virtualisation XEN
avec Linux CentOS
à l'Université Montpellier 2

CUME - 13 mars 2008

Marc FANGEAUD - CR2I UM2 -
Xen_CUME_080313

Quelques missions de l'équipe Réseau CR2I UM2

- Gestion du réseau fédérateur jusqu'au laboratoire
  - environ 110 switchs
  - 5600 machines
- Gestion des services "réseaux" (pas les applicatifs "gestion")
  - DNS
  - WWW (environ 170 sites hébergés)
  - FTP
  - Proxy
  - Courrier électronique (environ 5600 comptes)
  - ...
- → environ 40 serveurs
- → 2 administrateurs système/réseaux

Problématique Système et Sécurité

- 1 service / serveur → sécurité accrue
  (l'exploitation d'une faille de sécurité ne compromet qu'1 serveur)
- 1 service / serveur → maintenance facilitée
  (éviter des "surprises" lors de maj. système)
- MAIS :
  - 1 service / serveur → cher !
  - redondance : plusieurs serveurs pour 1 service → cher !
  - 1 service / serveur → coûteux en temps (si pas de solution de clonage)
  - 1 service / serveur → une machine peut passer son temps à ne rien faire
    (DNS, Radius, ... sur un Xeon Quad-Core ???)

Problématique Système et Sécurité

- une solution : la (para-)virtualisation
  - plusieurs serveurs virtuels / serveur physique (coût diminué)
  - NE PAS anticiper les besoins (augmentation de la puissance d'une machine virtuelle quand le besoin s'en fait sentir)
  - déploiement facilité (temps de mise en œuvre réduit)
  - fiabilité des services améliorée (plus rapide de "réinstaller" une machine virtuelle qu'une machine physique)

les techniques
de la (para-)virtualisation

- différents niveaux d'exécution sur un OS non-virtualisé :
  - 0 = système d'exploitation (système "hardware" = système "utile")
  - 1 = application

- 1 niveau d'exécution supplémentaire sur un OS virtualisé :
  - -1 = système d'exploitation hôte (système "hardware")
  - 0 = système d'exploitation invité/virtualisé (système "utile")
  - 1 = application

les techniques et fonctionnalités de la (para-)virtualisation

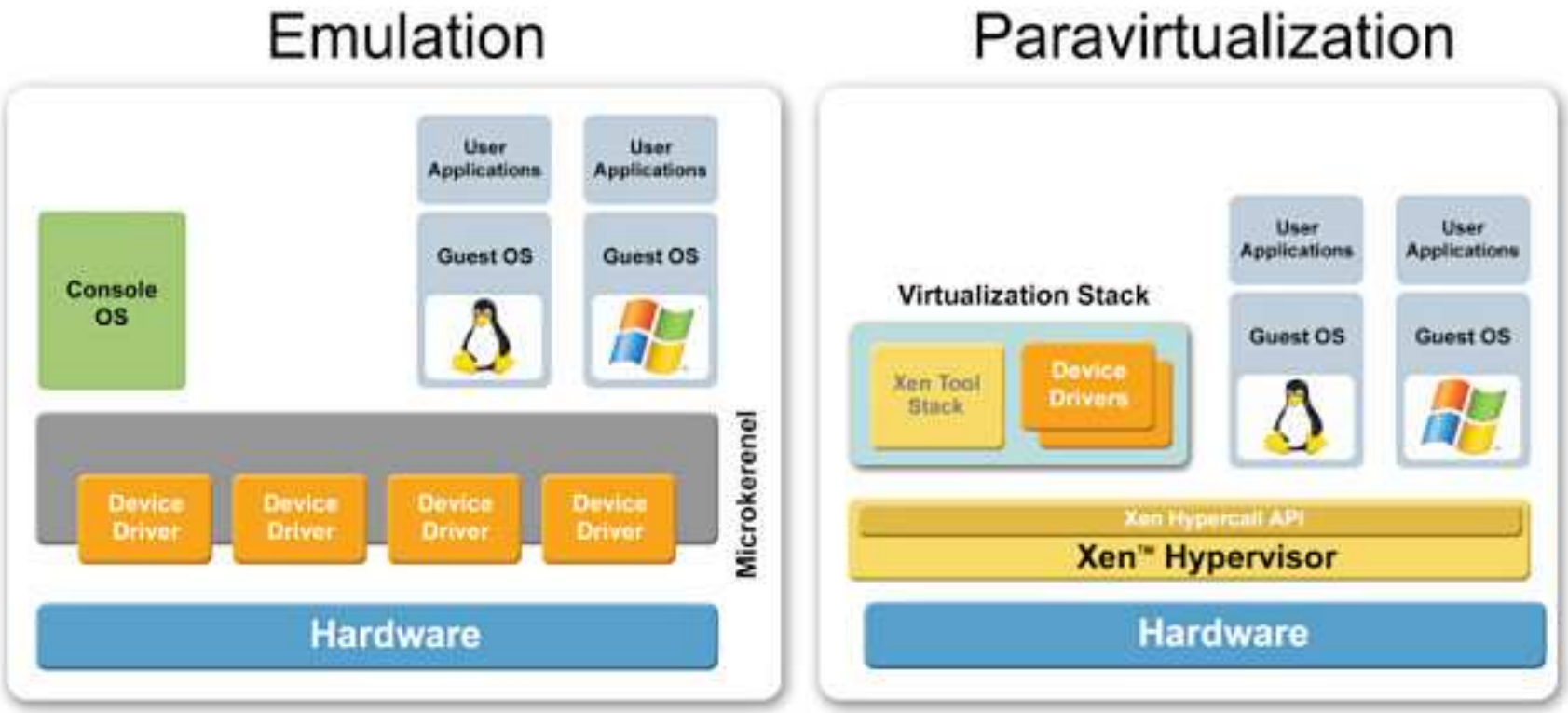
- → OS hôte :
  - Couche d'abstraction matérielle (réseau, mémoire, disque, ...)
  - Partitionnement, isolation et/ou partage des ressources physiques
  - Gestion des "images" des OS invités
  - Clonage, sauvegarde et restauration des "images" des OS invités
  - Démarrage, arrêt, gel, migration d'un OS invité d'une machine physique vers une autre

- → OS invité :
  - architecture matérielle connue différente de l'architecture matérielle réelle
  - Avec des pilotes
    - Génériques (virtualisation)
    - Adaptés à l'OS hôte (para-virtualisation)

