Système de stockage Ceph pour une infrastructure de virtualisation à haute disponibilité

Hervé Ballans

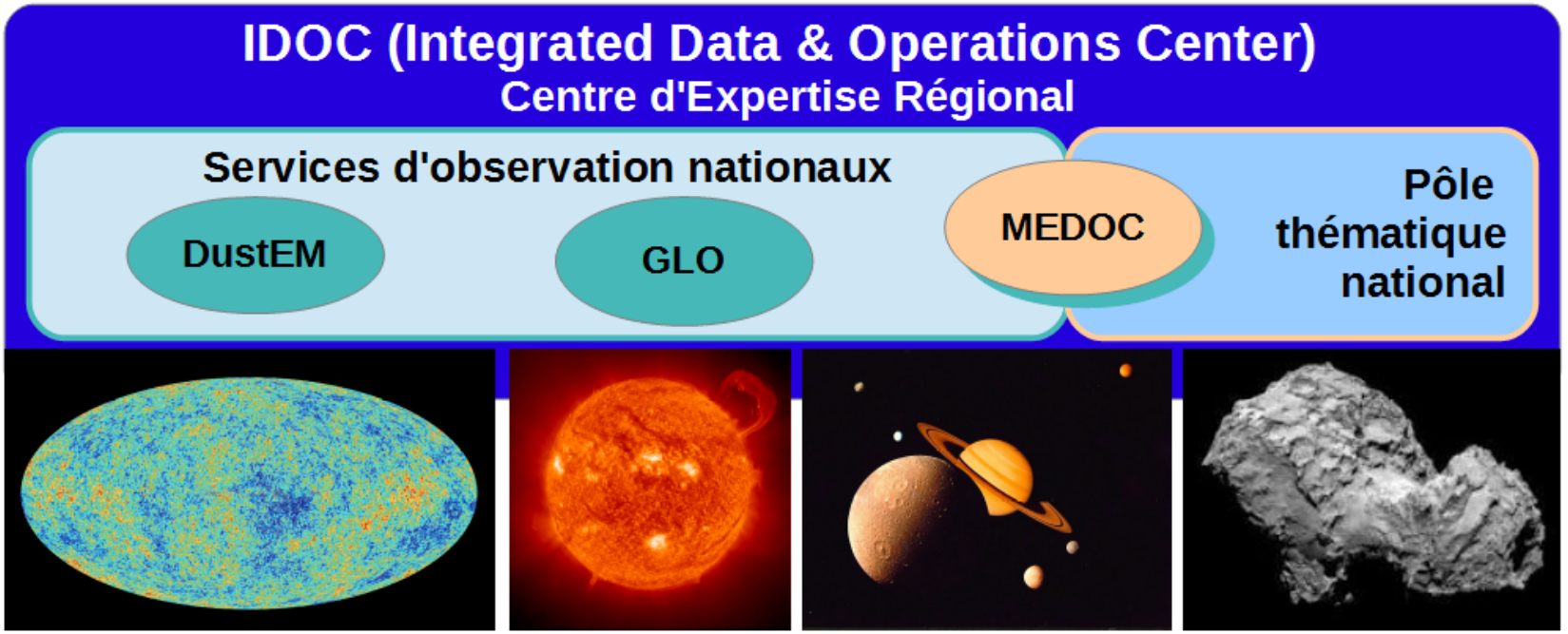
Journées Mathrice

14 *octobre* 2015

Plan de la présentation

Plan de la présentation

Centre de Données Spatiales



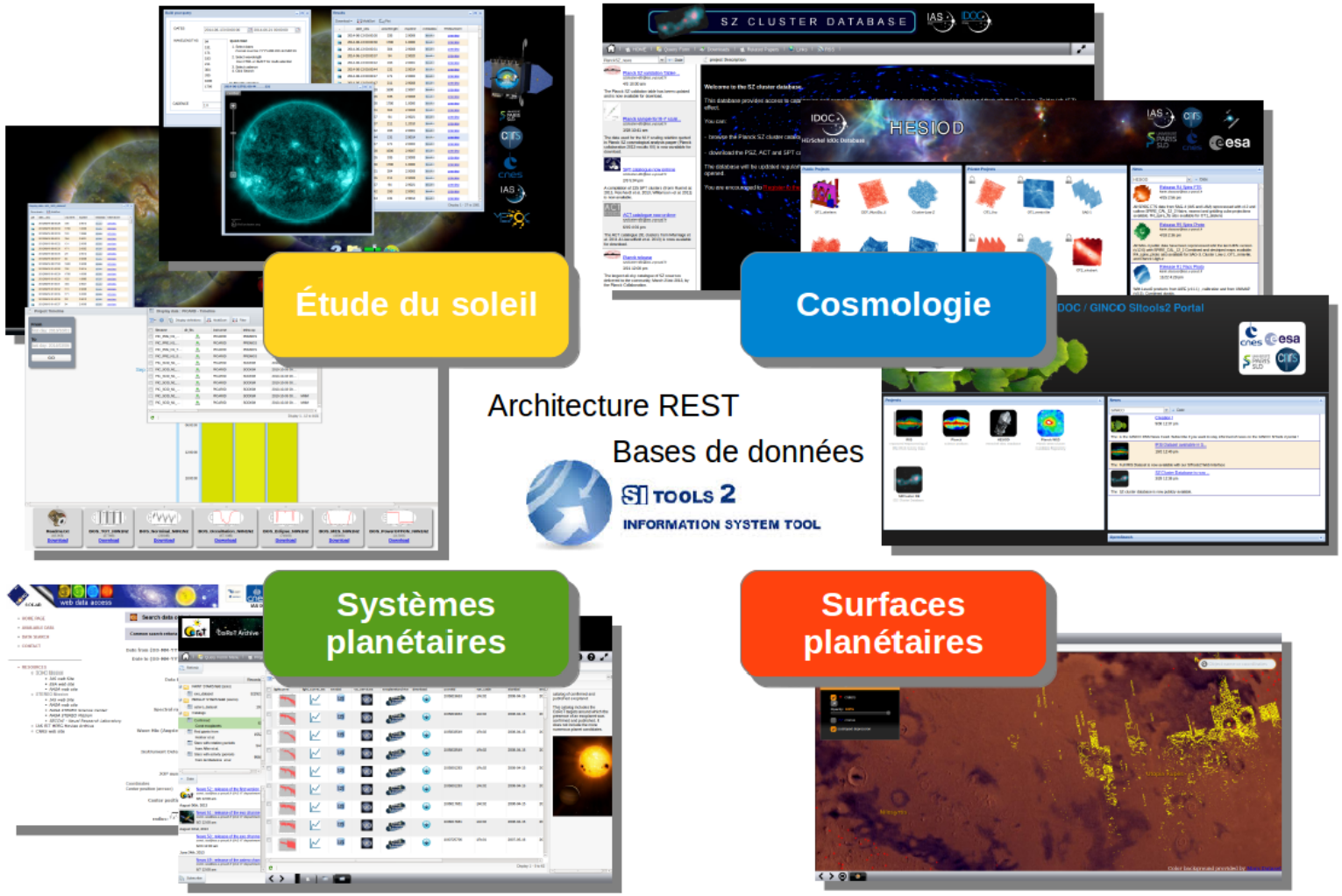
Données scientifiques
provenant d'observations de satellites dans l'espace

Données : 450 To

Métadonnées
Informations sur les données pour leur archivage

Bases de données : 600 Go

Interfaces d'accès aux données



Interfaces d'accès aux données

1 instance = 1 interface d'accès SITools2 (projet mission ou portail thématique)
1 instance = 1 virtual host apache
Toutes les instances (une dizaine) sont exécutées sur 2 serveurs différents (avec des versions différentes de SITools)
Problèmes rencontrés :

- disponibilité globale en cas de maintenance sur le serveur
- impact possible de la maintenance d'une instance sur les autres instances
- extensibilité (ressources, stockage,...)

Besoins

- Séparation des instances
  - 1 instance = 1 VM
- Disponibilité des services
  - Bascule automatique d'une VM en cas de problème sur un noeud
- Faible volumétrie initiale
  - 1 VM instance nécessite peu de stockage (serveur apache/SITools2)
  - Les données (gros volumes) sont stockées à part
- Extensibilité
  - Services et stockage

Besoins

Chiffres :

- une douzaine de VM
  - -> 9 serveurs SITools2
  - 1 serveur de cartographie (mapserver)
  - 1 serveur de base de données (PostgreSQL)
  - 1 serveur de logs centralisé (rsyslog)
- 8 Go RAM et entre 4 et 8 CPUs par VM
- Une dizaine de To net pour le stockage au total

Les snapshots et les backups se font en réseau sur un système de stockage distant (accessible en NFS)

Groupe de travail stockage distribué

Réflexion mutualisée avec d'autres laboratoires et constitution d'un groupe de travail (dans le cadre du Labex P2IO) depuis janvier 2015

- Retours d'expérience
  - Expertise Ceph (CEA)
- Plate-forme de test
  - Mise en œuvre d'une maquette
  - Configuration et tests sur Ceph
- Pérennisation
  - Documentation
- Consolidation
  - Validation d'un choix technologique pour les projets du Labex (VirtualData)

Virtualisation avec Proxmox

- Solution maîtrisée
  - Expertise dans l'administration d'un cluster proxmox (3 serveurs avec quorum)
  - 50 VMs
  - Stockage NFS vers système HITACHI
- Solution libre
  - Utilisation de Proxmox VE
- Based on Debian
  - Système homogène avec nos autres serveurs
- Large communauté
  - Très nombreux retours d'expérience
  - Références institutionnelles

