IBM SYSTEM p5
Virtualisation des ressources

IBM System p5

IBM eServer™ pSeries®

WAKE UP TO THE POWER.

Gilles Rigitano

FTSS IBM France et pays d'Afrique francophone

Gilles.Rigitano@fr.ibm.com

ON DEMAND BUSINESS™

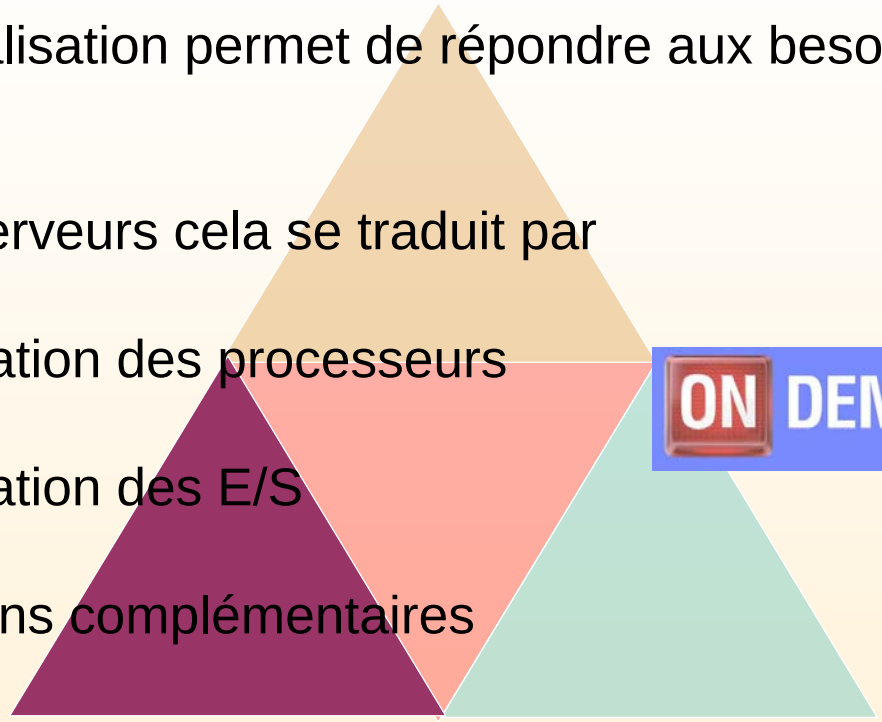
Auteur: alain.lechevalier@fr.ibm.com

Pourquoi la virtualisation ?

La virtualisation permet de répondre aux besoins *à la demande*

Au niveau des serveurs cela se traduit par

- La virtualisation des processeurs
- La virtualisation des E/S
- Des fonctions complémentaires



ON DEMAND BUSINESS™

Ce qui facilite leur adaptabilité

Evolution des technologies

Gamme Power5

2004

Virtualisation

Gamme Power4

2002

Partitionnement Dynamique

2001

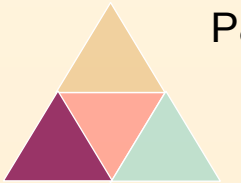
Partitionnement Logique
Processeur MultiCoeurs

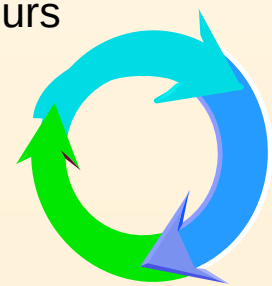
RS6000
SP2

2000

SMP
Pas de partition

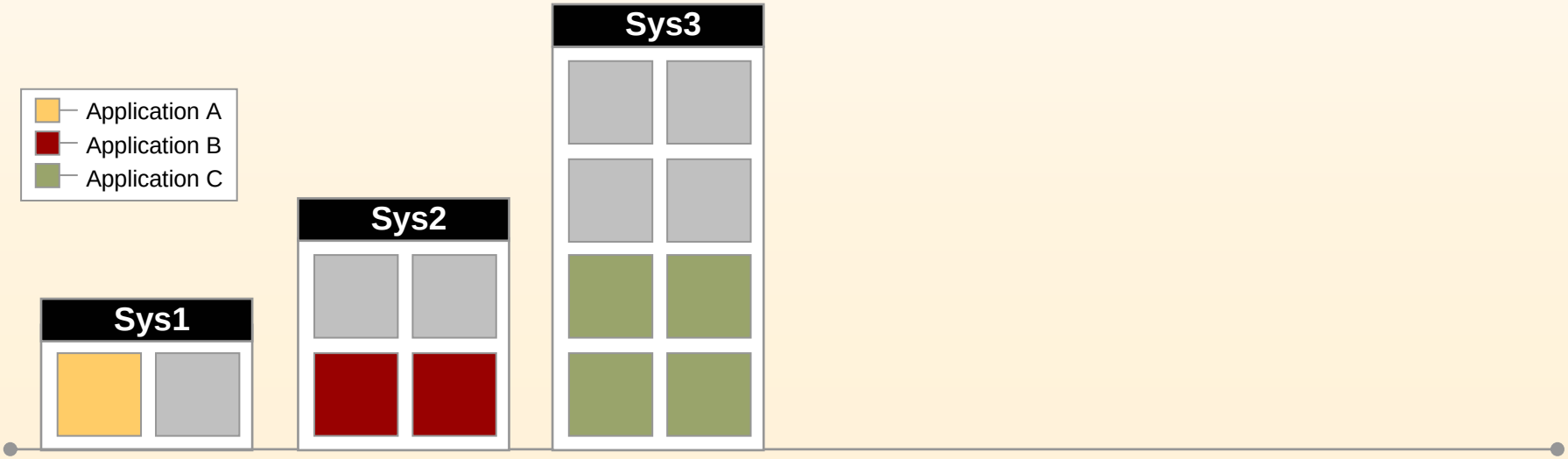
Hyperviseur





Du partitionnement …

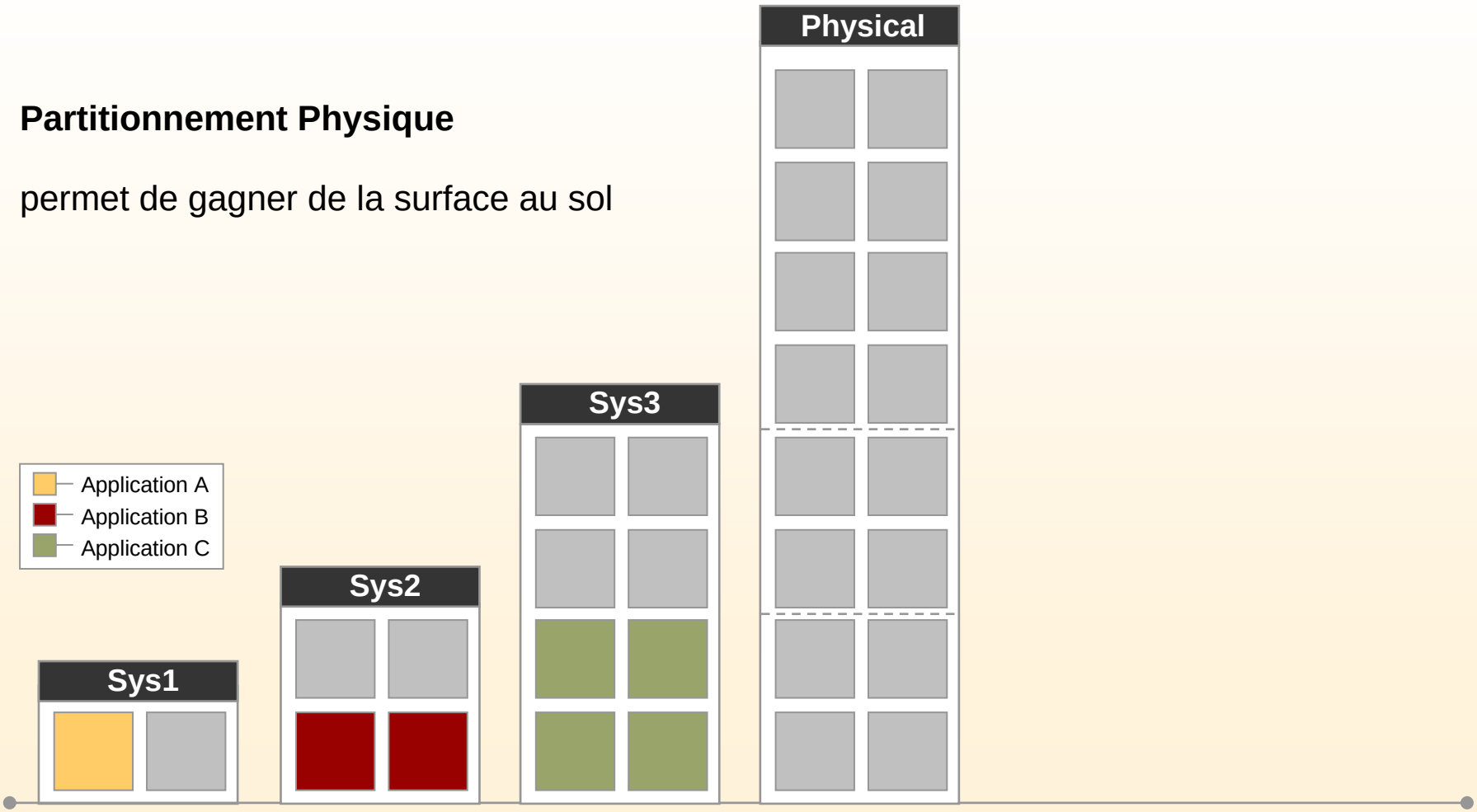
Classiquement, on trouve un système par application



