Sébastien Geiger
IPHC UMR7178 du CNRS

IN2P3
Institut national de **physique nucléaire**
et de **physique des particules**

Retour d'expérience de la plateforme de virtualisation sous Proxmox VE à l'IPHC

journée SysAdmin du 04/12/2014
à l'INRA de Toulouse

Sommaire

- Présentation de Proxmox
  - Historique des versions
  - Présentation des fonctionnalités
  - Mesure de performance
- Présentation de la plate-forme IPHC
  - Schéma de la solution
  - Migration de la version 2.x à 3.x
  - Comment déterminer les services à virtualiser
  - Supervision des services et des VMs
  - Liste des services virtualisés
  - Bilan
- Evolution des technologies de virtualisation

Proxmox

- Société autrichienne
  - Créée en 2004
  - Open Virtualization Alliance
  - Linux Foundation
  - Solution de virtualisation « Proxmox VE » GNU AGPL v3
  - Solution commerciale « Mail Gateway »
- Proxmox VE
  - Solution de virtualisation *"bare metal"* (KVM, OpenVZ)
  - Cluster 2 à 16 nodes
  - Fournit plus de 50 Virtual Appliances
  - Distribution Debian avec un kernel RedHat 2.6.32

Historique des versions

- Proxmox VE v1.x (2008)
  - Cluster 2 à 16 nodes
  - Live migration, template
- Proxmox VE v2.x (2012)
  - Mode HA (3 hosts et stockage partagé)
  - KSM, Zero live migration
  - Nouvelle interface basée sur ExtJs
- Proxmox VE v3.x (2013)
  - Live storage migration, Live backup, Firewall
  - Console Java, spice, html5
  - Support Ceph, ZFS, Open vSwitch
  - Changement concernant la distribution des mises à jour

Fonctions intégrées

- Live migration & Live Storage Migration
- Spice & Console Java ou HTML5
- Optimisation de la mémoire
- Mode Haute disponibilité
- Authentification et rôle

Live Migration

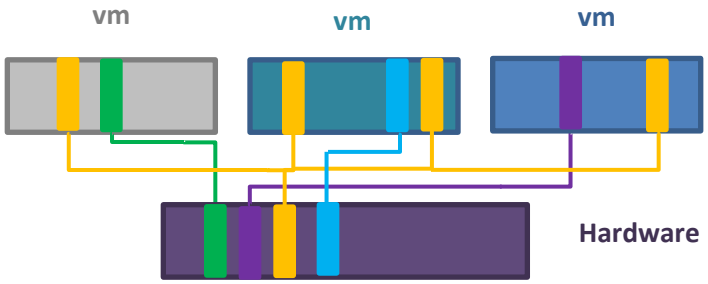
- Live Migration
  - Déplacement d'une VM en fonctionnement d'un nœud du cluster à un autre
  - Nécessite un stockage partagé
  - Temps de transfert < 40s et temps coupure < 1s
  - Pas de perte de connexion via ssh sur la VM
- Live Storage Migration
  - Déplacement des disques virtuels d'une zone de stockage à une autre
  - Fonctionne avec des VMs en fonctionnement ou à l'arrêt
  - Changement de format de disque virtuel
  - Prévoir l'espace disque pour contenir les disques sources et destinations

Console Java ou HTML5 & Spice

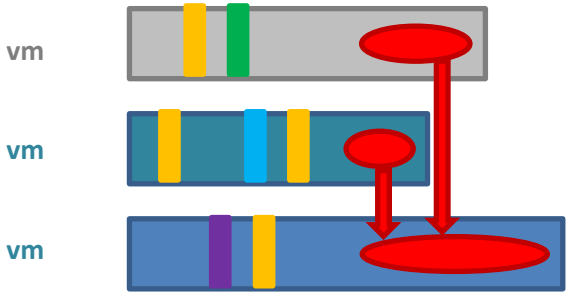
- Console
  - Accès à la console et au clavier de la VM
  - Idéale pour faire l'installation à distance
  - Console HTML5 fonctionne sans plugin
- Spice
  - Solution indépendante d'accès à distance
  - Support du copier/coller, USB, multimédia
  - Pilote pour Windows 7,8 ou Linux
  - Nécessite un client
  - Pas de support multi-utilisateurs, d'impression, mécanisme d'authentification en cours d'intégration
  - Alternative : utilisation de RDP sous Windows ou x2go sous Linux

