

[Proxmox] CEPH et HA

Introduction

La mise en place de **CEPH** vous permet d'avoir un disque partagé entre les noeuds de votre cluster Proxmox.

Cela sera utile notamment pour la haute disponibilité (**HA**) que l'on mettra en place dans un second temps.



Prérequis

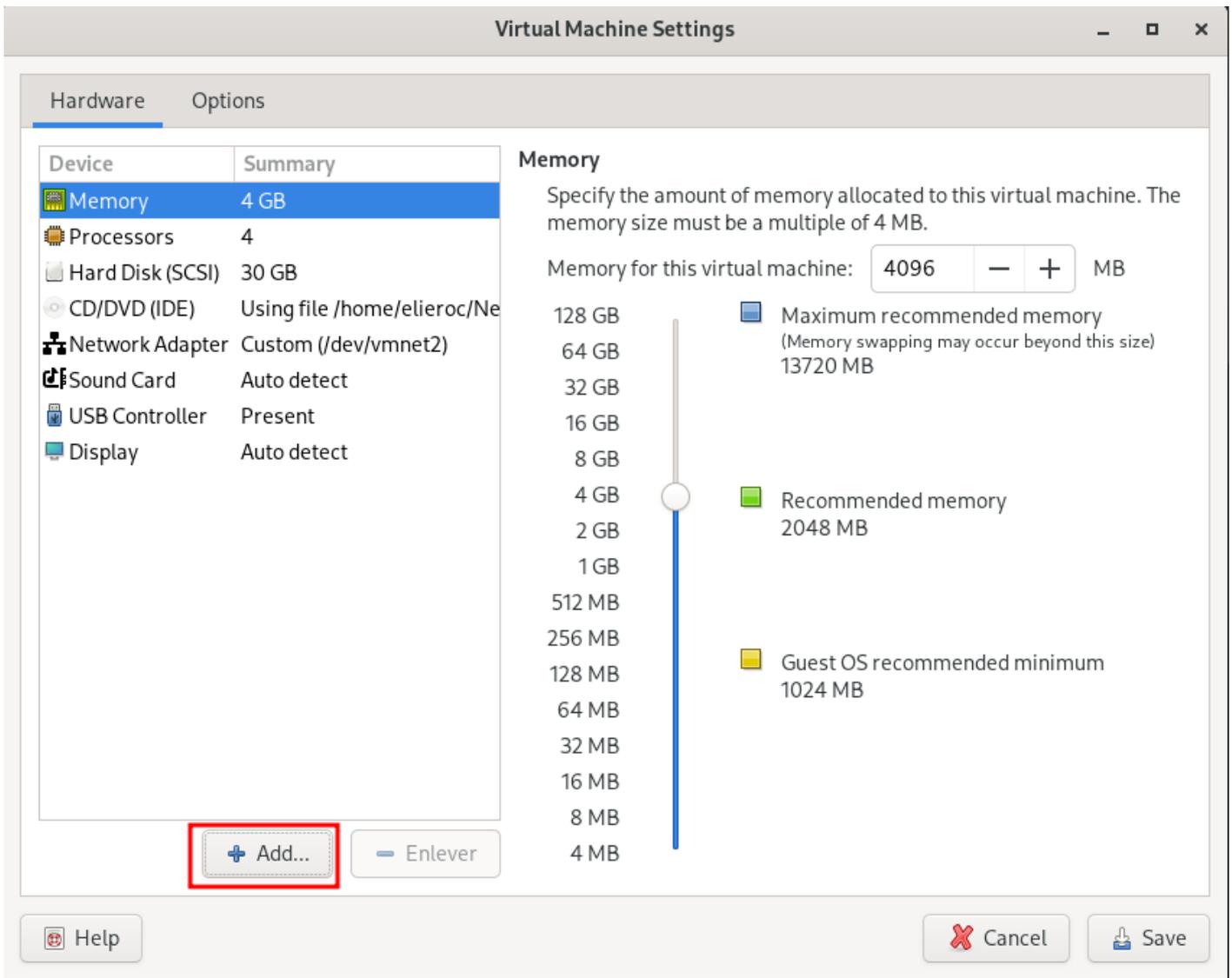
- [Installer un cluster Proxmox](#)

Installation de CEPH

Ajout des disques et carte réseaux CEPH

La première étape va être d'ajouter un disque partagé dans chacun des noeuds du cluster.

J'utilise un environnement **VMware** mais on peut très bien le faire avec des disques physiques sur des serveurs dans une baie :



Hardware Type



What type of hardware do you want to install?

vmware
WORKSTATION
PRO™

17

This wizard will guide you through the steps of adding new hardware to your virtual machine.

To begin, please select the type of hardware you want to add.

- Hard Disk**
- CD/DVD Drive
- Floppy Drive
- Network Adapter
- USB Controller Maximum limit reached
- Sound Card Maximum limit reached
- Parallel Port
- Serial Port
- Generic SCSI Device
- Trusted Platform Module The virtual machine must be encrypted and using UEFI firmware.

✖ Cancel

⏪ Back

⏩ Next

Select a Disk Type



What kind of disk do you want to create?



Virtual Disk Type

- IDE
- SCSI (Recommended)
- SATA
- NVMe

 Cancel

 Back

 Next

Select a Disk



Which disk do you want to use?

vmware
WORKSTATION
PRO™
17

Disk

- Create a new virtual disk
A virtual machine is composed of one or more files on the host file system, which will appear as a single hard disk to the guest operating system. Virtual disks can easily be copied or moved on the same host or between hosts.
- Use an existing virtual disk
Choose this option to reuse a previously configured disk.
- Use a physical disk (for advanced users)
Choose this option to give the virtual machine direct access to a local hard disk.

✖ Cancel

⏪ Back

⏩ Next

J'ajoute un disque d'une taille de **10Go** :

Specify Disk Capacity



How large do you want this disk to be?

vmware
WORKSTATION
PRO™

17

Disk Size

Maximum disk size (in GB): - +

Recommended size for Debian 12.x 64-bit: 20 GB

 Allocate all disk space now

Allocating the full capacity can enhance performance but requires all of the physical disk space to be available right now. If you do not allocate all the space now, the virtual disk starts small and grows as you add data to it.

 Store virtual disk as a single file Split virtual disk into multiple files

Splitting the disk makes it easier to move the virtual machine to another computer but may reduce performance with very large disks.

Cancel

Back

Next

Specify Disk File



Where would you like to store the disk file?



