ÉCOLE **NORMALE**
SUPÉRIEURE
DE **LYON**

Virtualisation en Open Source

Des ressources « dynamiques » aux ressources « statiques »

Retour d'expérience

« Nous vivons chaque jour dans des environnements virtuels définis par nos idées. »
(Michael Crichton)

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Virtualisation : Un terme approprié ?

- Opposition : « virtuel », pas « réel »
  - Virtuel : *virtus*, la « force » ou la « vertu » en latin
  - Réel : *res,* la « chose » en latin
- Opposition : « logique » & « physique »
- Dans l'usage courant (des informaticiens) :
  - « Émulation » par Qemu
  - « Isolation » par chroot, UML, vServer
  - « Abstraction » de bande pour l'archivage
  - LVM : volume « physique » et volume « logique »
  - Xen : Hôte Dom0 et DomU (Domaine utilisateur)

ÉCOLE **NORMALE**
SUPÉRIEURE
DE **LYON**

Le Centre Blaise Pascal au service de la recherche

Dryden Flight Research Center EC87 0182-14 Photographed 1987
X-29

Nasa X29

- Cellule de F5
- Moteur de F18
- Servos de F16
- Études
  - Flèche inversée
  - Incidence >50°
  - *« Fly-By-Wire »*

EQ au CBP : récupère & réutilise des composants

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Présentation CBP JRES 2012 - Distonet

- Au CBP, parmi les demandes récurrentes :
  - « Environnement de travail » personnalisé :
    - Pour l'utilisation de logiciels spécifiques
    - Pour exploiter les ressources de son poste personnel
    - Pour offrir aux étudiants/chercheurs un cadre générique
- Solution : une machine virtuelle comparable aux autres
  - VirtualBox Open Source « *... it is also the only professional solution that is freely available as Open Source Software under the terms of the GNU General Public License (GPL) version 2.* »
  - Système Debian Linux amd64 ou i386 Squeeze :
    - Tous les logiciels science-* installés plus les logiciels spécifiques...

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Présentation CBP
JRES 2012 – Distonet – Le socle

- Pourquoi ?
  - Solution (encore) sous contrôle
  - Liaison entre système hôte et VM
  - Espace de disque « à discrétion »
  - Administration à distance
- Comment ?
  - Autonome : système sur espace disque
  - Distante : démarrage par PXE NFSROOT ou iSCSI
- Liaison ?
  - Réseau direct
  - Réseau privé indirect (via passerelle OpenVPN)

ÉCOLE NORMALE SUPÉRIEURE DE LYON

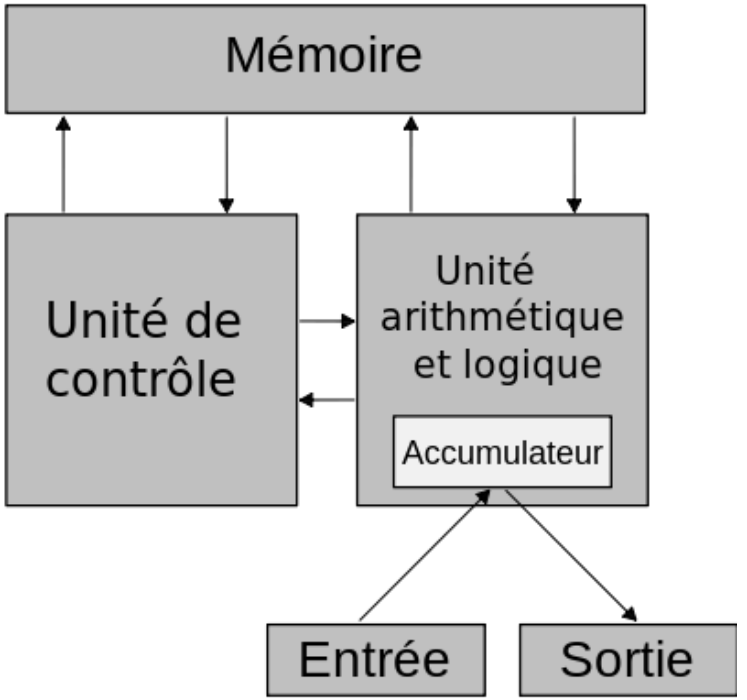
Virtualisation Open Source Pourquoi ? (source Wikipedia)

- Une « salle machine » dans une « machine »
  - Partager CPU/RAM/Disque/Périphériques
- Optimiser l'usage du matériel
- Installer, déployer, migrer plus facilement
- Mutualiser les ressources hôte
- Mettre à disposition des environnements matériels
- Mettre à disposition des environnements de tests
- Allouer des ressources à la demande
- Limiter les risques de dimensionnement

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Virtualisation Open Source Mais Quoi ?

