue May 22 13:13:10 2018)
lrm	node01 (idle, Tue May 22 13:13:14 2018)
lrm	node02 (idle, Tue May 22 13:13:14 2018)
lrm	node03 (active, Tue May 22 13:13:14 2018)

Resources

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

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

Health

Status

HEALTH_WARN

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

Status

Monitors		OSDs		PGs	
node01: ✓	node02: ✓	● In	○ Out	active+undersized:	60
node03: ✗		🟢 Up	4	active+undersized+degraded:	4
		🔴 Down	2		
Total: 6					

Performance

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?

Conclusion

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