Technologies informatiques récentes et Supports matériels

David Delavennat, Stéphane Aicardi

Centre de Génétique Moléculaire, Polytechnique

Rencontres *Mathrice* 2007

Plan

Plan

Technologies d'interconnexions

Historique

Nom	Année	Bits	Fréq.	B. P.
ISA	1984	16	8 MHz	16 Mo/s
PCI	1993	32	33 MHz	133 Mo/s
PCI-X	1999	64	133 MHz	1014 Mo/s
PCI-X QDR	2002	64	133 MHz	4056 Mo/s
AGP 1x	1997	32	66 MHz	266 Mo/s
AGP 8x	2002	32	66 MHz	2133 Mo/s
Ultra DMA ATA 100	2002	16	50 MHz	100 Mo/s
Ultra-320 SCSI	2002	16	160 MHz	320 Mo/s
USB	1996	-	-	1.5 Mo/s
Firewire (IEEE-1394)	1995	-	-	50 Mo/s

Technologies d'interconnexions

PCI express

Introduit en 2004, PCI-express est destiné en remplacer les bus locaux parallèles PCI et AGP par un bus série full-duplex à 250 Mo/s. Pour augmenter le débit, on met plusieurs liens série en parallèle.

Exemple : un lien PCI-express 16x monte à 4 Go/s.

Technologies d'interconnexions

HyperTransport

Anciennement **L**ightning **D**ata **T**ransport le bus HyperTransport est une technologie issue des laboratoires de recherche de DIGITAL EQUIPMENTS. Son développement a été poursuivi par AMD, IBM et nVidia. C'est un bus local série/parallèle plus rapide que le bus PCI tout en utilisant le même nombre de broches.

- HT 1 (2001) => 800 MHz, 6.4 Go/s en 16 bits.
- HT 2 (2004) => 1.4 GHz, 11.2 Go/s
- HT 3 (2006) => 2.6 GHz, 20.8 Go/s

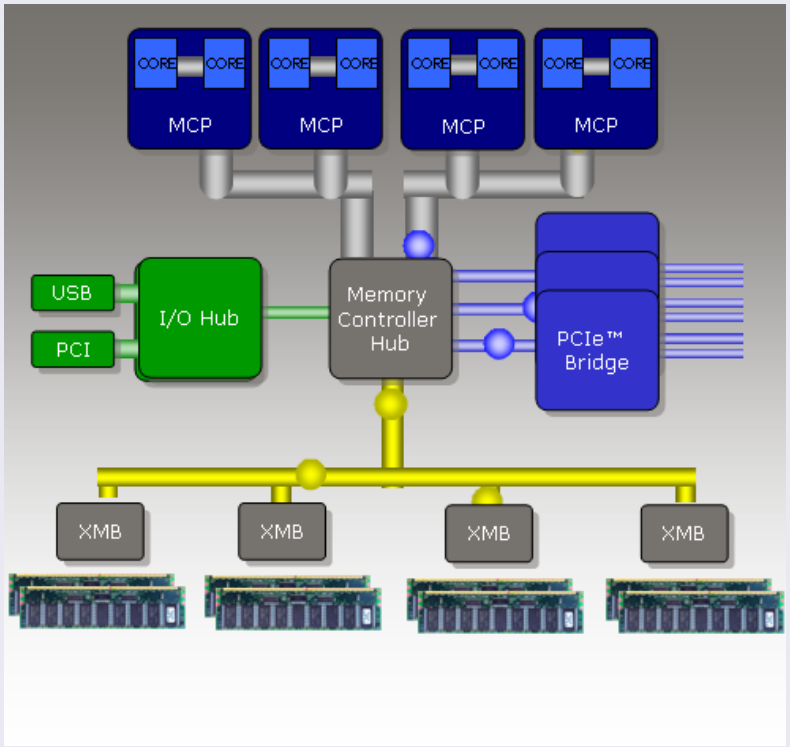
Technologies d'interconnexions

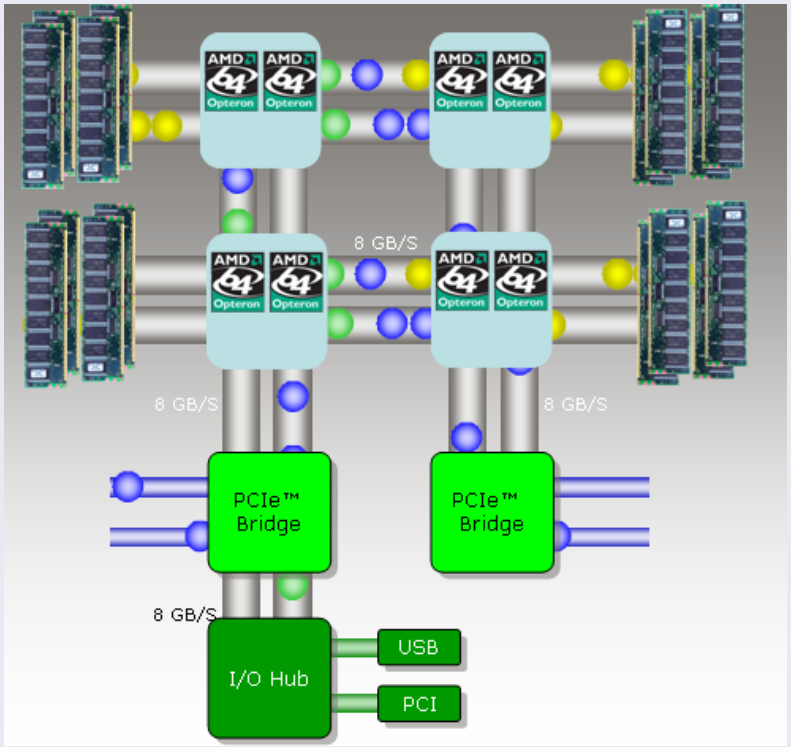
Apports de l'HyperTransport

- Remplacement du Front-Side Bus,
- Interconnexion de processeurs,
- Slots de connexion HyperTransport eXpansion (HTX)
- Depuis 3.0, interconnexion avec PCI-express
- HTX externe en prévision.

Technologies d'interconnexions

Nouvelle organisation interne





Technologies d'interconnexions

Infiniband

Technologie d'interconnexion externe. Infiniband utilise des liens série point à point full duplex et permet d'aggréger des liens pour augmenter la bande passante. Infiniband est populaire pour sa faible latence et son très haut débit (jusqu'à 12 Go/s théoriques).

Technologies d'interconnexions

S-ATA

Évolution de standard ATA avec une interconnexion série. Dans la dernière version, on arrive à 300 Mo/s.

La plupart des disques actuels supportent le Native Command Queuing (NCQ) : inspiré de SCSI, c'est la possibilité pour un contrôleur de réordonner les requètes pour optimiser les accès disques.

Technologies d'interconnexions

Serial Attached SCSI (SAS)

Évolution du standard SCSI avec une interconnexion série :

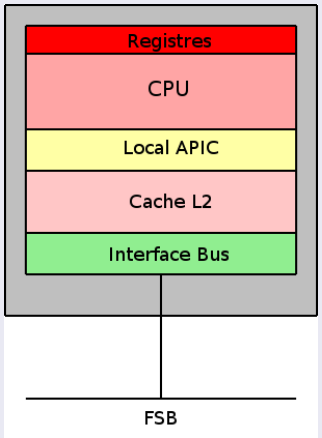
- Connexions point à point,
- plus de problèmes de terminaison,
- jusqu'à 16384 périphériques sur un bus,
- supporte les périphériques S-ATA,
- supporte beaucoup plus de types de périphériques que S-ATA, cables usqu'à 8m, etc.
- Multipath I/O

Plan

Le processeur de l'an 2000

Pentium 4

Avec une nouvelle microarchitecture conçue pour monter en fréquence, le Pentium 4 reste dans la continuité des processeurs Intel précédents.



Intel Architecture 64

Itanium

Développé conjointement par Intel et Hewlett-Packard et mis en oeuvre au sein des processeurs Itanium (2001) et Itanium 2 (2002). Cette architecture est fondée sur le modèle EPIC (Explicitely Parallel Instruction Computing).

Au lieu de systèmes couteux de réarrangement, prédiction, anticipation, l'Itanium mise sur plusieurs unités d'exécution parallèle et sur des instructions prévues pour être parallélisées (VLIW).

Intel Architecture 64

Instruction Level Parallelism

L'ILP est un ensemble de techniques pour favoriser l'exécution simultanée de (micro-)instructions :

- pipeline,
- exécution superscalaire,
- exécution dans le désordre,
- prédiction de branches,
- exécution spéculative,

Explicitely Parallel Instruction Computing

Contrairement au choix fait dans le Pentium 4, le parallélisme n'est pas déterminé dynamiquement par le processeur, mais au moment de la compilation.

AMD64

Motivation pour une nouvelle architecture 64 bits

Pour des raisons de licenses, AMD ne pouvait pas développer de processeur compatible IA-64. Le choix a été fait de développer sa propre microarchitecture 64 bits et un nouveau jeu d'instructions x86_64 qui étend celui du x86. Le nouveau processeur peut exécuter nativement du code 32 bits.

AMD64

Apports du x86_64

- tous les registres généraux passent en 64 bits et leur nombre est doublé,
- doublement des registres XMM,
- adressage virtuel sur 48 bits,
- adressage physique sur 40 bits (1 To),
- jeu d'instruction SSE, SSE2 puis SSE3,
- bit NX (No-Execute),
- suppression de fonctionalités obsolètes.

Extended Memory 64-bit Technology

La réponse d'Intel tarde

Suite au succès commercial des processeurs AMD64, Intel adapte à reculons son Pentium 4 au nouveau jeu d'instructions d'AMD. Cette extension est appelée EM64T et apparaît en même temps que d'autres évolutions (gravure 90nm, Hyper-Threading, etc.)

Un rattrapage progressif

- extensions de registres
- adressage sur 36 bits, puis sur 40 bits.
- bit XD (eXecute Disable)
- intégration native dans la microarchitecture Core

Extended Memory 64-bit Technology

La réponse d'Intel tarde

Suite au succès commercial des processeurs AMD64, Intel adapte à reculons son Pentium 4 au nouveau jeu d'instructions d'AMD. Cette extension est appelée EM64T et apparaît en même temps que d'autres évolutions (gravure 90nm, Hyper-Threading, etc.)

Un rattrapage progressif

- extensions de registres
- adressage sur 36 bits, puis sur 40 bits.
- bit XD (eXecute Disable)
- intégration native dans la microarchitecture Core

Extended Memory 64-bit Technology

La réponse d'Intel tarde

Suite au succès commercial des processeurs AMD64, Intel adapte à reculons son Pentium 4 au nouveau jeu d'instructions d'AMD. Cette extension est appelée EM64T et apparaît en même temps que d'autres évolutions (gravure 90nm, Hyper-Threading, etc.)

Un rattrapage progressif

- extensions de registres
- adressage sur 36 bits, puis sur 40 bits.
- bit XD (eXecute Disable)
- intégration native dans la microarchitecture Core

Extended Memory 64-bit Technology

La réponse d'Intel tarde

Suite au succès commercial des processeurs AMD64, Intel adapte à reculons son Pentium 4 au nouveau jeu d'instructions d'AMD. Cette extension est appelée EM64T et apparaît en même temps que d'autres évolutions (gravure 90nm, Hyper-Threading, etc.)

Un rattrapage progressif

- extensions de registres
- adressage sur 36 bits, puis sur 40 bits.
- bit XD (eXecute Disable)
- intégration native dans la microarchitecture Core

Multi Threading

Principe

Optimiser l'exécution de deux ou plusieurs threads par un processeur physique.

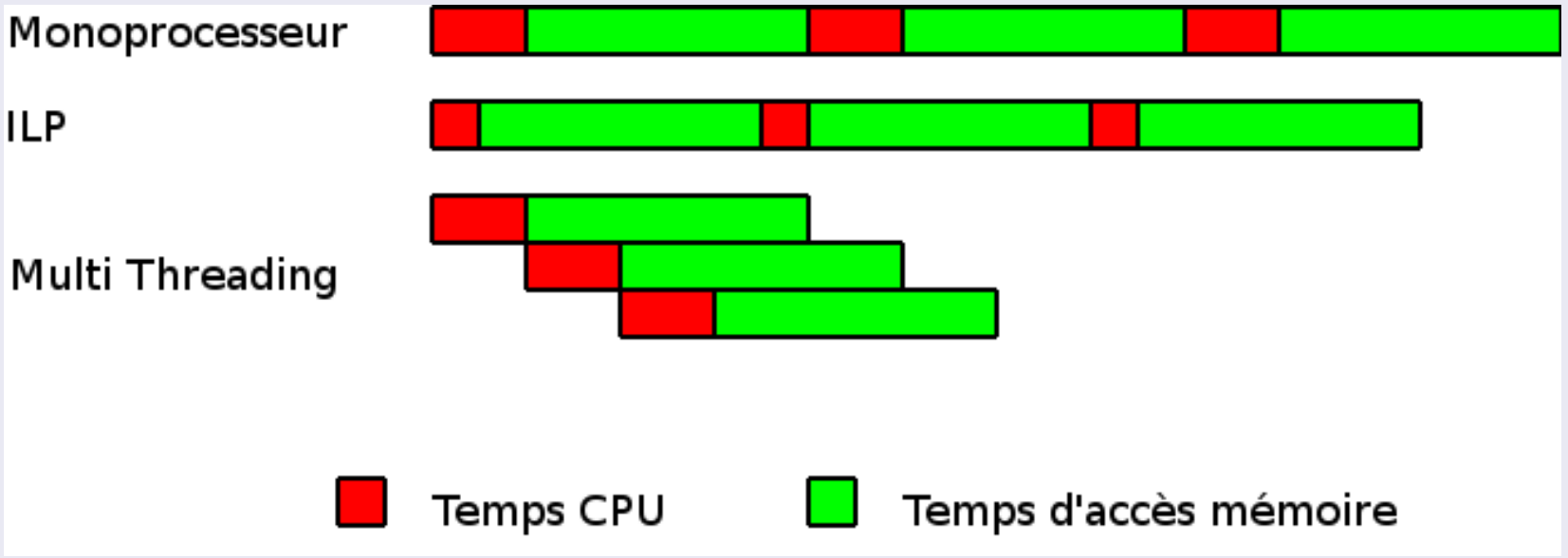
Quand un processeur classique exécute deux processus, il exécute pendant un certain temps le premier, stocke son état (pile, registres), charge l'état du second processus et exécute le code.