Proxmox HA

Au minimum 3 serveurs pour faire de la haute disponibilité avec Proxmox
Configuration du fencing avec IDRAC (ou IPMI)

Stockage distribué avec Ceph

- Scalabilité
- Solution libre
  - sous licence libre
- Projet actif
  - Mises à jours régulières et 1 version majeure tous les 3 mois
- Documentation complète
  - Site ceph.com et forums
- Large communauté
  - Très nombreux retours d'expérience
  - Références institutionnelles (dont CERN)

Proxmox intègre Ceph

Roadmap : http://pve.proxmox.com/wiki/Roadmap#Proxmox_VE_3.2

Proxmox VE 3.2

Released 10.03.2014: See Downloads

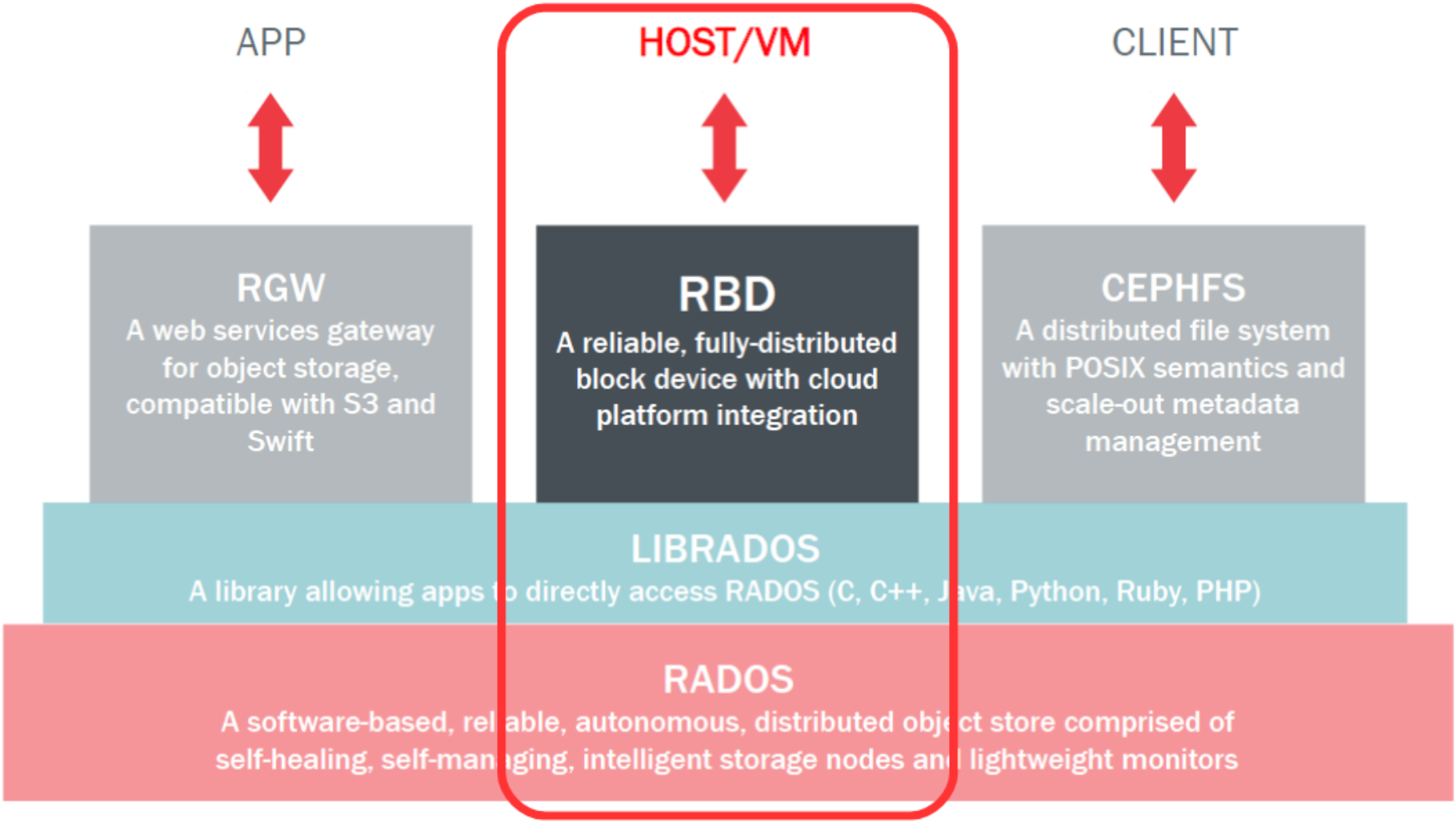
- improved SPICE support
  - spiceterm: console for OpenVZ and host
  - add new console option to datacenter.cfg (java applet vs. spice)
  - add multi-monitor support
  - GUI: use split-button to easily select SPICE or VNC
  - more details on http://pve.proxmox.com/wiki/SPICE
- update qemu to 1.7.0
  - add 'pvscsi' to the list of scsi controllers (emulate the VMware PVSCSI device)
  - add 'lsi53c810' to the list of scsi controllers (supported on some very old Windows NT versions)
  - add 'vmxnet3' to the list of available network card models (emulate VMware paravirtualized network card)
  - add drive option 'discard'
  - add support for new qemu throttling burst max parameters
  - improved live backup
- pve-kernel-2.6.32-27-pve: 2.6.32-121
  - update to vzkernel-2.6.32-042stab084.20.src.rpm
  - update e1000, igb, ixgbe, netxtreme2, megaraid_sas
  - include latest ARECA RAID drivers
  - update Broadcom bnx2/bnx2x drivers to 7.6.62
  - update aacraid to aacraid-1.2.1-30300.src.rpm
- Ceph Server (Technology Preview)
  - new GUI to manage Ceph server running on PVE nodes
  - more details on http://pve.proxmox.com/wiki/Ceph_Server
- added Open vSwitch support (Technology Preview)
- Optional 3.10 Kernel (based on RHEL7 beta, currently without OpenVZ support, for testing only)

Proxmox intègre Ceph

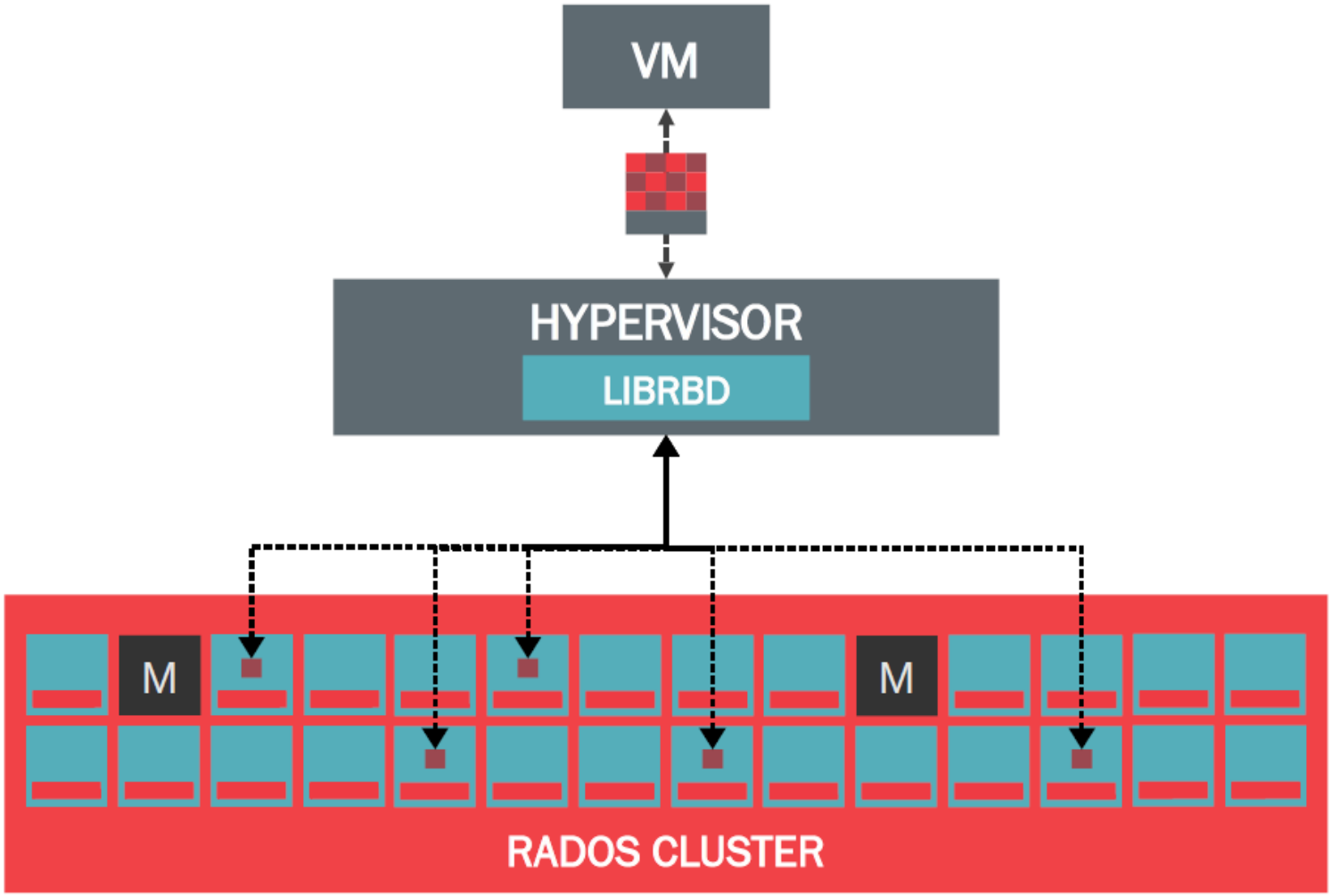
Support de Ceph client et serveur

- Ceph serveur pour la configuration et la gestion du stockage (démons OSD)
- Ceph client pour accéder au stockage pour les VMs

Accès disque en mode "bloc" : RADOS



Accès disque en mode "bloc" : librbd



Ceph MON

Il est conseillé d'installer un nombre impair de MON (pour éviter notamment le phénomène *"split-brain"*)
Sur ceph.com :
"It is advisable to run an odd-number of monitors but not mandatory. An odd-number of monitors has a higher resiliency to failures than an even-number of monitors...For an initial deployment of a multi-node Ceph cluster, it is advisable to deploy three monitors, increasing the number two at a time if a valid need for more than three exists."