Disk File

A 10 GB virtual disk be created using multiple disk files. The disk files will be automatically named based on this file name.

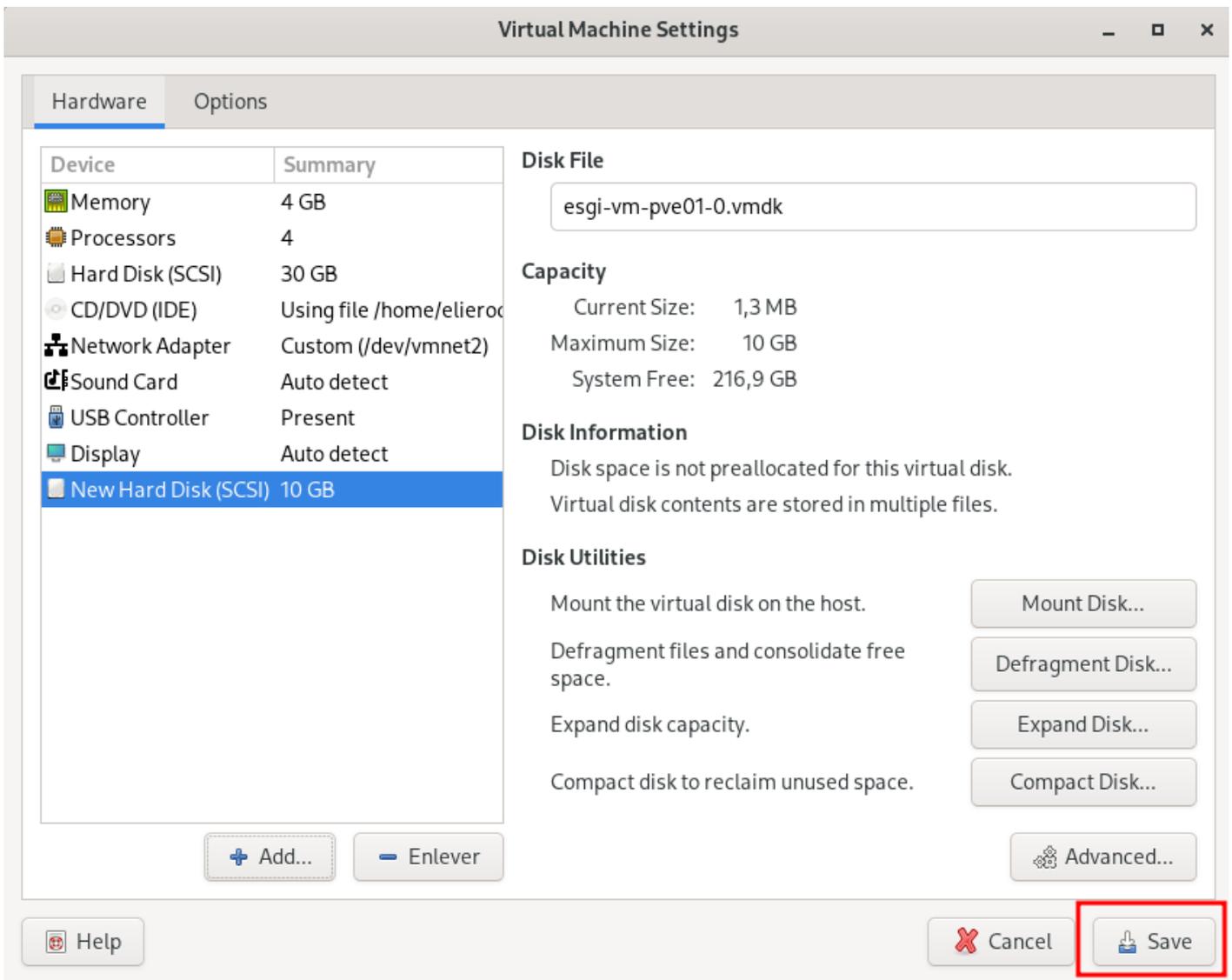
File name:

 Browse...

 Cancel

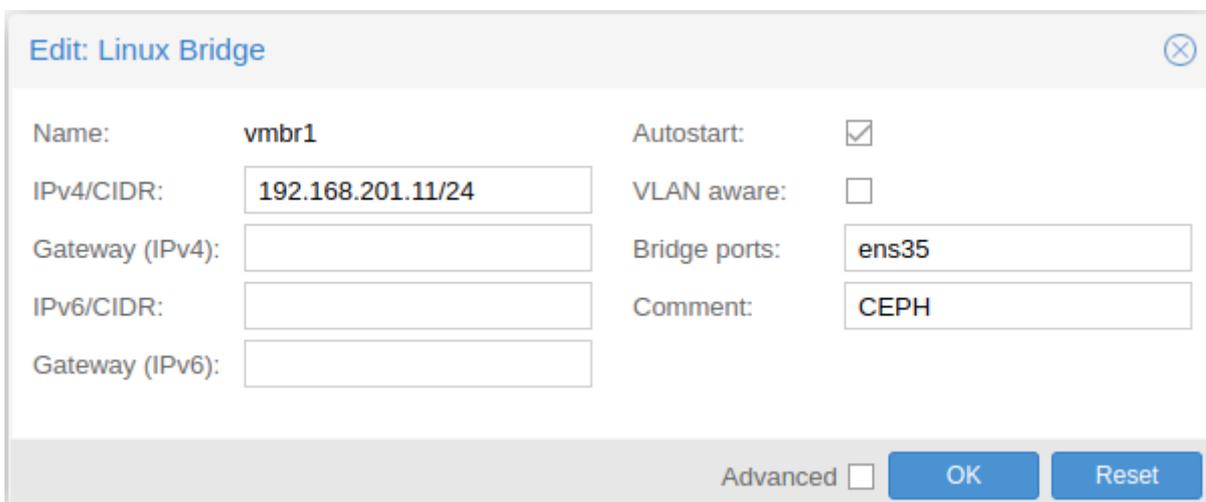
 Back

 Finish



Puis on réitère l'opération sur chacun des noeuds du cluster Proxmox.