Un processeur multi-thread peut stocker plusieurs états concurrents dans le coeur du processeur et partager le pipeline d'exécution entre plusieurs threads, ce qui accélère les changements de contexte.

On parle de Symetric Multi Threading (SMT) si le coeur du processeur peut exécuter plusieurs threads simultanément.

Multi Threading

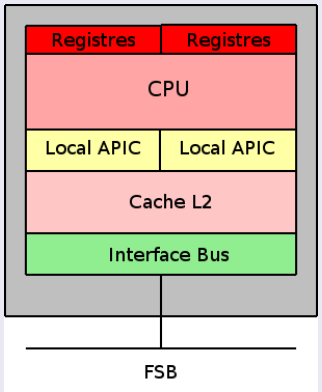
Principe



Hyper-Threading

Hyper-Threading

C'est un Multi Threading ajouté a minima par Intel au Pentium 4. Les queues internes et les caches sont partagés par les threads.



Gain/Coût

L'activation de l'Hyper-Threading dans le processeur Pentium 4 a représenté moins de 5% de la surface de la puce.

Défauts

La technologie HT n'est pas vraiment adaptée aux processeurs CISC (peu de registres) et particulièrement au P4 (pipeline très long) => pertes de performances dans certains cas.

Hyper-Threading : et maintenant ?

Verdict

AMD n'a jamais suivi.

Intel a abandonné l'Hyper-Threading avec la microarchitecture Core™.

Multi Threading : et maintenant ?

UltraSparcT1 ou Niagara

Au contraire, Sun a renforcé le Multi Threading dans le Niagara (2005) : 8 coeurs, 4 threads par coeur, 4 contrôleurs mémoires intégrés, pour seulement 72 W.

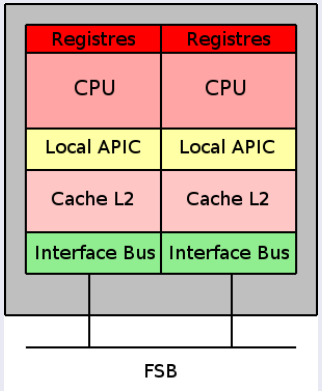
UltraSparcT2 ou Niagara 2

Dans le Niagara 2 à venir, on passe à 8 coeurs, 8 threads/coeur, 1 FPU/coeur, une unité de crypto/coeur, 4 contrôleurs mémoires FBDIMM intégrés, un PCIe X8 et deux ports 10G Ethernet pour 84 W.

Multi-coeurs

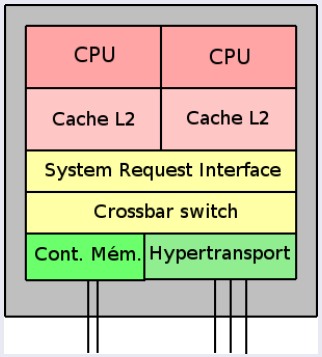
Pentium D

Intel sorti le premier un processeur bi-coeur : le Pentium D. Il s'agit en fait de deux Pentium 4 sur la même puce qui partagent le Front Side Bus.



AMD

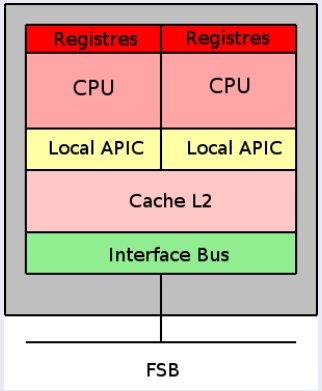
AMD a répliqué avec un processeur bi-coeur plus avancé et intégré. L'interconnexion entre les deux coeurs est interne.



Multi-coeurs

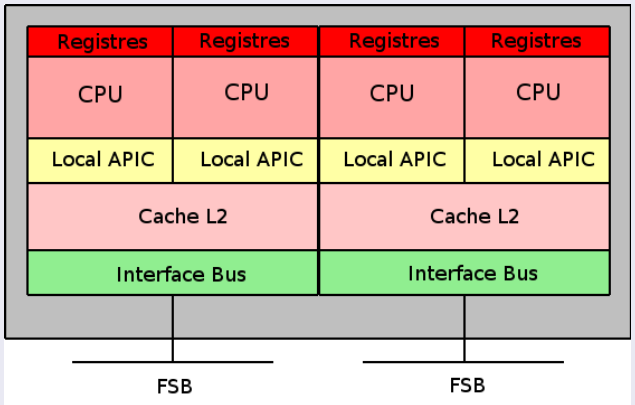
Core (2) duo

À partir du Core duo, Intel produit des processeurs plus intégrés



Quad-core Intel Xeon

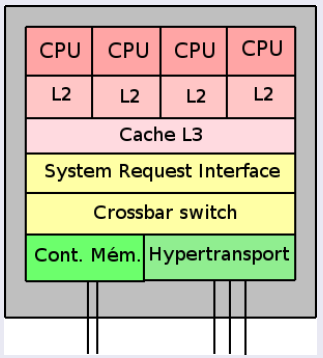
Mais pour gagner la course au quadri-coeur, Intel colle deux Core 2 duo côte à côte.



Multi-coeurs : en projet

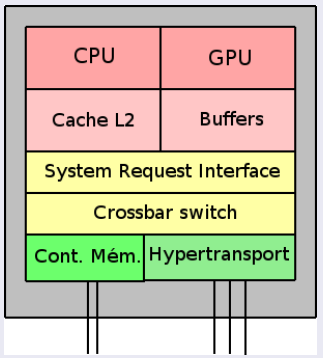
AMD quad-core

Les prochains Optérons seront gravés en 65nm. À noter un cache L3 partagé, une microarchitecture des coeurs revue (inspirée de l'Intel Core).

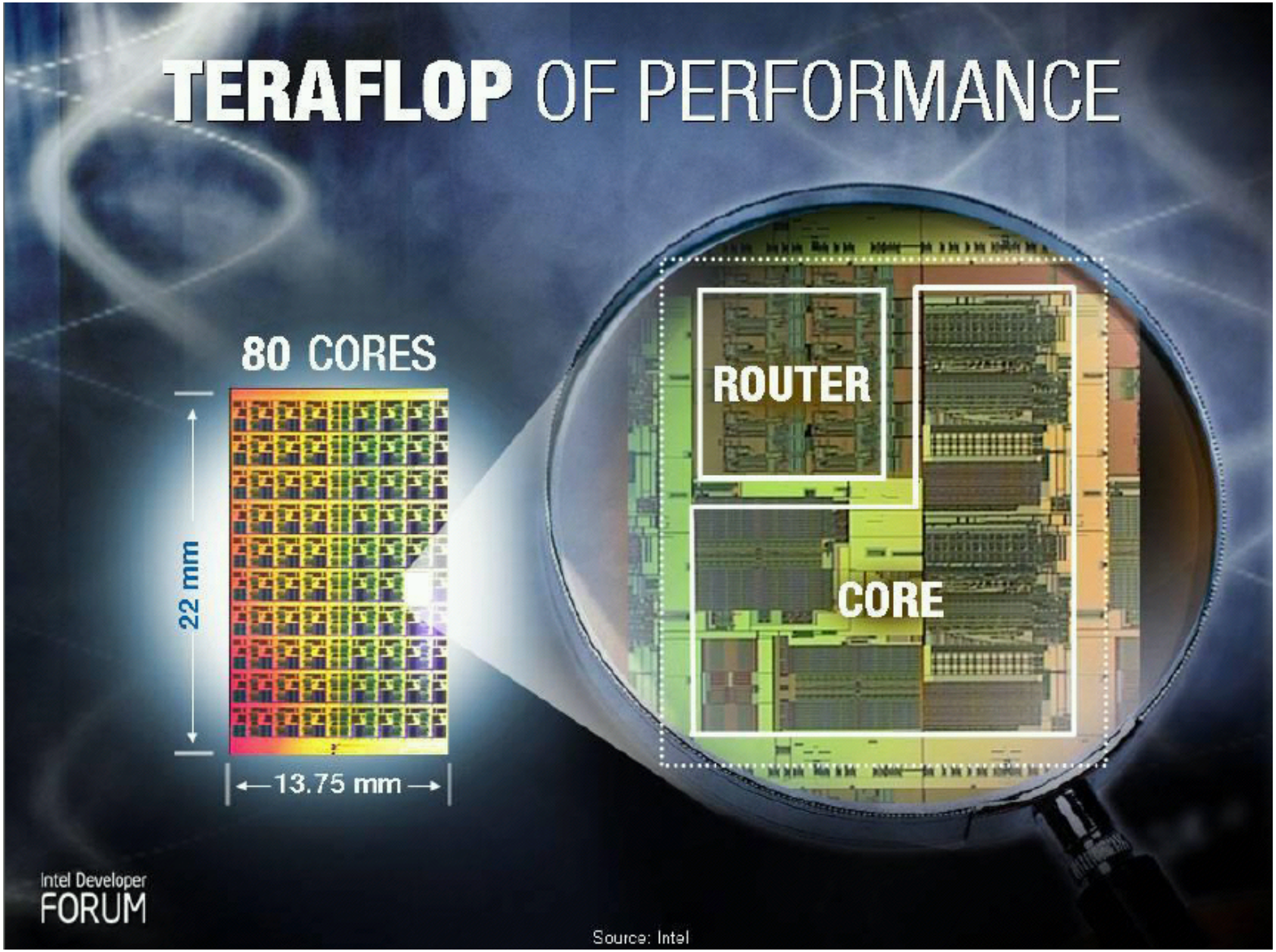


AMD Fusion

Suite au rachat d'ATI, AMD va produire des multi-coeurs mixtes CPU/GPU.



L'avenir ?


TERAFLOP OF PERFORMANCE
80 CORES
22 mm
13.75 mm
ROUTER
CORE
Intel Developer FORUM
Source: Intel

Mémoire

DDR3

Le successeur de la DDR2 devrait conserver le même format (240 pin), améliorer la bande passante (jusqu'à 10.6 GB/s) et la consommation énergétique, au prix d'une latence plus grande.

Fully-Buffered DIMM (FB-DIMM)

La mémoire elle-aussi a son évolution du bus parallèle vers le bus série ! La FB-DIMM est une DDR2 (ou DDR3) qui intègre un Advanced Memory Buffer (AMB). L'AMB sert d'intermédiaire entre le contrôleur mémoire et la mémoire et permet d'augmenter la fiabilité de la transmission. La communication entre le contrôleur mémoire et les AMB se fait sur un lien série full-duplex.

Mémoire

FB-DIMM : Avantages

- scalabilité,
- fiabilité,
- bande passante,
- indépendance du modèle de mémoire,
- simplification de la carte mère.

FB-DIMM : Défauts

- augmentation de la latence,
- dégagement de chaleur.

Énergie

Nouvelle préoccupation

Les Pentiums 4 étaient gourmands en énergie et chauffaient beaucoup. Le développement des portables et des fermes de calcul a motivé une progression des performances à énergie et chaleur dégagée constante.

Comment économiser ?

- Miniaturisation,
- Diminuer les voltages et les fréquences (SpeedStep, PowerNow, Cool'n'Quiet),
- Substrat (Silicon On Insulator),
- Changer de microarchitecture (Pentium 4 -> Core 2).

Énergie : tableau comparatif

Constr.	Modèle	Fréq.	Année	TDP
Intel	Pentium	75 MHz	1993	8 W
Intel	Pentium III	1 GHz	1999	29 W
Intel	Pentium 4HT	3.06 GHz	2002	81.8 W
Intel	Pentium D	3.2 GHz	2005	130 W
Intel	Core 2 Duo	2.67 GHz	2006	65 W
Intel	Core 2 Quad	2.67 GHz	2006	130 W
AMD	Athlon	1 GHz	2000	54.3 W
AMD	Athlon 64	2 GHz	2003	89 W
AMD	Athlon 64 X2	2.2 GHz	2005	89 W

Énergie : Dernières tendances

AMD

Les prochaines puces AMD permettront de réduire la fréquence ou d'arrêter indépendemment les coeurs non utilisés.

Intel Core 2

Le Core 2 peut même décider d'arrêter les parties du coeur qui ne sont pas utilisées.

Intel Core 2

Différences avec le Pentium 4

- micro-architecture dérivée du Pentium M,
- 64 bits natifs,
- pipeline plus court,
- trois unités arithmétiques par coeur,
- macro-fusion et micro-fusion,
- jusqu'à 6 micro-ops exécutées par cycle,
- jusqu'à 8 opérations flottantes par cycle,
- jusqu'à quatre coeurs,
- exécution optimisée pour camoufler la latence d'accès à la mémoire,
- SpeedStep affiné.

Plan

Virtualisation

Historique : 1966 - IBM 360

Virtualisation

Historique : 1966 - IBM 360

- IBM 360/67 - Premier et unique IBM 360 à mémoire virtuelle. Muni de registres associatifs et pouvant être équipé en biprocesseur.
- Logiciel TSS largement développé à Grenoble
- Système CP/CMS (Control Program/Cambridge Monitor System).
- CP était un hyperviseur gérant des machines virtuelles sous lequel on pouvait faire tourner indifféremment des CMS, des DOS et des OS !

Virtualisation

Historique : 1966 - IBM 360

- IBM 360/67 - Premier et unique IBM 360 à mémoire virtuelle. Muni de registres associatifs et pouvant être équipé en biprocesseur.
- Logiciel TSS largement développé à Grenoble
- Système CP/CMS (Control Program/Cambridge Monitor System).
- CP était un hyperviseur gérant des machines virtuelles sous lequel on pouvait faire tourner indifféremment des CMS, des DOS et des OS !