On choisit 3 moniteurs

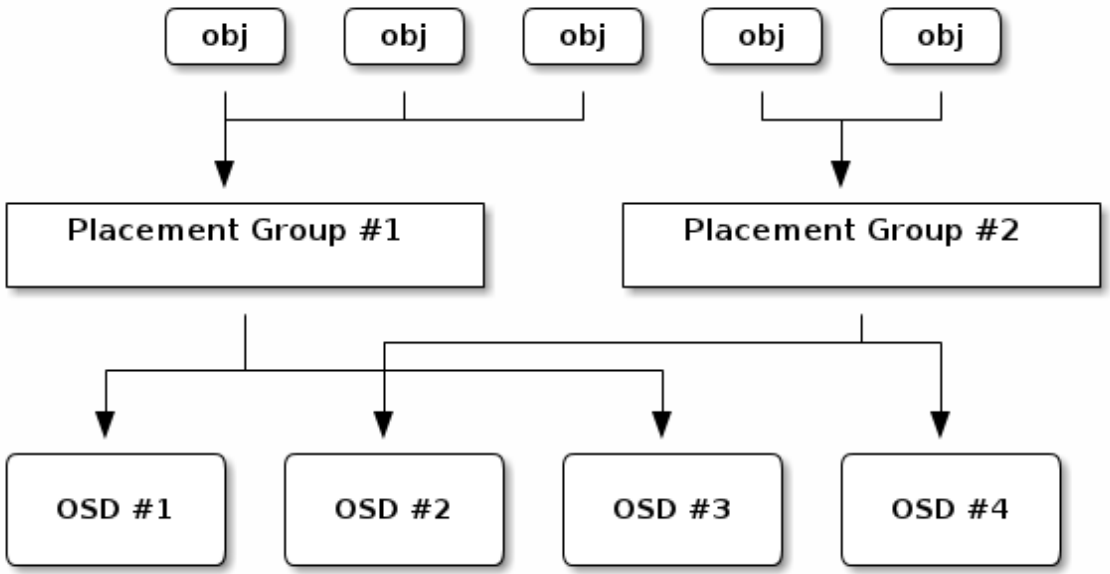
Ceph OSD

1 disque physique = 1 OSD

On choisit 2 réplicats

1 objet sera copié sur 2 OSD : un OSD primaire et un secondaire (selon un processus synchrone)

Placement Groups et Crush Map



La CRUSH Map contient la carte de réplication des données sur les OSDs et permet, pour un PG donné, de déterminer automatiquement les OSDs de stockage.

Plan de la présentation

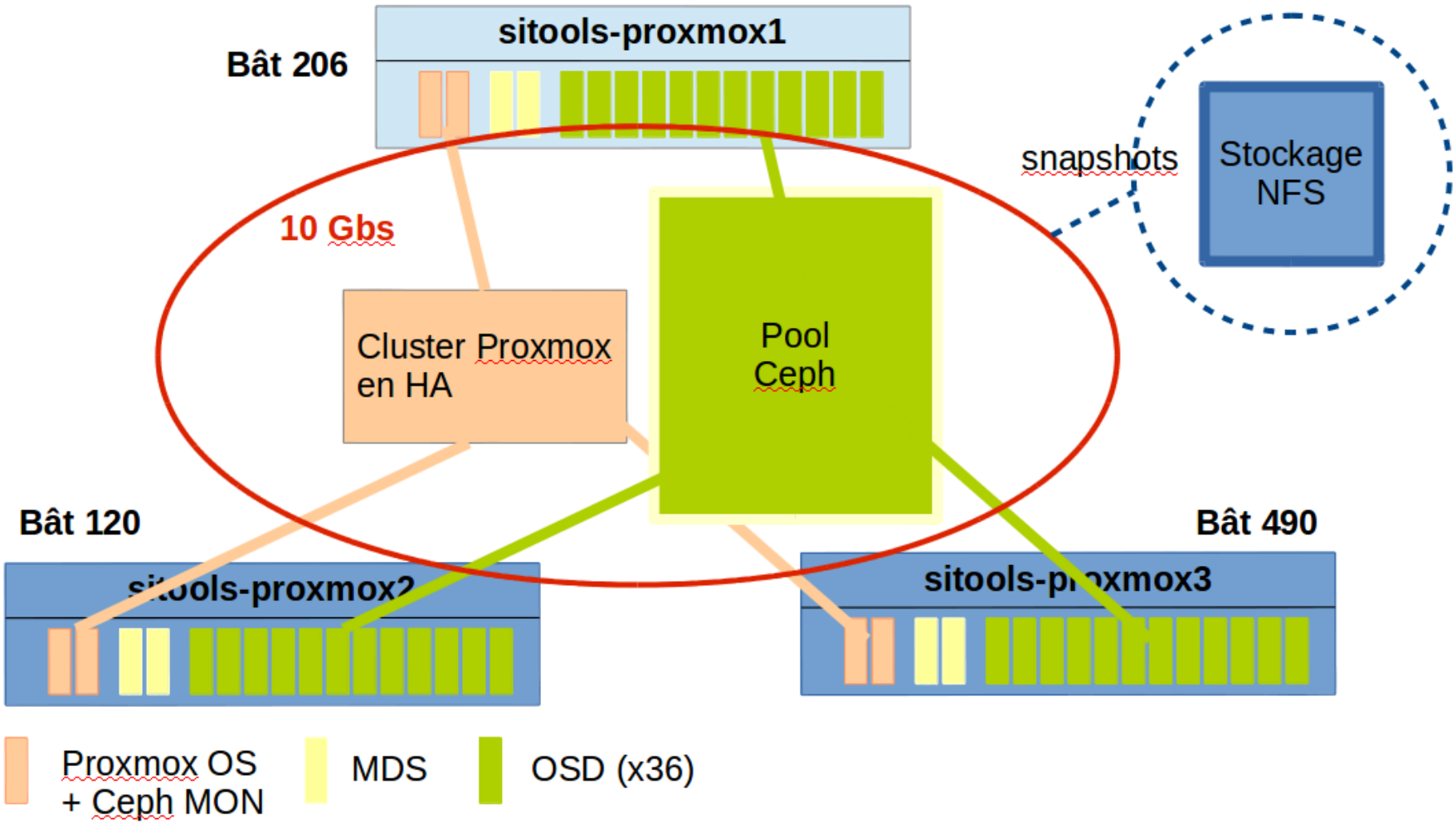
Proxmox et Ceph sur le même serveur ?

2 politiques :

- Sur des serveurs différents
- Sur les mêmes serveurs (voir doc de Proxmox)

Au vu des besoins évoqués précédemment, on choisit d'installer Proxmox et Ceph sur les mêmes serveurs, en prenant soin :

- de séparer les réseaux administration Proxmox & management des VMs du réseau de stockage Ceph
- de réserver les ressources nécessaires à Ceph (CPUs et mémoire)


Bât 206
sitools-proxmox1
snapshots
Stockage NFS
10 Gbs
Cluster Proxmox en HA
Pool Ceph
Bât 120
Bât 490
sitools-proxmox3
Proxmox OS + Ceph MON
MDS
OSD (x36)

Administration à distance

DELL R820 Facilement administrable à distance :

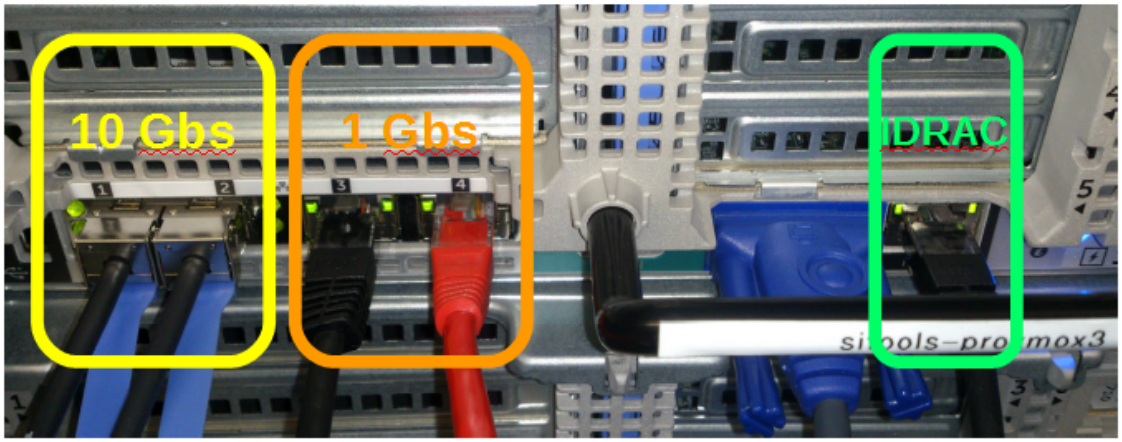
- IDRAC Enterprise 7 (voir annexe A)
  - Envoi d'alertes par mail
  - Mode console (https://*ip-idrac-serveur*/console)
- OpenManage (voir annexe B)
  - DELL Lifecycle controller
  - Administration du contrôleur PERC H710 via un navigateur (port 1311)