Du partitionnement …

Partitionnement Physique

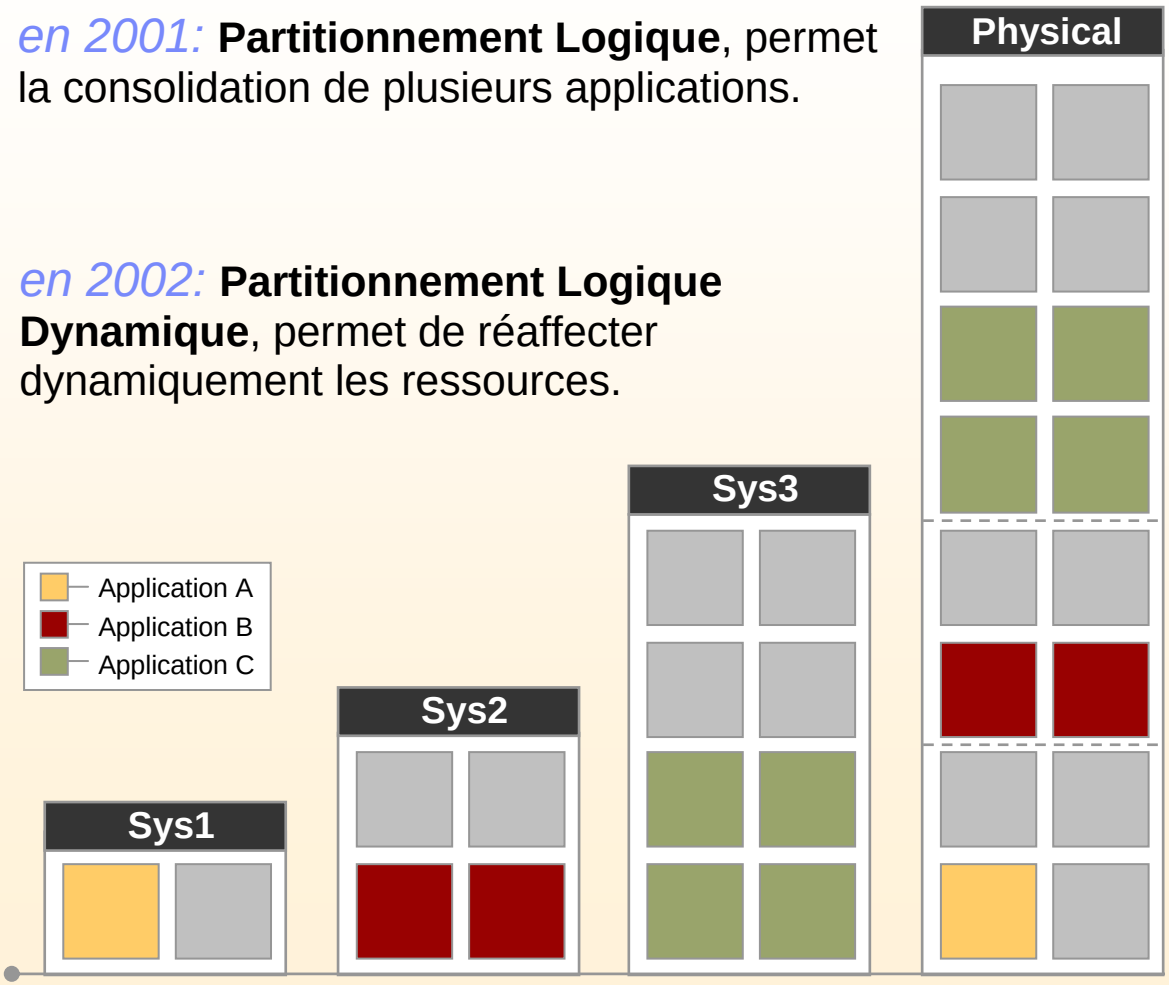
permet de gagner de la surface au sol



Du partitionnement …

en 2001: **Partitionnement Logique**, permet la consolidation de plusieurs applications.

en 2002: **Partitionnement Logique Dynamique**, permet de réaffecter dynamiquement les ressources.



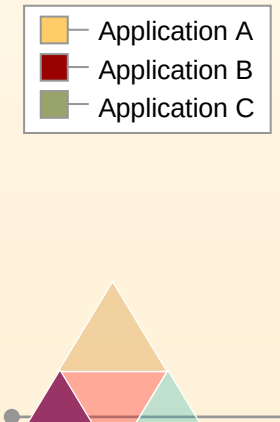
Du partitionnement … à la Virtualisation

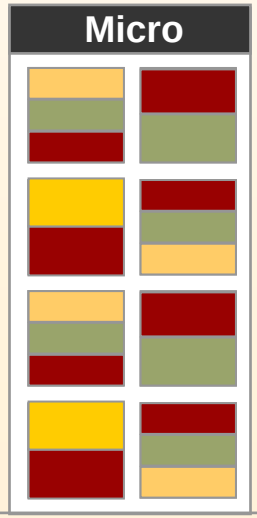
Partage de ressources du serveur entre les partitions

- Ajustement fin des ressources suivant les besoins

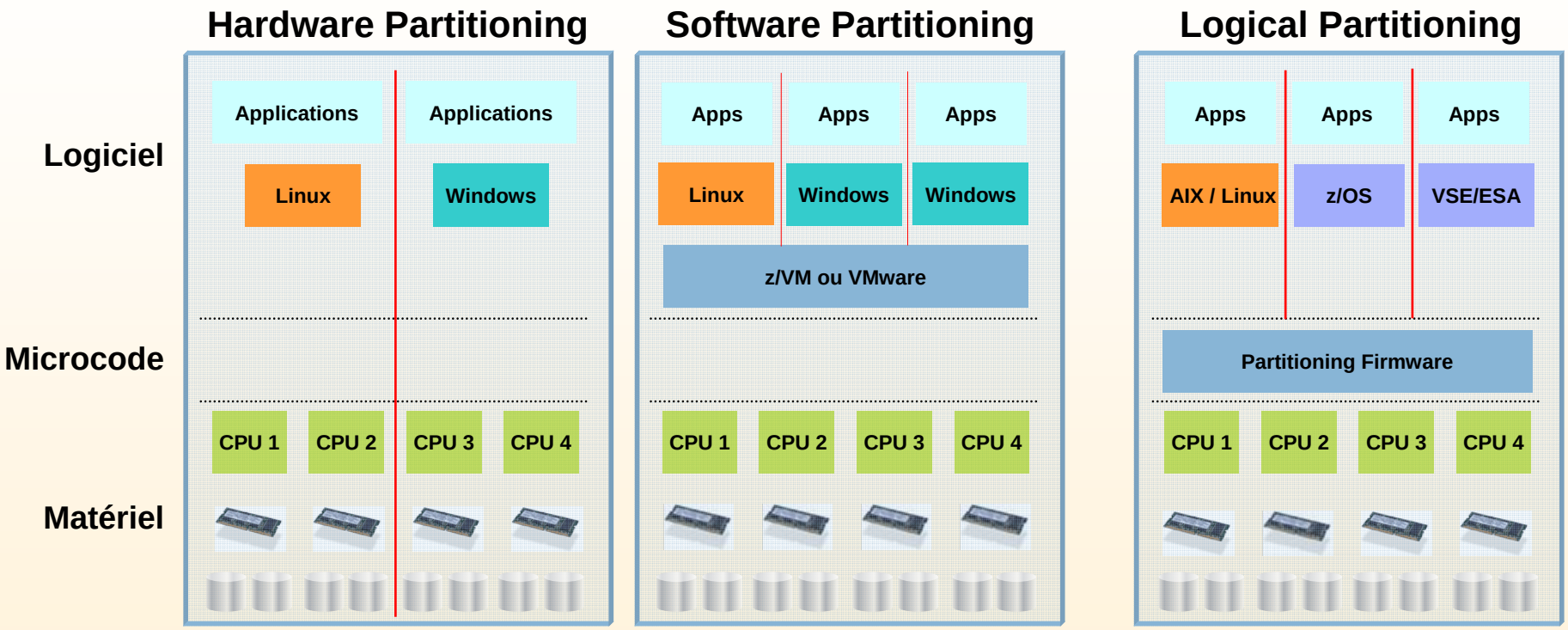
Adaptation automatique des partitions à la puissance demandée

- Pas de ressources inutilisées





Partitionnement et Virtualisation



- BladeCenter
- xSeries
- Sun Domain
- HP nPars

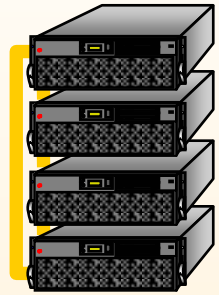
- z/VM sur zSeries
- VMware sur xSeries & BladeCenter
- HP vPars