Virtualisation

Historique : 1966 - IBM 360

- IBM 360/67 - Premier et unique IBM 360 à mémoire virtuelle. Muni de registres associatifs et pouvant être équipé en biprocesseur.
- Logiciel TSS largement développé à Grenoble
- Système CP/CMS (Control Program/Cambridge Monitor System).
- CP était un hyperviseur gérant des machines virtuelles sous lequel on pouvait faire tourner indifféremment des CMS, des DOS et des OS !

Virtualisation

Historique : 1966 - IBM 360

- IBM 360/67 - Premier et unique IBM 360 à mémoire virtuelle. Muni de registres associatifs et pouvant être équipé en biprocesseur.
- Logiciel TSS largement développé à Grenoble
- Système CP/CMS (Control Program/Cambridge Monitor System).
- CP était un hyperviseur gérant des machines virtuelles sous lequel on pouvait faire tourner indifféremment des CMS, des DOS et des OS !

Virtualisation

Loi de Grosh

la puissance réelle d'un ordinateur croît généralement bien plus vite que son coût $\Rightarrow$ d'importantes économies d'échelle sont possibles en allant vers le gigantisme.

Comment

- Mutualiser les ressources matérielles entre plusieurs machines.
- C'est déjà le cas pour le réseau, les disques..mais qu'en est il des processeurs et de la mémoire ?
- Il faut découpler les "machines" du matériel qui les exécutent.
- Une machine virtuelle, découplée du matériel, ne correspond pas nécessairement au matériel sous-jacent.

Virtualisation

Loi de Grosh

la puissance réelle d'un ordinateur croît généralement bien plus vite que son coût $\Rightarrow$ d'importantes économies d'échelle sont possibles en allant vers le gigantisme.

Comment

- Mutualiser les ressources matérielles entre plusieurs machines.
- C'est déjà le cas pour le réseau, les disques..mais qu'en est il des processeurs et de la mémoire ?
- Il faut découpler les "machines" du matériel qui les exécutent.
- Une machine virtuelle, découplée du matériel, ne correspond pas nécessairement au matériel sous-jacent.

Virtualisation

Loi de Grosh

la puissance réelle d'un ordinateur croît généralement bien plus vite que son coût $\Rightarrow$ d'importantes économies d'échelle sont possibles en allant vers le gigantisme.

Comment

- Mutualiser les ressources matérielles entre plusieurs machines.
- C'est déjà le cas pour le réseau, les disques..mais qu'en est il des processeurs et de la mémoire ?
- Il faut découpler les "machines" du matériel qui les exécutent.
- Une machine virtuelle, découplée du matériel, ne correspond pas nécessairement au matériel sous-jacent.

Virtualisation

Loi de Grosh

la puissance réelle d'un ordinateur croît généralement bien plus vite que son coût $\Rightarrow$ d'importantes économies d'échelle sont possibles en allant vers le gigantisme.

Comment

- Mutualiser les ressources matérielles entre plusieurs machines.
- C'est déjà le cas pour le réseau, les disques..mais qu'en est il des processeurs et de la mémoire ?
- Il faut découpler les "machines" du matériel qui les exécutent.
- Une machine virtuelle, découplée du matériel, ne correspond pas nécessairement au matériel sous-jacent.

Virtualisation

Loi de Grosh

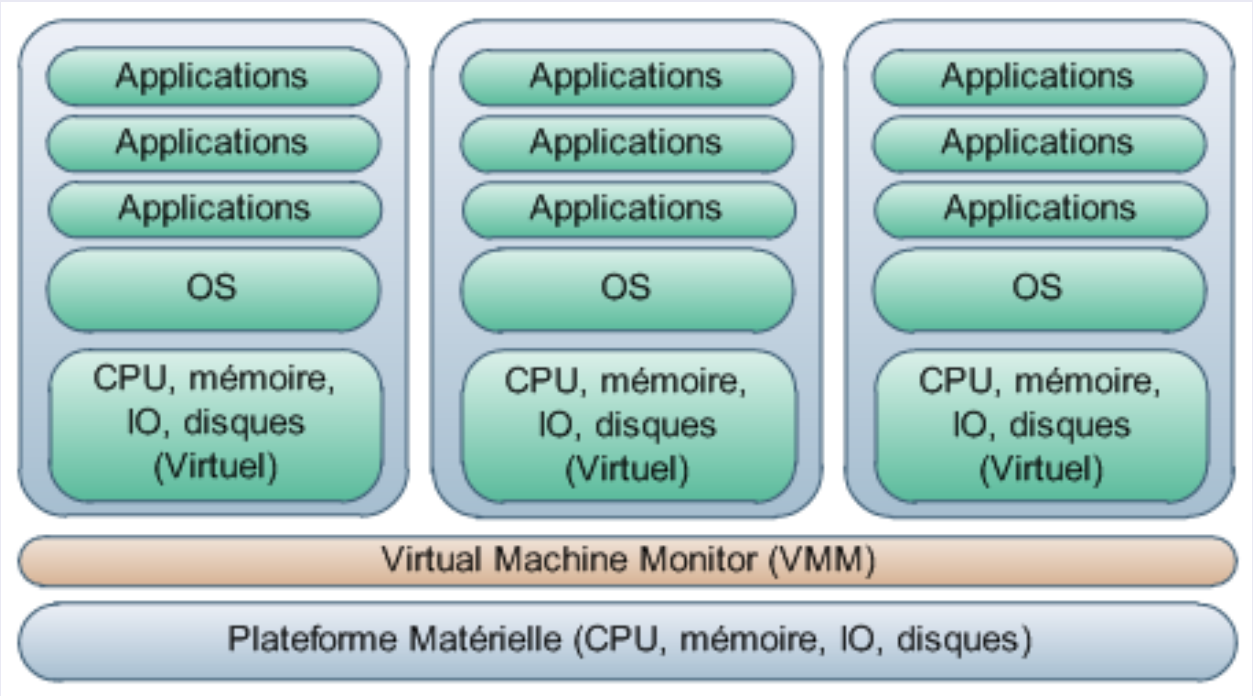
la puissance réelle d'un ordinateur croît généralement bien plus vite que son coût $\Rightarrow$ d'importantes économies d'échelle sont possibles en allant vers le gigantisme.

Comment

- Mutualiser les ressources matérielles entre plusieurs machines.
- C'est déjà le cas pour le réseau, les disques..mais qu'en est il des processeurs et de la mémoire ?
- Il faut découpler les "machines" du matériel qui les exécutent.
- Une machine virtuelle, découplée du matériel, ne correspond pas nécessairement au matériel sous-jacent.

Virtualisation

Machines virtuelles



Virtualisation

Qu'est ce que c'est ?

Virtual Machine Monitor (**VMM**). Fine couche logicielle permettant de virtualiser les ressources matérielles auprès des machines virtuelles (**VM**). La VMM doit permettre :

- un niveau de performance proche d'un système natif.
- l'isolation entre les VM (chaque VM est seule au monde sur le matériel).
- l'abstraction des interfaces matérielles de communication (disque IDE virtualisé en tant que disque SCSI).
- qu'une VM ne soit finalement rien d'autre qu'un fichier

Virtualisation

Qu'est ce que c'est ?

Virtual Machine Monitor (**VMM**). Fine couche logicielle permettant de virtualiser les ressources matérielles auprès des machines virtuelles (**VM**). La VMM doit permettre :

- un niveau de performance proche d'un système natif.
- l'isolation entre les VM (chaque VM est seule au monde sur le matériel).
- l'abstraction des interfaces matérielles de communication (disque IDE virtualisé en tant que disque SCSI).
- qu'une VM ne soit finalement rien d'autre qu'un fichier

Virtualisation

Qu'est ce que c'est ?

Virtual Machine Monitor (**VMM**). Fine couche logicielle permettant de virtualiser les ressources matérielles auprès des machines virtuelles (**VM**). La VMM doit permettre :

- un niveau de performance proche d'un système natif.
- l'isolation entre les VM (chaque VM est seule au monde sur le matériel).
- l'abstraction des interfaces matérielles de communication (disque IDE virtualisé en tant que disque SCSI).
- qu'une VM ne soit finalement rien d'autre qu'un fichier

Virtualisation

Qu'est ce que c'est ?

Virtual Machine Monitor (**VMM**). Fine couche logicielle permettant de virtualiser les ressources matérielles auprès des machines virtuelles (**VM**). La VMM doit permettre :

- un niveau de performance proche d'un système natif.
- l'isolation entre les VM (chaque VM est seule au monde sur le matériel).
- l'abstraction des interfaces matérielles de communication (disque IDE virtualisé en tant que disque SCSI).
- qu'une VM ne soit finalement rien d'autre qu'un fichier

Virtualisation

Qu'est ce que c'est ?

Virtual Machine Monitor (**VMM**). Fine couche logicielle permettant de virtualiser les ressources matérielles auprès des machines virtuelles (**VM**). La VMM doit permettre :

- un niveau de performance proche d'un système natif.
- l'isolation entre les VM (chaque VM est seule au monde sur le matériel).
- l'abstraction des interfaces matérielles de communication (disque IDE virtualisé en tant que disque SCSI).
- qu'une VM ne soit finalement rien d'autre qu'un fichier

Virtualisation

Architecture x86 historique

Il existe 4 niveaux de privilèges d'exécution du code binaire

- Ring 3 $\Rightarrow$ le mode **Utilisateur**
- Ring 2 $\Rightarrow$ inutilisé
- Ring 1 $\Rightarrow$ inutilisé
- Ring 0 $\Rightarrow$ le mode **Privilégié**. Il existe 17 instructions de niveau critique.

Virtualisation

Architecture x86 historique

Il existe 4 niveaux de privilèges d'exécution du code binaire

- Ring 3 $\Rightarrow$ le mode **Utilisateur**
- Ring 2 $\Rightarrow$ inutilisé
- Ring 1 $\Rightarrow$ inutilisé
- Ring 0 $\Rightarrow$ le mode **Privilégié**. Il existe 17 instructions de niveau critique.

Virtualisation

Architecture x86 historique

Il existe 4 niveaux de privilèges d'exécution du code binaire

- Ring 3 $\Rightarrow$ le mode **Utilisateur**
- Ring 2 $\Rightarrow$ inutilisé
- Ring 1 $\Rightarrow$ inutilisé
- Ring 0 $\Rightarrow$ le mode **Privilégié**. Il existe 17 instructions de niveau critique.

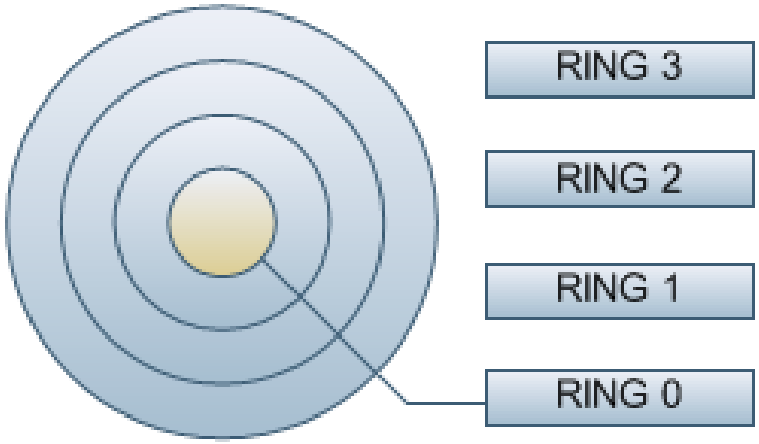
Virtualisation

Architecture x86 historique

Il existe 4 niveaux de privilèges d'exécution du code binaire

- Ring 3 $\Rightarrow$ le mode **Utilisateur**
- Ring 2 $\Rightarrow$ inutilisé
- Ring 1 $\Rightarrow$ inutilisé
- Ring 0 $\Rightarrow$ le mode **Privilégié**. Il existe 17 instructions de niveau critique.

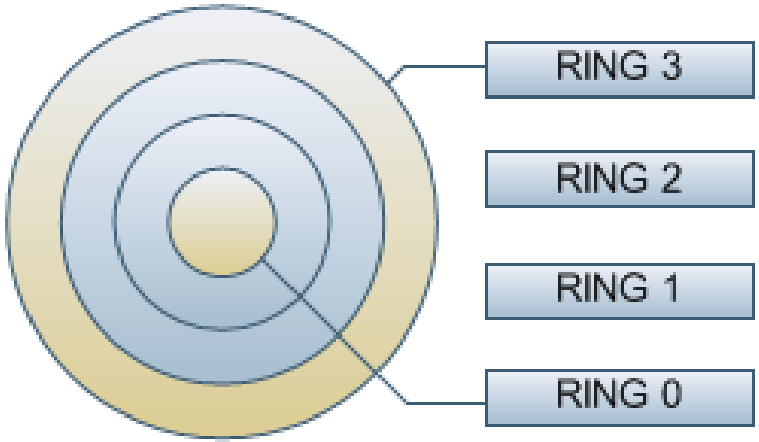
Virtualisation : x86



Ring 0

Il est utilisé par le noyau du système d'exploitation ainsi que par les pilotes pour effectuer les accès directes au matériel. Les instructions relatives à la gestion de la mémoire ainsi qu'aux interruptions sont toutes **Privilégiées**.

Virtualisation : x86



Ring 3

C'est à ce niveau de privilège que s'exécutent les applications utilisateurs.