Caratéristique techniques

Pour chaque serveur :

- 4 processeurs 6 coeurs, soit 24 CPU (soit 48 en hyperthreadés) - Xeon E5-4607
- 64 Go RAM RDIMM 1600MHz
- 2 X disques SAS 6Gb/s 15KTpm de 300Go (proxmox VE + Ceph MON)
- 2 X disques SSD SAS 6Gb/s de 200Go (Ceph MDS)
- 12 X disques SAS 6Gb/s 10KTpm de 600Go (Pool Ceph OSDs)
- 2 X contrôleurs Ethernet 10Gb SFP+ (dont 1 dédié au pool ceph)
- 2 X contrôleurs Ethernet 1 Gb (dont 1 dédié au cluster Proxmox VE)
- 1 X contrôleur IDRAC Enterprise

Caratéristiques techniques



Installation jumelée Proxmox + Ceph

Point de départ :

- ISO image : https://pve.proxmox.com/wiki/Downloads
- Documentation : https://pve.proxmox.com/wiki/Ceph_Server

Dédicaces :)

- Jeff Renaudat (IE CDD CNES/INSU)
- Jérémie Jacob (IE CSNSM/IN2P3)

Configuration du cluster Proxmox VE

- A partir de l'ISO, installation des noeuds
  -> très simple "à la Debian"
  https://pve.proxmox.com/wiki/Quick_installation
- Installation de Proxmox VE en cluster HA
  https://pve.proxmox.com/wiki/Proxmox_VE_2.0_Cluster
  En résumé, on créé la configuration cluster sur son premier nœud puis on intègre les autres nœuds
  Sur le noeud 1 :
  ```
  # pvecm create cluster
  ```
  Puis sur les autres noeuds :
  ```
  # pvecm add ip_noeud_1
  ```

Configuration du cluster Proxmox VE

`/etc/pve/cluster.conf`

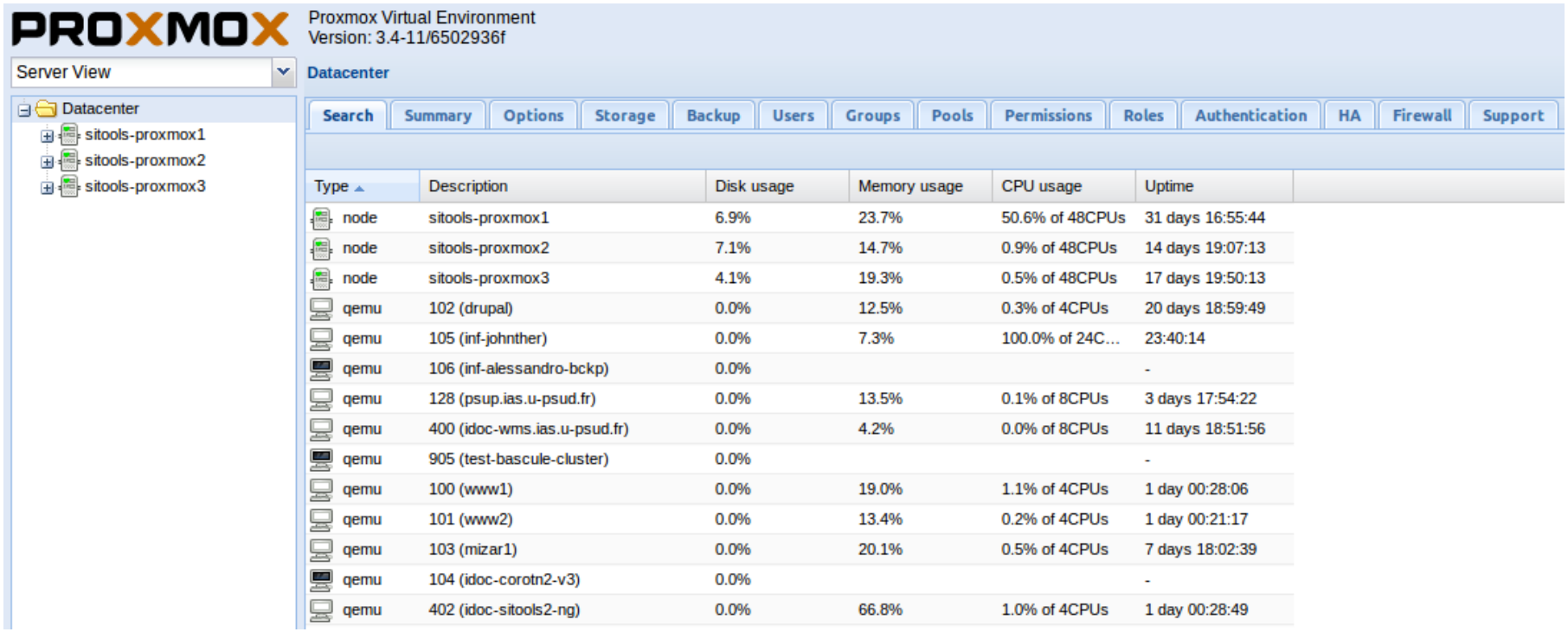
- Configuration du fencing -> via IDRAC ou IPMI
- Configuration de la politique de fail-over pour les VMs
- Mise à jour et validation du fichier de configuration :
  `# ccs_config_validate -v -f /etc/pve/cluster.conf.new`
  Note : à chaque modification, penser à incrémenter le numéro de série du fichier

Cluster Proxmox VE OK !

```
pveceph  pvecm    pvectl
root@sitools-proxmox1:~# pvecm status
Version: 6.2.0
Config Version: 32
Cluster Name: cluster
Cluster Id: 13364
Cluster Member: Yes
Cluster Generation: 2636
Membership state: Cluster-Member
Nodes: 3
Expected votes: 3
Total votes: 3
Node votes: 1
Quorum: 2
Active subsystems: 6
Flags:
Ports Bound: 0 177
Node name: sitools-proxmox1
Node ID: 1
Multicast addresses: 239.192.52.104
Node addresses: [redacted]
root@sitools-proxmox1:~#
root@sitools-proxmox1:~# pvecm nodes
Node  Sts   Inc   Joined               Name
   1   M   2588   2015-09-11 17:24:32  sitools-proxmox1
   2   M   2636   2015-09-28 15:12:59  sitools-proxmox2
   3   M   2628   2015-09-25 14:30:00  sitools-proxmox3
root@sitools-proxmox1:~#
```

Cluster Proxmox VE GUI

Accessible sur un navigateur via le port 8006 (par défaut)



Type	Description	Disk usage	Memory usage	CPU usage	Uptime
node	sitools-proxmox1	6.9%	23.7%	50.6% of 48CPUs	31 days 16:55:44
node	sitools-proxmox2	7.1%	14.7%	0.9% of 48CPUs	14 days 19:07:13
node	sitools-proxmox3	4.1%	19.3%	0.5% of 48CPUs	17 days 19:50:13
qemu	102 (drupal)	0.0%	12.5%	0.3% of 4CPUs	20 days 18:59:49
qemu	105 (inf-johnther)	0.0%	7.3%	100.0% of 24C...	23:40:14
qemu	106 (inf-alessandro-bckp)	0.0%			-
qemu	128 (psup.ias.u-psud.fr)	0.0%	13.5%	0.1% of 8CPUs	3 days 17:54:22
qemu	400 (idoc-wms.ias.u-psud.fr)	0.0%	4.2%	0.0% of 8CPUs	11 days 18:51:56
qemu	905 (test-bascule-cluster)	0.0%			-
qemu	100 (www1)	0.0%	19.0%	1.1% of 4CPUs	1 day 00:28:06
qemu	101 (www2)	0.0%	13.4%	0.2% of 4CPUs	1 day 00:21:17
qemu	103 (mizar1)	0.0%	20.1%	0.5% of 4CPUs	7 days 18:02:39
qemu	104 (idoc-corotn2-v3)	0.0%			-
qemu	402 (idoc-sitools2-ng)	0.0%	66.8%	1.0% of 4CPUs	1 day 00:28:49

La commande pveceph

For the use in the specific Proxmox VE architecture we use pveceph. Proxmox VE provides a distributed file system (pmxcfs) to store configuration files.
We use this to store the Ceph configuration. The advantage is that all nodes see the same file, and there is no need to copy configuration data around using ssh/scp. The tool can also use additional information from your Proxmox VE setup.
Tools like ceph-deploy cannot take advantage of that architecture.

```
# ls -l /etc/pve
```