Optimisation de la mémoire

- Kernel Samepage Merging



- Le Ballooning et l'Auto-Ballooning (KVM)



https://pve.proxmox.com/wiki/Dynamic_Memory_Management

Mode Haute disponibilité

- Principe
  - Redémarre une VM si elle est défaillante
  - Redémarre les VMs d'un nœud si celui-ci est défectueux
  - Empêche un nœud défaillant de rejoindre le cluster
- Prérequis
  - 3 nœuds minimum + un espace disque partagé
  - Dispositif d'empêchement (fencing device)
    - IPMI, UPS, BMC

Authentification et rôle

- Contrôle de l'accès au cluster par authentification
  - LDAP, Active Directory ou interne
- Rôles
  - Noacces : par défaut
  - Pvevmuser : backup, config.cdrom, console et power
  - PvevmAdmin: pvevmuser + config.* et migration
  - PveDatastore: ajout des templates et espaces disques
  - pveAdmin: PvevmAdmin + PveDatastore + audit
  - Admin: pveAdmin + connexion par ssh sur les noeuds
- Autorisation par groupe ou par personne sur chaque VM

Support commercial

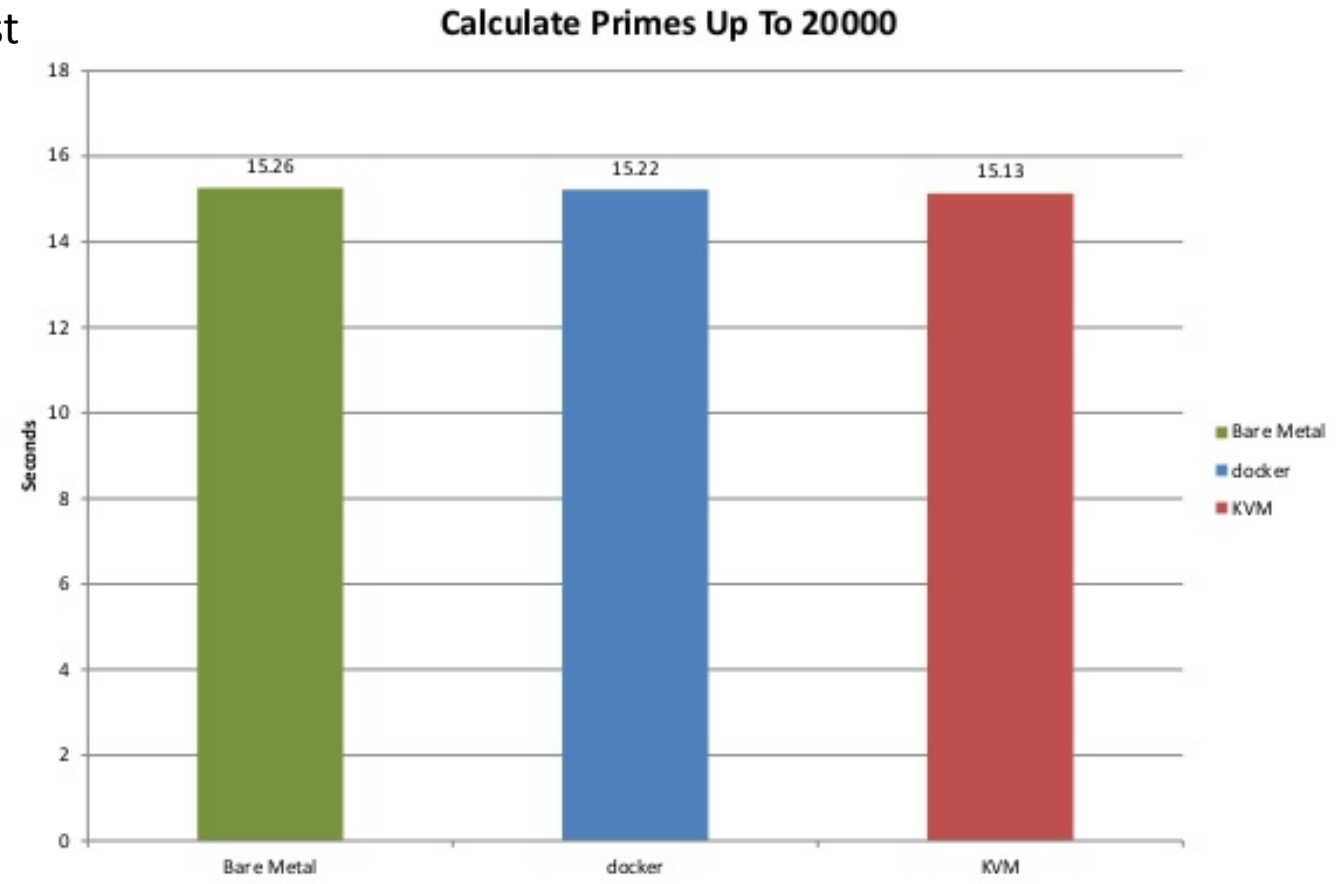
	SANS SUPPORT	COMMUNITY	BASIC	STANDARD	PREMIUM
prix	0€	€ 4,16 / CPU & mois	€ 16,58 / CPU & mois	€ 33,17 / CPU & mois	€ 66,33 / CPU & mois
Access au Repository Enterprise	Non	Oui	Oui	Oui	Oui
Mise à jour distribution	Oui	Oui	Oui	Oui	Oui
Support	via le forum communautaire	via le forum communautaire	via un portail dédié	via un portail dédié	via un portail dédié
tickets/an	0	0	3	10	Illimité
intervention à distance via ssh	Non	Non	Non	Oui	Oui