- LPAR sur zSeries, pSeries & iSeries

Les briques technologiques

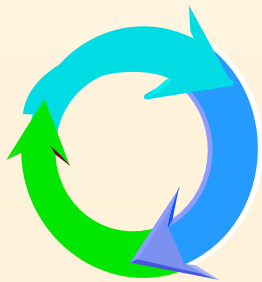
Le processeur Power5

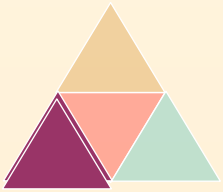
La gamme p5



Systems
Technologies

Hyperviseur





Processeur Power5

- Deuxième génération de processeur multi-cœurs
- Support du Simultaneous Multi Threading (SMT)
  - Meilleur utilisation des unités de traitement
- Gestion de la consommation électrique
- Gain de fiabilité
- Support du micro partitionnement
  - Le micro partitionnement fait partie intégrante du système

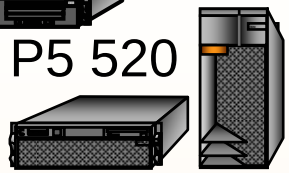
Disponible sur tout serveur Power5 et Blade JS21

P5-505/ p5-510

1-2 voies POWER5

20 partitions

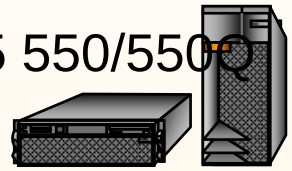
P5 520



P5 550/550Q

1-8 voies POWER5

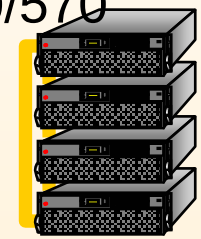
40 partitions



P5 560/570

2-16 voies POWER5

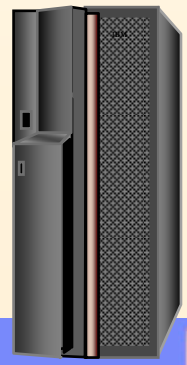
160 partitions



P5 590/595

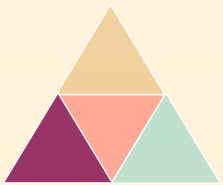
8-32 / 16-64 voies POWER5

254 partitions



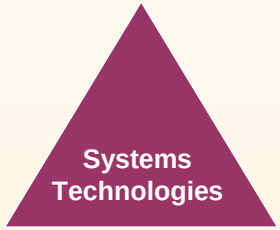
La Gamme

Virtualisation

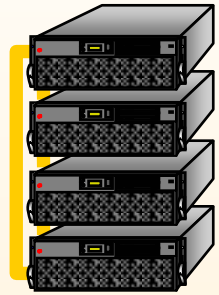


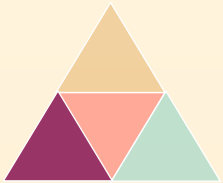
Les briques technologiques

Le processeur Power5

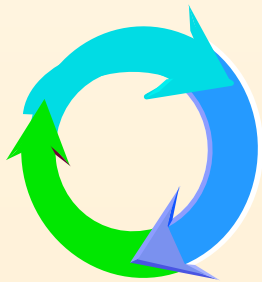


La Gamme P5





Hyperviseur



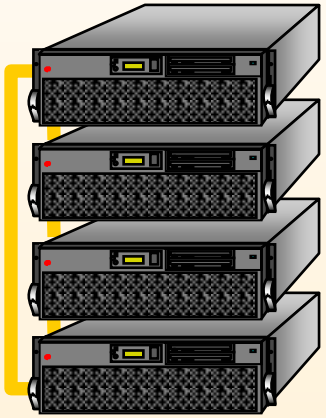
P5 Virtualisation

Tous les niveaux participent …

Systèmes d'Exploitation

- Les OS peuvent *redonner* les ressources processeur inutilisées

Aix5L version 5.3
Linux RHEL 4, SLES 9 ..
I5OS



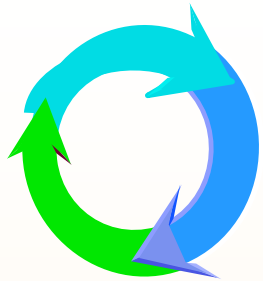
Hyperviseur (Firmware)

- Assure l'interface entre le matériel et sa représentation virtuelle

Matériel (Power5)

- Le processeur génère les intervalles de temps pour l'hyperviseur

PowerPC Hypervisor (PHYP)



- Nouvel hyperviseur pour les systèmes Power 5

Convergence avec les systèmes i5 (ex AS400)

Fonctions Passives :

Partitionnement mémoire, processeurs, I/O (mode power4)

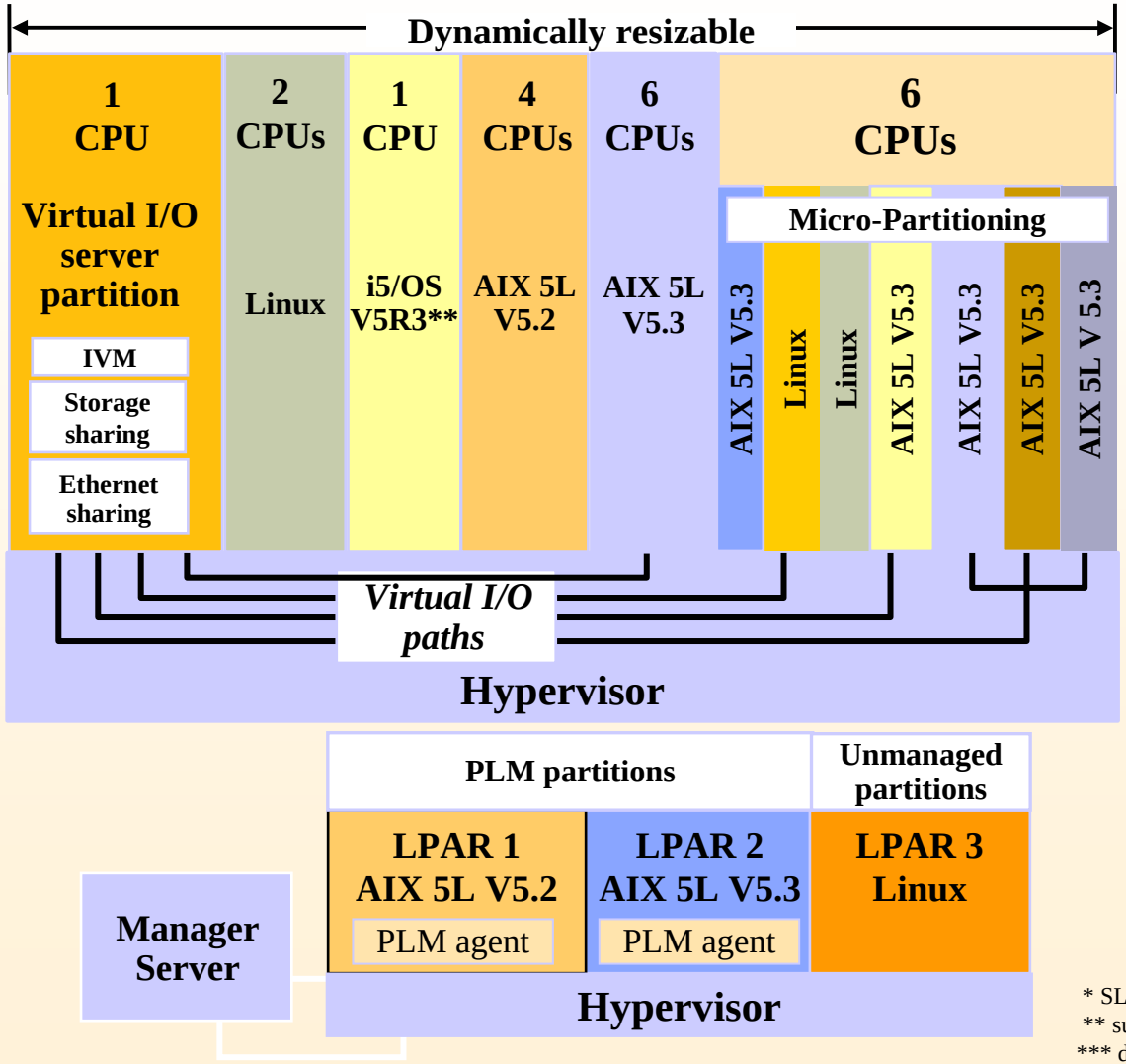
Partitions à processeurs dédiés

Fonctions Actives :

Notion de processeurs virtuels, I/O virtuelles (pour AIX5.3, Linux)

Partitions à processeurs partagés (micro partition)

Advanced POWER Virtualization pour IBM System p5



Micro-Partitioning

- Partage les processeurs entre de multiple partitions
- Minimum d'une partition à 1/10ème processeur
- AIX 5L V5.3, Linux*, or i5/OS**

Virtual I/O Server

- Ethernet partagé
- Partage des disques SCSI et Fiber Channel
- Partage des DVD-RAM et DVD-ROM
- Supporte les partitions AIX 5L V5.3 et Linux*