Isolateur

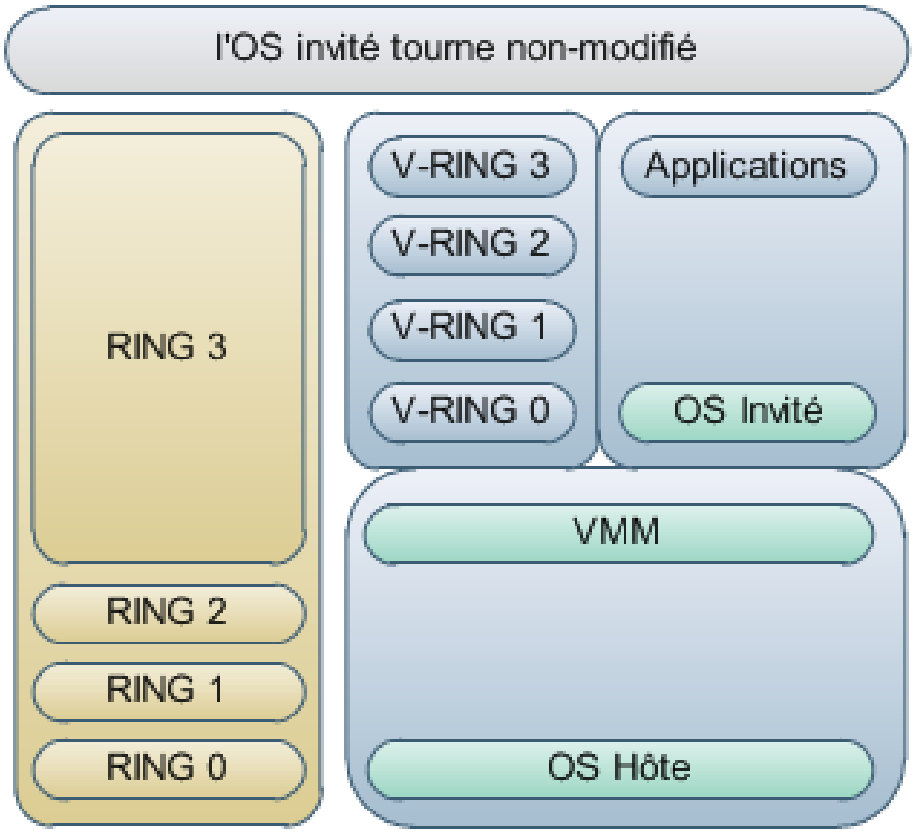
Intéret

- Isolation (faible) entre applications.
- Permet de faire tourner plusieurs instances d'une même application prévue pour n'authoriser qu'une instance.
- Peu consommateur en ressources.

Exemple

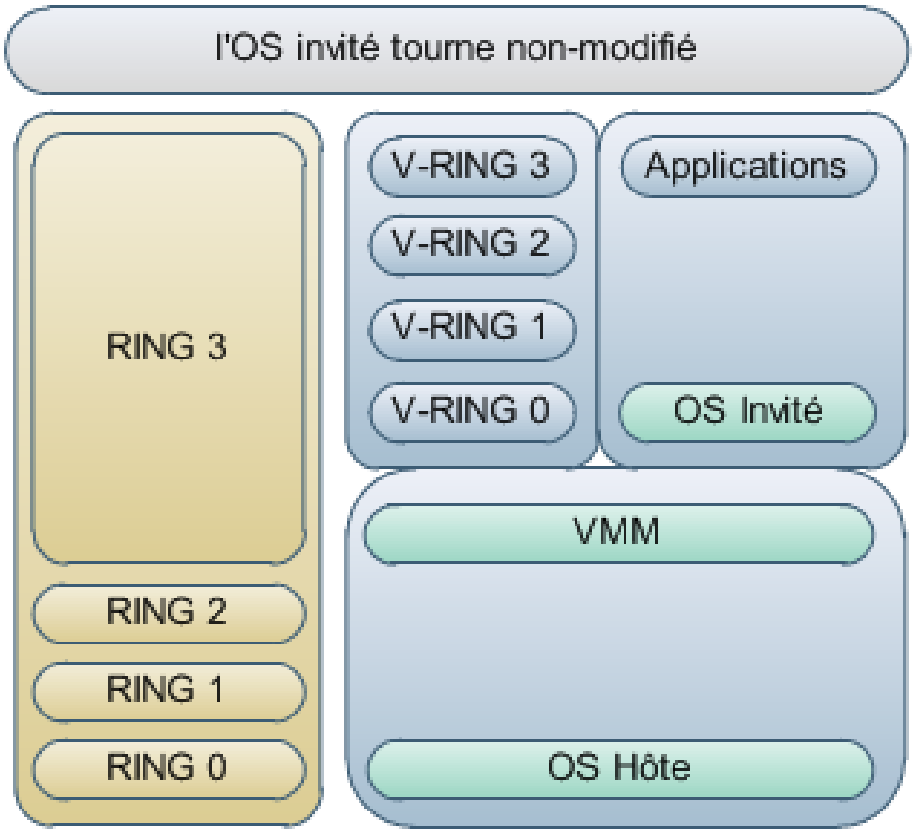
- chroot $\Rightarrow$ isolation d'espace de fichiers.
- BSD Jail & Linux-Vserver $\Rightarrow$ isolation des processus en espace utilisateur.
- OpenVZ $\Rightarrow$ isolation au niveau noyau sous Linux et Windows 2003

Virtualisation logicielle complète (cf VMWare)



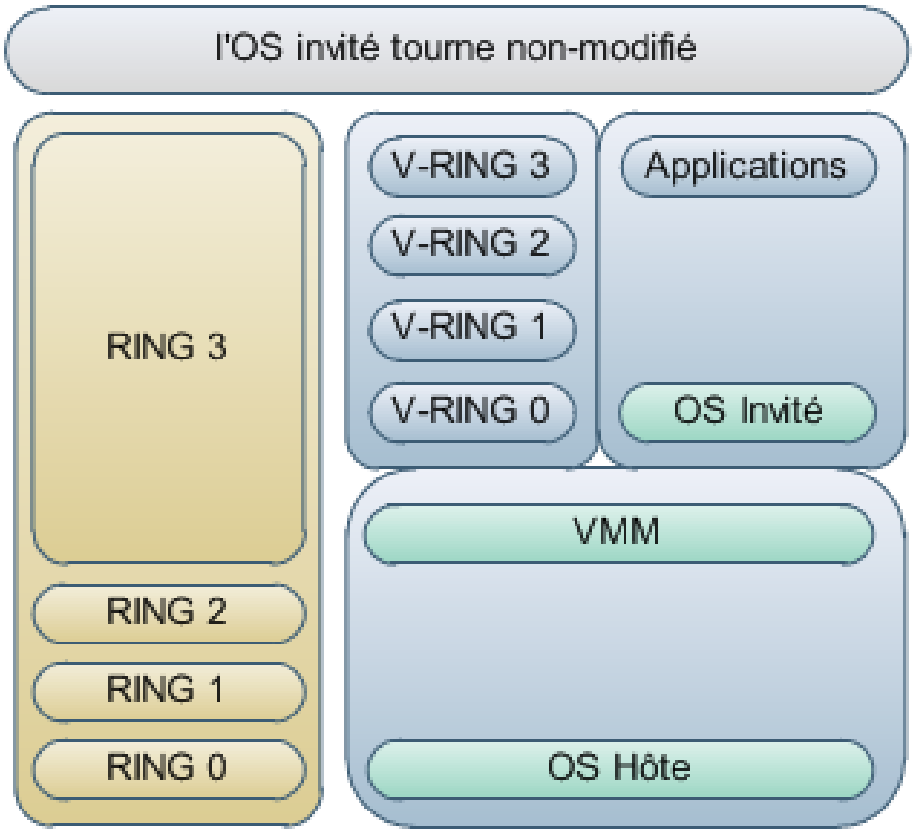
- Le système d'exploitation invité n'est pas modifié.
- Bonne isolation entre systèmes invités.
- Techniques d'optimisation logicielles ⇒ Recompilation dynamique.

Virtualisation logicielle complète (cf VMWare)



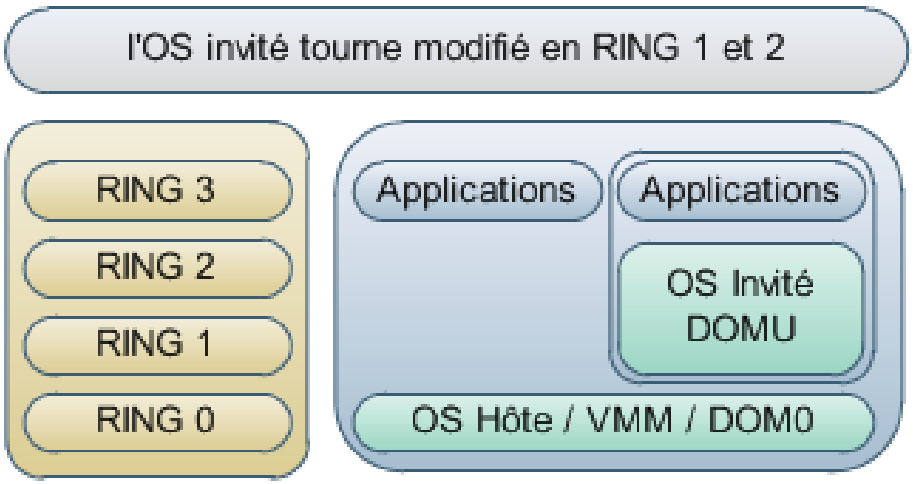
- Le système d'exploitation invité n'est pas modifié.
- Bonne isolation entre systèmes invités.
- Techniques d'optimisation logicielles ⇒ Recompilation dynamique.

Virtualisation logicielle complète (cf VMWare)



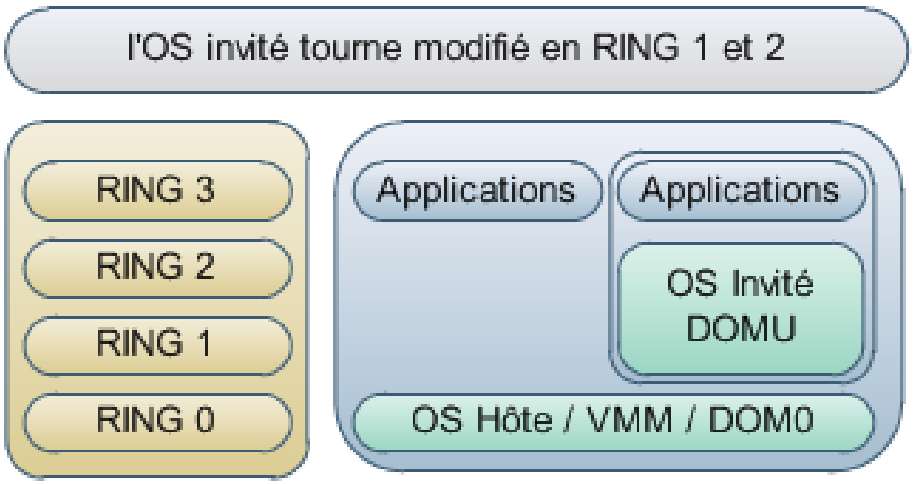
- Le système d'exploitation invité n'est pas modifié.
- Bonne isolation entre systèmes invités.
- Techniques d'optimisation logicielles $\Rightarrow$ Recompilation dynamique.

Para-Virtualisation logicielle (cf XEN)



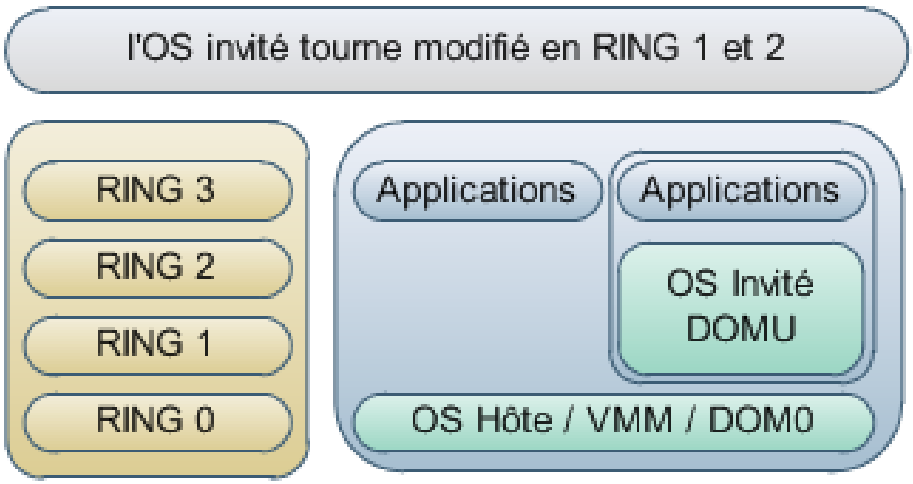
- Le système d'exploitation invité est modifié pour s'exécuter en RING 1 & 2 (cf DOMU).
- L'hyperviseur gère les interactions entre les systèmes d'exploitation Hôte et Invité (cf Hypercall).
- Le système d'exploitation Hôte est adapté à l'API XEN (cf DOM0).

Para-Virtualisation logicielle (cf XEN)



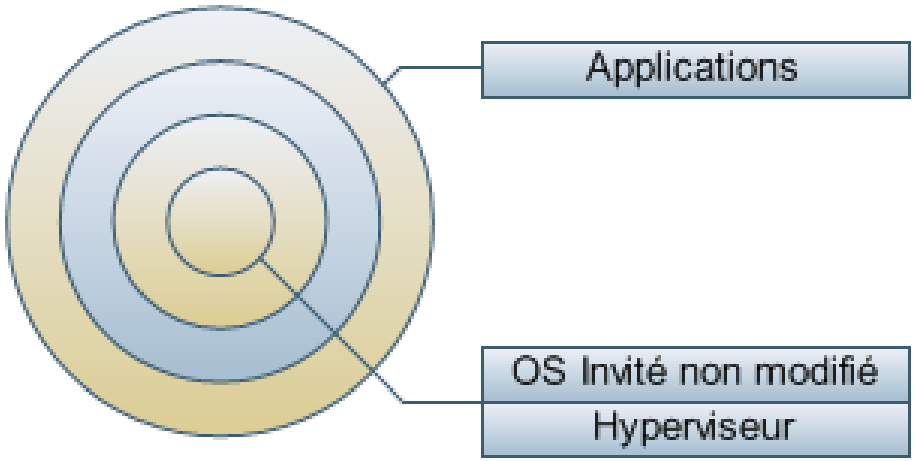
- Le système d'exploitation invité est modifié pour s'exécuter en RING 1 & 2 (cf DOMU).
- L'hyperviseur gère les interactions entre les systèmes d'exploitation Hôte et Invité (cf Hypercall).
- Le système d'exploitation Hôte est adapté à l'API XEN (cf DOM0).