- Modèle de Von Neumann (la base) :
  - Une UAL/UC (en fait, plein...)
  - Une mémoire (en fait « des »)
  - Des Entrées/Sorties
- Ressources « Internes »
  - Processeurs
  - Mémoire vive
- Ressources « Externes »
  - Espace de stockage
  - Périphériques



ÉCOLE NORMALE SUPÉRIEURE DE LYON

Virtualisation Open Source Mais Comment?

- Ressources Processeur/Mémoire
  - OpenSource : Qemu, Xen, KVM, VirtualBox
- Ressources Disques :
  - Simple : NFSROOT, iSCSI, AoE
  - Distribué : GlusterFS, XTreemFS, CephFS, ...
- Ressources matérielles
  - Interface réseau
  - Extension VT-d : (mouais...)
    - Cartes Infiniband : HPC
    - Cartes Vidéo : calcul GPU

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Virtualiser en Open Source ? Mais Pour Qui ?

- Administrateur
  - Une manière de simplifier son parc
  - Clonage facile
- Développeur/Intégrateur
  - Une manière de tester « au niveau machine »
  - Une multiplication de sa plate-forme
- Utilisateur
  - Mon environnement « à moi »
  - Un environnement « en plus » complètement intégré

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Au début, Chroot & Qemu...

- Déjà de la virtualisation (plutôt de l'isolation)
  - Plusieurs « systèmes » sur un noyau
  - Un système RHCE 2.1 avec tout antédiluvien
    - Impossibilité d'installer quoi que ce soit...
  - Des distributions Debian chrootées
- Journée Méso à l'ENS-Cachan en février 2006
  - 4 utilisations à prévoir en HPC :
    - Python comme langage scientifique
    - Processeurs graphiques
    - **Virtualisation pour l'apprentissage X>1 processeurs**
    - Algorithmes génétiques
- Mais, dès 2006, généralisation multi-cœurs

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Du Cecam de mi 2009... … au CBP de mi 2012

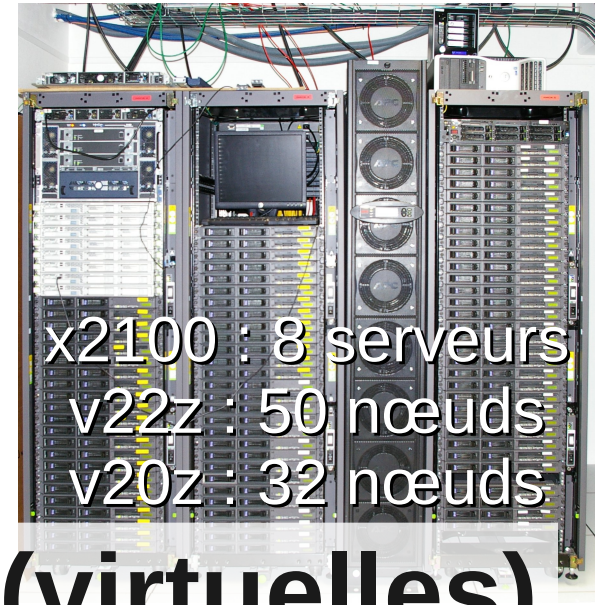
DNS
MTA
MUA

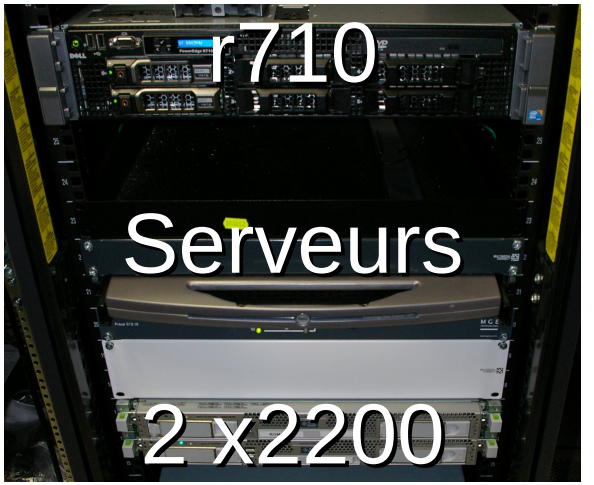
NFS
NIS
DHCP
XDMCP

WWW
8
sites

15
Neoware

8 P360





De 28 à 152(+37) machines (virtuelles)

