MISE EN PRODUCTION DE LA PLATE-FORME DE VIRTUALISATION SOUS PROXMOX V2

Le 17/12/2013

Sébastien Geiger IPHC IN2P3 CNRS

Sommaire

- Calendrier des actions
- Installation et configuration du matériel
- Test de fonctionnement
  KMS, live migration, configuration réseau, HA
- Migration des serveurs (P2V et V2V)
- Gestion des machines virtuelles
- Montée en charge
- Astuces
- Evolution
- Bilan

Calendrier

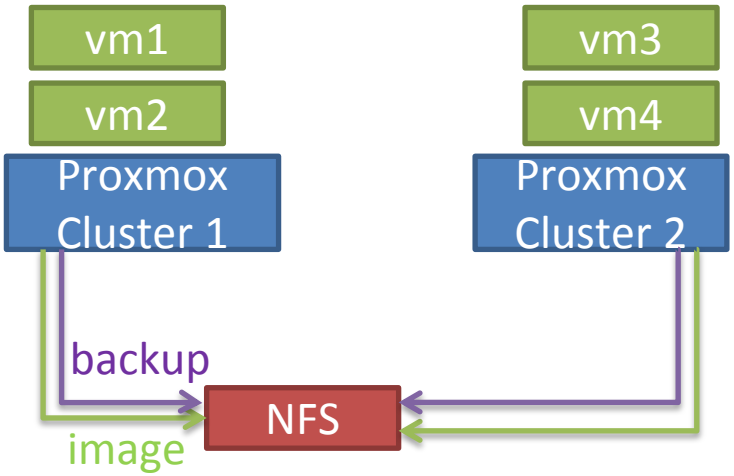
- Achat du matériel 07/2012
- Installation et configuration 08/2012
- Vérification des fonctionnalités 09/2012
- Migration des serveurs de 09/2012 à 03/2013
- Mise à jour
  - Proxmox 2.2 le 24.10.2012
  - Proxmox 2.3 le 03.01.2013
- Extension de la solution
  - Achat matériel (07/2013)
    blade 2500€ + R720 3500€ + 2*M620 7000€
  - Ajout nouveau matériel 10/2013
  - Définition plan de reprise

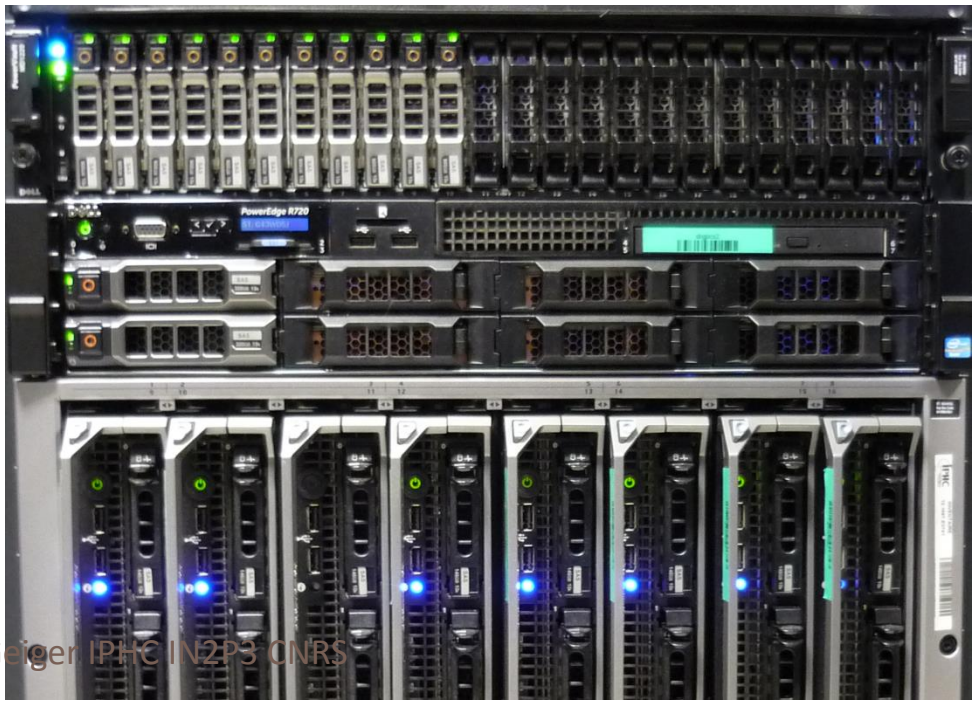
Installation de la solution

- Mise à disposition de 8 lames M600 pour réaliser 2 clusters de 4 nœuds.
- Installation du serveur NFS
  - Installation du système ScientificLinux 6.3
  - Configuration du réseau et du bonding
  - Configuration du raid6 de 6To avec disque de spare
  - Création de deux volumes de 3To sous ext4 sur un LVM
  - Export de l'espace disque
  - Configuration des outils de surveillance (Nagios , OpenManage)
- Installation des clusters Proxmox
  - Installation de Proxmox v2.1
  - Activation du support pour la virtualisation : grep -E 'vmx|svm' /proc/cpuinfo
  - Création des 2 clusters de 4 noeuds
  - Configuration des montages NFS
  - Configuration des autorisations d'accès et des groupes des vms
  - Configuration des règles de backup
- Rédaction de la documentation d'installation