Virtualisation vs ParaVirtualisation

- Virtualisation (émulation)
  - Pilotes génériques
  - Interception et traduction des appels système des OS invités
  - → CPU consommée
- ParaVirtualisation
  - Accès « direct » au matériel via des pilotes adaptés/spécifiques à l'hyperviseur

Virtualisation vs ParaVirtualisation



XEN – principes

- XEN = noyau Linux modifié + outils (hyperviseur)
  - si Intel VT ou AMD Pacifica,
    Xen 3.x peut héberger des OS invités non modifiés (QEMU)
    (virtualisation complète mais pas plus rapide)
  - sinon, les OS invités doivent être modifiés (Linux, NetBSD et OpenBSD)

  - Isolation complète entre les machines virtuelles
  - Performances pour les machines virtuelles proches d'un système natif

  - machine virtuelle = domaine
    - dom0 : domaine avec privilèges
    - domU : domaine sans privilèges (machines virtuelles invitées)

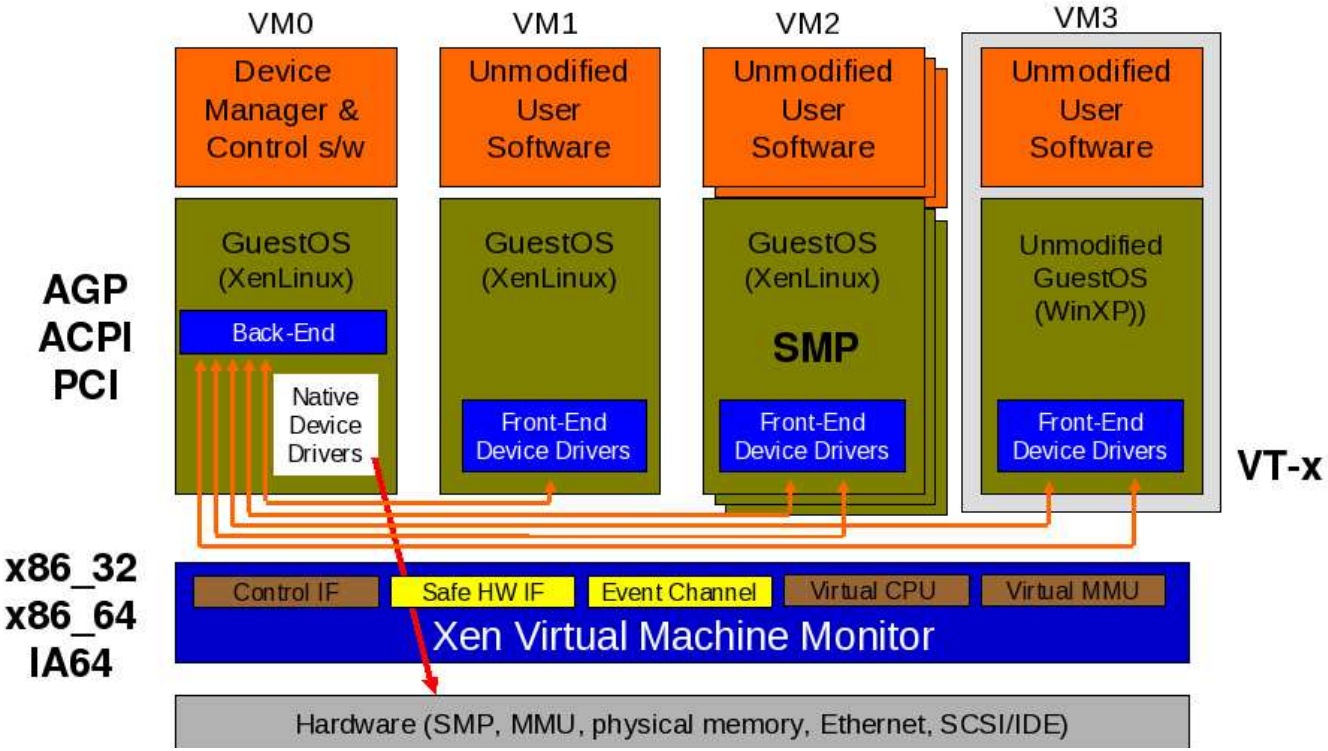
XEN – principes

- dom0 (domaine privilégié) + Hyperviseur
  - lancé automatiquement au boot
  - gestion du harware (via les pilotes standards du noyau Linux)
    → le seul à pouvoir interagir directement avec le matériel
  - Gestion du temps d'utilisation de la machine hôte par les domaines invités
  - gestion de la mémoire
  - Gestion des tâches d'administration du système
    via le daemon xend dans l'espace utilisateur
    (création, démarrage, arrêt, restauration ou migration des domaines)
  - accès natif au matériel (pilotes "backend")
  - Fonction de switch Ethernet virtuel

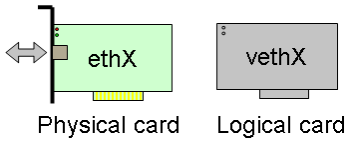
XEN – principes

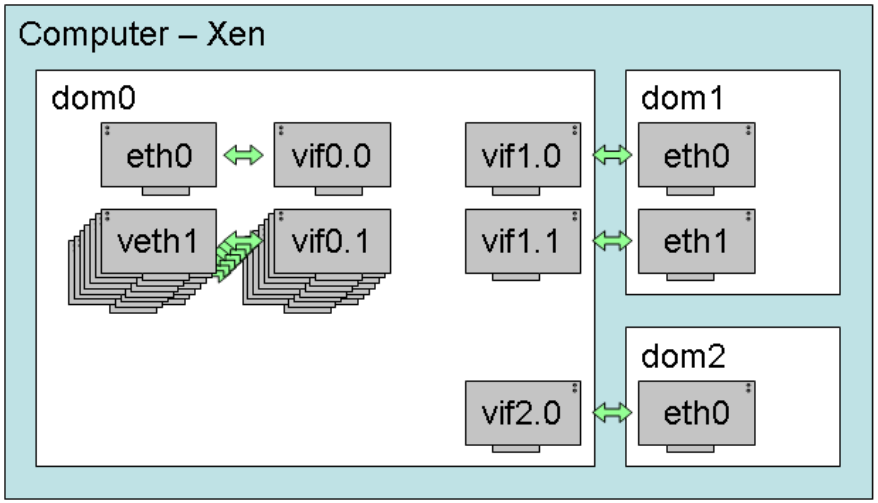
- domU (domaine non privilégié)
  - contrôlé(s) par le dom0
  - noyau modifié et pilotes virtuels ("frontend") qui communiquent avec les pilotes "backend" du dom0
    → noyau "xenifié" (sauf si Intel VT ou AMD Pacifica)
    - "xenblk" : mode bloc pour les disques
    - "xennet" : cartes réseaux virtuelles