Performance

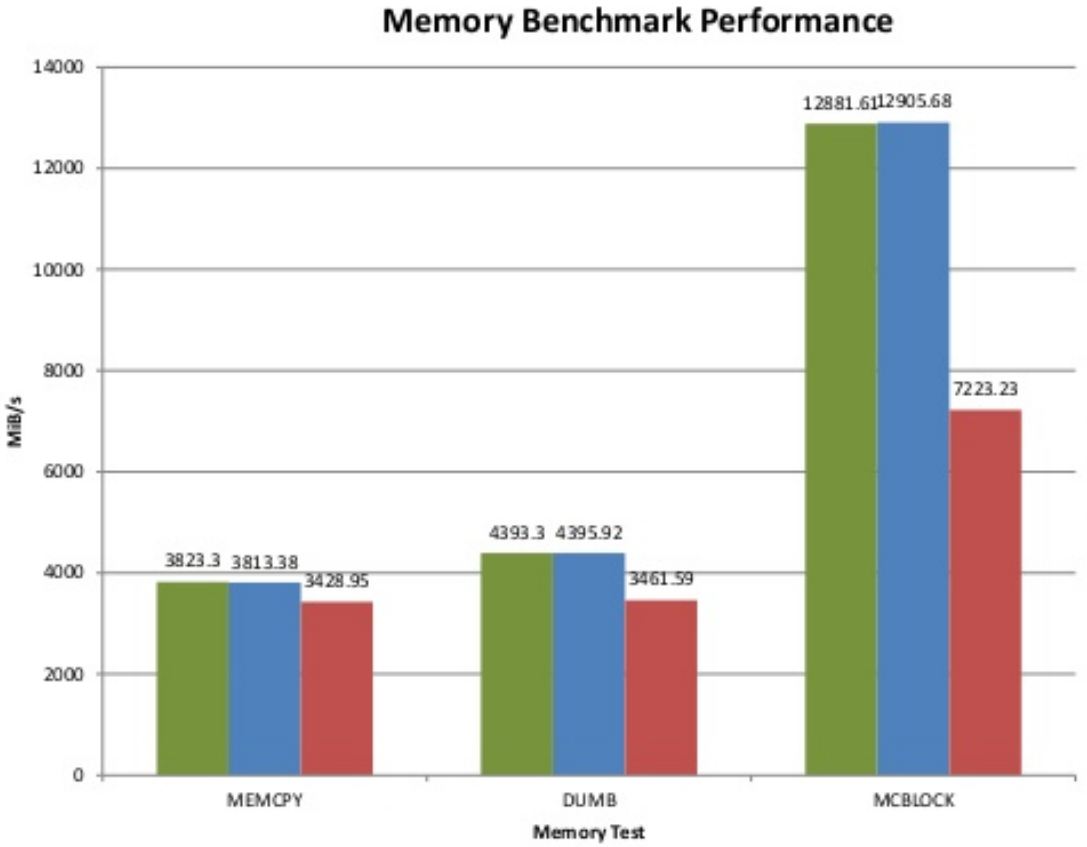
- Comparatif des performance entre une solution native, KVM et DOCKER http://fr.slideshare.net/BodenRussell/kvm-and-docker-lxc-benchmarking-with-openstack
- Comparatif VirtIO pour KVM http://jrs-s.net/2013/05/17/kvm-io-benchmarking/