Partition Load Manager

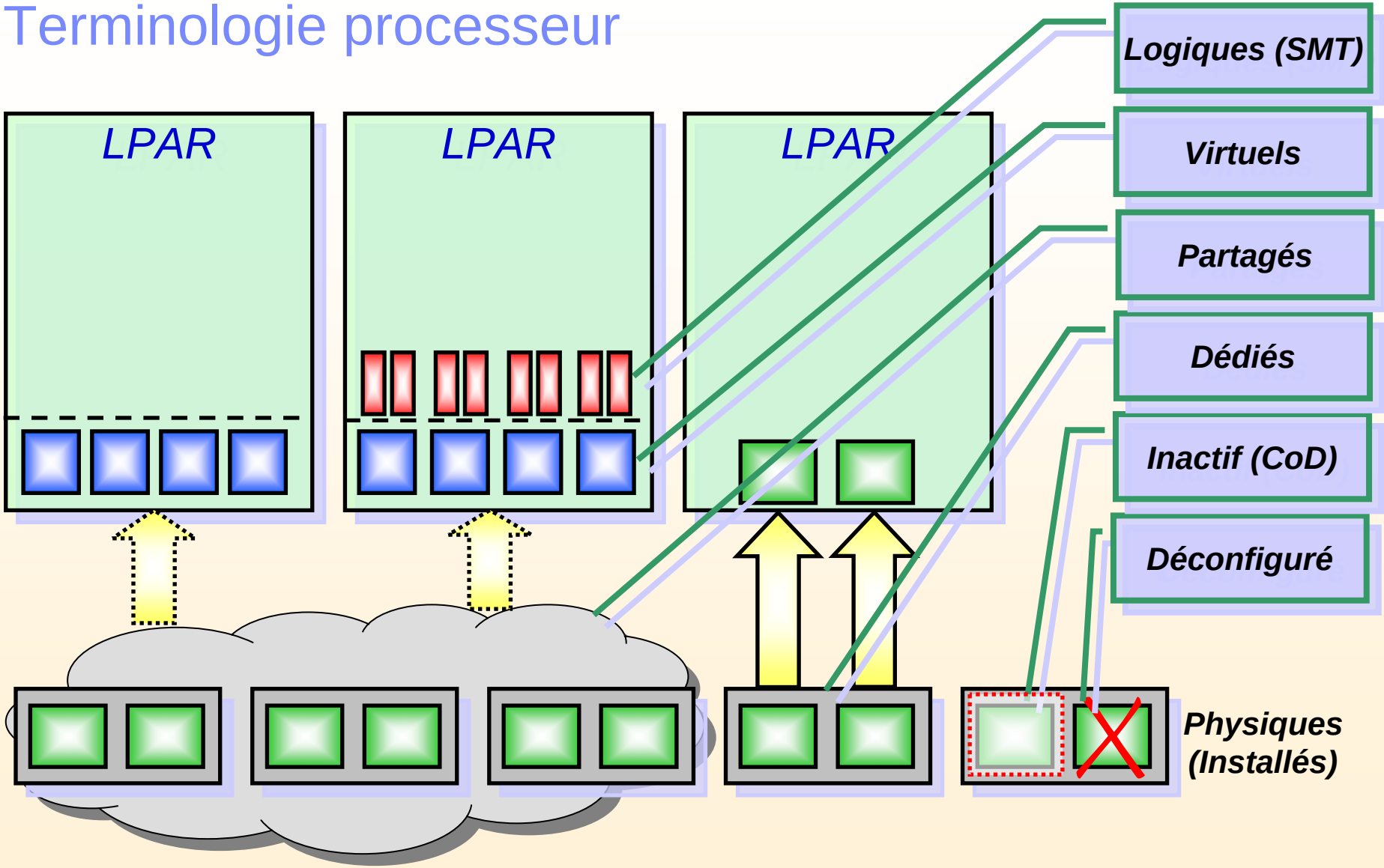
- Équilibre les demandes en processeur et mémoire

Géré grâce à la HMC ou IVM***

OPTION APV

* SLES 9 / RHEL AS 4 ou +
** sur p570 et p59x
*** du p505 au p560Q

Terminologie processeur



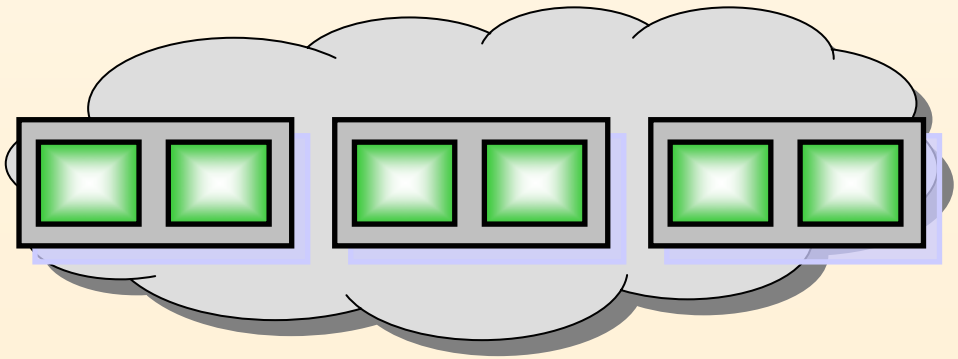
Micro-partitions : CE Capacity Entitlement

Partage de la puissance des processeurs

Le CE est l'unité de puissance des processeurs (1 processeur physique = 1 CE)

Les micro partitions vont chacune recevoir un nombre de CE représentant des fractions de processeurs.

Par exemple un pool de 6 processeurs offre 6 CE à partager entre les micro partitions
Un CE est divisible en 10 dixièmes



6 Processeurs physiques dans le pool

Micro-partitions : CE Capacity Entitlement

- **Partage des CE**
  - Chaque partition reçoit un CE (capacité d'exécution) égale au minimum à 1/10 de processeur physique.
  - Incréments par 1/100 jusqu'à la taille maximum du pool

- **mais AIX ne connaît que la notion de processeur ….**
  - Une partition va être constituée de 1 ou plusieurs processeurs virtuels qui *portent* la capacité d'exécution.

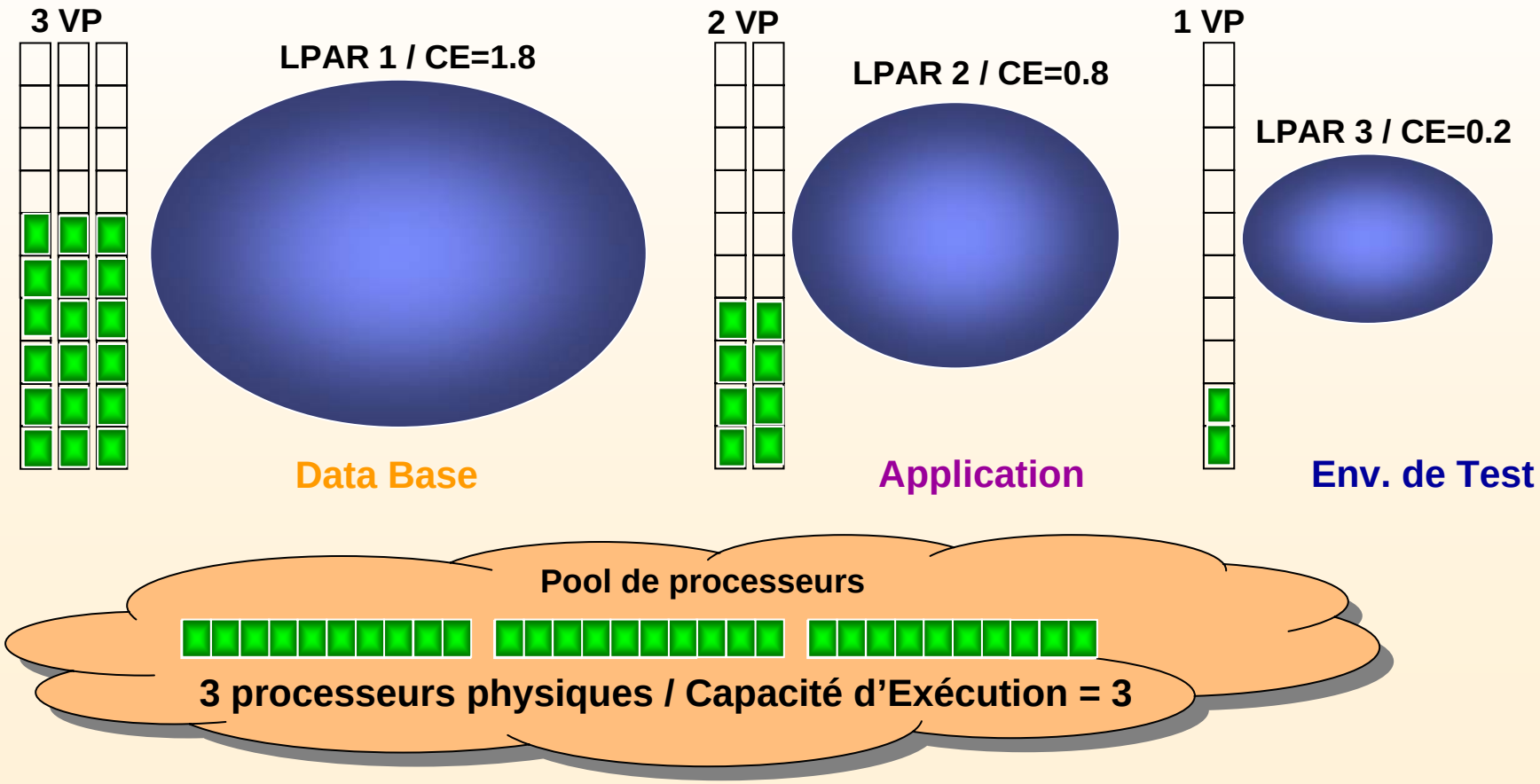
- L'affectation de la capacité d'exécution est indépendante de l'affectation de mémoire ou de slot I/O

Micro-partitioning : CE Capacité d'Exécution

3 processeurs dans le pool - **C**apacité d'**E**xécution du pool = 3.00 (3x10x0,1)

Partition 1 : **Data Base** CE=1.80, Virtual Proc = 3 (0,60 par processeur)
Partition 2 : **Application** CE=0.80, Virtual Proc = 2 (0,40 par processeur)
Partition 3 : **Env. de Test** CE=0.20, Virtual Proc = 1 (0,20 par processeur)

Total CE= 2.80, Total Virtual Proc = 6 (reste 0.20 CE disponible)



Processeurs virtuels

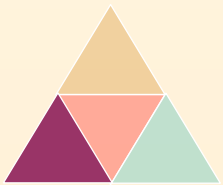
- *Une partition reçoit une « puissance processeur » exprimée en CE*
- *Les processeurs virtuels représentent cette puissance*

Mais quel est le lien entre les processeurs virtuels et les processeurs réels ?