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Du Cecam au CBP : de l'état des lieux à la migration

- Infrastructure complète sur Internet !
  - MTA (Postfix relais ouvert),
  - MUA (Cyrus)
  - DNS (Bind)
  - XDMCP
  - WWW (Apache/Tikiwiki)
  - NIS/NFS
  - Pare-feu
- Des OS antédiluviens : SuSE 9.2 (2004)

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Du Cecam au CBP : migration

- Un objectif
  - Refaire ce qui peut l'être
  - Conserver ce qui ne peut pas
- Un chemin
  - Séparer les services :
    - NFS, SMB/CIFS, WWW, SGBD, ...
  - Créer de nouveaux services
    - Serveurs Web externe
    - Serveurs de fichiers avec authentification établissement, SGBD
- Des moyens : seulement 2 serveurs Sun x2200
  - Recours à la virtualisation

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Le choix de Xen™

Connaissance/Simplicité

- Déjà une bonne expérience (2 ans au PSI)
- Facilité de création des DomU
  - Un « xen-create-image » pour chacun
  - Un LV de LVM pour chaque DomU (enfin 2, au moins)
- Problème :
  - Et pour les sites Web antédiluviens ?
- Possibilité de création de machine « HVM »
  - Création d'un LV de volume identique (exactement)
  - Copie bit à bit de l'image disque
  - Démarrage...

ÉCOLE NORMALE SUPÉRIEURE DE LYON

L'implémentation sous Xen

- Les images disques : volumes logiques
  - Un clonage pour dupliquer les machines
    - Utilisation pour la Forge à disposition du LIP
- Les ressources dédiées
  - Pas de nécessité directe
- Le réseau
  - Utilisation de ponts Ethernet pour les différents réseaux
- Efficacité :
  - I/O : très correct

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Pour plus de virtualisation

- Pourquoi ?
  - Avant la Squeeze, un futur compromis dans Debian
  - Un noyau très modifié
  - Une impossibilité de fournir des plates-formes *borderline*
    - Autres noyaux
    - Autres distributions
- Comment ?
  - Des machines « récentes » (avec VMX ou SVM)
  - Virt-manager comme GUI
  - Virsh pour la commande en ligne
  - Un manque : le « kvm-create-image »

ÉCOLE NORMALE SUPÉRIEURE DE LYON

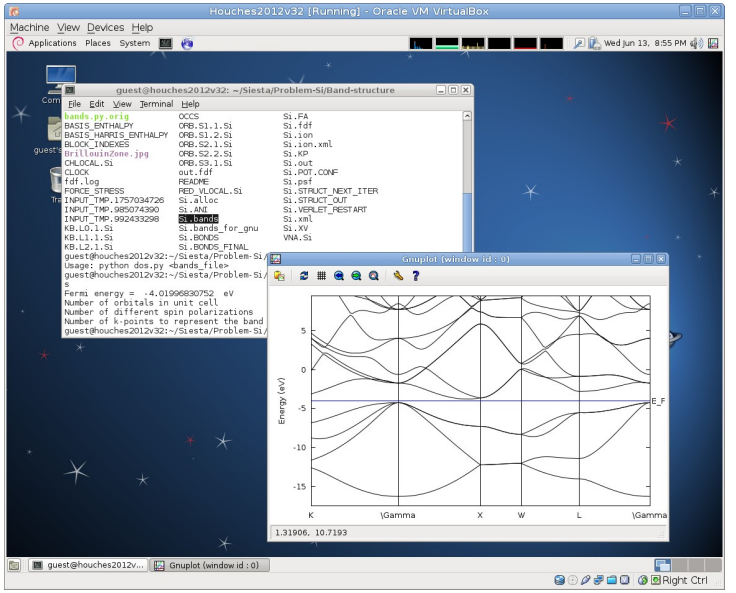
VirtualBox
La virtualisation pour tous

- Pourquoi ?
  - Une demande originelle : tourner Vasp sur MacOSX
  - Une proposition : faire tourner le code dans une VM
- Quoi ?
  - Création d'une image VirtualBox
  - Installation d'un environnement Debian scientifique
  - Installation des paquets complémentaires
  - Exportation/Importation de la VM
- Pour qui ?
  - Initialement 1 chercheuse, puis ses 4 collaborateurs