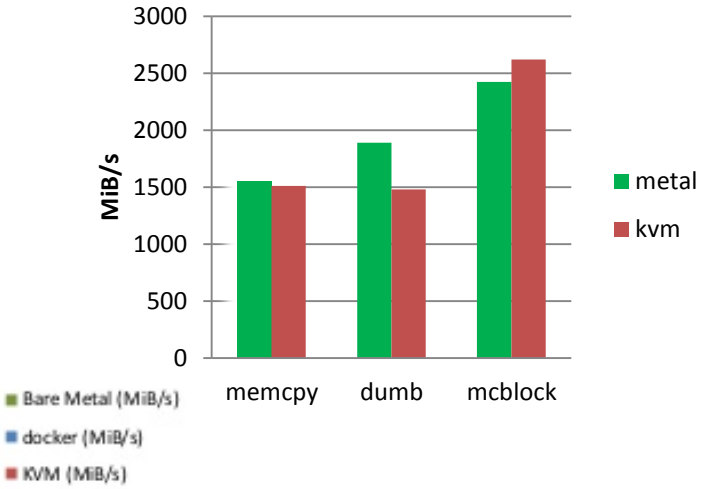
Guest performance : CPU

Linux Sysbench cpu test
VM :
2 vcpu
4Go RAM
20Go de disque



Guest performance : Memory

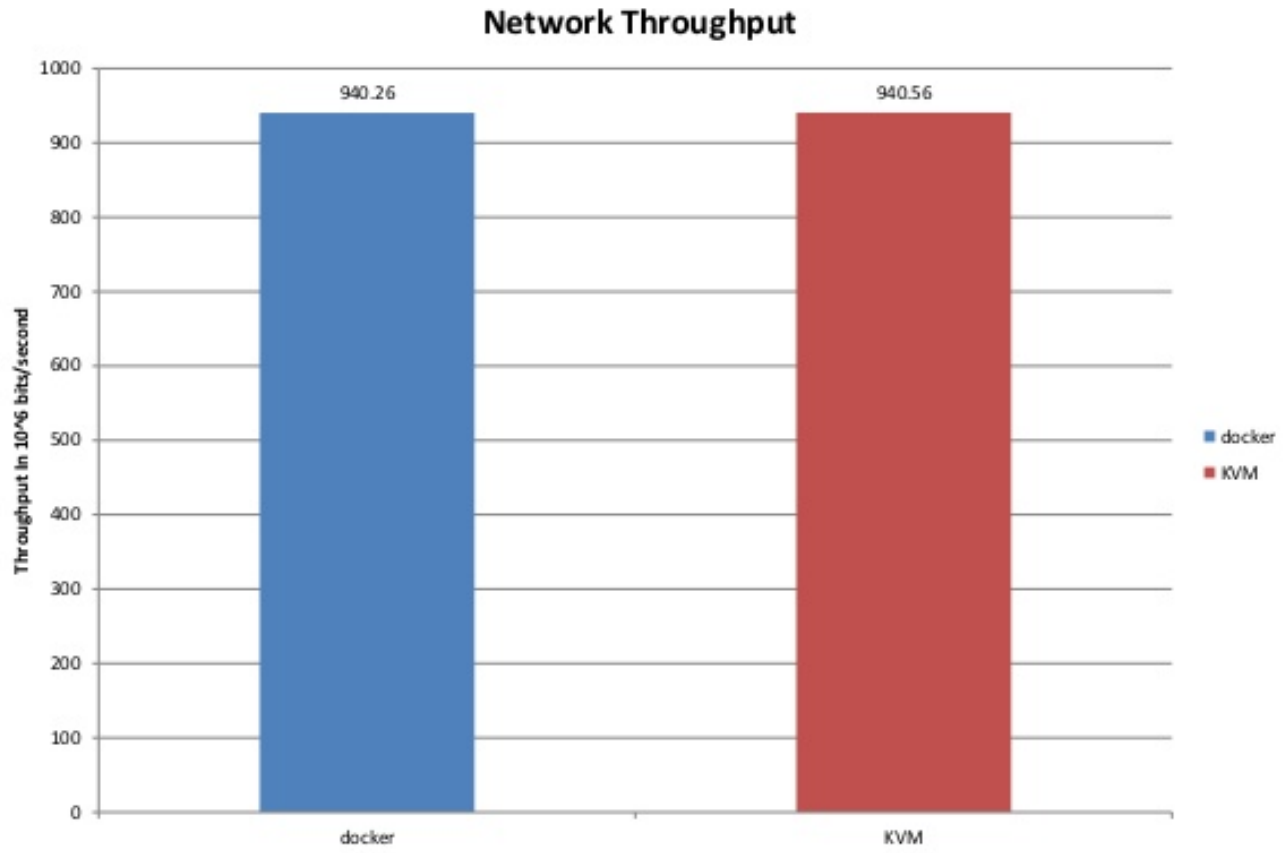