Un processeur virtuel est en fait un échantillon de temps de processeurs réel

C'est le rôle du ***dispatcher*** de répartir la puissance CPU réelle sur les processeurs virtuels

• Il utilise un intervalle de répartition de 10ms
  • Le POWER5 possède un dispositif de décrémentation qui génère une interruption toutes les 10 ms
  • Le temps minimum alloué à un processeur est de 1ms
  • 1 CE correspond en fait à 10ms de CPU dans un intervalle de 10ms



Optimisation de l'utilisation des ressources

Une partition peut être *bridée* ou *non-bridée*

- Bridée / Non-bridée (Capped / Uncapped)
  - Bridée: Les partitions sont strictement limitées à leur valeur de CE maximum définie.
  - Non-bridée: une partition peut utiliser des ressources **disponibles** dans le pool, à concurrence du *remplissage* des processeurs virtuels.
- Priorité (Capacity weight)
  - Prioritisation de l'affectation des ressources supplémentaires entre partitions.
  - Valeur 0-255

Participation du Système d'Exploitation

La virtualisation des processeurs permet de mieux utiliser les ressources

- Si une partition n'a pas besoin de ressource à un instant donné, le système d'exploitation rend (cède) son temps CPU

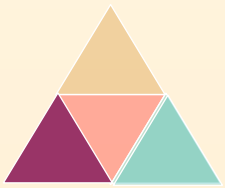
  Evite de perdre de la ressource processeur

  Comme par exemple une partition utilisant son CE à attendre une fin d'E/S

  Permet une meilleure utilisation du pool

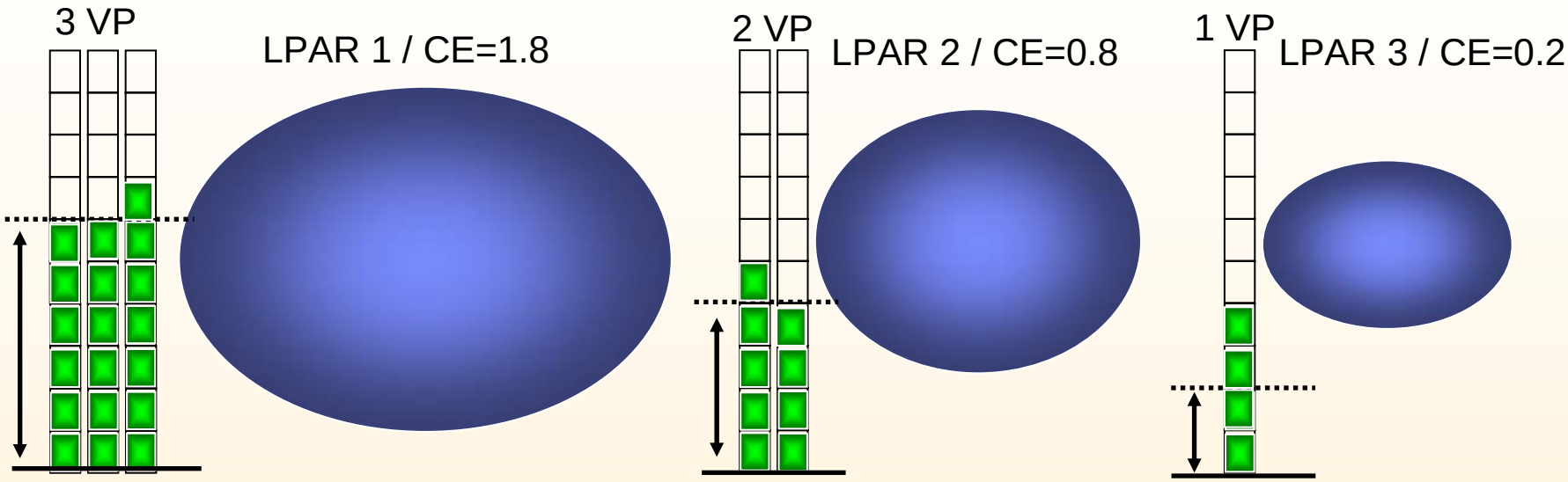
- Le temps peut être affecté à un autre processeur virtuel de la même partition si besoin

- En retour, le processeur virtuel est potentiellement réactivable dans le même intervalle de temps si nécessaire



Cet ajustement se fait toutes les 10 millisecondes !!

Micro-partitions: Ajustement des puissances



La virtualisation permet de faire varier en temps réel et d'une façon transparente la « puissance d'un processeur »

3 processeurs physiques / Capacité d'Exécution = 3

Virtual I/O Server: Entrees Sorties Virtuelles

Virtual I/O Server (VIOS)

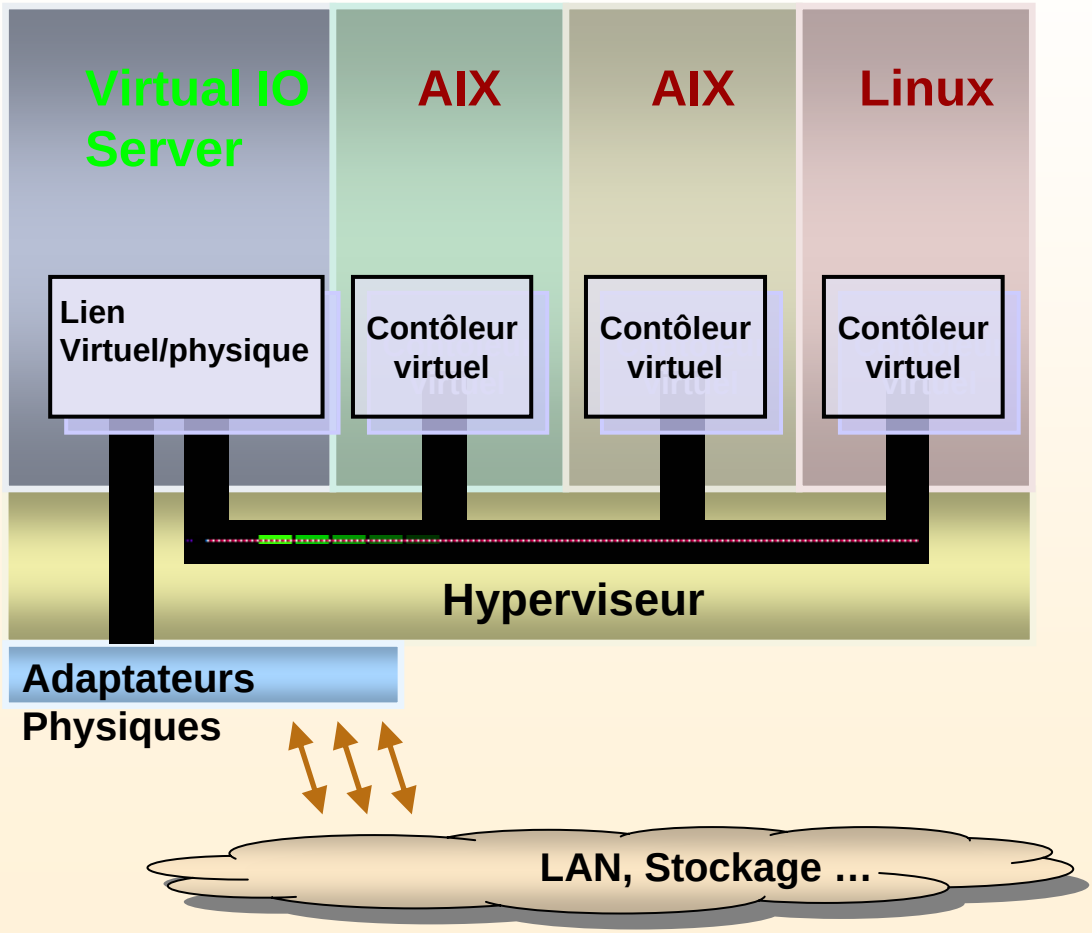
- **VIOS : partition supplémentaire prenant en charge la mutualisation d'E/S physiques**