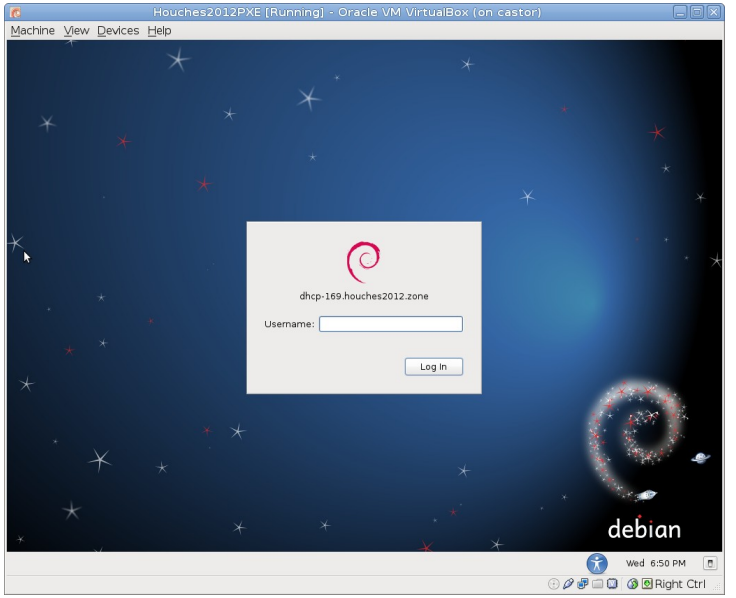
ÉCOLE NORMALE SUPÉRIEURE DE LYON

VirtualBox
La virtualisation pour tous

- **Pour où ? En standalone !**
  - Une machine virtuelle de 10Go
  - Environnement fonctionnel
  - Téléchargement de 3.5Go



- **Pour où ? Dans un réseau !**
  - Une machine virtuelle de 13Ko
  - Environnement complet
  - Téléchargement de 13Ko



ÉCOLE NORMALE SUPÉRIEURE DE LYON

Virtualiser le stockage
De XDMCP à NFSRoot

- **Avant** : postes Neoware ~ un terminal X...
  - 500 MHz, 128 Mo de RAM, Flash de 256 Mo
  - Support de WWW, XDMCP, RDP
- **Après** : postes Neoware ~ un client léger
  - Passage à 1 Ghz par OC, 1Go de RAM, Flash GPXE
  - Système complet par NFSRoot
- v1 : NFSRoot RO et RW par TmpFS
- v2 : NFSRoot RO et RW par TmpFS over AUFS

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Virtualiser le stockage : Du disque local au NFSROOT

- **Avant** : 24 nœuds SunFire v20z : un OS classique
  - 2 cœurs avec 4 Go de RAM
  - 2 disques de 146 Go : un OS, un **scratch**
  - 1 CentOS
- **Après** : un système unique, quoique
  - Démarrage par PXE
  - Système complet par NFSRoot
- V1 : NFSROOT RO et RW par TmpFS over AUFS
- V2 : NFSROOT RO et RW par TmpFS over NFS

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Virtualiser le stockage : Du disque local au NFSROOT

- Ca « scale » ?
  - De 24 v20z à 48+16 v22z+x41 : OK
- Et la persistance de données ?
  - V1 TmpFS
  - V2 AUFS+NFS
  - V3 AUFS+iSCSI
- Et le portage ?
  - École de Physique des Houches
  - PSMN : en cours pour Equip@Meso

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Virtualiser le stockage : Du disque local au iSCSI

- Le protocole :
  - Créer une partition dans un LVM ou ZFS et la formater
  - Installer un système par debootstrap et chroot
  - Dupliquer avec les outils puis activer les duplications
  - Partager les partitions par iSCSI
  - Importer le initrd
  - Déclarer pour chaque boot PXE chaque machine
- Ensuite : choisir un démarrage par PXE
- Bon, quelques ruses :
  - Utiliser un Label pour le root,

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Démarrage iSCSI
Quelles réalisations ?