XEN – principes



XEN – principes
(réseau)





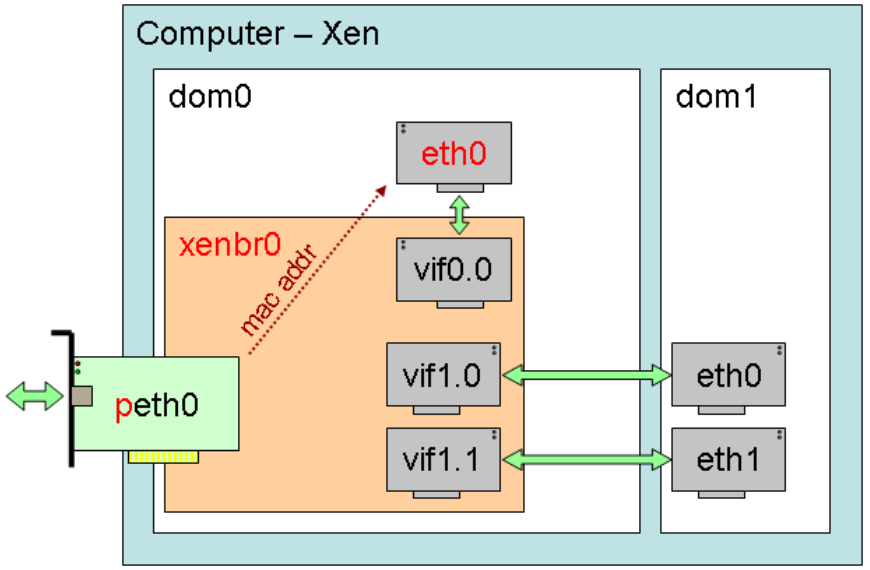
- réseau virtuel
  - paires d'interface virtuelles Ethernet utilisées par le dom0
    - une partie de chaque paire est dans le domU (eth0)
    - l'autre partie est dans le dom0 (vif<id_domaine>.0)

XEN – principes

(réseau)

- réseau virtuel
  - adresse MAC aléatoire pour chaque interface réseau virtualisée des domU
    (sauf si précisée dans le fichier de configuration de la machine virtuelle)
  - possibilité d'utiliser des VLANs
    - support du 802.1q par le dom0
    - un bridge par VLAN
    - les domU n'ont pas connaissance des VLANs
  - Le réseau (l'@IP) du domU peut être configuré depuis le dom0

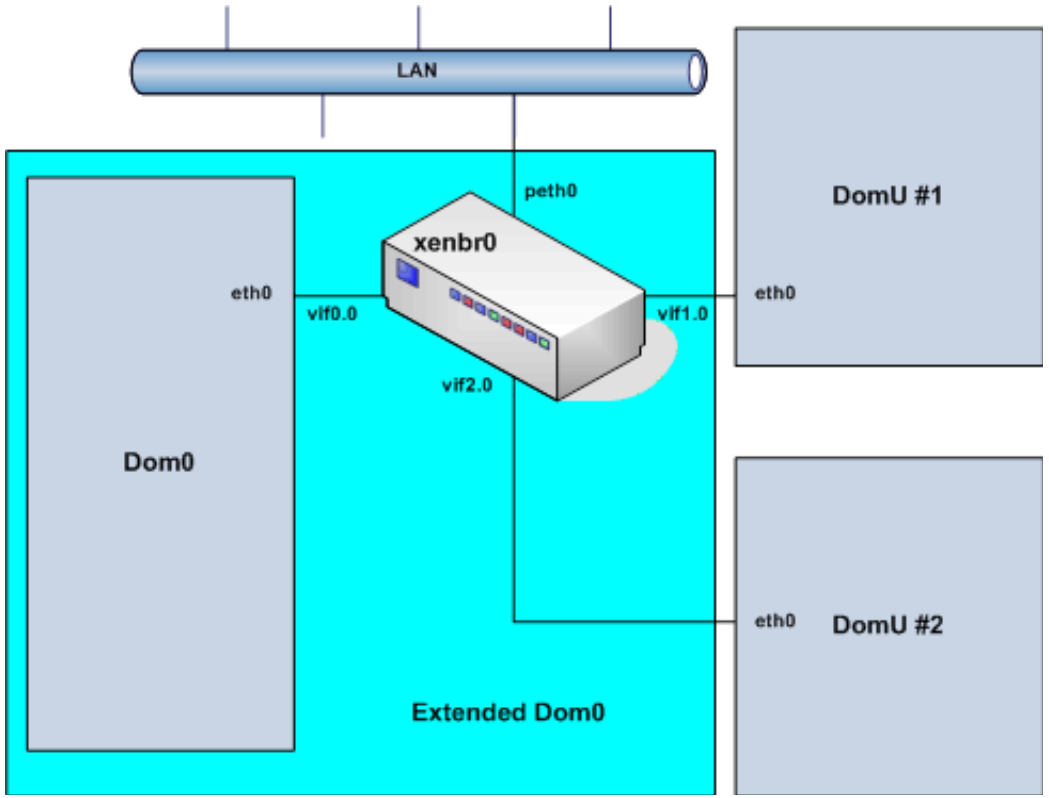
XEN – principes
(réseau)



- réseau virtuel
  - différents modes
    - Bridge
      - pont entre cartes virtuelles et carte physique
      - les vifX.Y n'ont PAS d'adresse IP
      - pas de contrôle au niveau du dom0 (ebtables ?)
      - eth0 du dom0 devient peth0, veth0 renommée en eth0, peth0 et vif0.0 attachées au bridge xenbr0
      - → attention au SpanningTree

XEN – principes
(réseau)

- réseau virtuel
  - différents modes
    - Bridge



XEN – principes
(réseau)

- réseau virtuel
  - différents modes
    - NAT
      - dom0 joue le rôle de passerelle pour les domU
      - règles iptables applicables à ces cartes sur dom0
    - Route
      - les vifX.Y ont pour adresse IP celle des cartes dans les domU
      - elles ne voient pas passer les paquets

XEN – principes

- Mémoire
  - quantité de mémoire donnée à un OS invité n'est pas définitive
    (modification de la quantité de mémoire attribuée à un système invité tout en continuant son exécution)
    (impossibilité de diminuer la quantité de mémoire d'une machine virtuelle en dessous de la quantité initialement donnée)