Matériel

- Stockage : Dell R720 + MD1220
  - Serveur NFS v3
- Nœuds de virtualisation
  - 8 lames M600 (8 core 2,5Ghz 16Go Ram)
  - Châssis M1000e



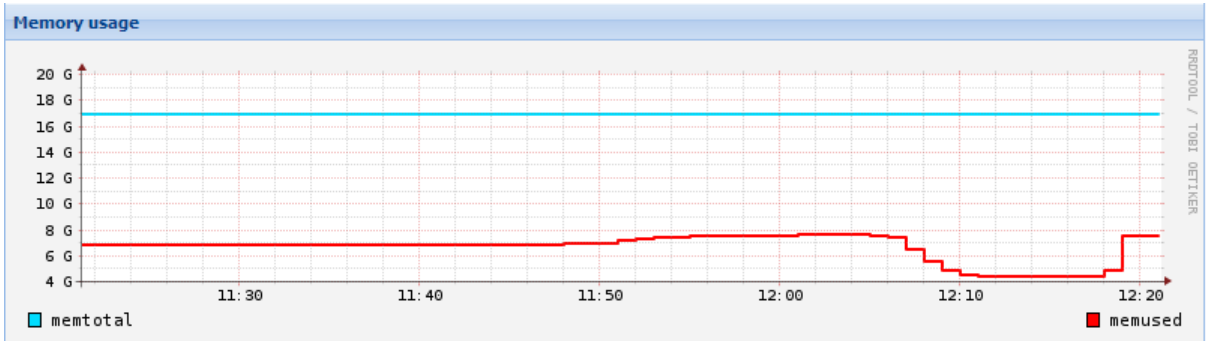


Vérification des fonctionnalités

- KSM (Kernel Samepage Merging)
- Live migration
- Configuration réseau
  - bonding
  - Support Multicast
- Mode HA

KSM (Kernel Samepage Merging)

- KSM permet au kernel Linux de partager les mêmes pages mémoires
- Par défaut Proxmox active le module ksm depuis la version 1.9
- # grep . /sys/kernel/mm/ksm/*
  - pages_shared : nombre de pages réellement utilisées par KMS
  - pages_sharing : nombre de pages globalement partagé
  - pages_volatile : les pages qui changent trop rapidement
- Utilisation avec 2 VM Linux et 2 VM Windows
  - 8Go sans KSM
  - 1Go après 6 minutes



Live migration

- Déplacement d'une VM en fonctionnement
  - Temps de transfert < 30s
  - Temps de coupure réseau < 20ms
  - Pas de perte de connexion via ssh sur la VM
- Réponse au ping pendant la tâche de migration

Reply from 193.48.86.223: bytes=32 time<1ms TTL=63

Reply from 193.48.86.223: bytes=32 time=43ms TTL=63

Request timed out.

Request timed out.