- 8 postes clients de la salle libre service
  - Precision 360 : Un master cloné 8 fois
- Equip@Meso : machines de constructeurs *diskless*
  - C6100 : master Cuda64 cloné 4x
  - SL250 : master Cuda64 cloné 1x
  - DL585 : master Squeeze64 cloné 1x
- Amalia : démonstrateurs Humanités Numériques
  - Sunfire v20z : master Squeeze4HS cloné 3x

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Petite conclusion CPU/RAM : Xen

- Les Plus :
  - Le support encore présent
  - La paravirtualisation (pas d'exigence de SVM ou VMX)
  - La possibilité de virtualisation matérielle
  - Le binding matériel
- Les Moins :
  - Un noyau très modifié fonctionnant parfois « mal »
  - Un TTY à modifier pour le mode console
  - Le noyau de l'invité est le noyau de l'hôte
  - Uniquement x86 et x86_64

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Petite conclusion CPU/RAM : KVM

- Les Plus :
  - Machine virtualisée complètement (pas noyau hôte)
  - Le support natif dans le noyau et les distributions
  - La réservation matérielle (pas toujours fonctionnelle)
- Les Moins :
  - Uniquement pour SVM et VMX « inside »
  - Création de la machine
  - Réservation matérielle capricieuse
  - x86 et x86_64 (mais portage ARM ...)

ÉCOLE NORMALE SUPÉRIEURE DE LYON

CPU/RAM : VirtualBox

- Les Plus :
  - Le côté multi-plateforme :
  - Une très bonne gestion des import/export
  - L'importation « simple » de composants USB
  - Le « *passthrough* » matériel (il paraît...)
- Les Moins :
  - Le support propriétaire de certains composants
  - Des performances « bizarres » de composants réseau
  - Uniquement x86 et x86_64

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Stockage : iSCSI

- Facilité de mise en œuvre
  - Côté Client
    - Uniquement OpenISCSI
    - Démarrage par PXE, puis montage disque distant
    - Authentification CHAP, TCP wrappers possible
  - Côté Serveur (cible ou « target »)
    - ISCSItarget (module noyau) ou TGT (userland)
    - Déclaration très simple
    - Attention aux nouvelles déclarations & leur chargement
- Inconvénients
  - Un petit bug en 3.2 (pas en 2.6.32) sur OpenISCSI

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Stockage : NFSRoot

- Facilité de mise en œuvre
  - Côté Client
    - Uniquement NFS client
    - Démarrage par PXE puis montage NFSROOT
  - Côté Serveur (cible ou « target »)
    - Partage NFS standard
    - Filtrage par adresse IP (ou MAC)
    - Déclaration PXE
- Inconvénients
  - Du NFS, donc tous les inconvénients de NFS...
  - Pas de manipulation de certaines types de partitions

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Conclusion
La virtualisation : oui mais...

Ressources dynamiques

- On gagne :
  - Flexibilité
  - Administration
- On perd :
  - Performances : 1 à 20%
  - Prédictibilité

Ressources statiques

- On gagne :
  - Flexibilité
  - Sécurité (déport)
- On perd :
  - Importance réseau
  - Sécurité (média)

ÉCOLE NORMALE SUPÉRIEURE DE LYON

Perspectives

- Futur en calcul scientifique ?
  - Convergence Poste Utilisateur/Nœud de Calcul
    - Le poste deviendra un nœud de calcul (nuage local)
    - L'environnement HPC sera celui du poste
    - Stockage distribué
  - Émergence ARM :
    - Massivement parallèle : plusieurs centaines de processeurs
    - Massivement distribué : plusieurs milliers de terminaux
- Finalement, le gagnant, c'est l'utilisateur !

ÉCOLE **NORMALE**
SUPÉRIEURE
DE **LYON**

Droits

- Iconographie
  - Nasa Dryden : X-29
  - Wikipedia : machine de Von Neumann
  - Nec : TX7
  - Xen : logo
  - Linux : KVM
  - Oracle : VirtualBox