    → balloon driver

XEN – principes

- CPU
  - VCPU = Virtual Central Process Unit
  - 1 processeur = 1 VCPU
    (1 thread = 1 VCPU)
    (1 coeur = 1 VCPU)

XEN - administration

- gestion des ressources CPU
  - Ajout/suppression à la volée de CPU virtuelles
    - ```
      cat /proc/cpuinfo |grep 'processor'|wc –l
      1

      xm  vcpu-set  domU  4

      cat /proc/cpuinfo |grep 'processor'|wc –l
      4
      ```
  - Affectation des CPU virtuelles aux CPU physiques
    - ```
      xm  vcpu-pin  domU  0  1
      ```

XEN - administration

- Gestion des ressources mémoire
  - `xm mem-max # Sets the maximum amount of memory # for a domain`
  - `xm mem-set # Sets the current memory usage # for a domain`
- Gestion des ressources disque
  - `xm block-attach # Creates a new virtual block device`
  - `xm block-detach # Destroys a domain’s virtual # block device`
- Gestion des ressources réseaux
  - `xm network-attach # Creates a new network device`
  - `xm network-detach # Destroys a network device`

XEN - administration

- gestion des domU
  - `xm create domU` # créer et démarrer un domU
  - `xm shutdown domU` # arrêter un domU
  - `xm destroy domU` # arrêter (BRUTALEMENT) un domU
  - `xm reboot domU` # redémarrer un domU
  - `xm pause domU` # mise en pause du domaine
consommation mémoire
pas de consommation CPU
  - `xm unpause domU`
  - `xm save domU hiberfile.sys` # "hibernation" du domaine
aucune consommation mémoire/CPU
  - `xm restore hiberfile.sys` #

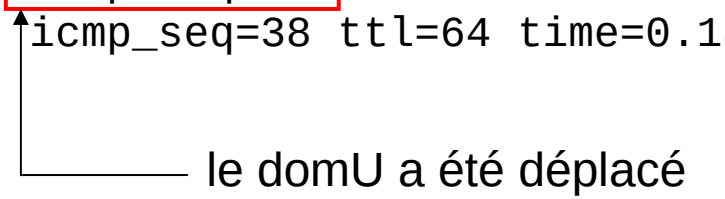
XEN - administration

- gestion des domU
  - `xm migrate --live domU dom0B` `# transfert d'un domU # d'un dom0 # vers un autre dom0`
    - déplacement de la mémoire non utilisée en écriture
    - pause du domU (60 à 300ms)
    - déplacement de la mémoire utilisée en écriture
    - réveil du domU
    - → connexions réseau ouvertes conservées !
    - → les 2 dom0 doivent avoir accès à l'image disque du domU (SAN) et doivent avoir la même version de Xen

XEN - administration

- gestion des domU
  - `xm migrate --live domU dom0B`
  - `ping domU`

```
64 bytes from 162.38.101.222: icmp_seq=30 ttl=64 time=0.169 ms
64 bytes from 162.38.101.222: icmp_seq=31 ttl=64 time=0.176 ms
64 bytes from 162.38.101.222: icmp_seq=37 ttl=64 time=0.176 ms
64 bytes from 162.38.101.222: icmp_seq=38 ttl=64 time=0.187 ms
```



XEN - administration

- État des domU
  - `xm uptime`
  - `xm top`
  - `xm list`
  - `xm vcpu-list`
  - `xm network-list`
  - `xm block-list`
  - `xm dmesg`
  - `xm log`

XEN - administration

- Mise à jour d’un domU

```
lvcreate --snapshot --size 128m  --name  domU-snap /dev/vglocal/domU
xm create -c domUsnap
…
xm shutdown domUsnap
dump/restore
xm shutdown domU
lvremove / dev/vglocal/domU
lvrename / dev/vglocal/domU-snap / dev/vglocal/domU
xm create -c domU
```

XEN - administration

- Sauvegarde d’un domU

```
lvcreate --snapshot \
         --size 128m \
         --name  domU-snap /dev/vglocal/domU
```

```
dump 0ufz9 - /dev/vglocal/domU-snap \
         | ssh –i sshkey cr2i@dom0B restore xf - .
```