Objectif : économiser des slots PCI

- **Important dans les environnements comportant de nombreuses partitions**

Virtualisation : Réseau, Stockage

Supporte les partitions AIX 5L V5.3 et Linux



Virtual I/O Server : Virtual Ethernet

Réseau inter-partition interne basé sur la mémoire

- Les paquets sont copiés entre les LPAR

Communications inter-partition sans adaptateur réel

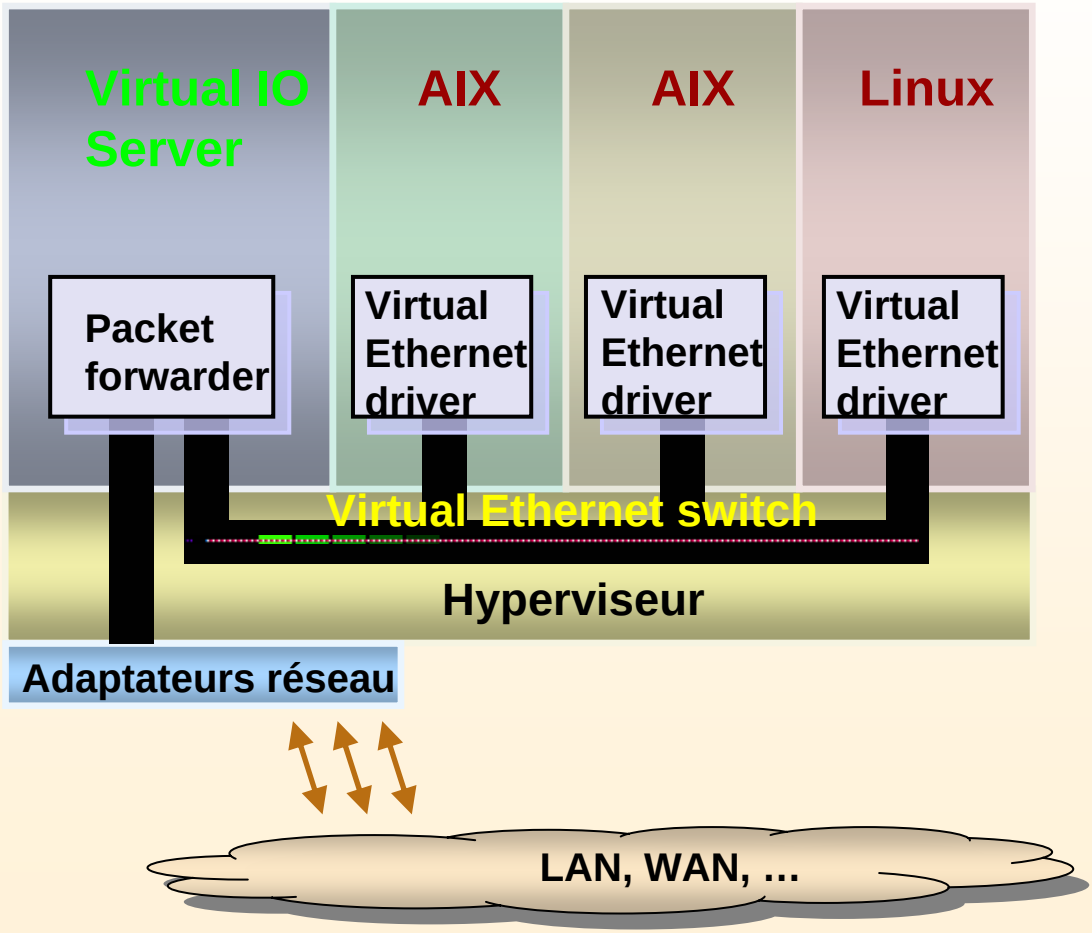
- Supporte les connexions HMC et NIM

Supporte plusieurs protocoles (IPv4, IPv6, ICMP)

Se configure comme un réseau standard Ethernet

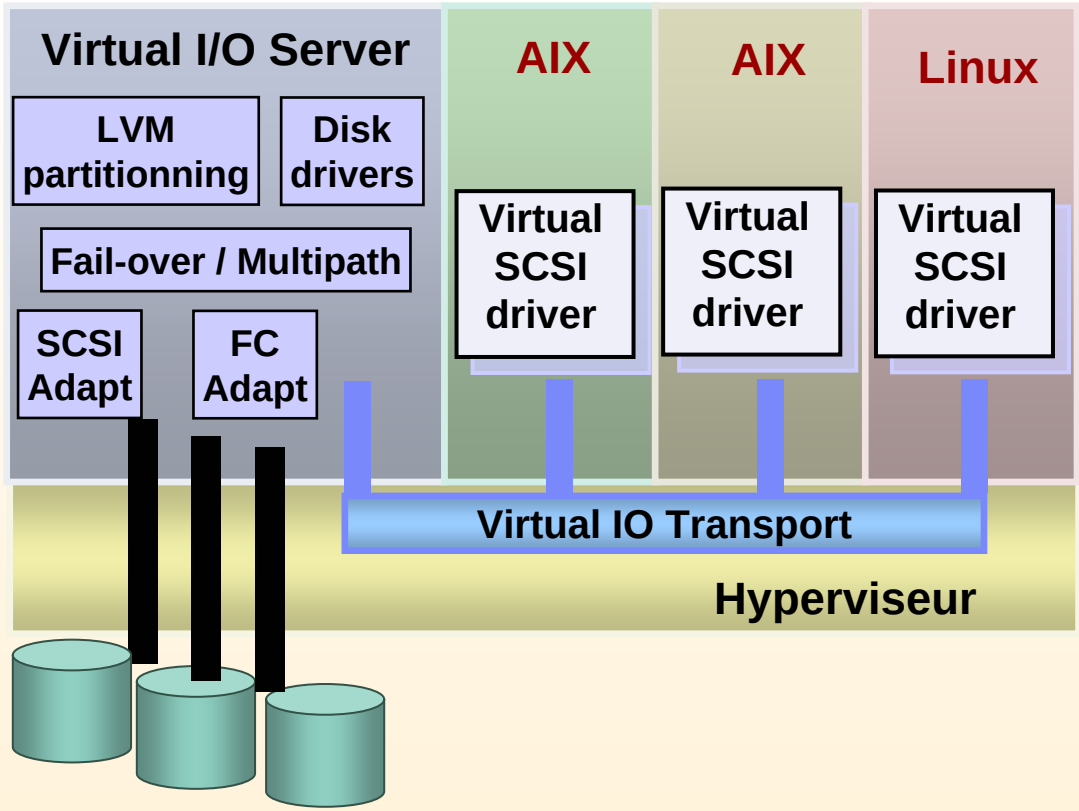
Jusqu'à 16 virtual Ethernet par adaptateur

Supporte les partitions AIX 5L V5.3 et Linux



Virtual I/O Server: Virtual SCSI

Partage d'adaptateurs SCSI et Fibre Channel disques



Optimise le nombre d'adaptateurs

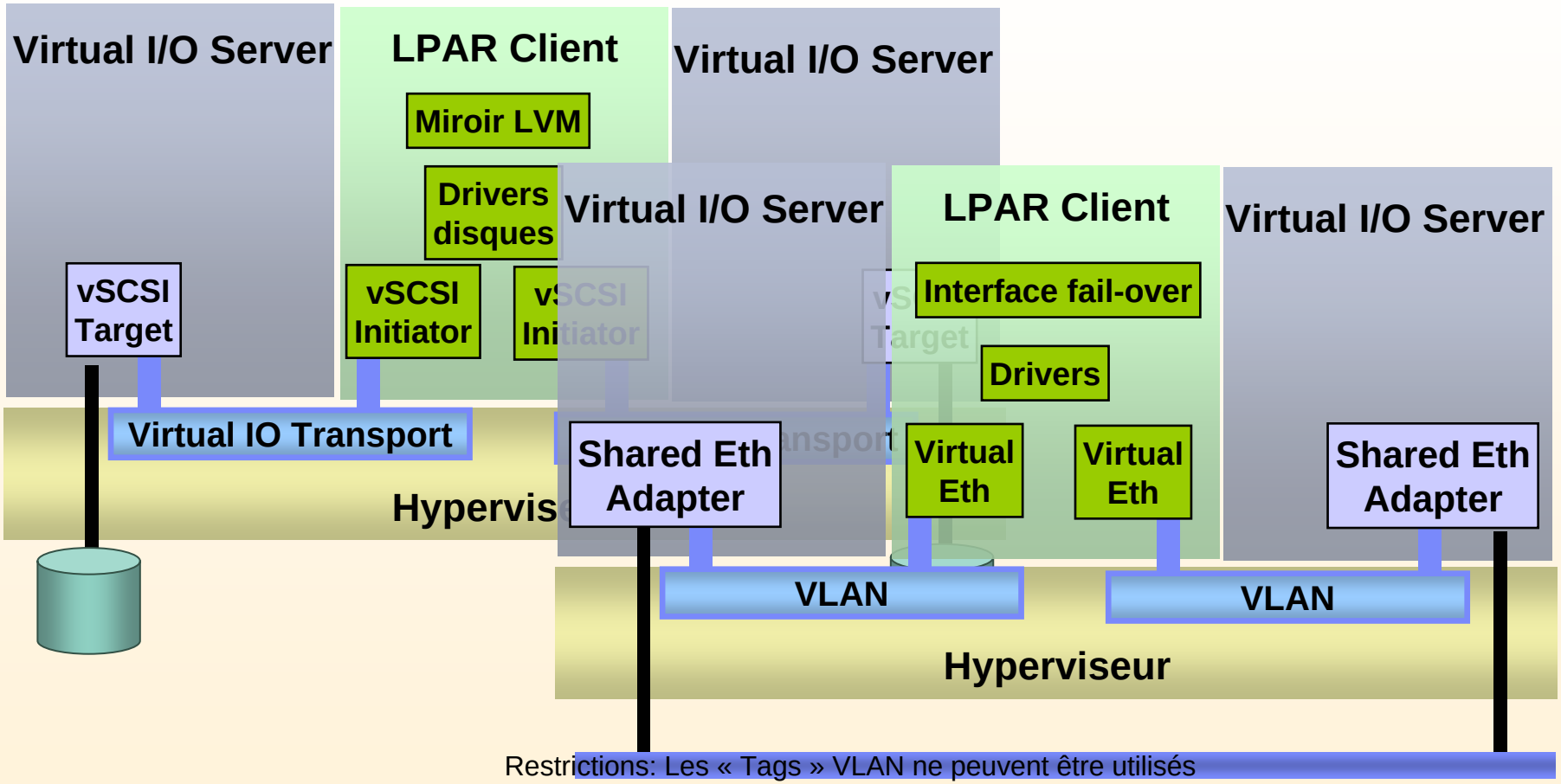
- Réduit le nombre de tiroirs I/O
- Évite le manque de slot E/S