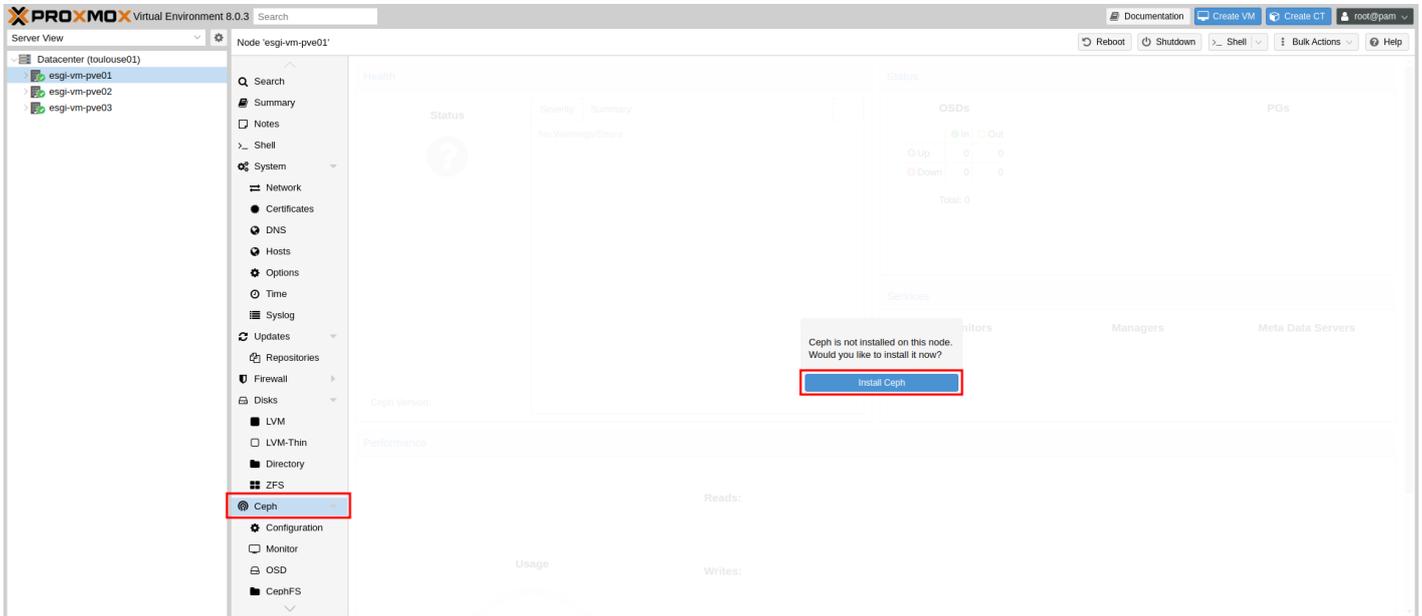
Désormais, il faut ajouter une carte réseau sur chaque noeud pour le **réseau CEPH** (192.168.201.0/24) et ajouté un Linux Bridge sur chacun des noeuds Proxmox :



N'oubliez pas d'appliquer la configuration réseau en appuyant sur **Apply configuration**.

Installation de CEPH sur les noeuds

Une fois les VMs redémarrées avec leur nouveau disque et nouvelle carte réseau, rendez-vous sur le **Node 1** dans l'onglet et cliquer sur **Installer CEPH** :



Puis on peut lancer l'installation de la version actuelle de CEPH (**quincy** dans notre cas) :

Ceph?

"Ceph is a unified, distributed storage system, designed for excellent performance, reliability, and scalability."

Ceph is currently **not installed** on this node. This wizard will guide you through the installation. Click on the next button below to begin. After the initial installation, the wizard will offer to create an initial configuration. This configuration step is only needed once per cluster and will be skipped if a config is already present.

Before starting the installation, please take a look at our documentation, by clicking the help button below. If you want to gain deeper knowledge about Ceph, visit ceph.com.

Hint: The enterprise repository is enabled, but there is no active subscription!