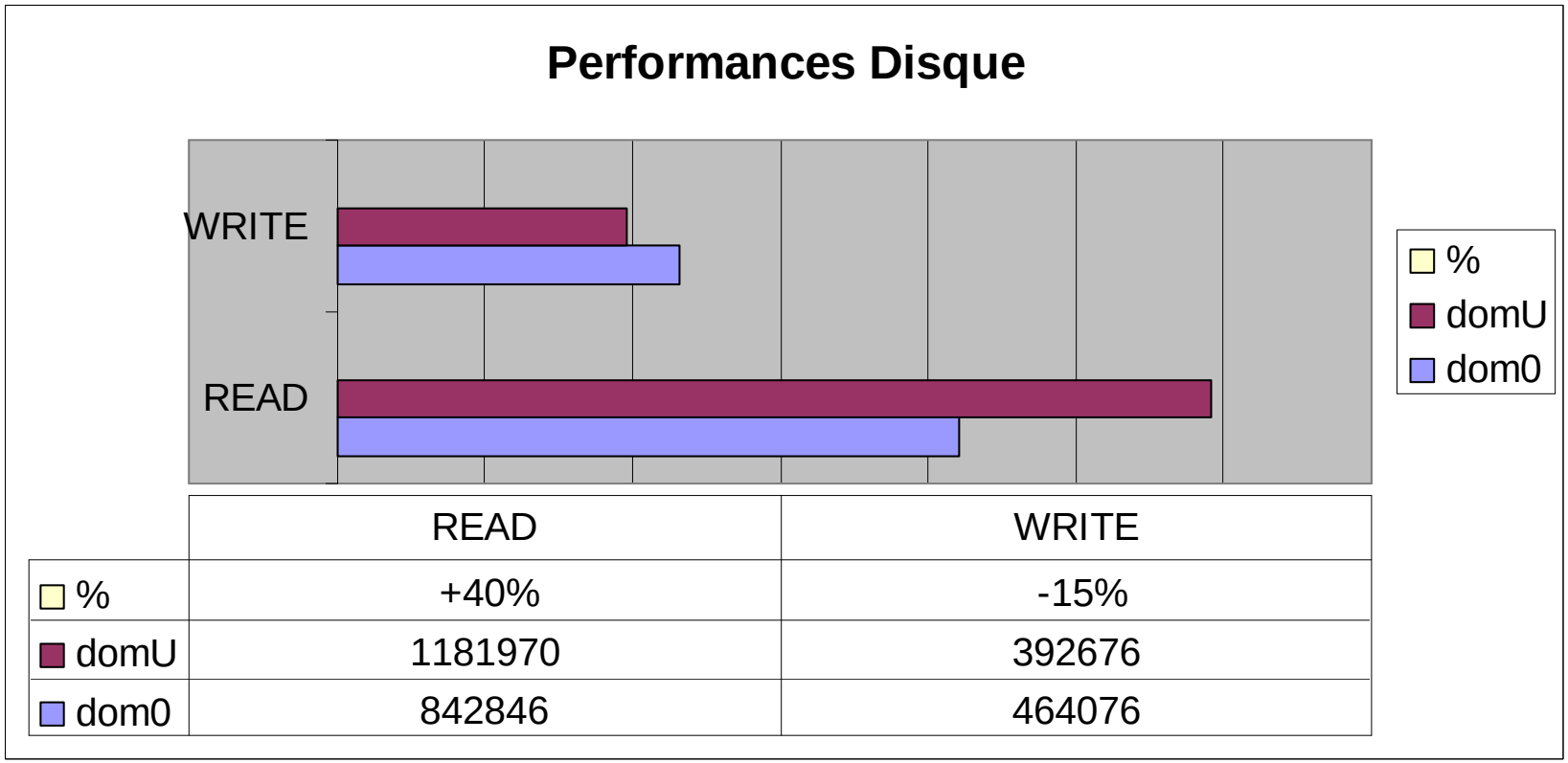
XEN - performances

- Re-démarrage d'un domU
  - `xm  shutdown domU -w ; xm create domU`

```
64 bytes from 192.168.1.222: icmp_seq=68 ttl=64 time=0.158 ms
64 bytes from 192.168.1.222: icmp_seq=113 ttl=64 time=1496 ms
64 bytes from 192.168.1.222: icmp_seq=114 ttl=64 time=496 ms
64 bytes from 192.168.1.222: icmp_seq=115 ttl=64 time=0.190 ms
```