Supporte les partitions AIX 5L V5.3 et Linux

- CPU dédiés ou partagés

VIOS: Les configurations sécurisées supportées

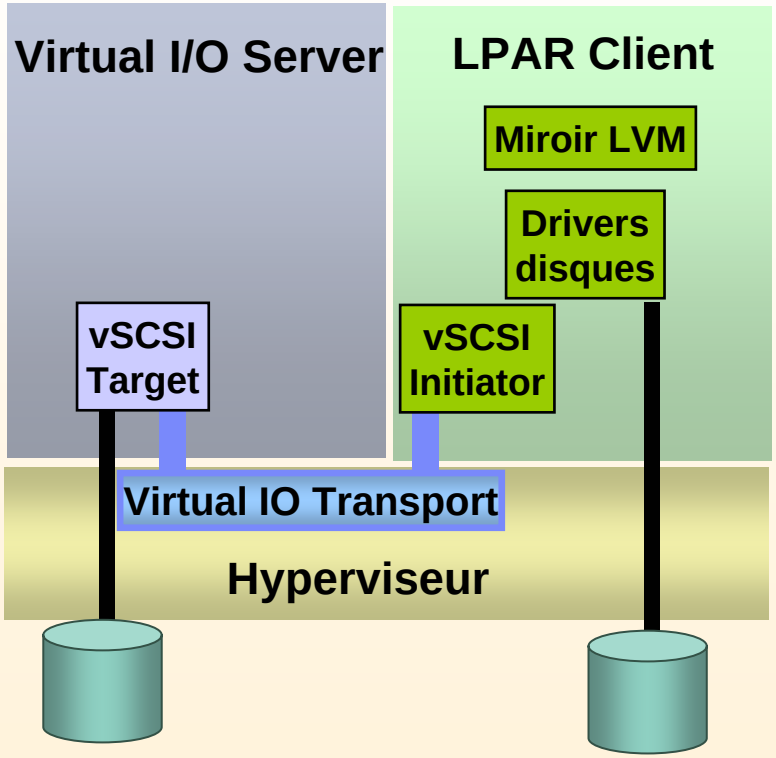
Sécurisation par le client

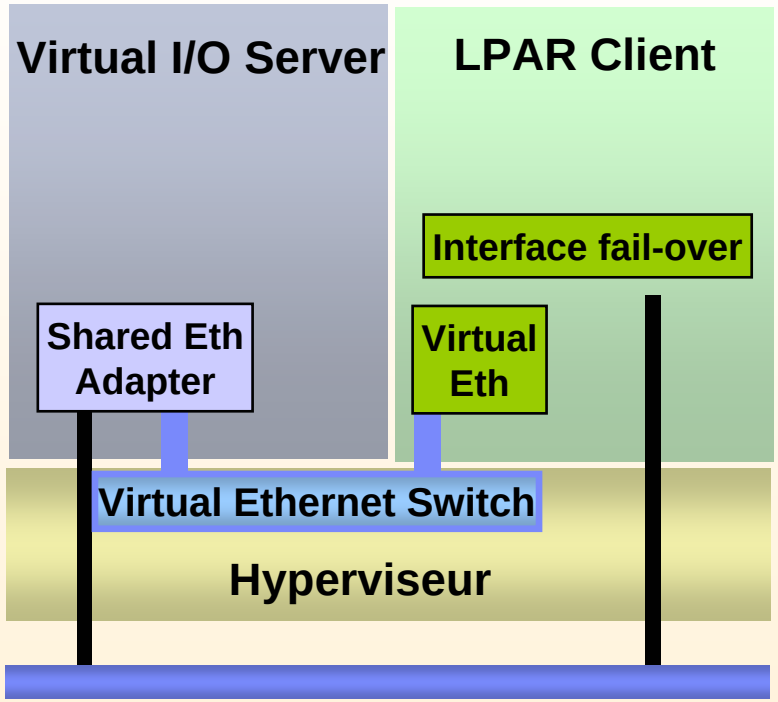


Protection contre un arrêt du VIOS
Sécurisation de l'adaptateur et/ou du disque effectuée par la partition cliente (LVM, NIB ...)

VIOS: Les configurations sécurisées supportées

Virtuel /réel

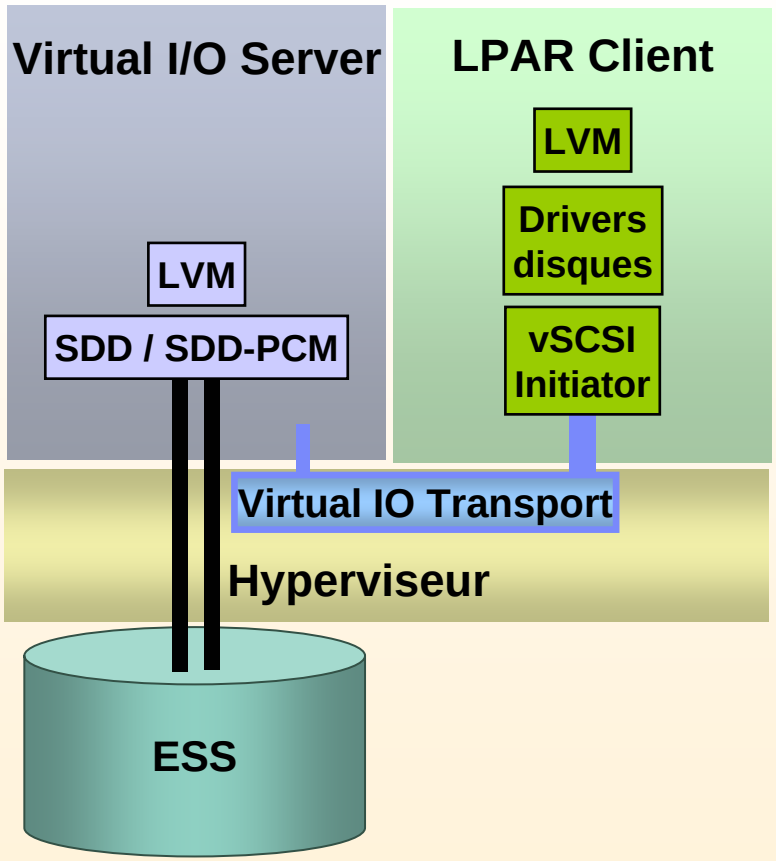


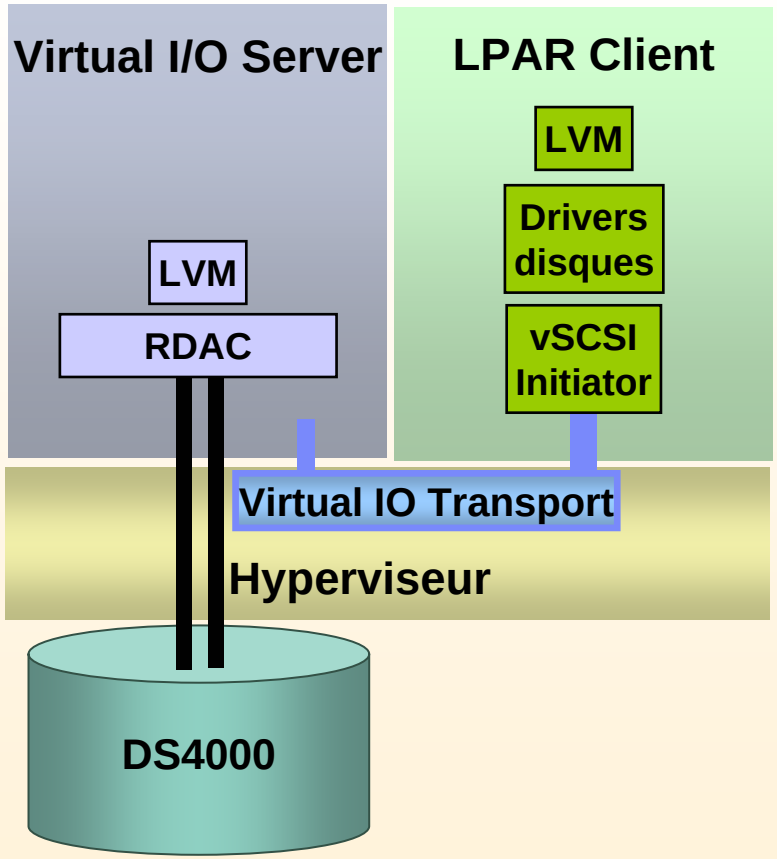


Protection contre un arrêt du VIOS
Sécurisation de l'adaptateur et/ou du disque effectuée par la partition cliente (LVM, NIB ...)