Ceph in the cluster: Could not detect a ceph installation in the cluster

Ceph version to install: quincy (17.2) ▼

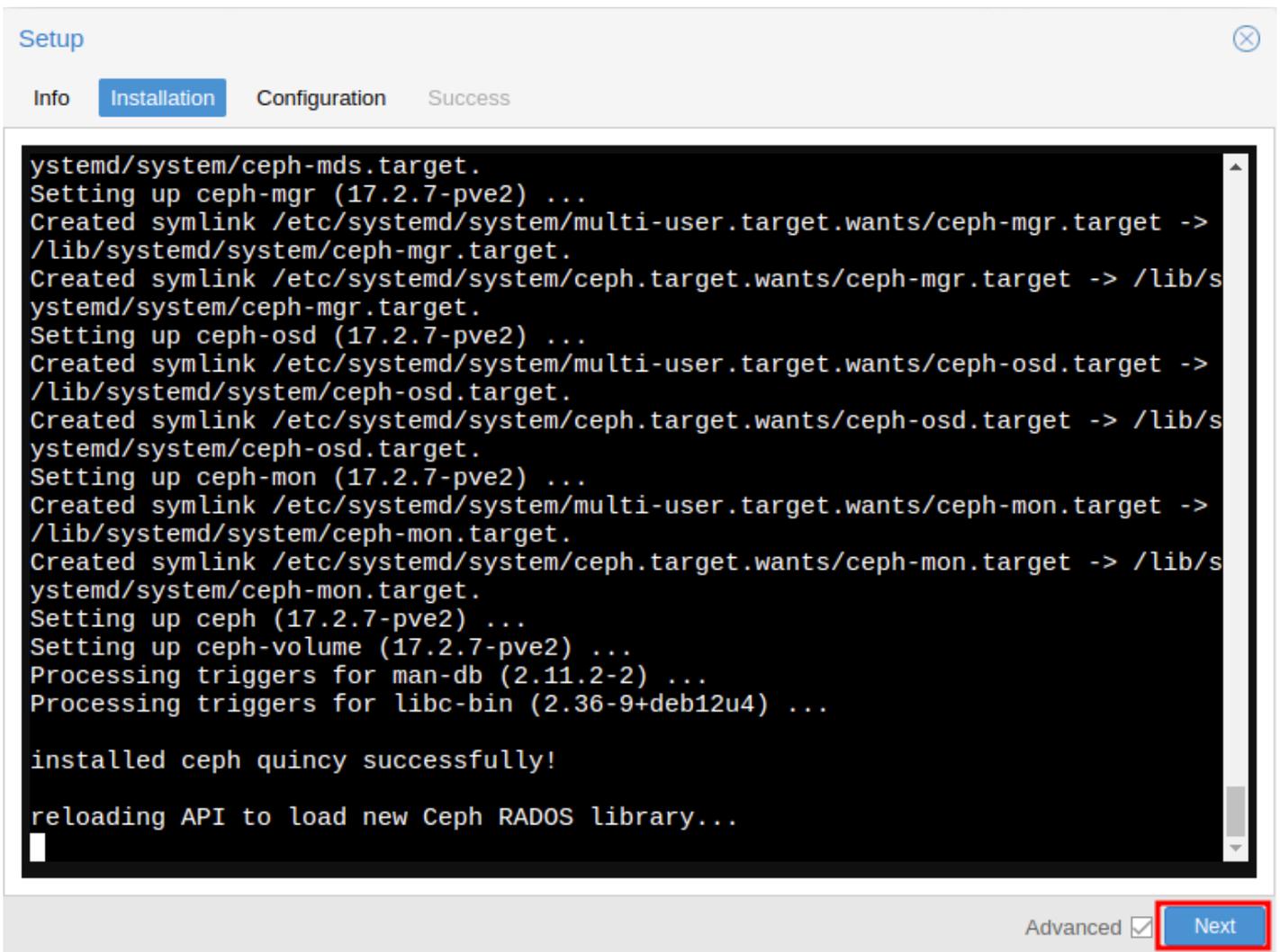
Repository: Enterprise (recommended) ▼

Help

Advanced

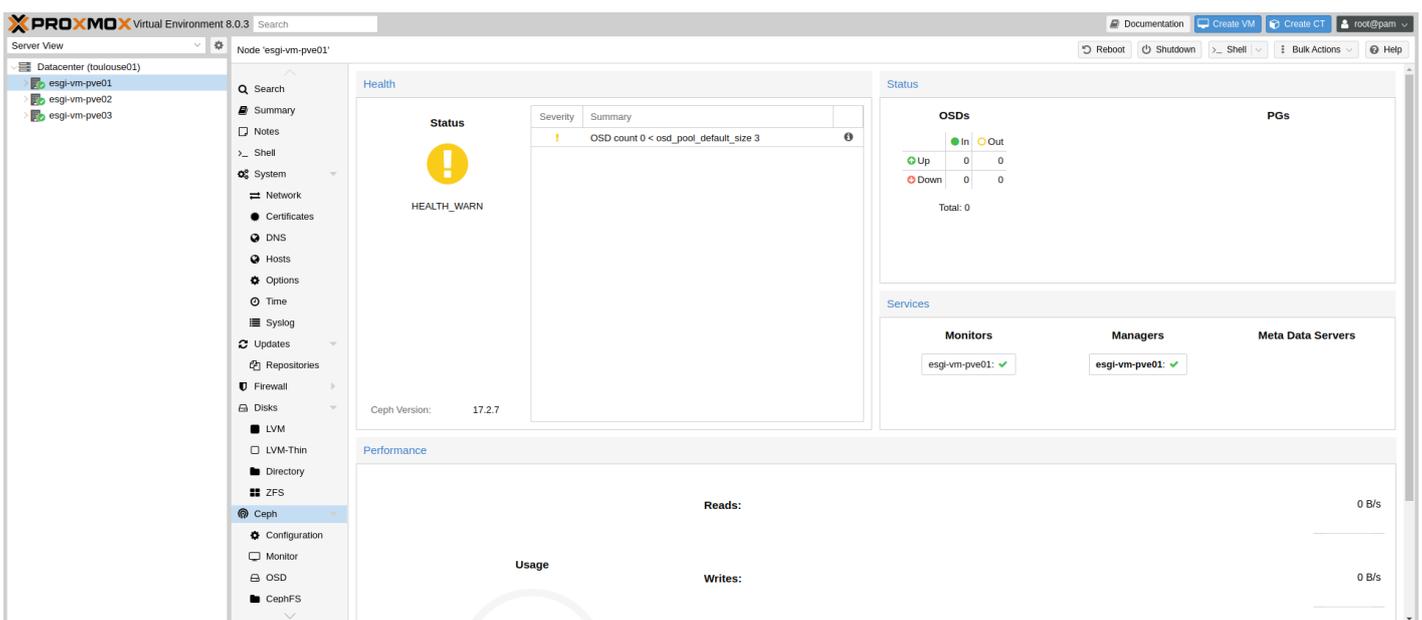
Start quincy installation

Une fois l'installation terminée, on clique sur **Next** :



L'écran de configuration de CEPH devrait s'ouvrir et vous devrez sélectionner l'interface du réseau CEPH.

Vous devriez voir cette configuration :



Pour le moment, l'état de santé n'est pas bon puisque nous avons pas de stockage CEPH défini.

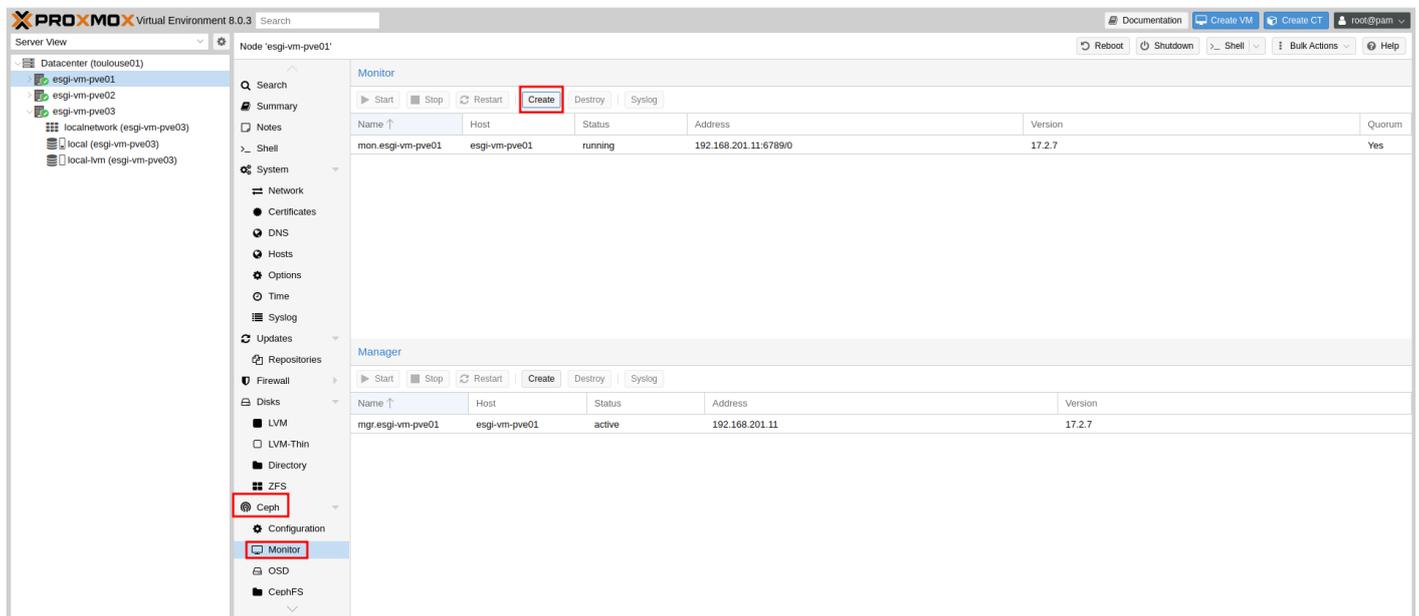
Maintenant, il faut réitérer l'installation de CEPH sur chacun des noeuds du cluster.