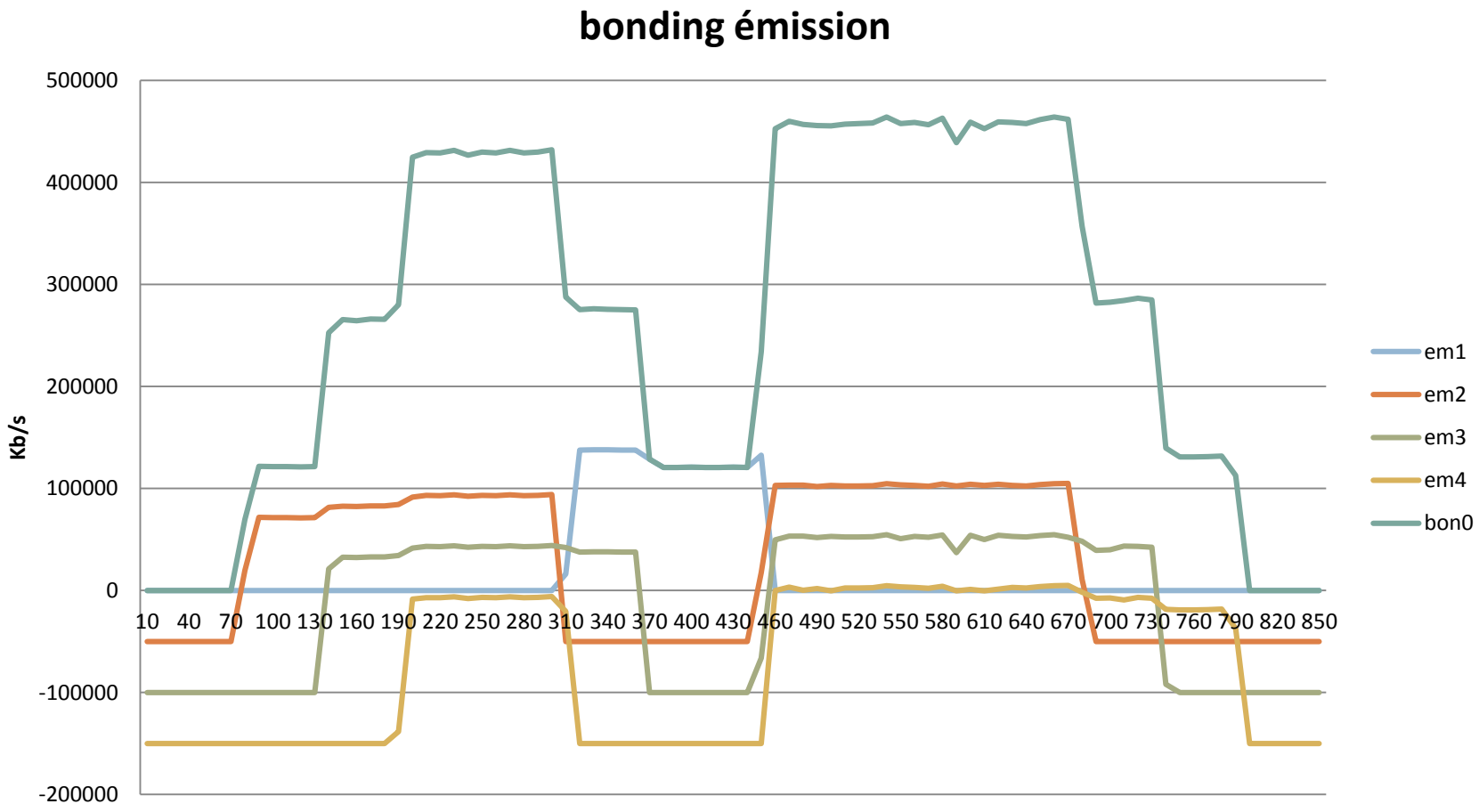
Reply from 193.48.86.223: bytes=32 time<1ms TTL=63