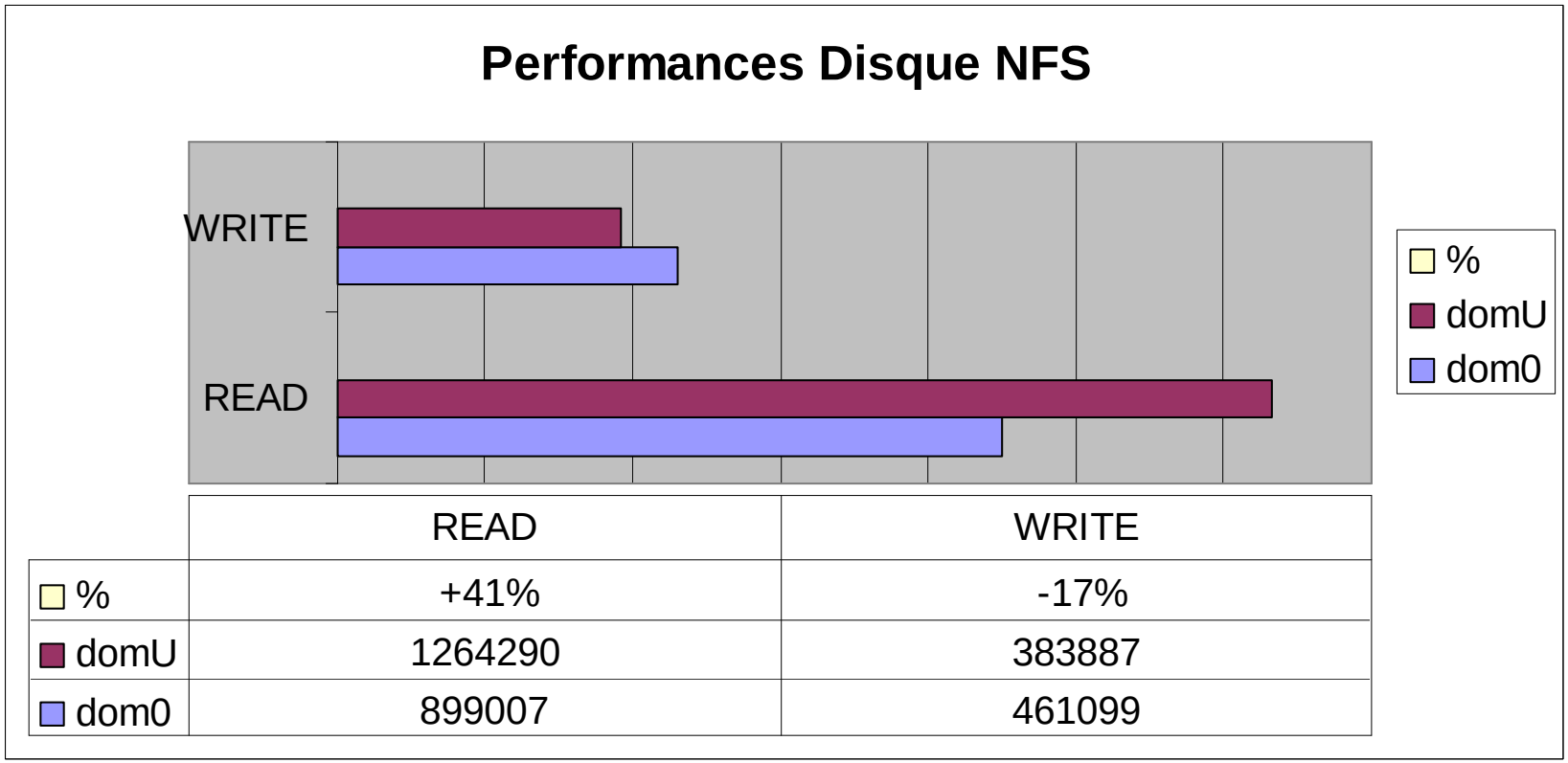
  - → 47 secondes

XEN - performances



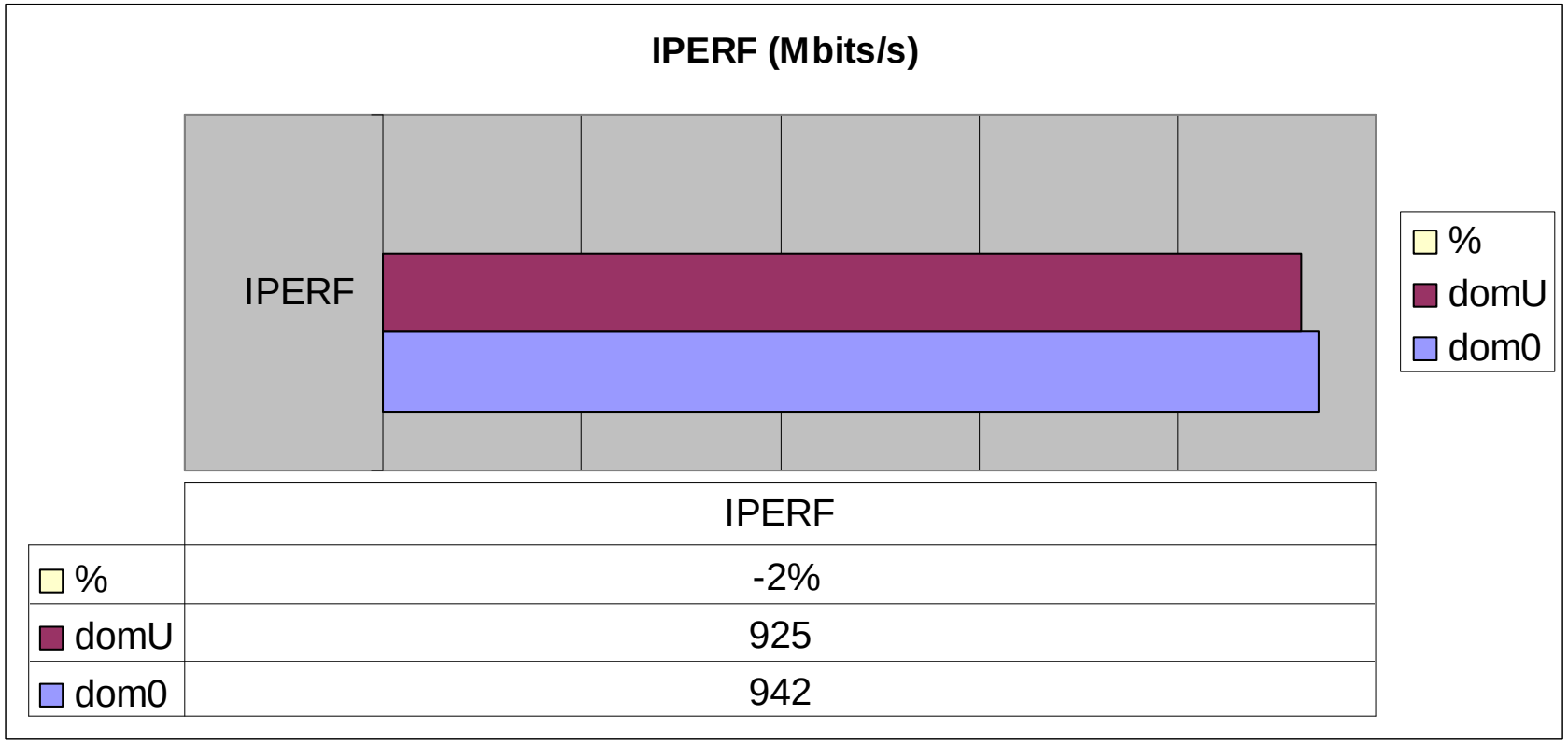
	READ	WRITE
%	+40%	-15%
domU	1181970	392676
dom0	842846	464076

XEN - performances



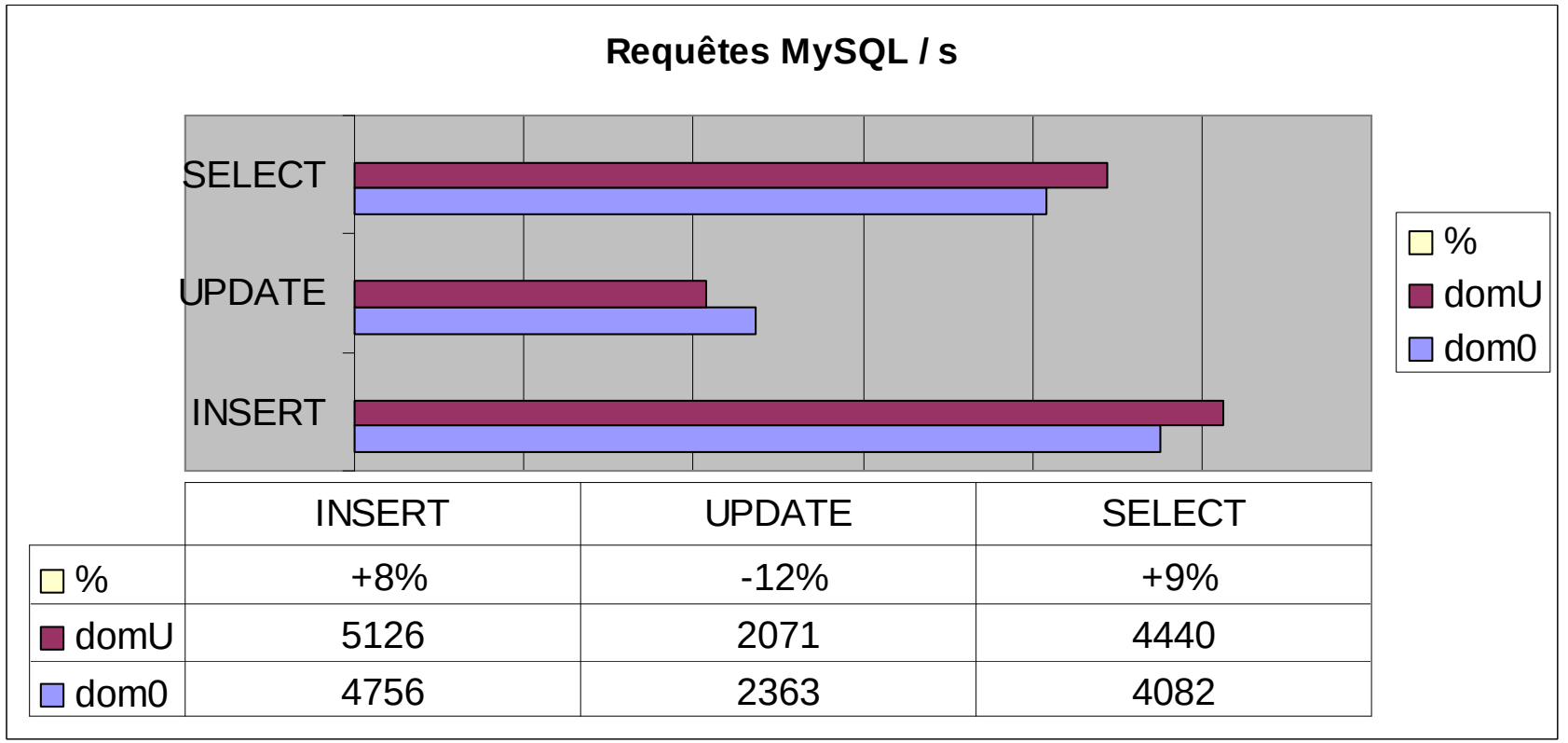
	READ	WRITE
%	+41%	-17%
domU	1264290	383887
dom0	899007	461099