Para-Virtualisation logicielle (cf XEN)



- Le système d'exploitation invité est modifié pour s'exécuter en RING 1 & 2 (cf DOMU).
- L'hyperviseur gère les interactions entre les systèmes d'exploitation Hôte et Invité (cf Hypercall).
- Le système d'exploitation Hôte est adapté à l'API XEN (cf DOM0).

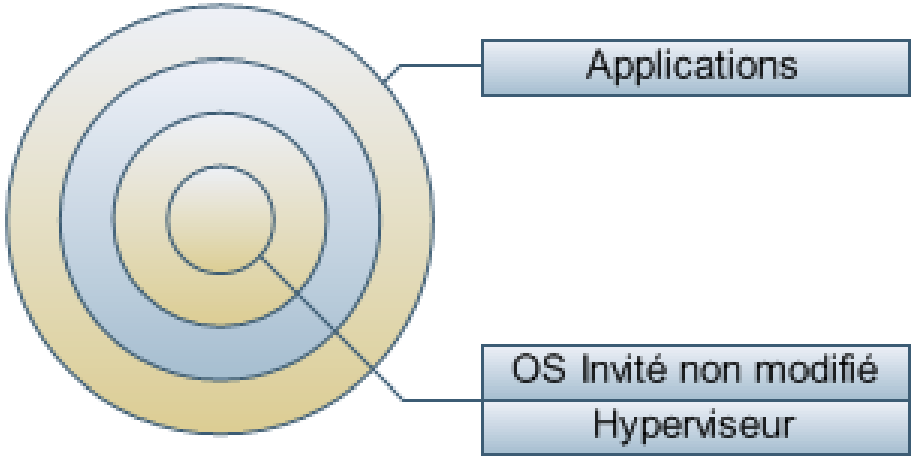
Virtualisation matérielle : x86



Processeur avec intructions de virtualisations

- Des intructions matérielles supplémentaires gérent la virtualisation.
- Les systèmes d'exploitation invités tournent en mode invité et n'ont pas à être modifiés.
- L'hyperviseur contrôle les accès au matériel.

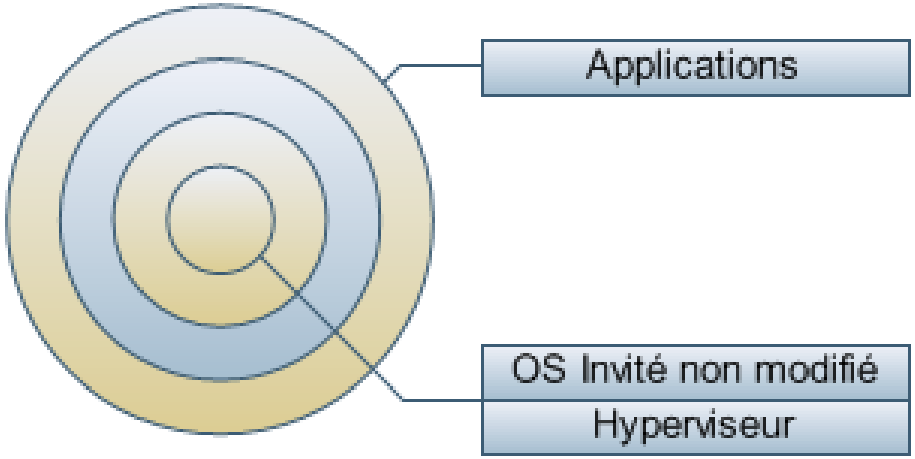
Virtualisation matérielle : x86



Processeur avec intructions de virtualisations

- Des intructions matérielles supplémentaires gérent la virtualisation.
- Les systèmes d'exploitation invités tournent en **mode invité** et n'ont pas à être modifiés.
- L'hyperviseur contrôle les accès au matériel.

Virtualisation matérielle : x86



Processeur avec intructions de virtualisations

- Des intructions matérielles supplémentaires gérent la virtualisation.
- Les systèmes d'exploitation invités tournent en **mode invité** et n'ont pas à être modifiés.
- L'hyperviseur contrôle les accès au matériel.

Support matériel

Intel VT

- nouveau mode d'exécution, baptisé VMX.

VMX

Il comporte un niveau racine (root), correspondant à des rings inférieurs à 0, et un niveau normal, correspondant aux anciens rings de 0 à 3. L'hyperviseur fonctionne en mode VMX racine, avec le niveau de contrôle le plus élevé. Les systèmes d'exploitation invités fonctionnent sur le ring 0 du mode VMX normal. Ils occupent bien l'emplacement pour lequel ils ont été conçus. Plus besoin de modification des invités, ni de translation binaire (en fait, celle-ci reste nécessaire pour d'autres fonctions).

Support matériel : exemple avec AMD-V

Instruction VMMRUN

Au boot un processeur AMD64 démarre en mode x86_32 afin de maintenir la compatibilité avec les Systèmes d'exploitation 32 bits. Le boot-loader d'un système d'exploitation 64 bits exécute une instruction qui bascule le processeur en mode x86_64. De manière similaire un processeur cablé avec Pacifica démarre en mode Invité (toutes les fonctions de virtualisation additionelles inactives) jusqu'à ce qu'une VMM compatible ne soit démarrée. La commande VMMRUN bascule alors le processeur en mode Hôte.

Support matériel : exemple avec AMD-V

Structure VMCB

En mode Hôte la VMM stocke dans la structure de données Virtual Machine Control Block les informations relatives au matériel associé à un système invité (processeur(s), blocks mémoire(s), entrées/sorties). Une fois la structure VMCB définie la VMM bascule le processeur en mode Invité, passe la main à l'OS invité associé qui démarre (se considérant seul au monde sur le matériel) executant son code privilégié en RING0 et son code applicatif en RING3. En coulisse la VMM monitor l'execution et intercepte les instruction privilégiées. Quand une instruction demande un accès à une ressource définie dans la VMCB le processeur bascule en mode Hôte. La VMM gère alors la requéte.

Support matériel : exemple avec AMD-V

Instruction VMMCALL

- La virtualisation est faites pour être invisible aux OS invités.
- Cependant, il serait appréciable qu'un OS invité qui arrive à saturation mémoire ou disque puisse demander plus de ressources.
- VMMCALL est une instruction privilégiée permettant aux OS invités de négocier des ressources avec la VMM.

Support matériel : exemple avec AMD-V

Instruction VMMCALL

- La virtualisation est faites pour être invisible aux OS invités.
- Cependant, il serait appréciable qu'un OS invité qui arrive à saturation mémoire ou disque puisse demander plus de ressources.
- VMMCALL est une instruction privilégiée permettant aux OS invités de négocier des ressources avec la VMM.

Support matériel : exemple avec AMD-V

Instruction VMMCALL

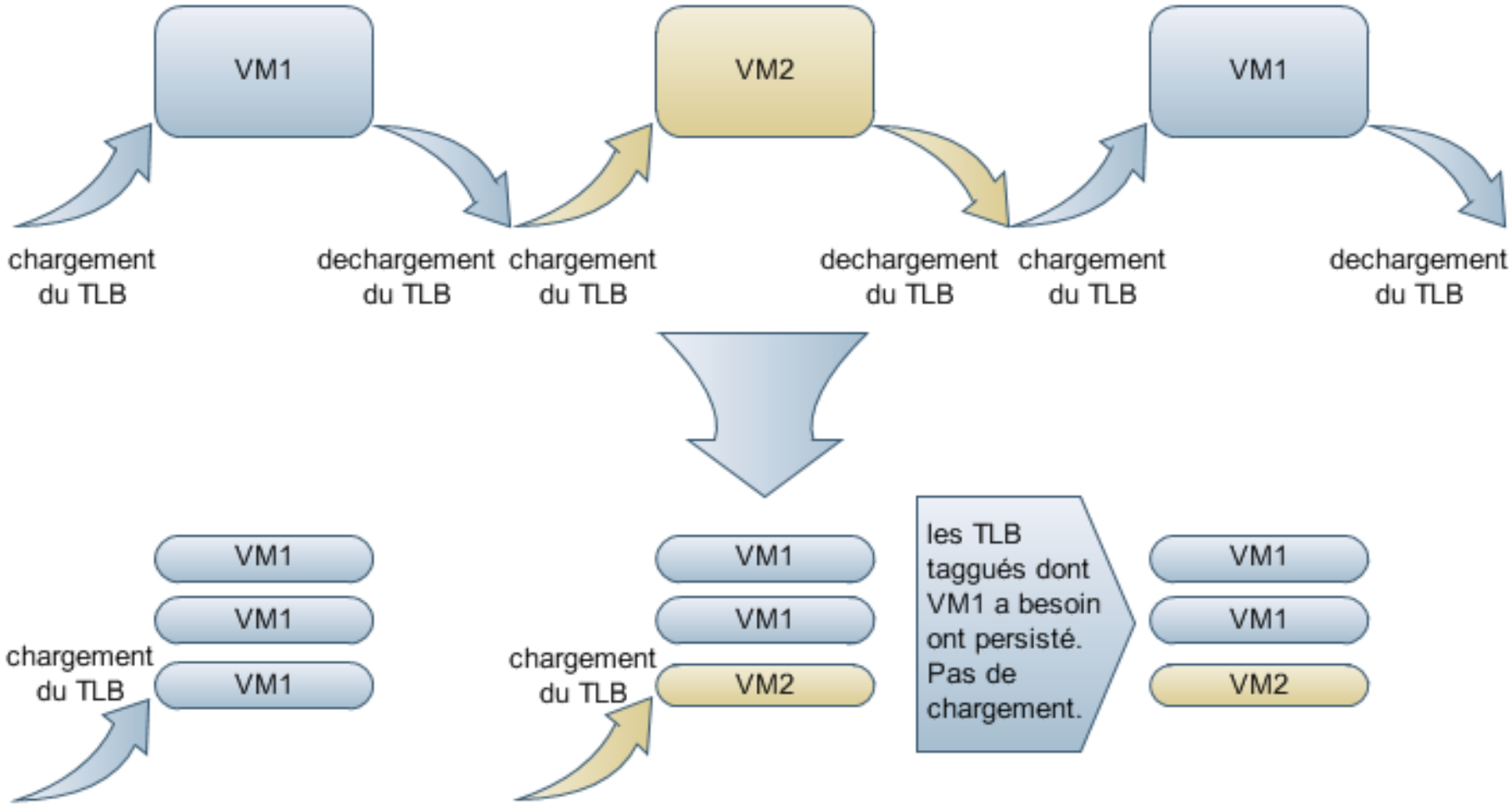
- La virtualisation est faites pour être invisible aux OS invités.
- Cependant, il serait appréciable qu'un OS invité qui arrive à saturation mémoire ou disque puisse demander plus de ressources.
- VMMCALL est une instruction privilégiée permettant aux OS invités de négocier des ressources avec la VMM.

Support matériel : exemple avec AMD-V

Tagged TLB

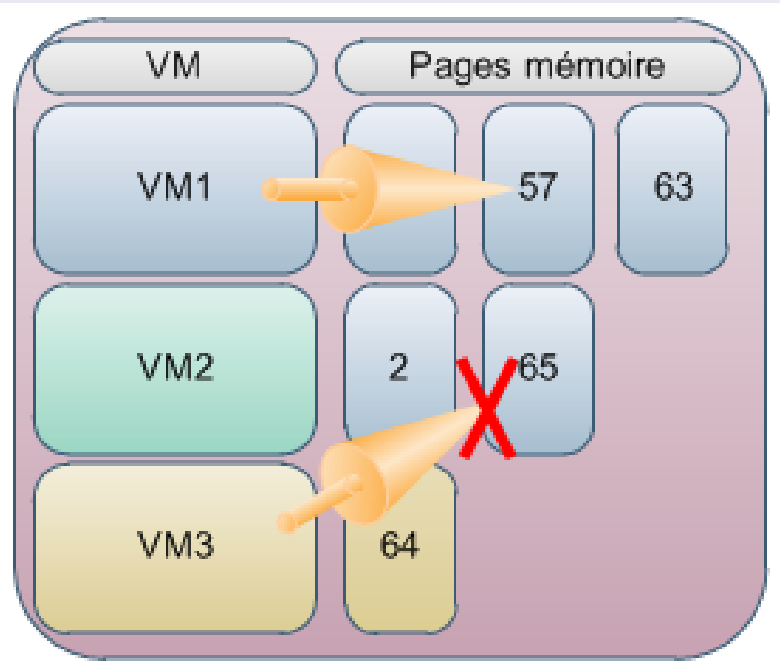
Le Translation Look-aside Buffer est une table qui contient les références entre addresses réelles et virtuelles des pages mémoire récemment accédées. TLB taggué signifie que l'on ajoute une référence (Address Space IDentifier) à la machine virtuelle associée aux données.