Reply from 193.48.86.223: bytes=32 time<1ms TTL=63

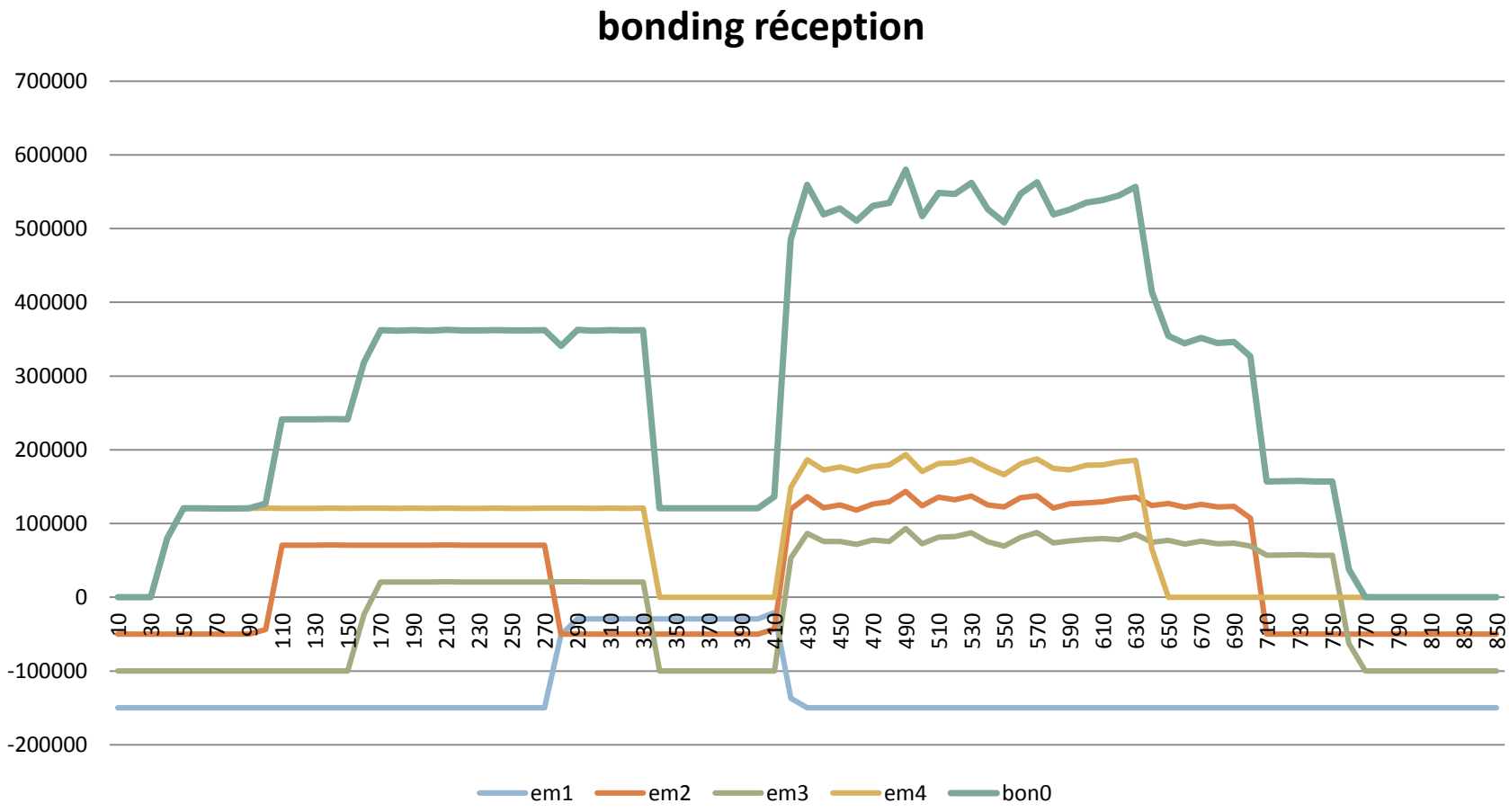
Configuration réseau et bonding

- Configuration
  - 4 liens sont agrégés sur le serveur de stockage.
  - mode 4 : 802.3ad
    - augmente la bande passante et la tolérance de panne
    - implique que le switch gère le 802.ad et les interfaces soient compatibles mii-tool et/ou ethtool
- Mode de test
  - Iperf –s sur le serveur de stockage et iperf -c serveur -t 600 sur les nœuds
  - lancement de trois clients décalés de 60s
  - arrêt de 1, puis 2, puis 3 interfaces sur les 4 toutes les 60s
  - mesure du débit avec sar –n DEV
  - vérification des erreurs de transfert

Configuration réseau et bonding



Configuration réseau et bonding



Configuration du Multicast

- L' « igmp snooping » est une fonction qui permet d'optimiser la diffusion des trames multicast en observant le trafic IGMP sur un switch.
- Fonction activée par défaut sur le matériel Cisco
- Cette fonction peut filtrer des informations échangées entre les nœuds de virtualisation. Il est conseillé de la désactiver

  # conf t

  # no ip igmp snooping
- Test multicast
  - Utiliser omping depuis les différents nœuds
    omping node1 node2 node3
  - Utiliser ssmpingd et asmping entre deux nœuds.
    - From node1 : ssmpingd
    - Then from node2 : asmping 224.0.2.10 ip_for_node1

Configuration en Unicast

- Problèmes de stabilité
- 4 nœuds maximum
- Limite l'ajout à chaud d'un nouveau nœud
- Passage en Unicast
  - Ajouter transport="udpu" dans /etc/pve/cluster.conf.new
    <cman keyfile="/var/lib/pve-cluster/corosync.authkey" **transport="udpu"**/>
  - Activer les modifications via l'interface de gestion
  - Ajouter tous les nœuds dans /etc/hosts et rebooter
  - Avant d'ajouter un nouveau nœud mettre à jour /etc/hosts sur tous les nœuds

Mode Haute Disponibilité

- Prérequis pour le mode HA
  - 3 nœuds minimum
  - Espace de stockage partagé (NFS, NAS, SAN)
  - La configuration des « fencing devices »
- Redémarrer une vm automatiquement
  - Si elle est arrêtée ou crachée
  - Si le nœud de virtualisation s'arrête ou ne répond plus
- Test
  - Arrêt d'une vm en HA
  - Arrêt forcé d'un des nœuds du cluster
    - Redémarrage de la vm sur un autre nœud
    - Pas de conflit après le retour du nœud arrêté