XEN - performances



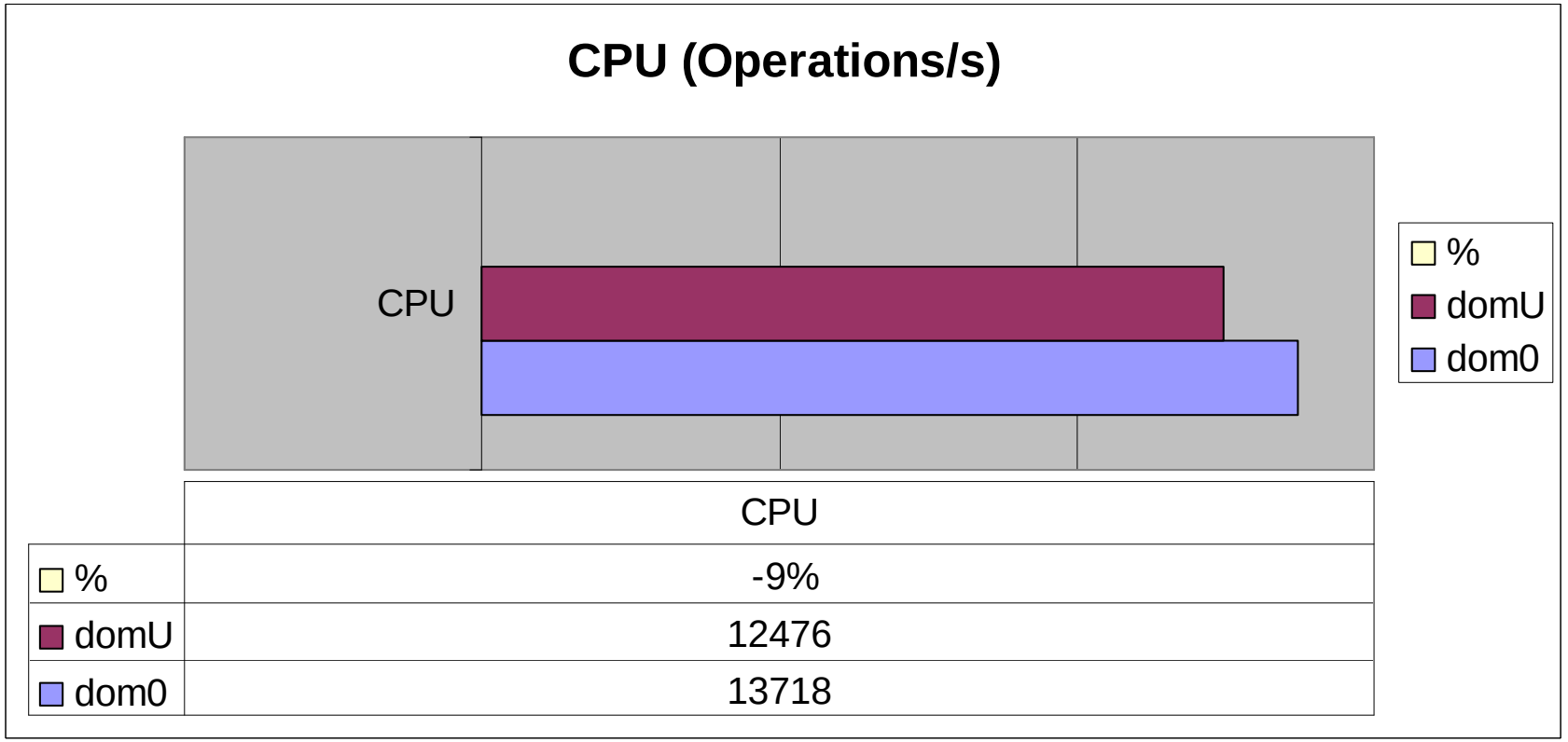
	IPERF
%	-2%
domU	925
dom0	942

XEN - performances



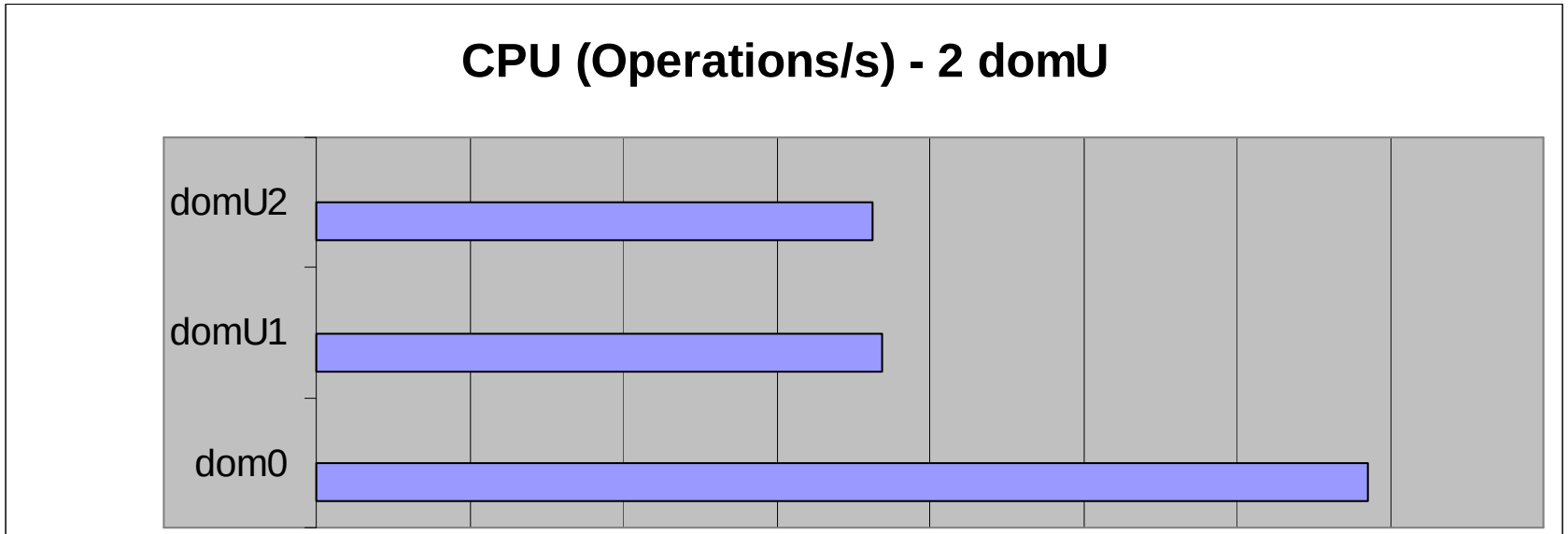
	INSERT	UPDATE	SELECT
%	+8%	-12%	+9%
domU	5126	2071	4440
dom0	4756	2363	4082