Support matériel : exemple avec AMD-V



Support matériel : exemple avec AMD-V

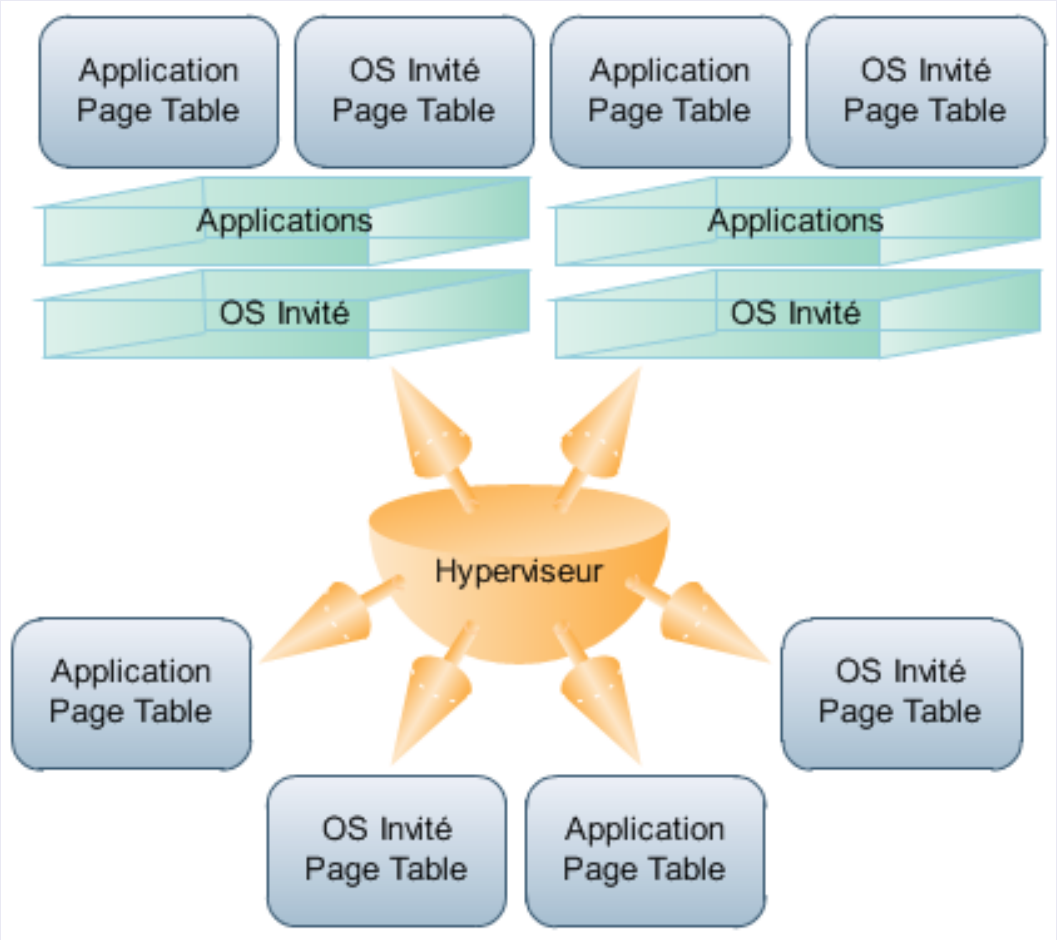
DEV



Le Device Exclusion Vector permet à la VMM de savoir immédiatement si un accés à une page mémoire est légitime ou non.

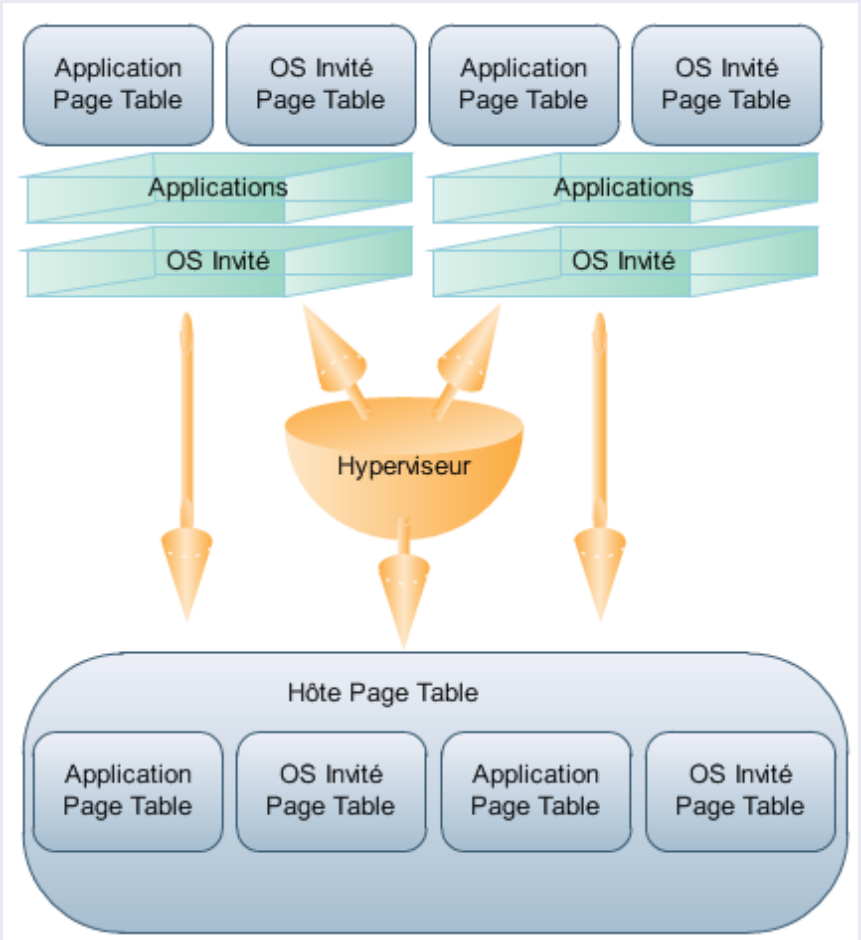
Support matériel : exemple avec AMD-V

AMD-V : Shadow Pages



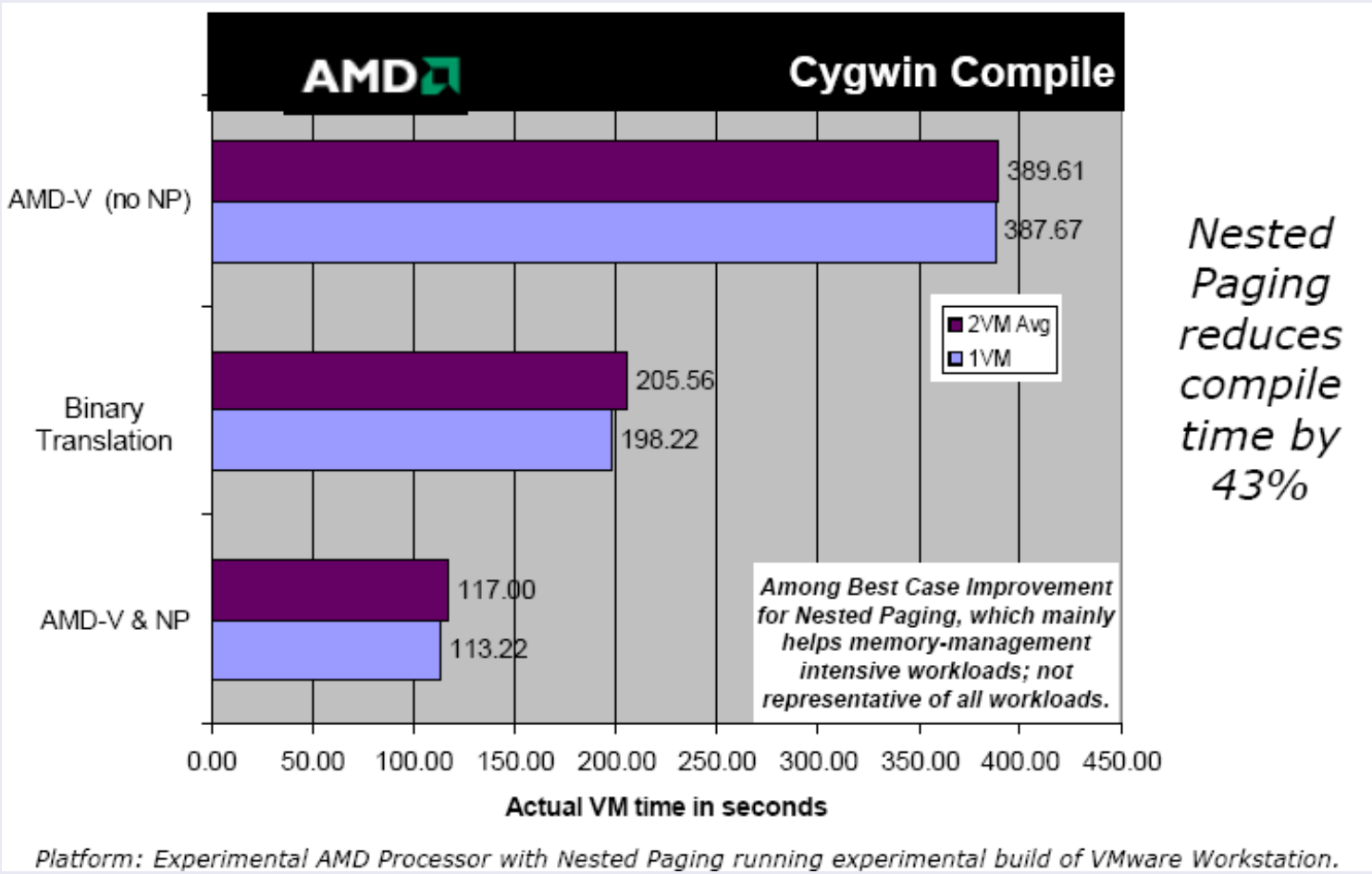
Support matériel : exemple avec AMD-V

AMD-V : Nested Pages



Support matériel : exemple avec AMD-V

AMD-V : Nested Pages



Plan

Stockage mutualisé

Stockage

Définition :

conservation d'une donnée qui varie ou non dans le temps.

Exemple :

tout type de fichiers.

Cycle de vie :

- la donnée est stoquée de manière fiable tant que son obsolescence n'est pas décidée.

Archivage

Définition :

conservation d'une donnée qui ne varie plus dans le temps.

Exemple :

résultat d'un calcul ou d'une acquisition.

Cycle de vie :

- la donnée **doit** être conservée 1 fois dans sa version définitive.
- la donnée **doit** être exclue des sauvegardes complètes.

Sauvegarde

Définition :

stockage d'une donnée qui varie dans le temps et dont on veut garder des instantannés.

Exemple :

versions successives d'un article.

Cycle de vie :

- la donnée **peut** être conservée en **N** exemplaires.
- la donnée **peut** être incluse dans les processus de sauvegardes complètes.

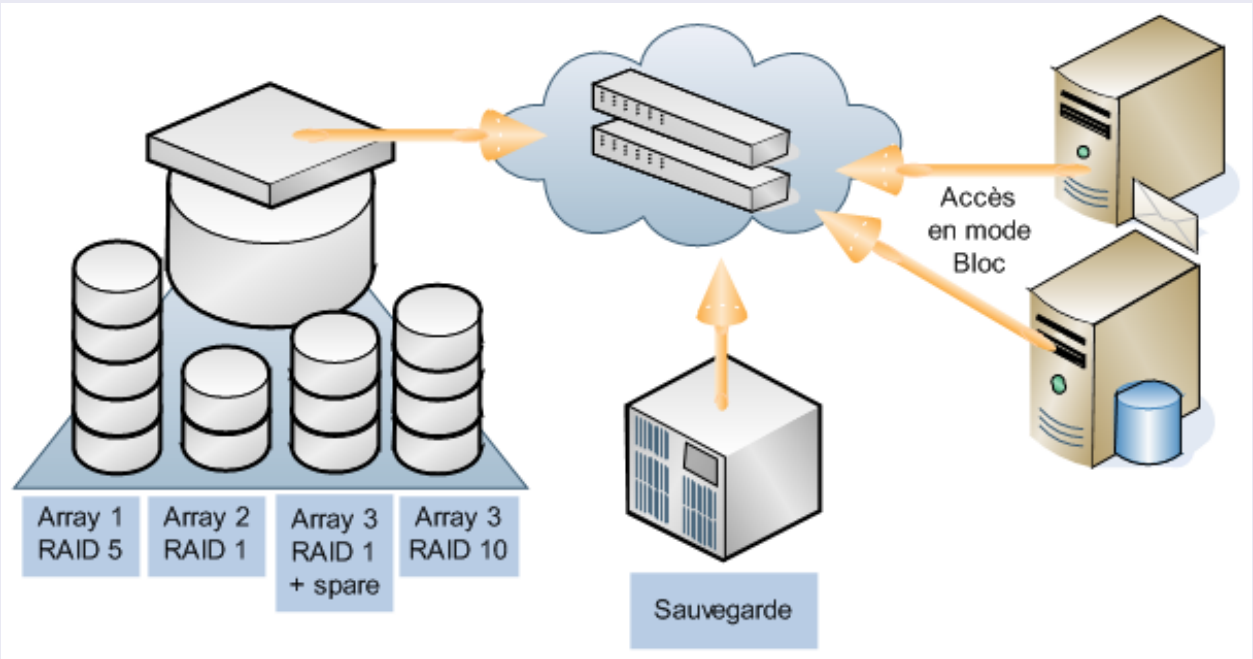
File Storage Technologies

Storage Area Network

L'**Espace Réseau de Stockage** est historiquement accessible aux travers de carte HBA Fiber Channel et depuis peu démocratisé via iSCSI. Les **SAN** présentent à leur clients des **disques virtuels**, accessibles en **mode bloc**, considérables par ceux-ci comme leurs propres disques locaux. Ils utilisent pour se faire des **unités logiques** (LUN) et du **zoning** (autorisation d'accés aux volumes). Les clients doivent gérer le système de fichiers présents sur les unités logiques.