Configuration du pool RBD Ceph

- Création du réseau dédié à Ceph (sur carte 10 Gbs)
  `/etc/network/interfaces`
  Par exemple sur le réseau privé 10.10.10.0/24
- Installation des paquets Ceph (version Firefly dans le cas de Proxmox 3.4)
  `pveceph install --version firefly`
- Sur le premier nœud, création du fichier initial de configuration de ceph
  `pveceph init --network 10.10.10.0/24`
  génère le fichier `/etc/pve/ceph.conf`

Configuration du pool RBD Ceph

- Création du moniteur Ceph (MON) sur tous les noeuds
  `pveceph createmon`
- Création des OSDs sur chaque disque dédié au stockage des données
  Par exemple si on a 12 disques des stockage sur notre serveur :
  `for i in {c..n}; do pveceph createosd /dev/sd$i; done`
  - partitionne le disque
  - créé le système de fichier
  - démarre le démon OSD
  - Rajoute l'OSD dans la crush map
- Création du pool Ceph Pool rbd avec (dans notre cas) 2048 pgs

Configuration du pool RBD Ceph

```
/etc/pve/storage.cfg
rbd:  mon-pool-ceph
        monhost 10.10.10.1 ; 10.10.10.2 ; 10.10.10.3
        pool rbd
        content images
        username admin
```

Configuration du client Ceph

Pour que Proxmox VE puisse stocker les VMs sur le pool rbd, il reste juste à partager la clé de stockage sur tous les noeuds :

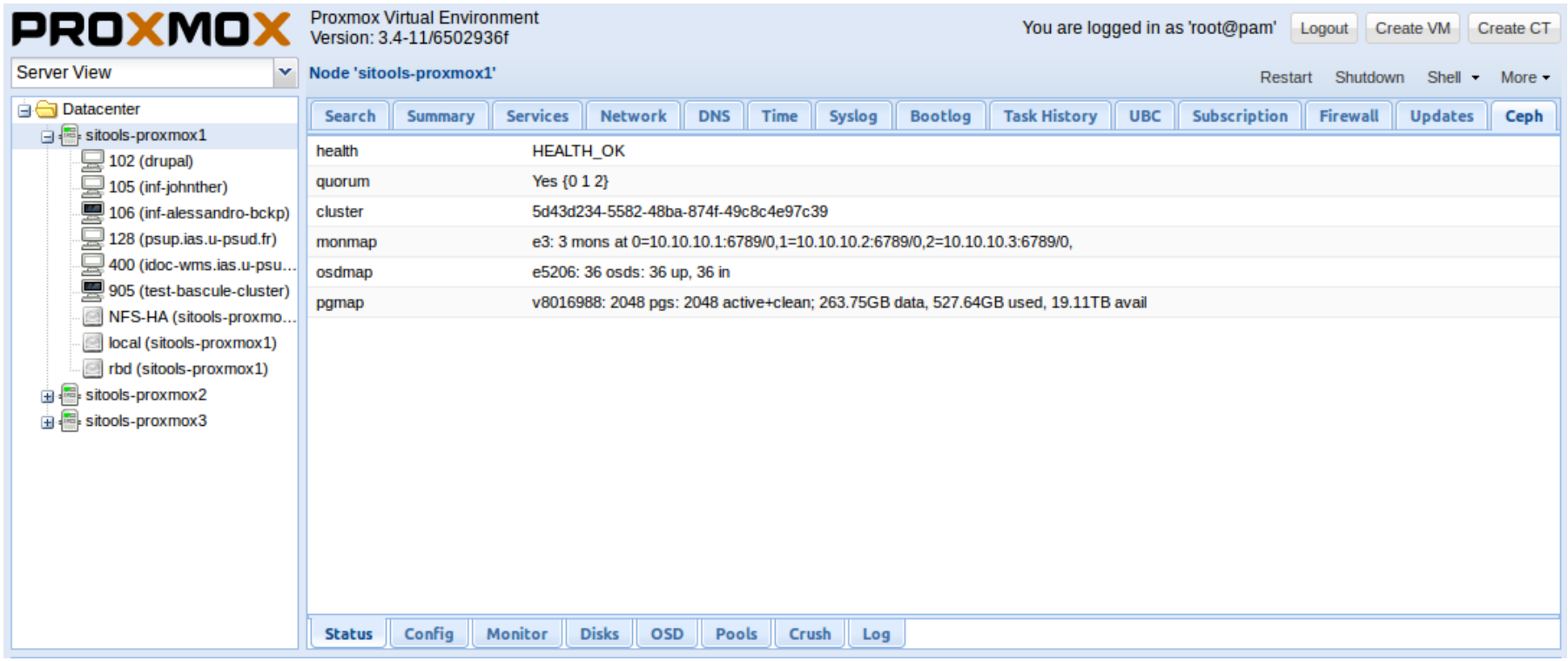
```
cp /etc/ceph/ceph.client.admin.keyring
                    /etc/pve/priv/ceph/mon-pool-ceph.keyring
```

Note : attention au nom de fichier = nom du pool + *.keyring*

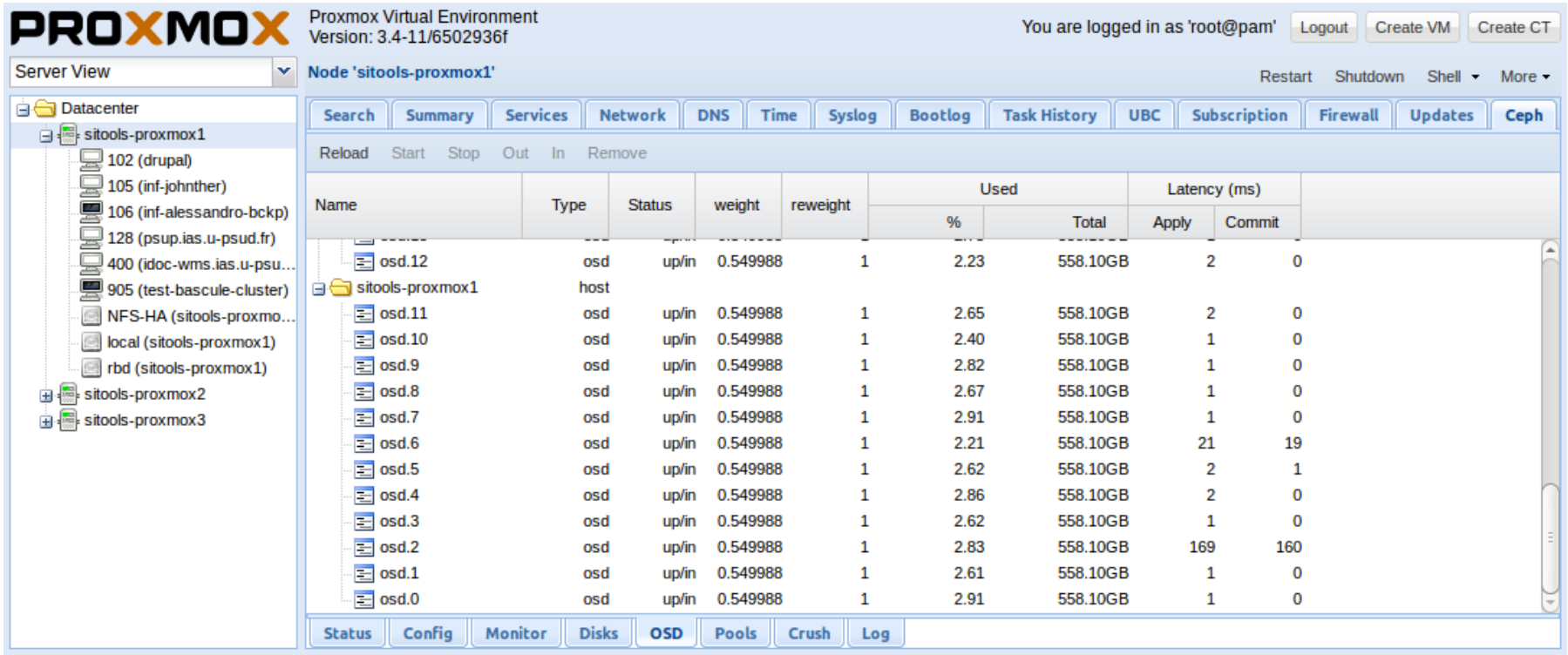
Pool Ceph OK !

```
root@sitools-proxmox1:/etc/pve# ceph status
    cluster 5d43d234-5582-48ba-874f-49c8c4e97c39
     health HEALTH_OK
     monmap e3: 3 mons at {0=10.10.10.1:6789/0,1=10.10.10.2:6789/0,2=10.10.10.3:6789/0}, election epoch 694,
     osdmap e5206: 36 osds: 36 up, 36 in
      pgmap v7990247: 2048 pgs, 1 pools, 263 GB data, 68620 objects
            527 GB used, 19564 GB / 20091 GB avail
                2048 active+clean
  client io 20381 B/s wr, 1 op/s
```