Linux mbw - Memory BandWidth benchmark moyenne sur 10 exécution

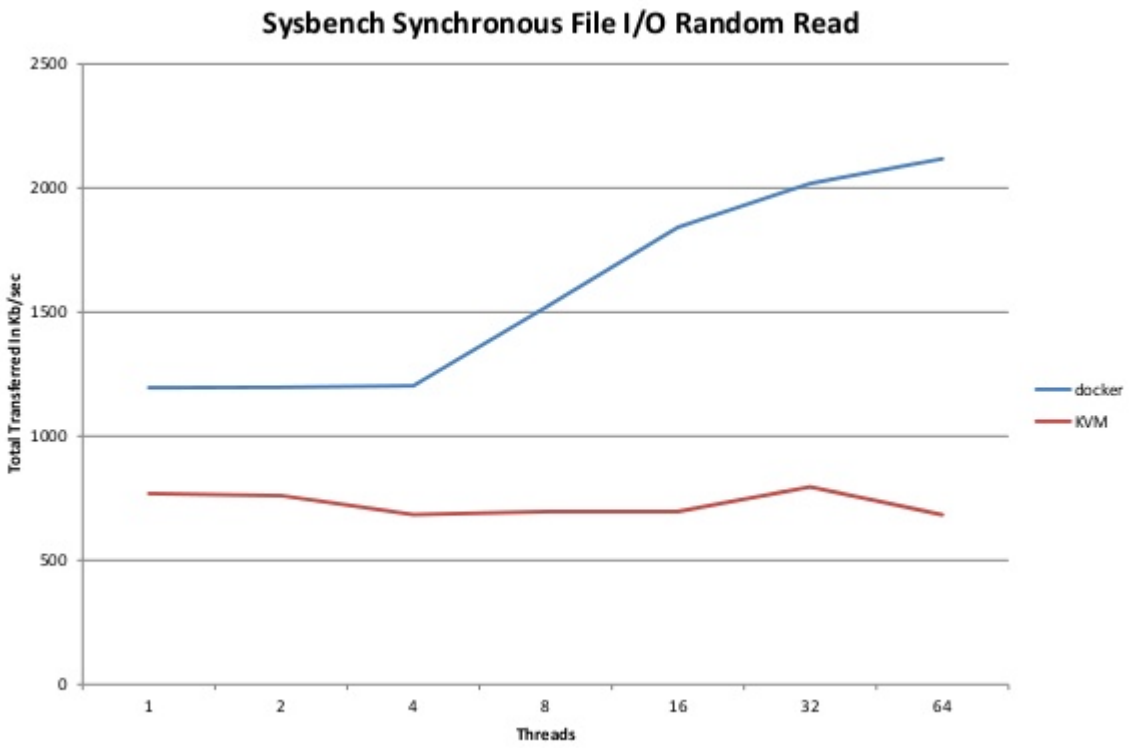
MEMCPY: fct libc DUMP: copy itérative MCBLOCK : memcpy par blocs