File Storage Technologies

Storage Area Network



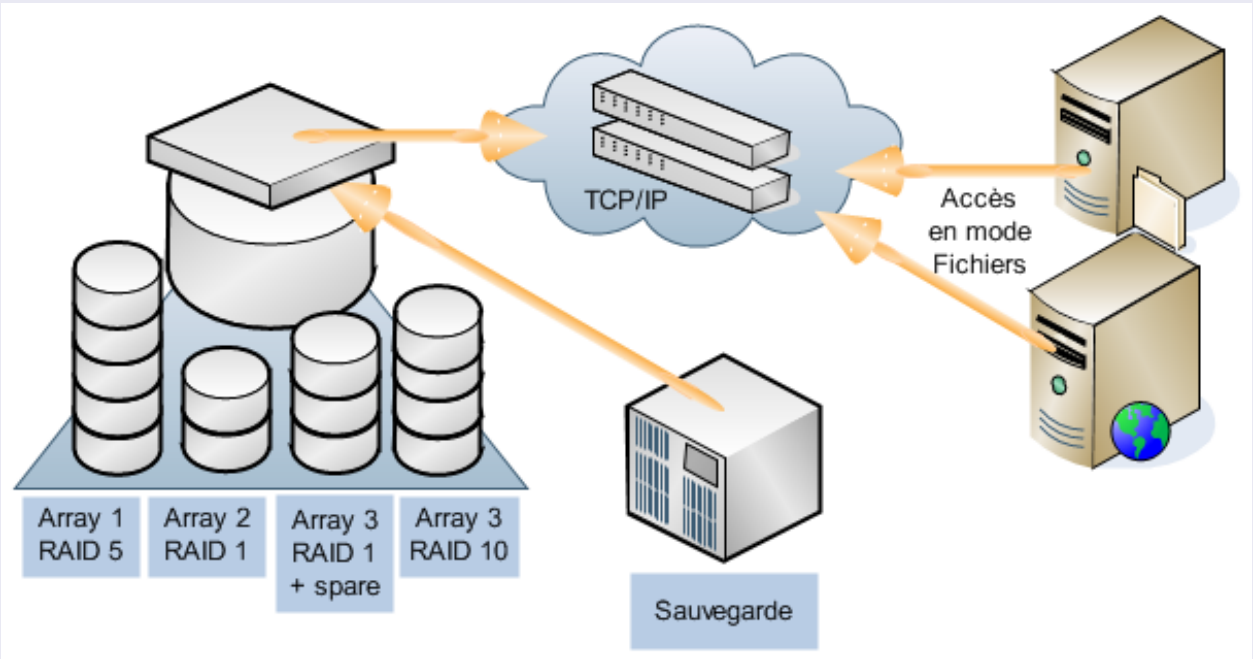
File Storage Technologies

Network Attached Storage

Un serveur NAS partage des données en **mode fichiers** avec ses clients via des protocoles réseaux comme **NFS**, **CIFS**, **AFP**. Les clients n'ont pas à connaître le système de fichier de stockage, uniquement les protocoles d'accès.

File Storage Technologies

Network Attached Storage



File Storage Technologies

RAIN

L'architecture **RAIN** (Reliable|Redundant|Random Array of Inexpensive|Independant Nodes) est au départ un sujet de recherche partant d'une vraie réflexion d'informatique théorique pour s'appliquer aux applications critiques des entreprises. Les chercheurs voulaient développer un modèle informatique distribué pour le stockage à base de **composants standards**. Le sujet a été étudié aux états-unis par **CalTech** (l'Institut de Technologie de Californie), le laboratoire **JPL** (Jet Propulsion) de la **NASA** et par le **DARPA** (Defense Advanced Research Projects Agency, département de la défense).

File Storage Technologies

RAIN

- Distribution entre les noeuds assurée par des assemblages dits **Maximum Distance Separable Array Codes** permettant de calculer une répartition des données et d'assurer le recouvrement en cas de défaillance d'un élément de la chaîne.
- **Auto-reconfiguration** en cas de panne d'un constituant, d'un ajout ou d'un retrait d'un noeud du cluster.
- Aucune limite du nombre de noeuds.

File Storage Technologies

RAIN

- Distribution entre les noeuds assurée par des assemblages dits **Maximum Distance Separable Array Codes** permettant de calculer une répartition des données et d'assurer le recouvrement en cas de défaillance d'un élément de la chaîne.
- **Auto-reconfiguration** en cas de panne d'un constituant, d'un ajout ou d'un retrait d'un noeud du cluster.
- Aucune limite du nombre de noeuds.

File Storage Technologies

RAIN

- Distribution entre les noeuds assurée par des assemblages dits **Maximum Distance Separable Array Codes** permettant de calculer une répartition des données et d'assurer le recouvrement en cas de défaillance d'un élément de la chaîne.
- **Auto-reconfiguration** en cas de panne d'un constituant, d'un ajout ou d'un retrait d'un noeud du cluster.
- Aucune limite du nombre de noeuds.

File Storage Technologies

RAIN

- La couche RAIN fournit un **mécanisme d'équilibre de charge** au sein du cluster pour les requêtes entrantes et sa philosophie de redondance permet d'accepter plusieurs défaillances de plusieurs éléments de la configuration : noeud, interface ou lien réseau, switch, stockage ou noeud complet.
- Côté configuration, il peut être envisagé de déporter certains noeuds et de mixer des liens LAN et WAN pour rigidifier le cluster.

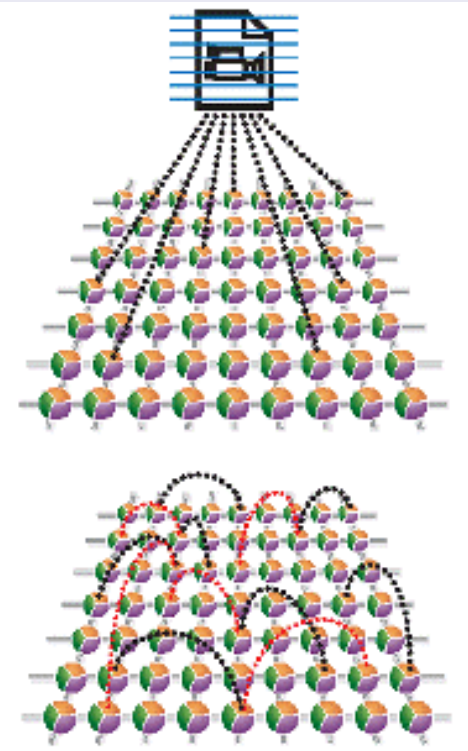
File Storage Technologies

RAIN

- La couche RAIN fournit un **mécanisme d'équilibre de charge** au sein du cluster pour les requêtes entrantes et sa philosophie de redondance permet d'accepter plusieurs défaillances de plusieurs éléments de la configuration : noeud, interface ou lien réseau, switch, stockage ou noeud complet.
- Côté configuration, il peut être envisagé de déporter certains noeuds et de mixer des liens LAN et WAN pour rigidifier le cluster.

File Storage Technologies

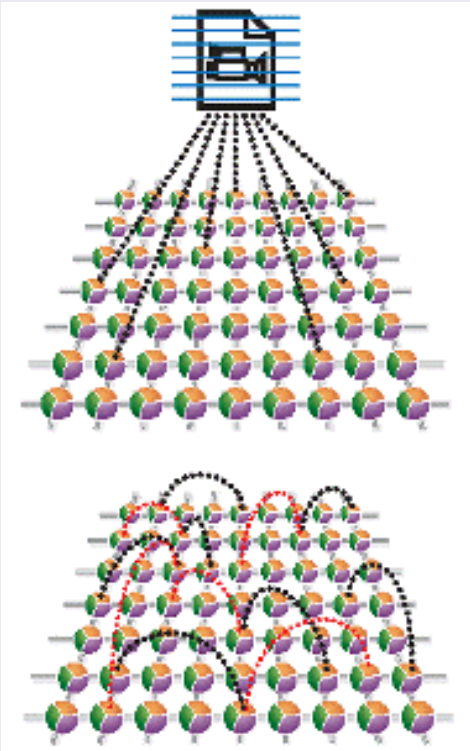
RAIN



- Gestion de fichiers, pas de disques.
- Evolutivité multi-dimensionnelle : performance, capacité, redondance
- Donnée éclatée/répliquée/distribuée entre plusieurs entités.
- Redondance assurée avec une finesse importante contrairement à une approche massive type RAID, coûteuse en reconstruction, qui fonctionne au niveau volume ou lun.

File Storage Technologies

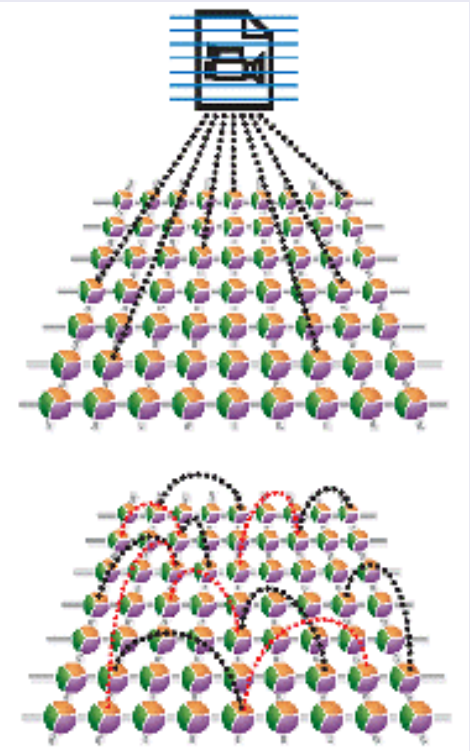
RAIN



- Gestion de fichiers, pas de disques.
- Evolutivité multi-dimensionnelle : performance, capacité, redondance
- Donnée éclatée/répliquée/distribuée entre plusieurs entités.
- Redondance assurée avec une finesse importante contrairement à une approche massive type RAID, coûteuse en reconstruction, qui fonctionne au niveau volume ou lun.

File Storage Technologies

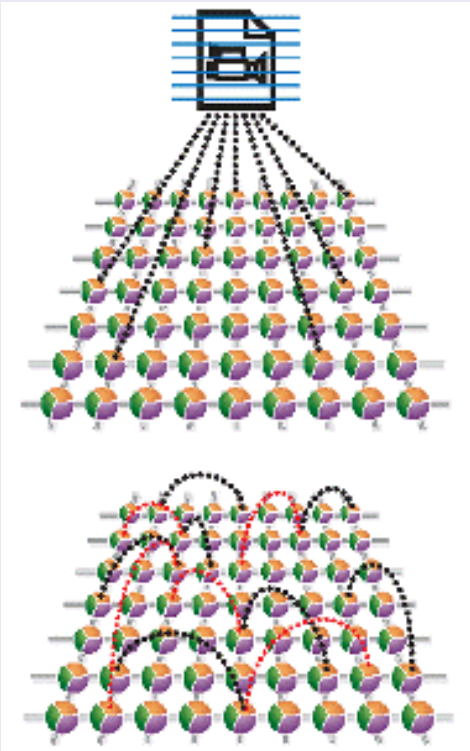
RAIN



- Gestion de fichiers, pas de disques.
- Evolutivité multi-dimensionnelle : performance, capacité, redondance
- Donnée éclatée/répliquée/distribuée entre plusieurs entités.
- Redondance assurée avec une finesse importante contrairement à une approche massive type RAID, coûteuse en reconstruction, qui fonctionne au niveau volume ou lun.

File Storage Technologies

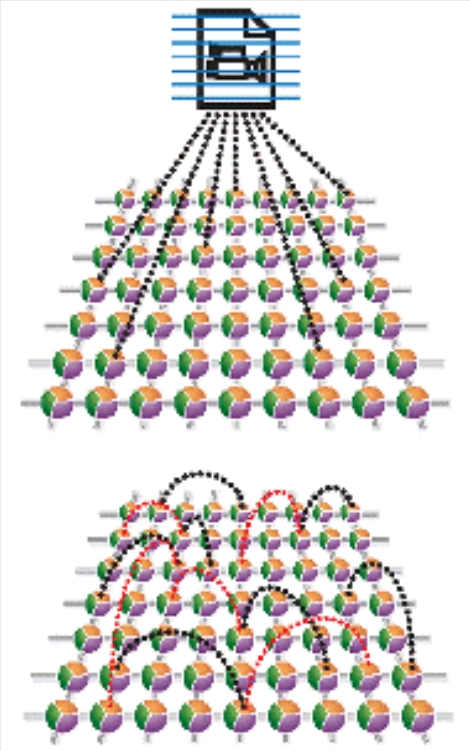
RAIN



- Gestion de fichiers, pas de disques.
- Evolutivité multi-dimensionnelle : performance, capacité, redondance
- Donnée éclatée/répliquée/distribuée entre plusieurs entités.
- Redondance assurée avec une finesse importante contrairement à une approche massive type RAID, coûteuse en reconstruction, qui fonctionne au niveau volume ou lun.

File Storage Technologies

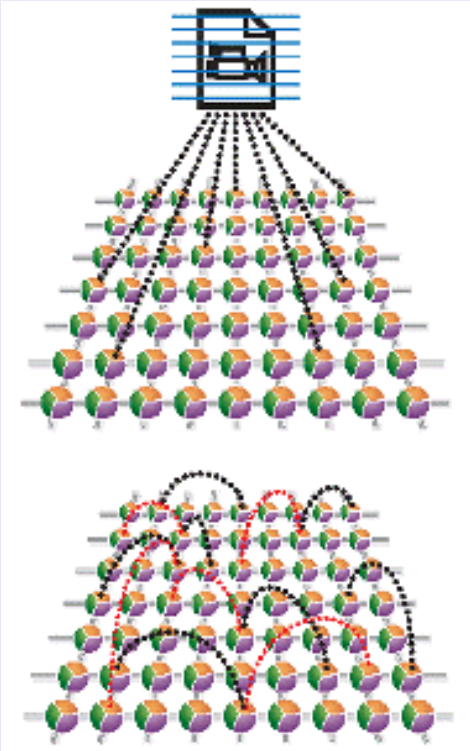
RAIN



- Répartition des fragments de fichiers sur plusieurs unités permettant d'aggréger la bande passante, de contrôler et limiter le temps de traitement.
- Petit et gros fichiers ne seront pas divisé dans le même nombre de fragments.
- Les datas ne peuvent plus vivre avec les meta-datas comme les systèmes de fichiers de conception "ancienne" le faisaient.