Migration des serveurs

- Serveurs physiques sous Linux
- Serveurs virtuels sous Vmware
- Serveur physique sous Windows
- Pilote de paravirtualisation
  - pour Linux
  - pour Windows

Migration des serveurs

- Serveur physique sous Linux
  - Pas de vielles machines
  - Distribution ScientificLinux 6.2 x64
  - Kernel 2.6.32 supporte par défaut des modules virtIo (disque et net)
  - Uniformisation des distributions

Migration des serveurs

- Serveur Virtuel sous VMware
  - Désinstaller les VMware Tools
  - Rajouter le support des disques IDE au système
    - Linux: reconfigurer les LVM
    - Windows voir la KB314082
  - Arrêt du serveur et conversion de l'image disque
    Sous Proxmox utiliser des disques de taille fixe.
    Si le disque VMware est dynamique, le convertir en disque de taille fixe avec l'outil vmware-vdiskmanager. Ensuite convertir celui-ci en disque au format qcow2
    # qemu-img convert myVMwFlatImage-pve.vmdk -O qcow2 myVMwFlatImage-pve.qcow2
  - sous Proxmox, définir une nouvelle VM, et remplacer le disque virtuel avec le fichier qcow2
  - Reconfiguration les périphériques (virtio, Net, vidéo)

Migration des serveurs

- Serveur physique sous Windows
  - Préparer le serveur en rajoutant les pilotes IDE Standard
    - vérifier la présence Atapi.sys, Intelide.sys, Pciide.sys, et Pciidex.sys dans le répertoire %SystemRoot%\System32\Drivers.
    - Activer les pilotes IDE au démarrage avec le fichier Mergeide.reg (voir la procédure complète depuis http://support.microsoft.com/kb/314082/)
  - Configurer une nouvelle VM sur Proxmox
    - Définir le nombre de CPU, la RAM et une carte réseau en mode NAT
    - Définir le contrôleur disque en IDE avec le même nombre de disques mais de taille de 1Go
  - Depuis le poste de capture
    - Installer VMware Vcenter Converter Standalone Client
    - Démarrer une nouvelle capture et sélectionner
      Select source type : powered-on machine
      specify the powered-on machine : A remote machine, puis fournir les informations de connexion
      Select destination type: VMware Workstation or other VMware virtual machine
      Select VMware Product: VMware Workstation 8.0.x
      Name: Saisir le nom du serveur
      Select a location for the virtual machine: définir un répertoire de destination
    - Image disque
      Choisir une taille fixe pour l'image disque. Il est possible de redéfinir la taille des disques virtuels à la hausse ou à la baisse

Migration des serveurs

- Transférer l'image disque vmdk vers Proxmox par scp dans /mnt/pve/datastore/images/vmid/
- Convertir l'image vmdk en qcow2 depuis un des nœuds du cluster Proxmox
  # qemu-img convert myVMwFlatImage-pve.vmdk -O qcow2 myVMwFlatImage-pve.qcow2
- Remplacer les disques temporaires par les disques qcow générés
- Démarrage de la machine virtuelle clonée
  - détection de périphérique (réseau, carte vidéo, contrôleur IDE, souris)
  - redémarrer le serveur
  - vérifier si les logiciels ou services fonctionnent correctement
  - désinstaller les anciens pilotes raid ou scsi
  - désinstaller l'agent de capture vmware
  - désinstaller des outils de surveillance matérielle (OpenManage)
- Synchronisation des documents via la connexion NAT
- Arrêt du serveur physique et transfert de l'ip vers la VM
- Passage de la connexion réseau en mode bridge
- Vérification de la reprise des services et de la montée en charge

Migration des serveurs

- Pilote de paravirtualisation
- Pour Linux
  - Kernel 2.6.32 supporte par défaut les modules virtIo (disque, net et balloon)
- Pour Windows

  Pour augmenter les performances réseaux et d'accès disque, il est conseillé d'utiliser les pilotes virtIo fournis sous forme d'ISO depuis http://alt.fedoraproject.org/pub/alt/virtio-win/stable/
  - Arrêter la VM
    - Ajouter un nouveau disque virtIO temporaire (1Go) depuis l'interface de gestion
    - Ajouter le cdrom des drivers virtIO pour Windows
  - Démarrer la VM et ajouter le pilote virtio pour reconnaitre le disque temporaire
  - Arrêter la VM
    - Supprimer le disque temporaire virtIO (1Go)
    - Changer le type de disque de boot en virtio
  - Démarrer la VM, Windows boote sans problème et sans écran bleu
  - Pour plus d'information voir : http://pve.proxmox.com/wiki/Paravirtualized_Block_Drivers_for_Windows