Configuration des Moniteurs et Managers

Désormais ce qui est recommandé, c'est de configurer chacun des noeuds avec le rôle **Monitor** et **Manager** pour qu'on puisse avoir de la surveillance du pool même si un des noeud tombe.*

Par défaut, le Node 1 possède déjà les rôles Monitor et Manager.

On commence par aller sur le **Node 1** dans **CEPH > Moniteur** puis **Créer** :



The screenshot shows the Proxmox VE interface for Node 'esgi-vm-pve01'. The left sidebar has 'Ceph' and 'Monitor' highlighted. The main panel shows the 'Monitor' configuration with a 'Create' button highlighted in red. Below it, a table lists the existing monitor:

Name	Host	Status	Address	Version	Quorum
mon.esgi-vm-pve01	esgi-vm-pve01	running	192.168.201.11:6789/0	17.2.7	Yes

Below the Monitor section, the 'Manager' section is visible with a 'Create' button highlighted in red. A table below it shows the existing manager:

Name	Host	Status	Address	Version
mgr.esgi-vm-pve01	esgi-vm-pve01	active	192.168.201.11	17.2.7

On sélectionne d'abord le **Node 2** :



The screenshot shows the 'Create: Monitor' dialog box. The 'Host' dropdown menu is open, and 'esgi-vm-pve02' is selected and highlighted with a red box. A 'Create' button is visible at the bottom right.

Puis le **Node 3** :

Create: Monitor

Host:

Create

Et on fait de même pour le rôle **Manager** qui est juste en dessous avec nos **Node 2 et 3** :

The screenshot shows the Proxmox VE 8.0.3 interface. On the left, a tree view shows the server hierarchy: Datacenter (toulouse01) > esgi-vm-pve01 > local (esgi-vm-pve03). The main panel displays the configuration for Node 'esgi-vm-pve01'. The 'Monitor' role is configured with three nodes:

Name	Host	Status	Address	Version	Quorum
mon.esgi-vm-pve01	esgi-vm-pve01	running	192.168.201.11:6789/0	17.2.7	Yes
mon.esgi-vm-pve02	esgi-vm-pve02	running	192.168.201.12:6789/0	17.2.7	Yes
mon.esgi-vm-pve03	esgi-vm-pve03	running	192.168.201.13:6789/0	17.2.7	Yes

The 'Manager' role is configured with one node:

Name	Host	Status	Address	Version
mgr.esgi-vm-pve01	esgi-vm-pve01	active	192.168.201.11	17.2.7

Create: Manager

Host:

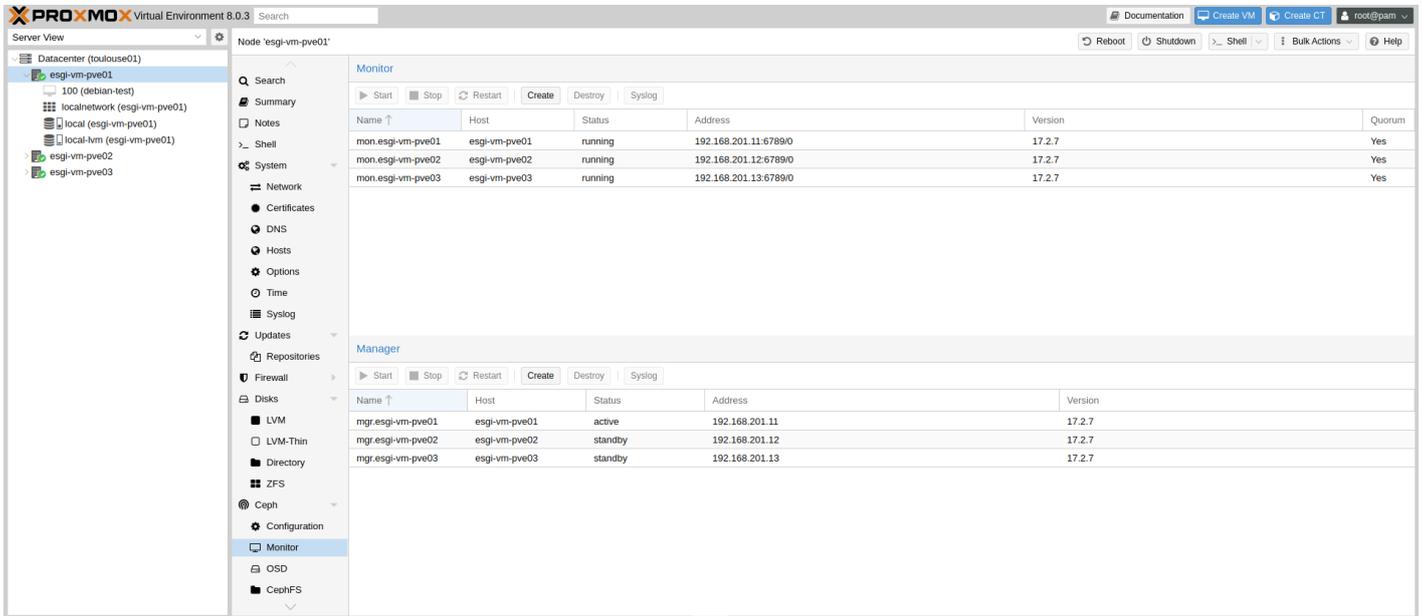
Create

Create: Manager

Host:

Create

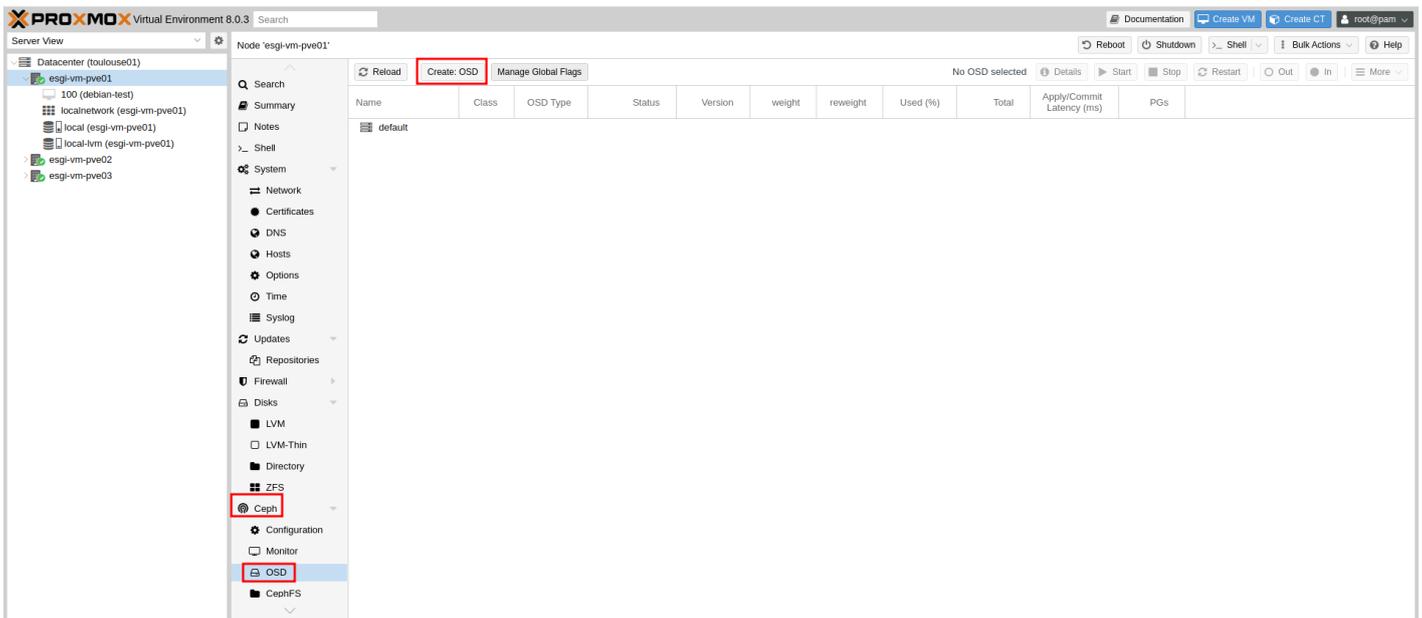
Au final vous devriez avoir cette configuration :



Création des OSDs

Les OSDs sont les disques partagés CEPH.

On va commencer par configurer un OSD sur le **Node 1** en allant dans **CEPH > OSD** puis **Create OSD** :



On peut sélectionner le disque sur lequel on souhaite installer CEPH (**/dev/sdb** dans notre cas) :

Create: Ceph OSD

Disk: DB Disk:
 DB size (GiB):

Note: Ceph is not compatible with disks backed by a hardware RAID controller. For details see [the reference documentation](#).

Advanced

Réitérer l'opération sur chacun des noeuds du cluster.

Vous devriez avoir la configuration suivante :

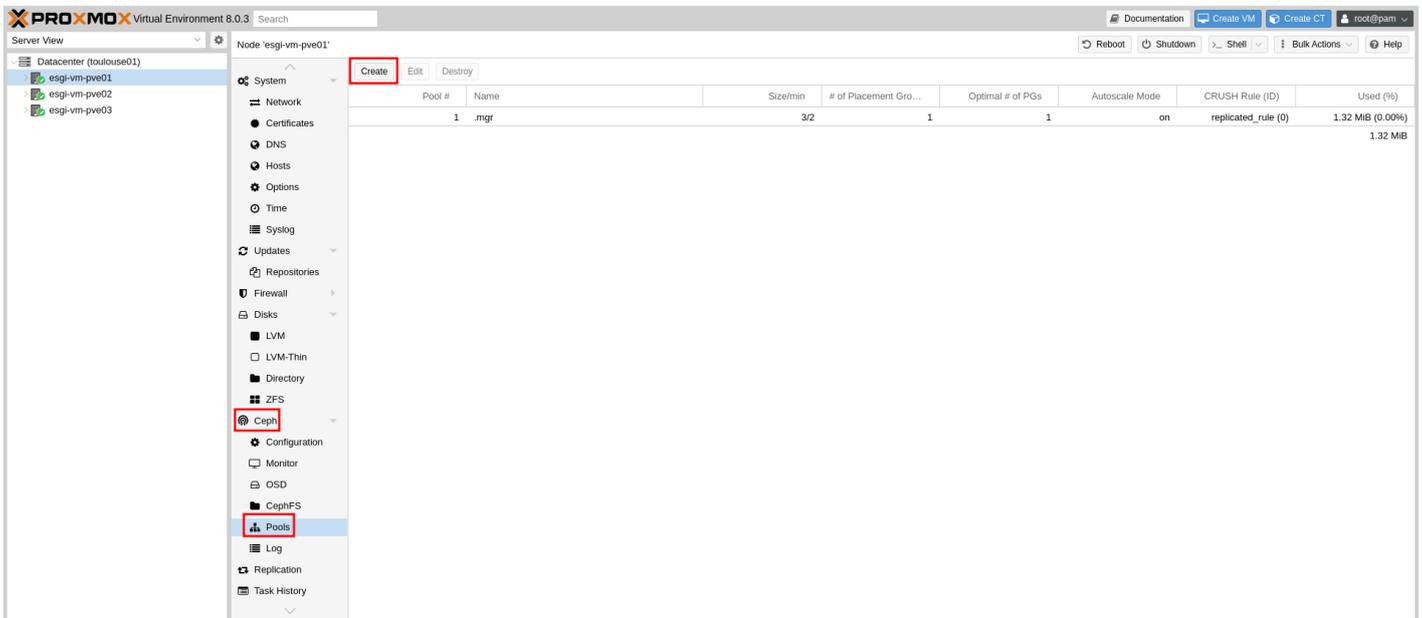
The screenshot shows the Proxmox VE interface with the Ceph OSD configuration for three nodes. The table below represents the data visible in the interface.