Proxmox/Ceph GUI



Proxmox/Ceph GUI

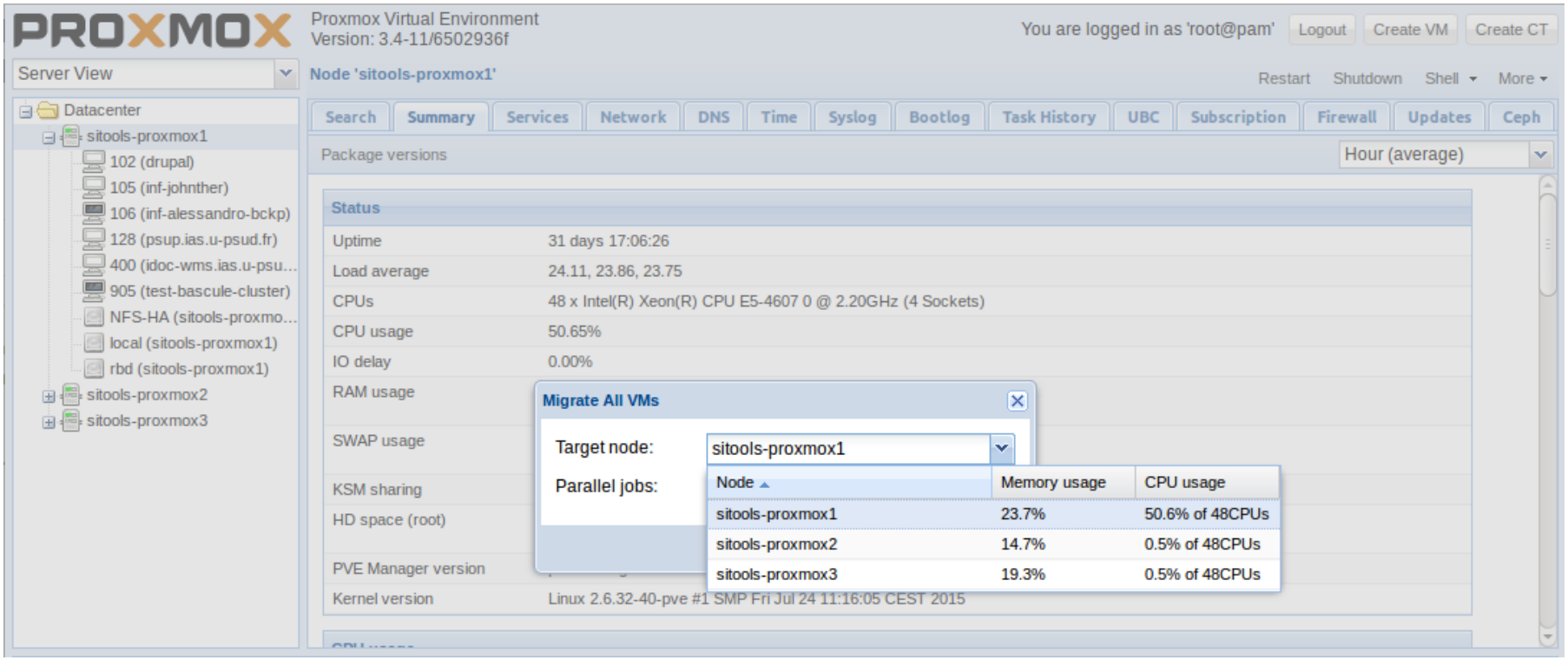


Name	Type	Status	weight	reweight	Used %	Used Total	Latency (ms) Apply	Latency (ms) Commit
osd.12	osd	up/in	0.549988	1	2.23	558.10GB	2	0
sitools-proxmox1	host							
osd.11	osd	up/in	0.549988	1	2.65	558.10GB	2	0
osd.10	osd	up/in	0.549988	1	2.40	558.10GB	1	0
osd.9	osd	up/in	0.549988	1	2.82	558.10GB	1	0
osd.8	osd	up/in	0.549988	1	2.67	558.10GB	1	0
osd.7	osd	up/in	0.549988	1	2.91	558.10GB	1	0
osd.6	osd	up/in	0.549988	1	2.21	558.10GB	21	19
osd.5	osd	up/in	0.549988	1	2.62	558.10GB	2	1
osd.4	osd	up/in	0.549988	1	2.86	558.10GB	2	0
osd.3	osd	up/in	0.549988	1	2.62	558.10GB	1	0
osd.2	osd	up/in	0.549988	1	2.83	558.10GB	169	160
osd.1	osd	up/in	0.549988	1	2.61	558.10GB	1	0
osd.0	osd	up/in	0.549988	1	2.91	558.10GB	1	0

Plan de la présentation

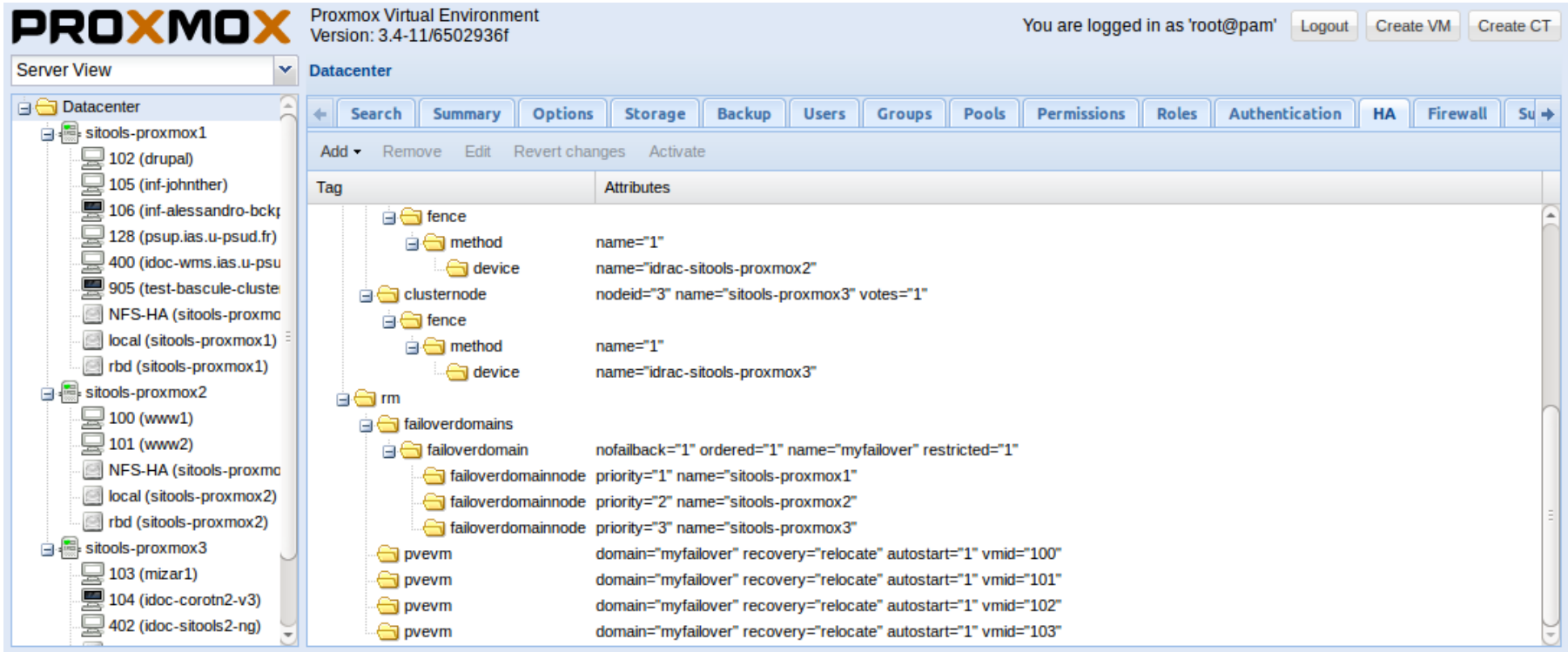
Migration manuelle

Scénario : maintenance programmée dans l'une des salles machines



Migration automatique

Scénario : arrêt d'urgence d'un des noeuds (shutdown -h now)
Les VMs configurées dans cluster.conf basculent automatiquement



Arrêt brutal d'un noeud : comportement de Proxmox VE

Coupure électrique brutale en salle machine
Problème : l'IDRAC du nœud n'est plus accessible et aucune décision automatique n'est prise au niveau du cluster !
Les VMs du nœud HS peuvent être remontées manuellement sur un autre nœud en déplaçant leur fichier de configuration
Exemple sur proxmox1 si le proxmox3 est planté :

```
mv /etc/pve/nodes/sitools-proxmox3/qemu-server/*
                    /etc/pve/qemu-server/
```

Arrêt brutal d'un noeud : comportement de Ceph

1/3 des disques sont absents
Après un temps court d'observation (5 minutes), les MON en ligne décident de reconstruire la crush map avec les disques toujours présents
Aucun impact sur le fonctionnement des VMs.

```
2015-10-01 15:43:31.380428 mon.0 10.10.10.1:6789/0 876272 : [INF] osdmap e4792: 36 osds: 24 up, 24 in
2015-10-01 15:43:31.436218 mon.0 10.10.10.1:6789/0 876273 : [INF] pgmap v7490304: 2048 pgs: 1 inactive, 3 active,
 1 degraded+remapped, 83 active+recovery_wait, 696 active+clean, 7 active+recovering, 1234 a
ctive+degraded, 23 active+degraded+remapped+wait_backfill; 257 GB data, 344 GB used, 13049 GB / 13394 GB avail; 6
652 B/s wr, 2 op/s; 48641/134143 objects degraded (36.261%); 199 MB/s, 50 objects/s recoveri
ng
```