Serveurs virtualisés

- 21 VMs en production, 5 VMs en pré-test
  - 10 serveurs Windows (contrôleur de domaine, accès au bureau à distance, serveurs antivirus, serveur d'impression, PXE, Ms SharePoint)
  - 6 serveurs Linux (outils de supervision, Owncloud, Annuaire LDAP, base de données Mysql, apache)
  - 4 Images des systèmes de référence pour le déploiement par PXE
  - Outils de compilation de circuit électronique (Xilinx)
  - 5 Serveurs de pré-production pour tester des logiciels ou des scripts de déploiement

Gestion de VMs

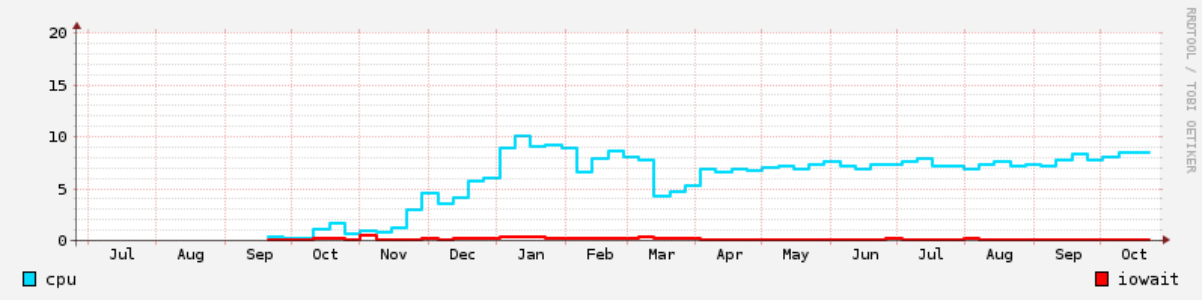
- Pour chaque VM, un administrateur est nommé. Celui-ci aura le rôle Pvevmuser lui permettant pour cette VM :
  - l'accès à la console distante
  - l'arrêt et le redémarrage
  - la sélection de l'image contenue dans le CD-Rom
- Les administrateurs de la plateforme ont un accès global à tous les paramètres
  - la création et la configuration des VMs
  - la configuration des règles de backup et la réalisation des opérations de restauration
  - la migration en live des VMs sur un autre nœud
  - la connexion par ssh sur les nœuds
  - la mise à jour de la solution

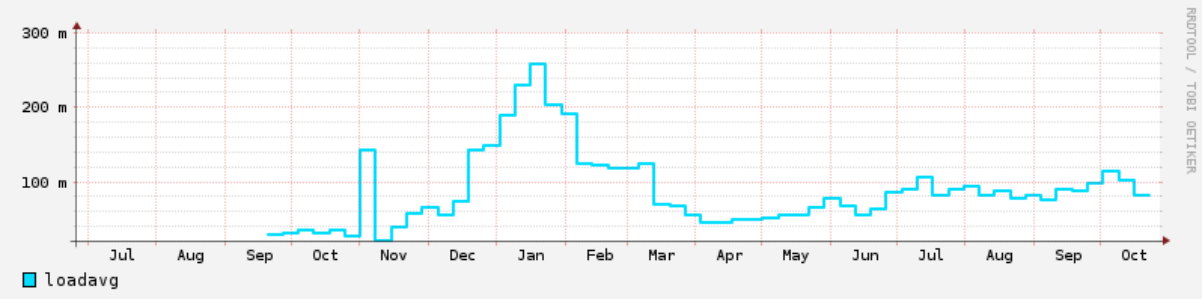
Backup et restauration

- Proxmox fournit un gestionnaire de sauvegarde et de restauration
  - Sauvegarde complète en live, mode suspendu, arrêt démarrage
  - Pas de sauvegarde incrémentale
  - Définition des créneaux horaires pour la sauvegarde
- Politique de sauvegarde : Les VMs en production sont sauvegardées intégralement une fois par semaine avec un maximum de 2 historiques par VM.
- Les VM qui intègrent des fichiers utilisateurs sont sauvegardées sur bande avec les clients de sauvegarde traditionnelle.
- Il est bon d'avoir un miroir des backups des VMs sur un autre espace que celui du serveur NFS