File Storage Technologies

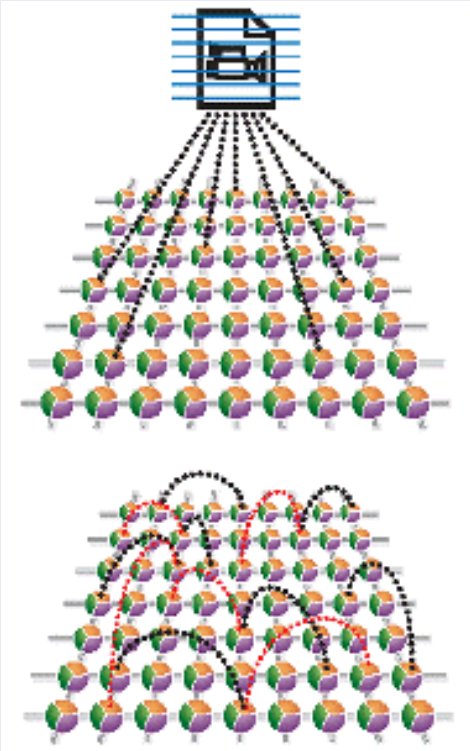
RAIN



- Répartition des fragments de fichiers sur plusieurs unités permettant d'aggréger la bande passante, de contrôler et limiter le temps de traitement.
- Petit et gros fichiers ne seront pas divisé dans le même nombre de fragments.
- Les datas ne peuvent plus vivre avec les meta-datas comme les systèmes de fichiers de conception "ancienne" le faisaient.

File Storage Technologies

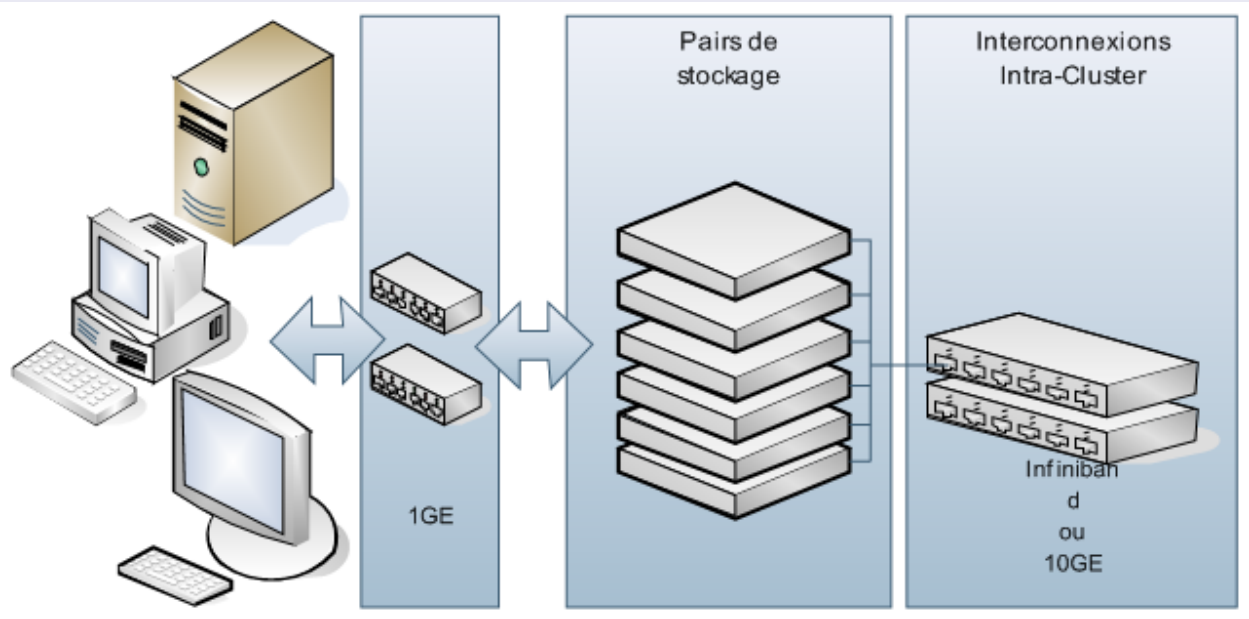
RAIN



- Répartition des fragments de fichiers sur plusieurs unités permettant d'aggréger la bande passante, de contrôler et limiter le temps de traitement.
- Petit et gros fichiers ne seront pas divisé dans le même nombre de fragments.
- Les datas ne peuvent plus vivre avec les meta-datas comme les systèmes de fichiers de conception "ancienne" le faisaient.

File Storage Technologies

RAIN



File Storage Technologies

CELLULAR

Architecture distribuée à base de matériel standard potentiellement hétérogène.

- cellule ⇒ unité de stockage potentiellement hétérogène (disques, bandes...)
- cercle ⇒ espace de stockage virtualisé, potentiellement distribué sur plusieurs sites
- domaine ⇒ regroupement logique de cellule (gestion des authorisations d'accès)

File Storage Technologies

CELLULAR

Architecture distribuée à base de matériel standard potentiellement hétérogène.

- cellule ⇒ unité de stockage potentiellement hétérogène (disques, bandes...)
- cercle ⇒ espace de stockage virtualisé, potentiellement distribué sur plusieurs sites
- domaine ⇒ regroupement logique de cellule (gestion des authorisations d'accès)

File Storage Technologies

CELLULAR

Architecture distribuée à base de matériel standard potentiellement hétérogène.

- cellule ⇒ unité de stockage potentiellement hétérogène (disques, bandes...)
- cercle ⇒ espace de stockage virtualisé, potentiellement distribué sur plusieurs sites
- domaine ⇒ regroupement logique de cellule (gestion des authorisations d'accès)

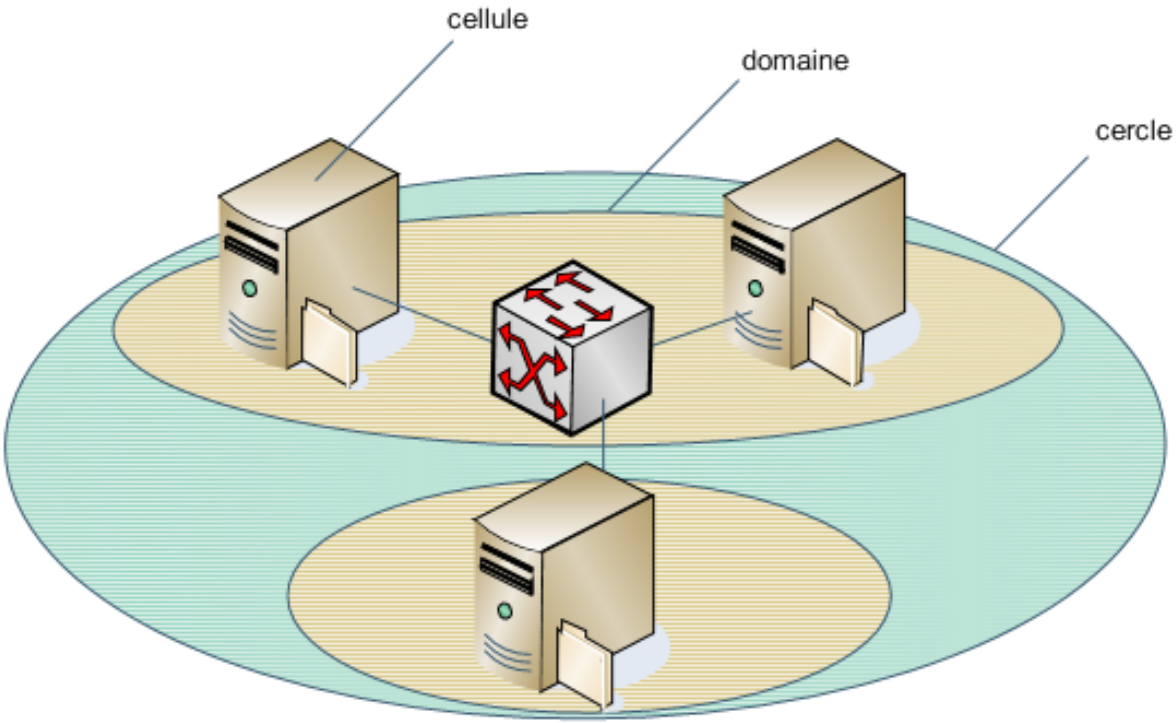
File Storage Technologies

CELLULAR

Architecture distribuée à base de matériel standard potentiellement hétérogène.

- cellule ⇒ unité de stockage potentiellement hétérogène (disques, bandes...)
- cercle ⇒ espace de stockage virtualisé, potentiellement distribué sur plusieurs sites
- domaine ⇒ regroupement logique de cellule (gestion des authorisations d'accès)

File Storage Technologies

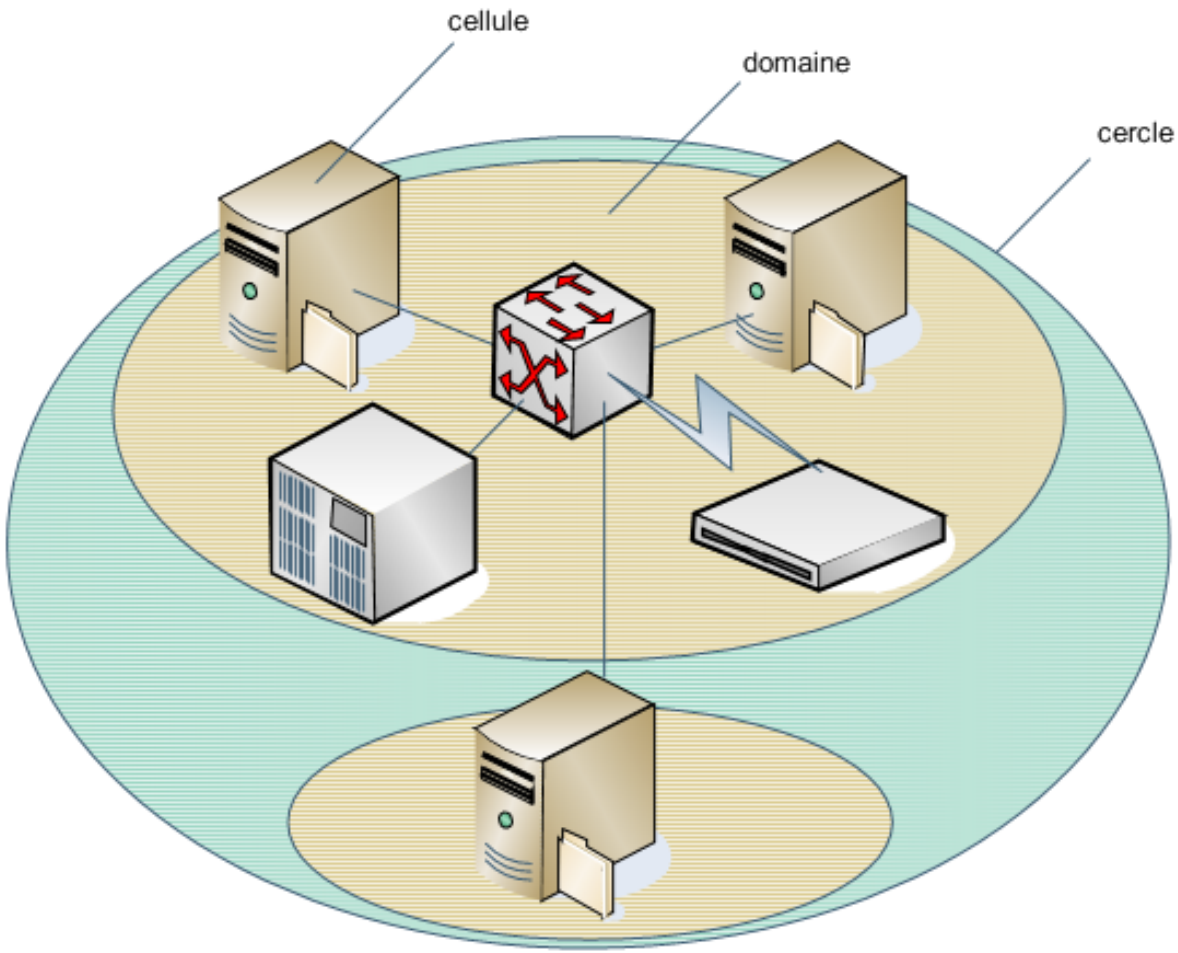


File Storage Technologies

CELLULAR

Comme dans l'architecture RAIN, l'**augmentation** du volume de stockage se fait par **ajout dynamique** de cellules supplémentaires aux cercles. Aucune cellule n'occupe de role maitre.

File Storage Technologies



Sauvegarde

NDMP

Network **D**ata **M**anagement **P**rotocol est un protocole de communication ouvert crée entre autre par Network Appliance pour assurer la sauvegarde des appliances NAS sur lesquels, par définition, on n'install pas de client de sauvegarde traditionnel. NDMP permet notemment un accès standardisé (à terme plus besoin de client propriétaire, il y'a une proposition de projet Google SoC 2007 NetBSD) et optimisé aux données dans un but de sauvegarde.

Sauvegarde

VTL

Une **V**irtual **T**ape **L**ibrary est un système de stockage incluant un serveur, une grappe de disques et un logiciel émulant une bande magnétique à partir de ces disques.

- diminution de la fenêtre d'ouverture.
- optimisation de l'espace utilisé sur les bandes.
- temps d'accès au données réduits.
- émulation de plusieurs lecteurs de bandes magnétiques
- les bandes servent à l'archivage final des données. On parle alors de **D2D2T** (Disk to Disk to Tape)

Sauvegarde

Volume Shadow-Copy Service

Service fourni par Microsoft depuis Windows 2003 (également disponible sous Samba > 3.0.3) permettant la prise d'instantané des volumes NTFS. L'**accès** aux versions antérieurs des fichiers se fait **par l'utilisateur**, également au travers de l'API CIFS, directement dans l'explorateur Windows via l'onglet *Propriétés/Versions Précédentes*.

Sauvegarde

Volume Shadow-Copy Service