Guest performance : Network



Moyenne sur 5 exécutions de Netperf entre le host et le guest

Guest performance : File I/O Random Read

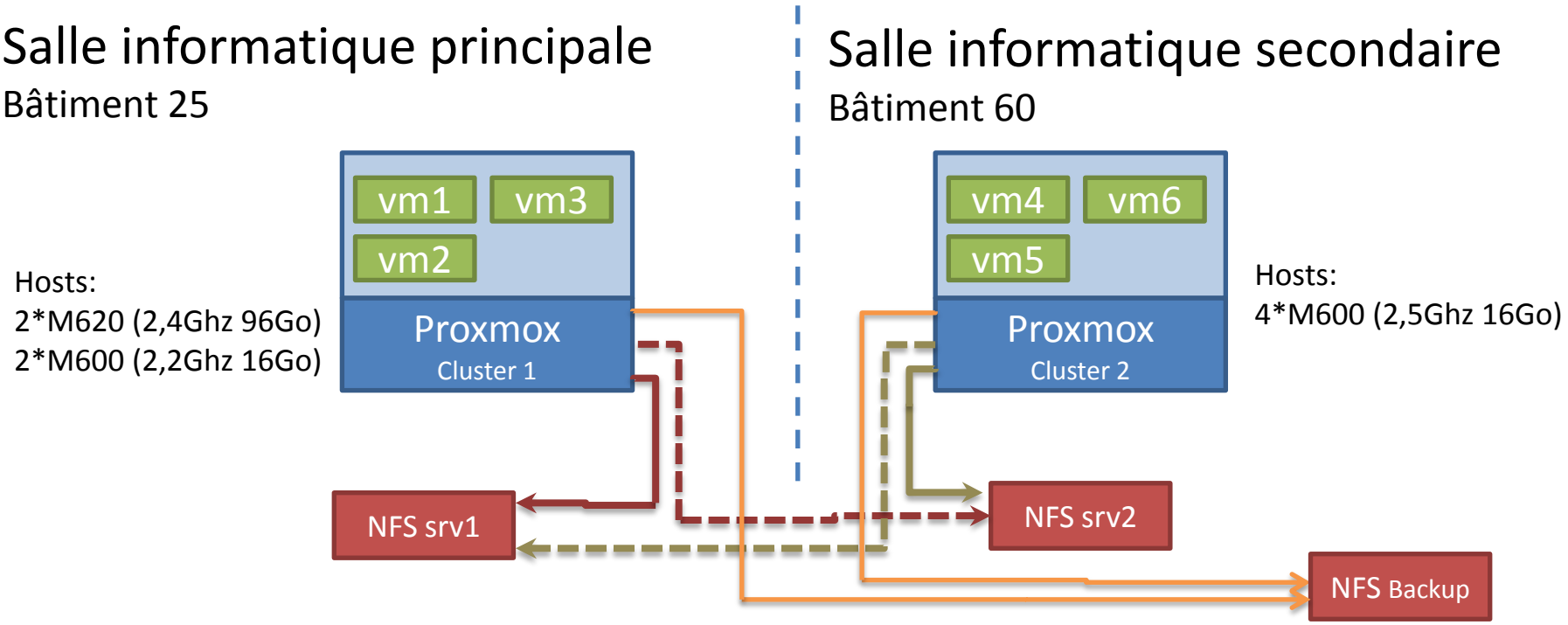


Sysbench lecture synchrone avec différents threads
KVM: disque **sans cache activé** + virtIO
Docker : AUFS

Présentation plate-forme IPHC

- Plusieurs serveurs physiques en fin de support
- Première phase
  - Achat du matériel 07/2012
  - Mise en production 09/2012
    - 2 clusters de 4 nodes sous proxmox 2.1
    - 1 serveur NFS
  - Virtualisation des serveurs physiques
  - Mise à jour v2.2 puis v2.3
- Deuxième phase
  - Achat du matériel 03/2014 (châssis M1000e + R720)
  - Mise en production 06/2014 (2eme serveur NFS avec PRA)
  - Mise à jour 3.x

Schéma de la solution



Data Store :

- principal cluster1
- secondaire cluster1
- backup
- principal cluster2
- secondaire cluster2