Name	Class	OSD Type	Status	Version	weight	reweight	Used (%)	Total	Apply/Commit Latency (ms)	PGs
esgi-vm-pve03	hdd	bluestore	up / in	17.2.7	0.0098	1.00	2.84	10.00 GiB	3 / 3	0
esgi-vm-pve02	hdd	bluestore	up / in	17.2.7	0.0098	1.00	2.84	10.00 GiB	0 / 0	0
esgi-vm-pve01	hdd	bluestore	up / in	17.2.7	0.0098	1.00	2.84	10.00 GiB	0 / 0	0

Création du pool de stockage

L'objectif va être de créer un groupe (pool) d'OSDs qui sera utilisé par nos VMs.

Pour cela, on se rend sur le **Node 1** dans **CEPH > Pools** puis **Create** :



On choisit un **nom** ainsi que la **taille** du pool qui correspond au nombre d'OSD présent dans le pool :

Create: Ceph Pool

Name: **store-ceph-01** PG Autoscale Mode: **on**

Size: **3** Add as Storage:

Min. Size: **2** Target Ratio: **0.0**

Crush Rule: **replicated_rule** Target Size: **0** GiB

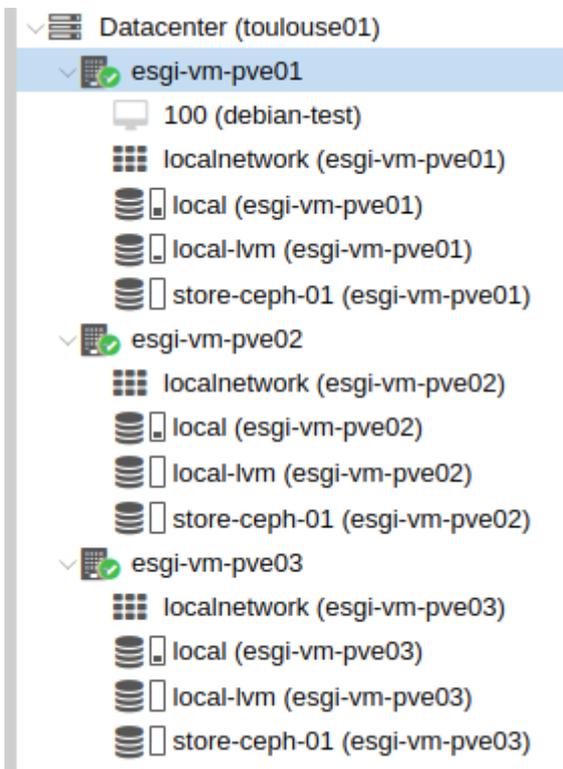
of PGs: **128** Target Ratio takes precedence.

Min. # of PGs: **0**

Help Advanced Create

La **taille minimale** (Min Size) signifie que le pool continuera de fonctionner avec seulement 2 noeuds sur 3 disponibles.

Le stockage CEPH devrait apparaître sur chacun des noeuds :



Installation de la Haute Disponibilité (HA)

Création de la VM pour la HA

Nous allons créer une VM qui prendra en charge la **HA**, ce qui signifie qu'elle sera automatiquement répliquée sur le une autre noeud si le noeud principal de VM venait à devenir défaillant.

La migration ne sera effective que si les lecteurs **CD** sont supprimés des VMs sinon elle échouera systématiquement.

Pour cela nous allons créer une VM de manière standard mais nous allons l'installer sur le **stockage CEPH** :

Create: Virtual Machine

General OS System **Disks** CPU Memory Network Confirm

scsi0

Disk Bandwidth

Bus/Device: SCSI 0 Cache: Default (No cache)

SCSI Controller: VirtIO SCSI single Discard:

Storage: store-ceph-01 IO thread:

Disk size (GiB): 5

Format: Raw disk image (raw)

+ Add

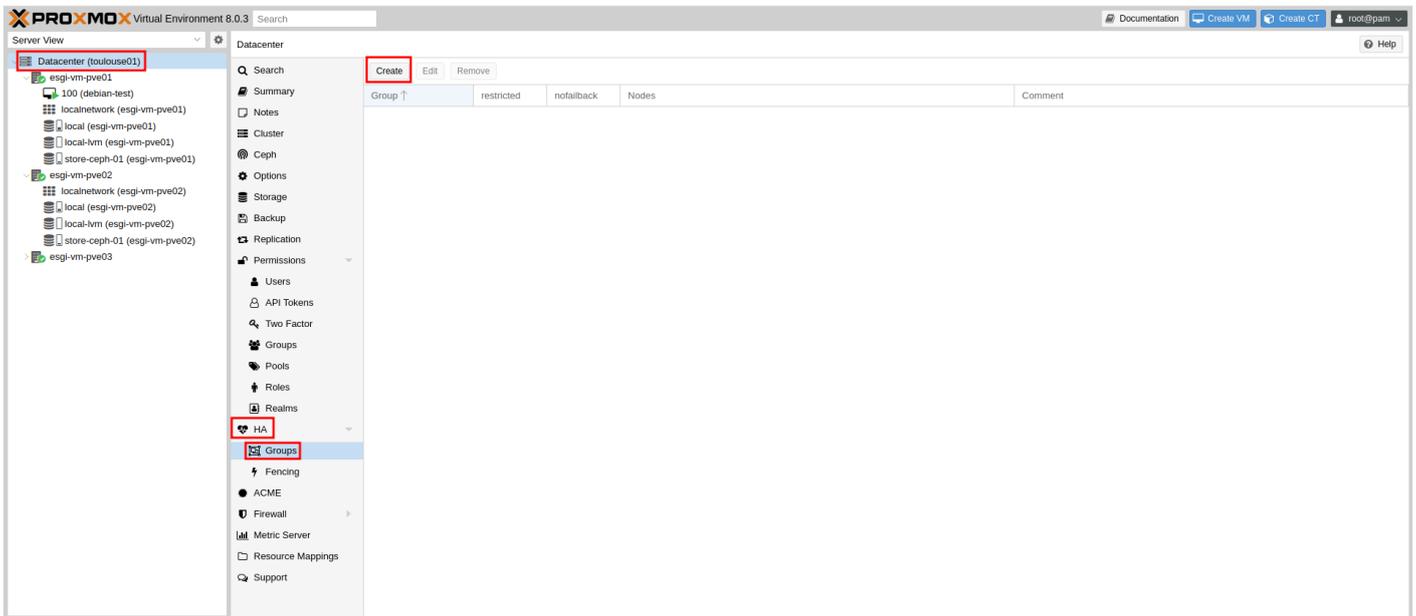
Help Advanced Back Next

Vous pouvez créer et installer votre VM normalement, la configuration de la HA viendra après ces étapes.