Charge des différents nœuds

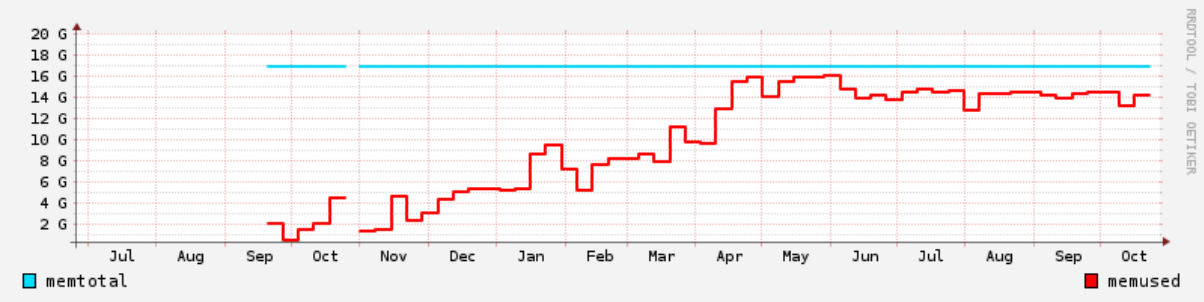
- Evolution cpu, load, mémoire, net au fur et à mesure de l'ajout de nouvelles machines



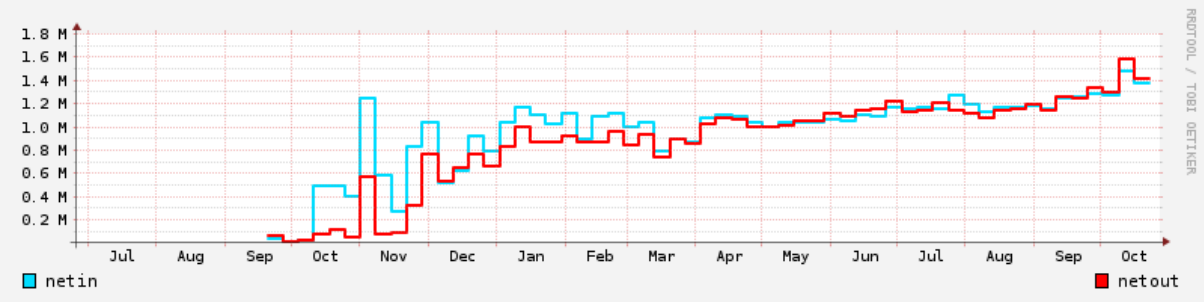


Charge des différents nœuds

- Répartition des VMs

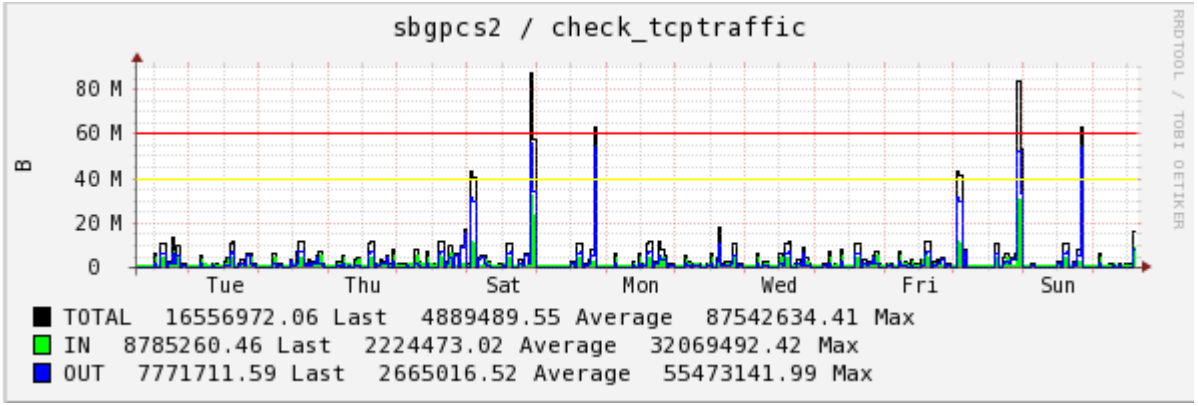


- Trafic induit par les sauvegardes



Trafic du Serveur NFS

- Evolution du trafic du serveur NFS du 11-25/11/2013
- Vendredi backup FULL de toutes les VMs Windows
- Samedi et dimanche backup des VMs par Proxmox
- En semaine backup classique (rsync ou arcserv) des VMs
- La charge en journée est relativement faible



Astuces

- Démarrage automatique des VMs
  - Après un démarrage du cluster Proxmox, il est possible de définir les VMs à démarrer automatiquement ainsi que l'ordre de démarrage avec les paramètres suivants :
  - Démarrage au boot : oui ou non
  - Ordre de démarrage : ordre de priorité pour le démarrage. 1 est la priorité la plus haute
  - Délais de démarrage : délais à respecter avant de démarrer une autre vm.
  - Temps d'arrêt : temps après lequel la VM est arrêtée de force si elle ne répond pas au signal d'arrêt. Par défaut c'est 180 secondes pour les VMs sous KVM