Stratégie de backup :
1 backup / semaine
2 historiques / VM

Migration de la version 2.x à 3.x

- Contraintes
  - Ne pas arrêter les VMs ou avoir une coupure de service
  - Live migration ne fonctionne pas entre 2.x et 3.x
- 2 solutions
  - Mise à jour
    - Etre en version 2.3 pour passer 3.0 puis faire les mises à jour vers la version 3.3
    - Espace disque nécessaire pour la mise à jour de la distribution (passage de Squeeze vers Wheezy)
  - Réinstallation du Node
    - Suppression, ré-installation, ajout du node dans le cluster
    - Reconfiguration des spécificités :
      Fence device, clés ssh d'administration, configuration réseau
  - Documentation :
    https://pve.proxmox.com/wiki/Upgrade_from_2.3_to_3.0

Comment déterminer les services à virtualiser

- Dépend des ressources disponibles sur la plateforme
  - VMs par défaut via KVM:
    - 4vcpu / 4Go Ram / 40 Go disque système
    - Réseau 1Gbs Max
  - VMs Extra via KVM ou OpenVZ
    - 8vcpu / 16Go Ram / 40 Go disque système
    - Réseau 1Gbs Max
    - Montage NFS si besoin d'accès aux zones de stockage
- Dépend de l'usage
  - Taux d'utilisation CPU/RAM/Accès R/W
  - Affichage 3D, redirection du son, Accès périphériques USB/série…
- Auditer le serveur physique avec la solution de supervision
  - Connaître le taux de charge, les périodes d'activité

Conversion d'un serveur physique en virtuel

- Capture du serveur physique
  - Ajout des pilotes IDE / réseau Intel E1000
  - Capture à chaud du système du serveur physique : VMware vCenter Converter, Virt-p2v, Snapshot+Robocopy
  - Conversion du disque virtuel au format qcow2 ou Raw
- Configuration d'une nouvelle VM
- Reconfiguration du système
  - Supprimer les outils de gestion (openmanage, agent de backup), les pilotes de périphériques non utilisés
  - Ajout des pilotes de paravirtualisation
- Test de charge
- Synchronisation des données utilisateurs
- Basculement de la configuration réseau

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

Gestion des VMs

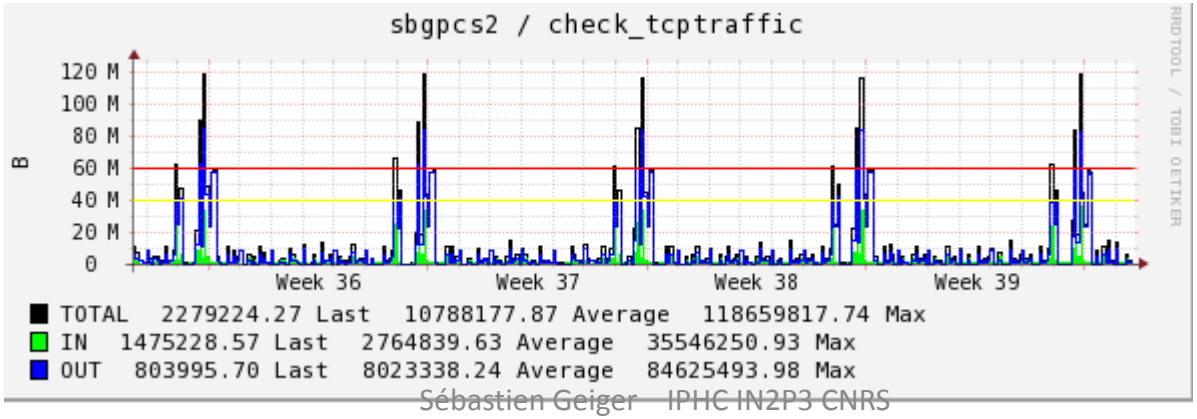
- Placement des VMs
  - Regroupement des VMs par OS pour l'optimisation de la surallocation de la mémoire
  - Equilibrage entre les 2 clusters (75%/25%)
- Allocation des Ressources
  - Extension possible des disques à chaud
  - Adapter les ressources CPU/RAM/Réseau/bande passante
- Plan de reprise d'activité
  - Répartir les services critiques sur les 2 clusters
  - Avoir les ressources pour fonctionner en mode dégradé
  - S'assurer que la documentation soit à jour, accessible et comprise de tout le monde -;)