Configuration de la HA sur le cluster

Désormais, on va configurer la HA sur le cluster.

Pour cela, rendez-vous dans **Cluster > HA > Groups** puis **Create** :



Ensuite, donnez un nom à votre groupe de VM du **Node 1** pour sa configuration HA et définissez les priorités :

Create: HA Group

ID: restricted:
nofailback:

Comment:

<input checked="" type="checkbox"/>	Node ↑	Memory usage %	CPU usage	Priority
<input checked="" type="checkbox"/>	esgi-vm-pve01	55.2 %	2.4% of 4 CPUs	3
<input checked="" type="checkbox"/>	esgi-vm-pve02	57.3 %	4.0% of 2 CPUs	2
<input checked="" type="checkbox"/>	esgi-vm-pve03	57.9 %	3.5% of 2 CPUs	1

Plus la priorité est haute, plus les VMs appartenant à ce groupe auront tendance à se rendre sur le noeud correspondant.

On peut maintenant créer les groupes HA pour les **Nodes 2 et 3** :

Edit: HA Group

ID: pool-ha-02 restricted:

nofailback:

Comment:

<input checked="" type="checkbox"/>	Node ↑	Memory usage %	CPU usage	Priority
<input checked="" type="checkbox"/>	esgi-vm-pve01	55.4 %	2.6% of 4 CPUs	1
<input checked="" type="checkbox"/>	esgi-vm-pve02	57.4 %	3.7% of 2 CPUs	3
<input checked="" type="checkbox"/>	esgi-vm-pve03	57.8 %	3.6% of 2 CPUs	2

Edit: HA Group

ID: pool-ha-03 restricted:

nofailback:

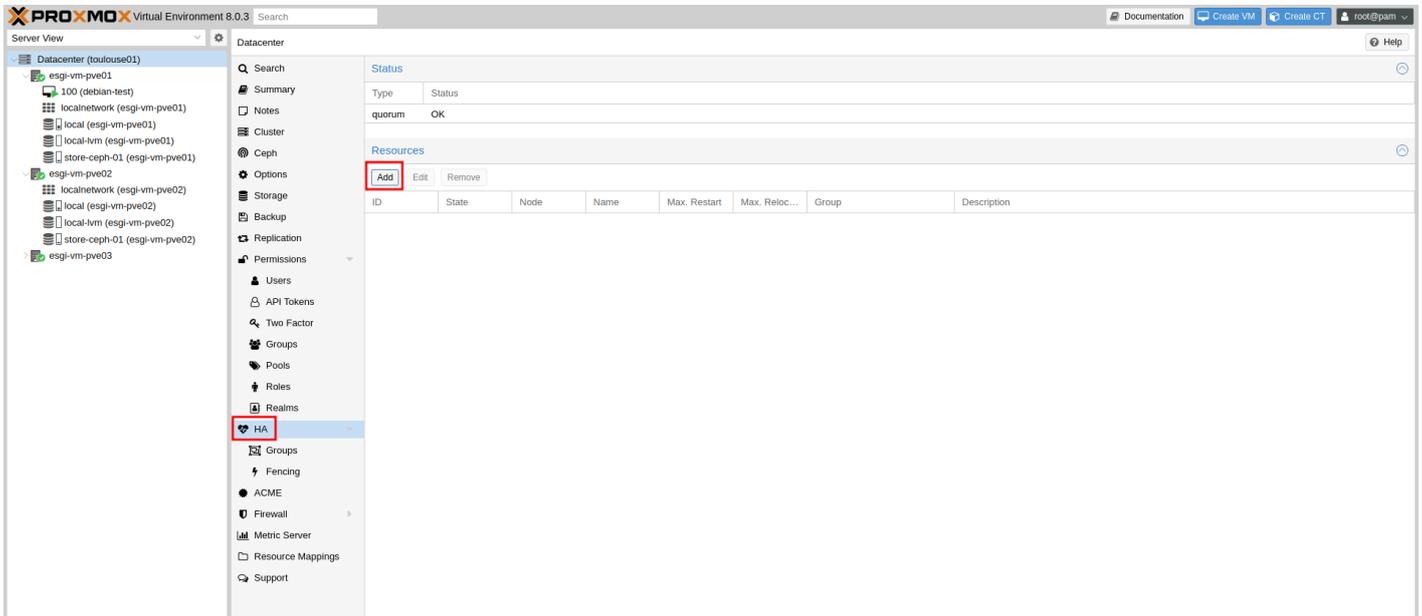
Comment:

<input checked="" type="checkbox"/>	Node ↑	Memory usage %	CPU usage	Priority
<input checked="" type="checkbox"/>	esgi-vm-pve01	55.5 %	2.3% of 4 CPUs	2
<input checked="" type="checkbox"/>	esgi-vm-pve02	57.4 %	4.0% of 2 CPUs	1
<input checked="" type="checkbox"/>	esgi-vm-pve03	58.2 %	3.0% of 2 CPUs	3

Ajout des ressources pour la HA

L'objectif va être de définir les VMs qui devront utiliser notre configuration HA.

Pour cela, rendez-vous dans **HA** Puis **Add** :



On peut sélectionner notre VM **debian-test** (mettez la VM où vous souhaitez activer la HA), sélectionner le groupe HA que l'on souhaite ici **pool-ha-01** et cliquer sur **Add** :

Add: Resource: Container/Virtual Machine

VM: **100** Group: **pool-ha-01**

Max. Restart: **1** Request State: **started**

Max. Relocate: **1**

Comment:

Add

La HA est désormais fonctionnelle sur votre VM.

On peut le tester en arrêtant le **Node 1** et la VM doit être migrée sur le **Node 2**.

Revision #6

Created 20 February 2024 10:54:21 by Elieroc

Updated 20 February 2024 16:44:24 by Elieroc