Crash d'un ou de plusieurs disques durs

Les PGs sont dans un état "degraded"

```
root@sitools-proxmox3:/var/log/ceph# ceph health
HEALTH_WARN 100 pgs degraded; 100 pgs stuck unclean; recovery 3299/133734 objects degraded (2.467%); 1/36 in osds are down

root@sitools-proxmox3:/var/log/ceph# ceph status
    cluster 5d43d234-5582-48ba-874f-49c8c4e97c39
     health HEALTH_WARN 100 pgs degraded; 100 pgs stuck unclean; recovery 3299/133734 objects degraded (2.467%); 1/36 in osds are down
     monmap e3: 3 mons at {0=10.10.10.1:6789/0,1=10.10.10.2:6789/0,2=10.10.10.3:6789/0}, election epoch 690, quorum 0,1,2 0,1,2
     osdmap e4672: 36 osds: 35 up, 36 in
      pgmap v7481600: 2048 pgs, 1 pools, 257 GB data, 66867 objects
            514 GB used, 19577 GB / 20091 GB avail
            3299/133734 objects degraded (2.467%)
                 100 active+degraded
                1948 active+clean
  client io 17658 B/s wr, 7 op/s
```

Crash d'un ou de plusieurs disques durs

Le remplacement du disque HS se fait en 3 étapes :

- Configuration du nouveau disque sur le R820
  - via OpenManage (voir Annexe B)
  - Consiste à re-créer un disque virtuel en RAID0
- Création du nouvel OSD
  - `pveceph createosd`
- Suppression de l'OSD correspondant au disque HS
  - `ceph osd crush remove osd.xx`
  - `ceph auth del osd.xx`
  - `ceph osd rm xx`

Perte du réseau ceph sur un noeud

Même comportement pour ceph que dans le cas où le noeud est éteint.
Aucun impact sur les VMs

pgs dans l'état "stale"

Inconsistence de la crush map !
erreur de type :
`"requests are blocked > 32 sec; x osds have slow requests"`
La commande `ceph health detail | grep stale`
liste les pgs qui posent problèmes

pgs dans l'état "stale"

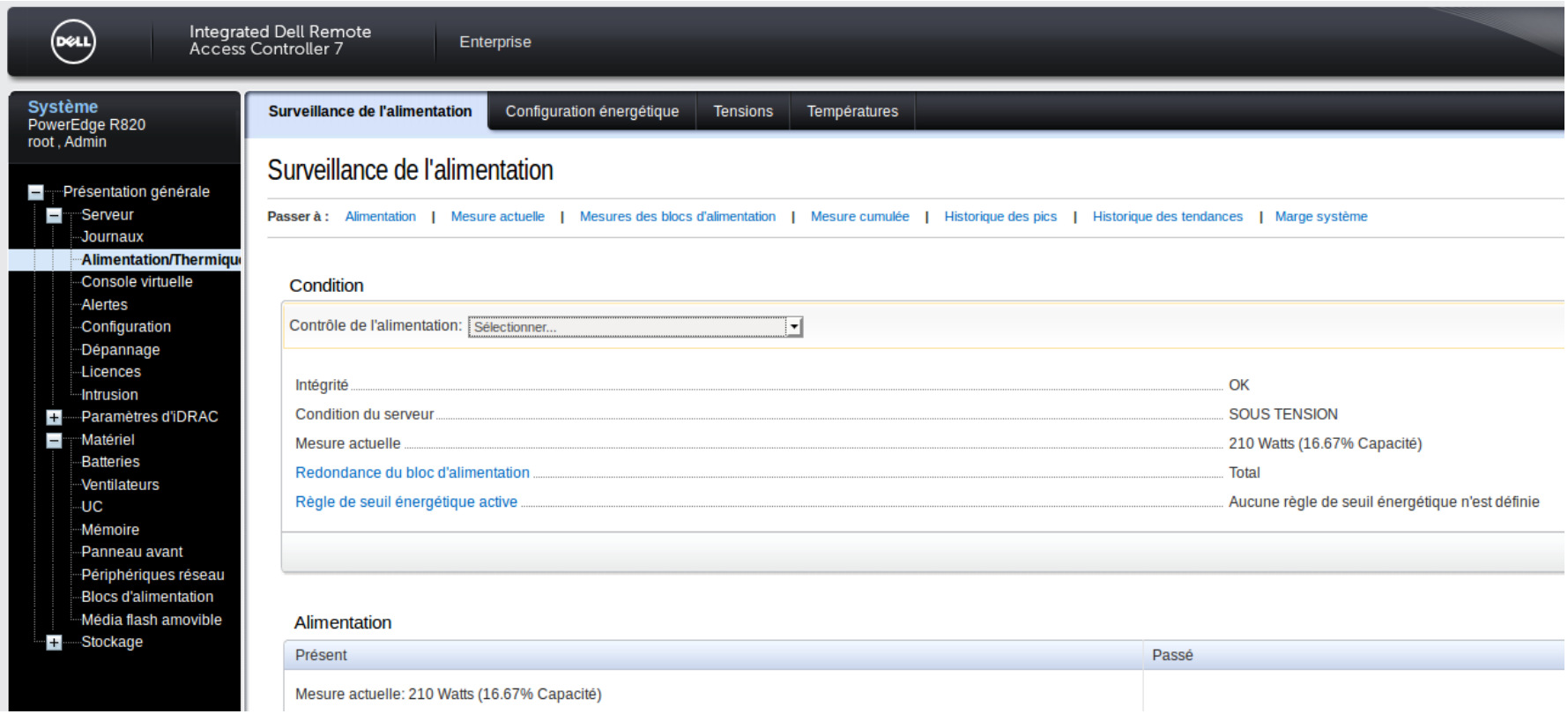
Manips à faire :

- forcer la reconstruction des pgs : `ceph pg force_create`
- stopper le(s) osd(s) associé(s) aux pgs : `/etc/init.d/ceph stop osd.xx`
- forcer la sortie du(es) osd(s) du pg : `ceph osd lost xx --yes-i-really-mean-it`
- redémarrer le(s) osd(s) : `/etc/init.d/ceph start osd.xx`