Backup|Restauration des VMs

- Backup
  - 1 fois par semaine avec 2 historiques par VM
  - VMs de service (ldap, antivirus, inventaire, impression, images)
    - Très peu de changement au niveau des configurations
  - VMs avec des données utilisateurs
    - Utilisation de l'agent de sauvegarde (idem serveur physique)
- Restauration
  - Depuis la solution de virtualisation, puis via l'agent de sauvegarde si nécessaire
- Snapshot
  - utilisé parfois avant une mise à jour système
  - Permet de revenir rapidement en arrière
  - Risque: toutes les applications ne supportent pas le retour en arrière (DC,LDAP,SGBD)

Supervision des services et des VMs

- Nagios pour Unix ou SCOM pour Windows
  - Surveillance du fonctionnement des services et du matériel
  - Collecte des données de performance
  - Graphique sur 4h,25h,1mois,1an, 5ans
- Services virtualisés
  - Mêmes règles, avec les mêmes outils
  - Collecte de nouveaux indicateurs
    - charge des VMs, charge des nœuds de virtualisation
    - Vitesse de lecture / écriture, temps d'accès, bande passante



Liste des services virtualisés

- Actuellement 23 VMs en production, 5 VMs en pré-prod
  - 10 serveurs Windows (contrôleur de domaine, accès au bureau à distance, serveurs antivirus, serveur d'impression, PXE, MS SharePoint)
  - 8 serveurs Linux (outils de supervision, Owncloud, Annuaire LDAP, base de données Mysql, apache, Tomcat, puppet)
  - 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
- Avenir
  - Solution de messagerie de l'IPHC
    - frontal (greylist, spam) + mailbox (smtps, imaps, pops)
    - webmail sous SOGo
  - Serveur de stockage, hébergement web
  - Machines de calcul / simulation sous Windows
  - Interaction avec le cloud sous OpenStack de l'IPHC

Bilan

- 23 VMs en production et 5 VMs en pré-prod
- Virtualisation sous Proxmox VE
  - Produit complet , stable et supporte la montée en charge
  - Nécessite de se documenter (HA, live migration, Backup)
  - Suivre les évolutions de Proxmox VE (passage à la version 3.x)
  - Supporte le monde Linux / Windows
- Gain
  - Simplification de l'administration courante
  - Mode de fonctionnement sans support, ou licence commerciale
  - Diminution des coûts, consommation électrique, encombrement
    - Avant : place : 10 serveurs (16U) + 2tours ; Energie : 3,4Kw
    - Après : place : 8*M6x0 (10U)+2*R720 (6U) ; Energie : 1,6Kw
  - Solution extensible

Evolution des technologies de virtualisation

- KVM
  - Page Delta Compression for Live Migration
  - Ajout vCPU et vDisque à chaud
  - Single Root I/O Virtualization (SR-IOV)
    (accès direct aux périphériques réseau du Node)
  - Hyper-V: support VMs de génération 2
    (périphériques synthétiques, boot UEFI et iSCSI)
  - VmWare: support des open-vm-tools
    (utilisation des périphériques vmware)
- OpenVZ
  - Maintenu par Parallels=>Cloud Server
  - Utilisation de LXC via les commandes vzctl
- VDI
  - Accès via SPICE
    (support des périphériques USB, du son, du copier/coller, …)

Annexe

- Proxmox
  http://pve.proxmox.com/wiki/Main_Page
- Vidéos tutoriels
  http://www.youtube.com/user/ProxmoxVE
- Du bon usage de la virtualisation de serveurs
  https://2011.jres.org/archives/124/index.htm
- network optimization topics for virtualized environments
  https://access.redhat.com/documentation/en-US/Red_Hat_Enterprise_Linux/6-Beta/html-single/Virtualization_Tuning_and_Optimization_Guide/index.html#chap-Virtualization_Tuning_Optimization_Guide-Networking
- virtio-win packages that are supported on Windows operating systems
  http://www.windowsservercatalog.com/results.aspx?text=Red+Hat&bCatID=1282&avc=10&ava=0&OR=5&=Go&chtext=&cstext=&csttext=&chbtext
- http://jrs-s.net/2013/05/17/kvm-io-benchmarking/