Astuces

- Arrêt des VMs
  - Proxmox utilise l'interface ACPI pour envoyer un signal aux VMs à éteindre. Cela nécessite l'activation du support de l'ACPI au niveau du système.
  - Sous Linux vérifier que le daemon ACPI est installé et démarré.
  - Sous Windows, autoriser l'arrêt de la machine via une GPO par :
    Computer Configuration\Windows Settings\Security Settings\Local Policies\Security Options\Shutdown: Allow system to be shut down without having to log on.

Astuces

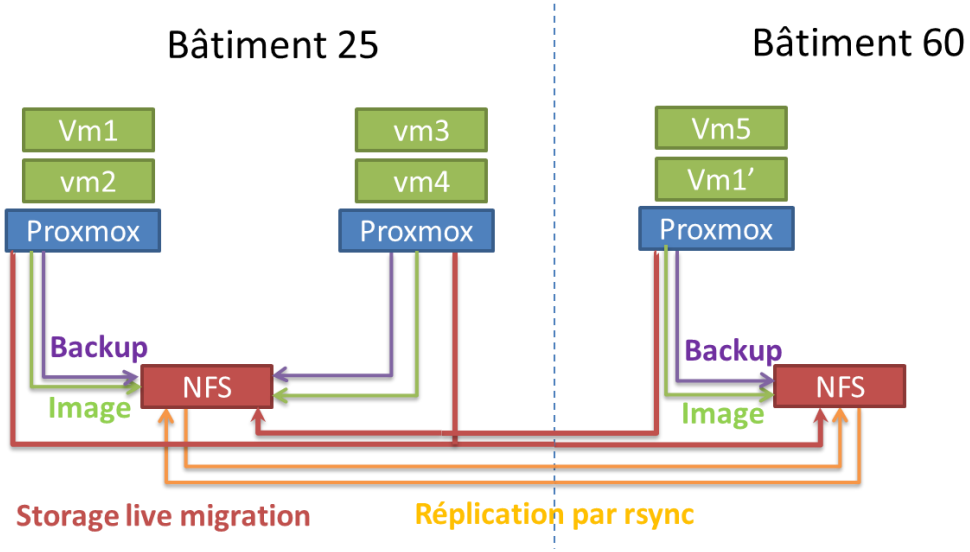
- Par défaut la commande « du » affiche la taille utilisée d'un disque virtuel sur l'espace de stockage. Pour avoir la taille effective, il faut rajouter l'option « --apparent-size »

```
# du -h
5.7G    .
# du -hs --apparent-size
41G     .
```

- Gestion des VMs en ligne de commande
  - qm list, migrate, start, shutdown
- Vérifier l'état du cluster
  - pvecm status, pvecm nodes, /var/log/syslog

Evolution

- Plan de reprise entre le bâtiment 25 et le bâtiment 60
  - Installation 2ème serveur de stockage au bâtiment 60
  - Installation des lames et configuration du Cluster numéro 3
  - Configuration de la fonction « Storage live migration »
  - Réplication des backups des VMs du bâtiment 25
  - Plan de reprise des VMs via une restauration du backup



Evolution Proxmox 3.1

- Fonctionnalités

  Live storage migration, Templates, Spice, GlusterFS storage

- Changement de politique
  - Mise à jour Enterprise Repository / No-Subscription Repository
    http://forum.proxmox.com/threads/15742-Details-about-the-new-pve-no-subscripton-repository
  - Prix par CPU et par mois
    Community 4€ Basic 17€ Standard 33€ Premium 66€
    http://www.proxmox.com/proxmox-ve/pricing
- Arrêt du support Proxmox 2.x

Bilan

- 21 VMs en production et 5 VMs en pré-test
- Virtualisation sous Proxmox
  - Produit complet , stable et supporte la montée en charge
  - Nécessite de bien se documenter
  - Suivre les évolutions de Proxmox (passage à la version 3.x)
- Gain en salle serveur
  - Place :10 serveurs (16u) + 2tours + 1DLT contre 5u (1/2 blade)+ 4u serveur NFS
  - Energie : 3,4Kw contre 1,6Kw
  - Performance : les anciens serveurs avaient 6 ans (gain en RAM et CPU). Exemple pour la VM de compilation, un gain de 50% a été mesuré.
  - Solution extensible, il reste de la place pour d'autres VMs
- Si besoin de monter en charge
  - Transfert des VMs sur la solution Cloud sous OpenStack de l'IPHC.