Plan de la présentation

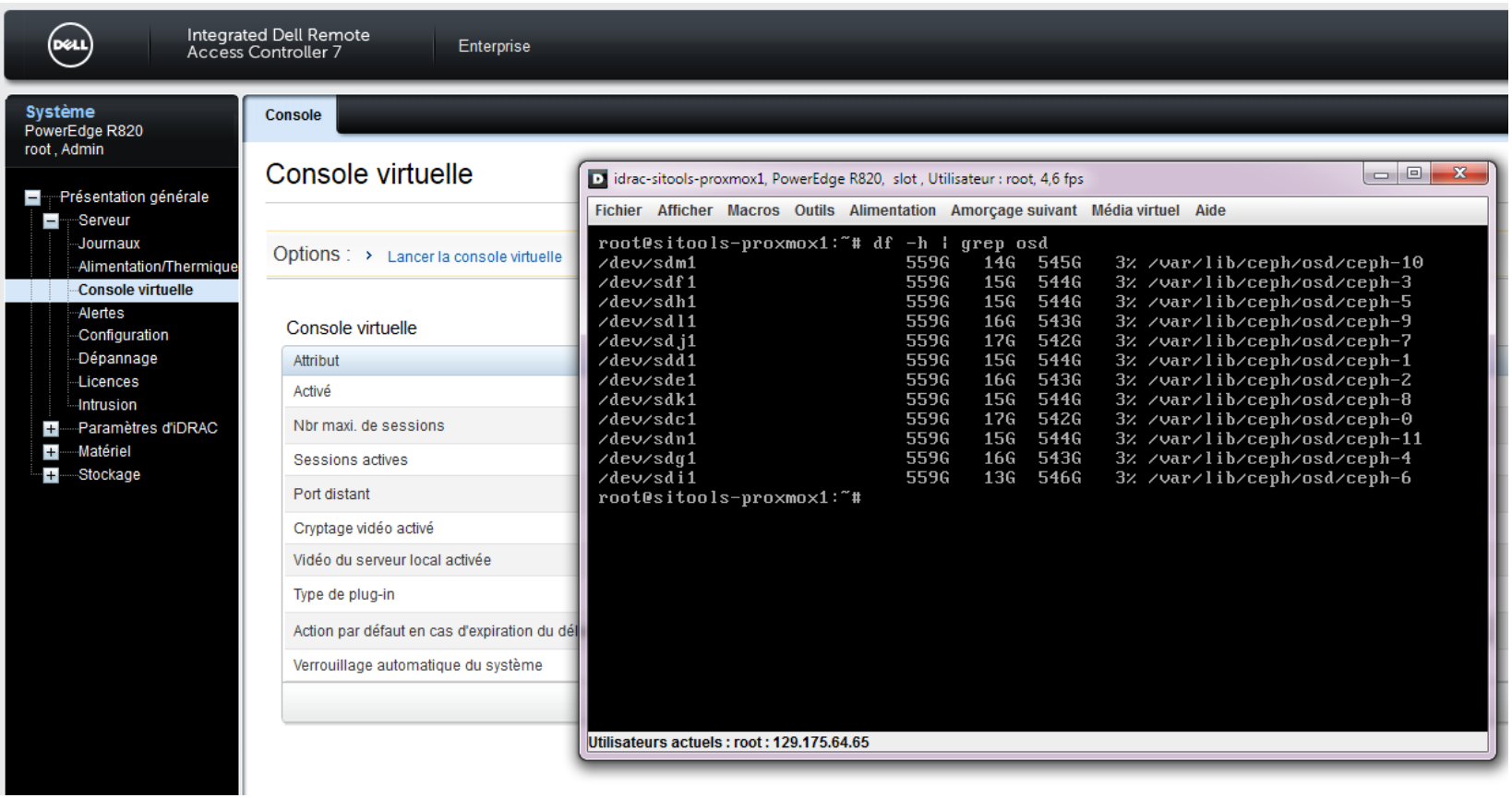
Accès par l'interface web

https://*ip-idrac-serveur*



Accès par la console virtuelle

https://*ip-idrac-serveur*/console

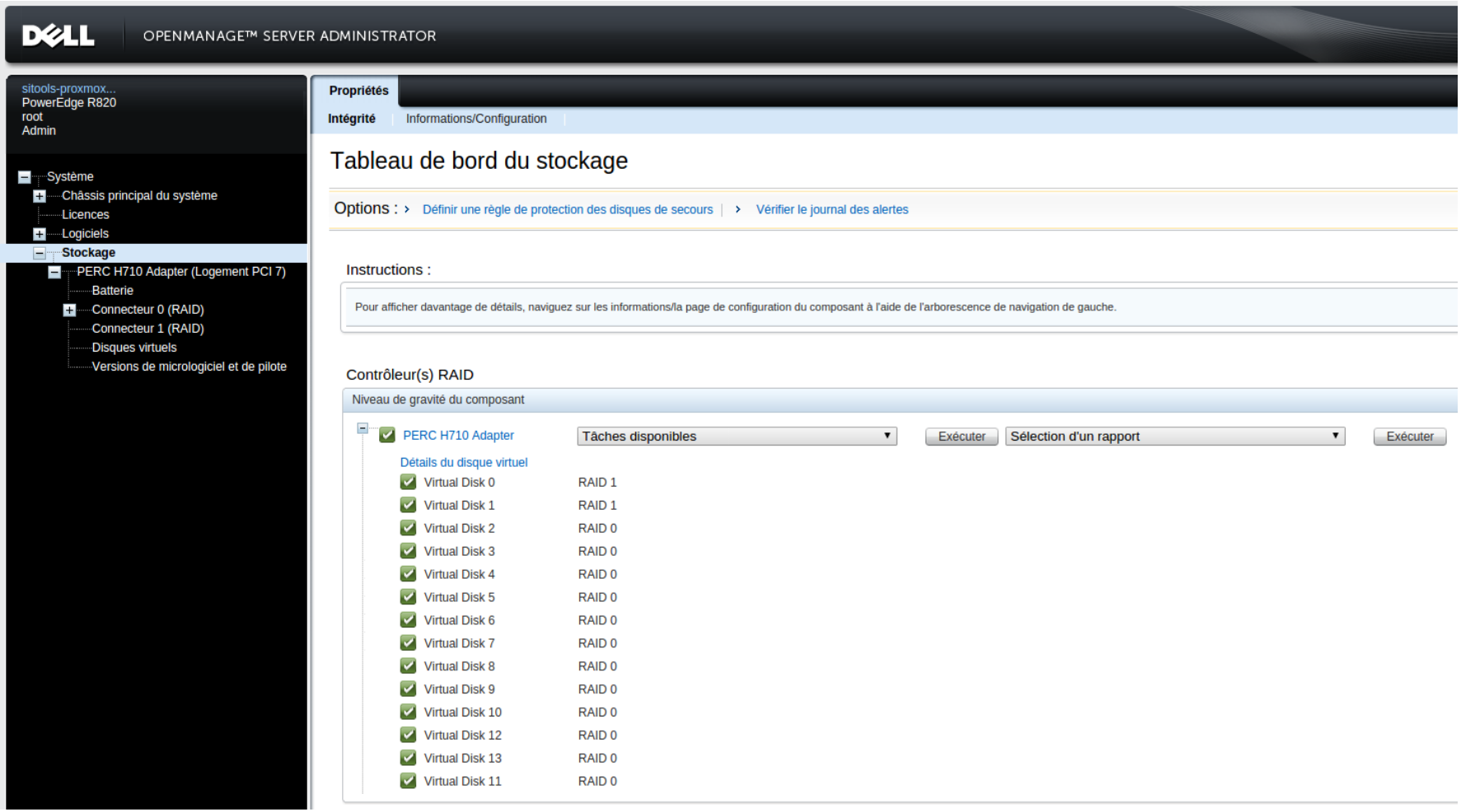


Alertes par mail

Exemple : interface réseau du pool ceph du serveur3 déconnectée

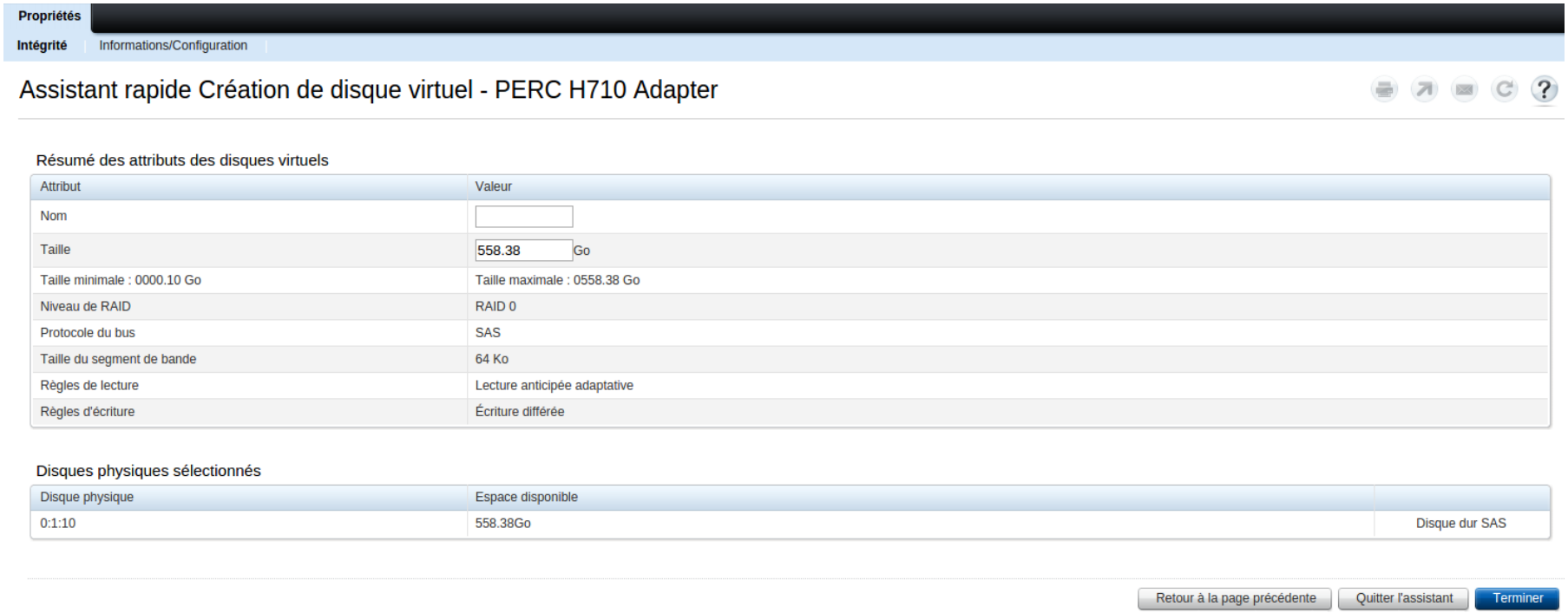
```
System Host Name: sitools-proxmox3
Event Message: The NIC Integrated 1 Port 2 network link is down.
Date/Time: Thu Oct 01 2015 15:38:30
Severity: Warning

Detailed Description: The network link is down. Either the network cable is not connected
or the network device is not working.
Recommended Action: Verify that the network port is enabled and if the port has
Activity/Speed LEDs, that they are lit. Check the network cable, network cable
connections, and the attached network switch.
Message ID: NIC100

System Model: PowerEdge R820
Service Tag: FNXPJ32
Power State: ON
Operating System: Linux
System Location: Slot 1 (2 U)
```

https://*nom-serveur*:1311



Remplacement d'un disque

Création d'un disque virtuel (configuré en RAID0)



Sortie officielle le 6 octobre 2015

Nouveautés :

- Support natif des containers LXC
- Nouveau gestionnaire de Haute Disponibilité
- Proxmox HA Simulator
- Ceph Hammer
- DRBD9
- Amélioration de la console NoVNC
- ...

Ceph Hammer (v0.94)

https://ceph.com/releases/v0-94-hammer-released/

Nouveautés :

- Amélioration des performances (RADOS cache tiering, RBD object maps,...)
- Améliorations sur CephFS, mais :

Important: CephFS currently lacks a robust 'fsck' check and repair function. Please use caution when storing important data as the disaster recovery tools are still

Références sur Proxmox

- http://www.iphc.cnrs.fr/IMG/pdf/2014.proxmox.inra.pdf
- http://xstra.u-strasbg.fr/lib/exe/fetch.php?media=doc:2012-05-31.proxmox-2.0.pdf

Références sur Ceph

- Sécurité des données
  - http://cargo.univ-brest.fr/membres/ressources/journee-thematiques/cargoday4-protection-des-donnees/pm-01-lt-prolland-ceph-security.pdf
- Ceph et Proxmox
  - http://www.jamescoyle.net/how-to/1213-ceph-storage-on-proxmox
- Retours d'expérience
  - https://2013.jres.org/archives/48/paper48_article.pdf