XEN - performances



	CPU
%	-9%
domU	12476
dom0	13718

XEN - performances

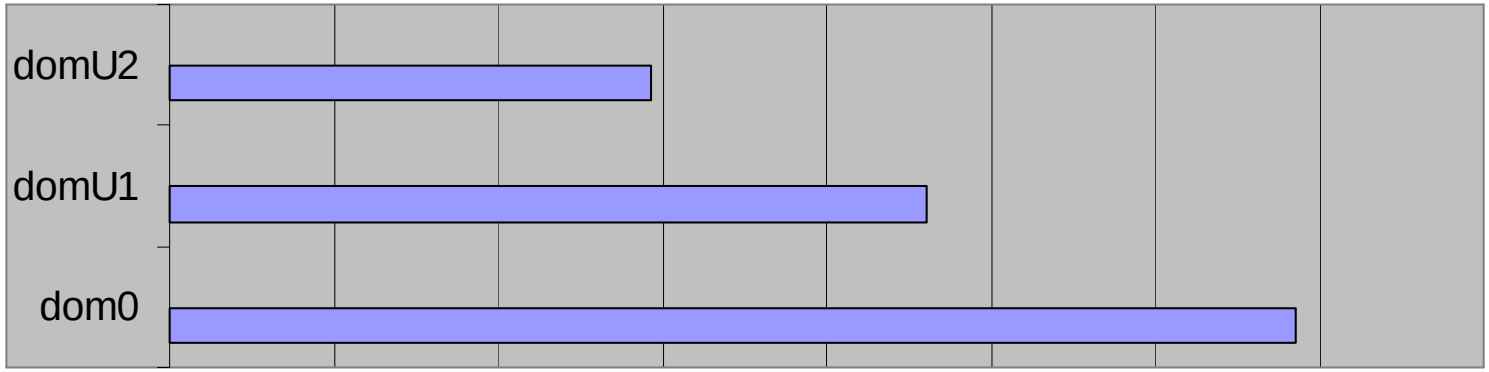


	dom0	domU1	domU2
%		-46%	-47%
CPU	13718	7363	7256

XEN - performances

CPU (Operations/s) - 2 domU

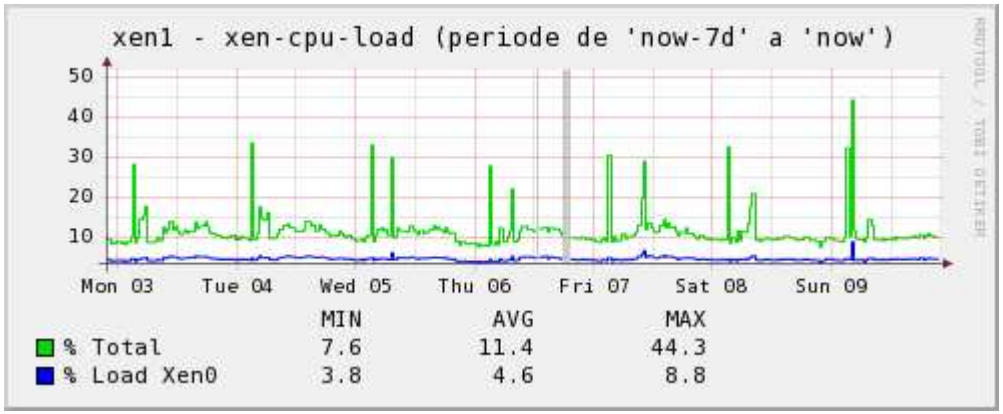
(xm sched-credit -d domU1 -w 512)



	dom0	domU1	domU2
%		-33%	-57%
CPU	13718	9208	5848

XEN - performances

- **1 serveur « syslog »**
- **1 serveur « DNS »**
- **1 serveur « FlexLM »**
- **1 serveur « Radius »**



- **1 serveur physique non virtualisé utilise 15% de sa capacité**

XEN – pour & contre

- Limitations
  - Berkeley DB est/était incompatible
  - les librairies threadées sont à désactiver (fonctionnent, mais plus lentement)
  - ne pas activer NTP dans les domU
    (ce n'est pas un bug, mais une fonctionnalité : domU synchronisé avec dom0)
    (ou modifier /proc/sys/xen/independent_wallclock)
  - le plantage du dom0 (software ou hardware) entraîne le plantage de plusieurs domU (!)

XEN – pour & contre

- Outils d'administration basiques (OpenSource Xen)
  - `xm` (en mode texte)
  - quelques outils de monitoring en mode graphique
  - « assistant » de création d'images virtuelles et de fichiers de configuration ?
  - affichage de la configuration d'un domU ?
  - → Citrix XenServer ?

XEN – pour & contre

- « faiblesse » de l’ordonnanceur
  - Basé sur le « poids » des vCPU / domU
  - Pas de « QoS » I/O réseau
  - Pas de « QoS » I/O disque

XEN – pour & contre

- Erreurs de manipulation possible
  - 2 (ou plusieurs) dom0 avec les images domU partagées (SAN)
    - `[dom0A] # xm create domU1`
      `[dom0B] # xm create domU1`

      → possible
      → pas longtemps …
    - → `lvchange --available ey /dev/vgsan/domU`

      → LiveMigration ?

XEN – pour ou contre ? pour !

- Fiabilité
  → Très peu (2 ou 3) de « `xm destroy domU` » en 1 an d'utilisation
- Perte de performances minime
- Facilité de déploiement
- Facilité d'augmentation de puissance de traitement
- Facilité de Reprise d'Activité
- Stress diminué !

Des questions ?