VIOS: Les configurations IBM sécurisées supportées

Sécurisation des adaptateurs du VIOS

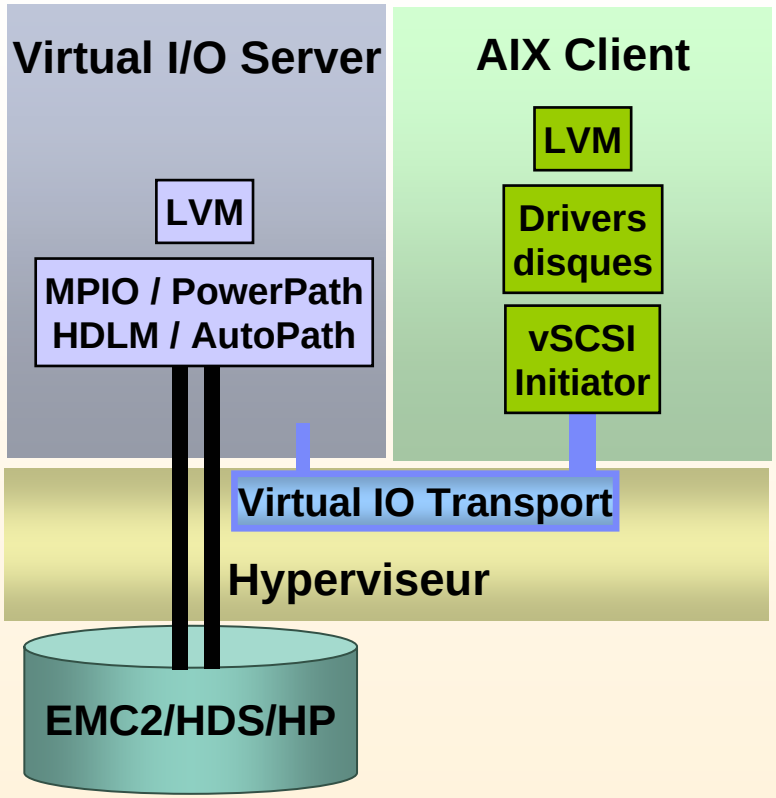




ESS: SDD ou SDD-PCM
DS4000: RDAC

VIOS: Baies EMC2/HDS/HP sécurisées supportées

Sécurisation des adaptateurs du VIOS



Précautions

- Consulter les CSA
- Pas de reprise de disques existants
- Client Linux, HACMP et GPFS exclus
- EMC2 demande un RPQ

Client

- AIX 5.3 ML1 ou +
  - patch IY70082, IY70148, IY70336
- Attachement simultané Direct et via VIOS possible
- boot SAN possible
  - de préférence vSCSI dédié

VIO Server

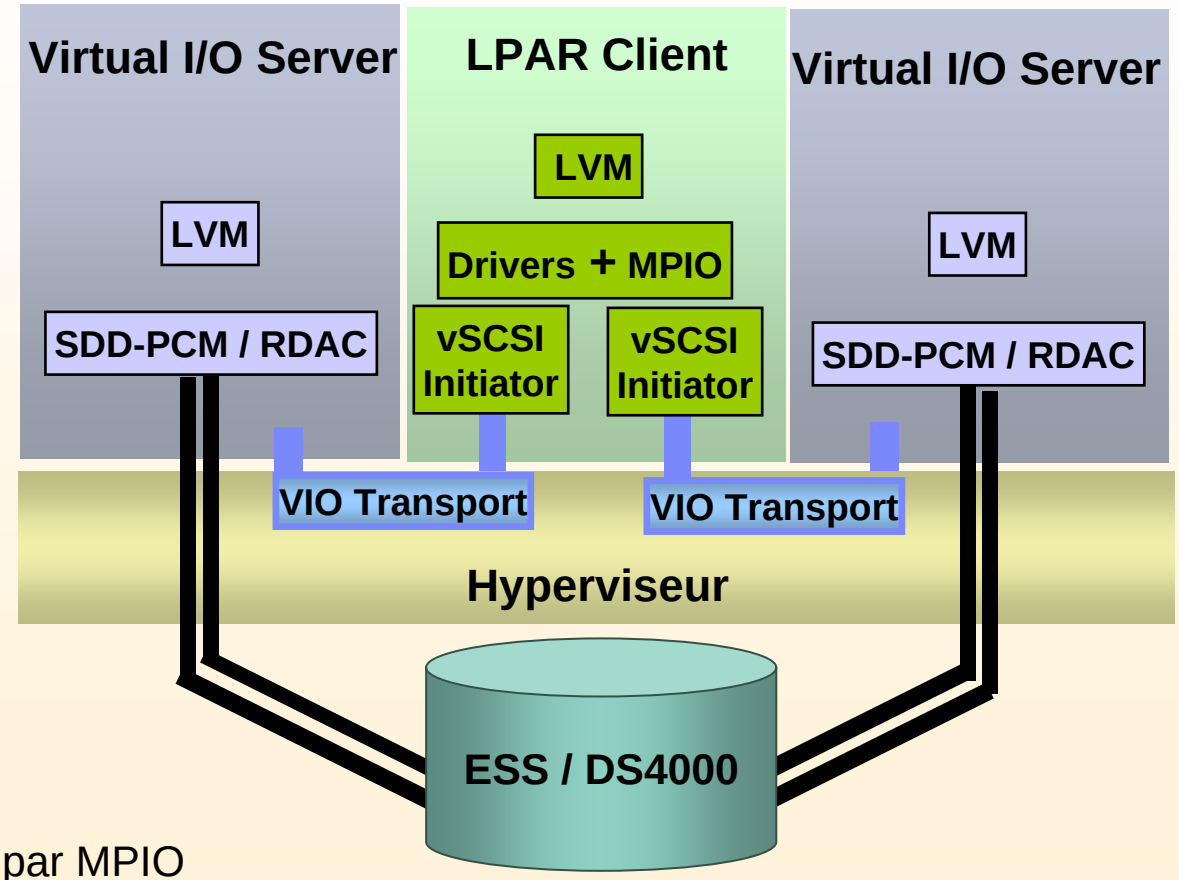
- FixPack 6.2 ou +
- installer l'ODM fournie par EMC2/HDS/HP
- PowerPath 4.4/ AutoPAth 4.4.1/ HDLM 5.4.2/ MPIO
  - MPIO implique UDID (Unique Device Identifier)

Modèles supportés

- Symmetrix série 8000, DMX 800, 1000, 2000, 3000
- Avec PP: CLARiiON CX300, 400, 500, 600, 700
- HDS 9910, 9960, 9970, USP 100, 600, 1100
- HP XP48, 128, 512, 1024

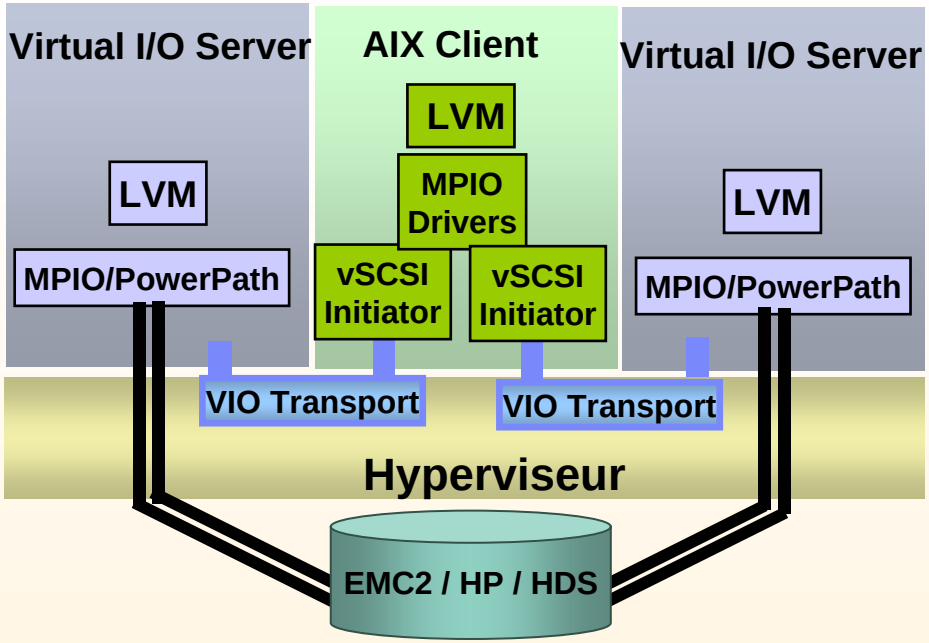
VIOS: Les configurations IBM sécurisées supportées

Sécurisation du VIOS et des adaptateurs



Client: Protection par MPIO
VIOS: protection par SDD-PCM sur ESS et par RDAC sur DS4000

VIOS: Baies EMC2/HDS/HP sécurisées possibles



Client

- AIX 5.3 ML1 ou +
  - patch IY70082, IY70148, IY70336
- Attachement simultané Direct et via VIOS possible
- Exclure rootvg (pas de SAN boot)

VIO Server

- FixPack 6.2 ou +
- installer l'ODM fournie par le constructeur
- Autopath 4.4.1/ HDLM 5.4.2 / PowerPath 4.4/ MPIO
  - MPIO implique UDID (Unique Device Identifier)
- MPIO: fail-over et load-balancing possible
- chdev no_reservation sur la LUN

Précautions

- Consulter les CSA
- Pas de reprise de disques existants
- Client Linux, HACMP et GPFS exclus
- EMC2 demande un RPQ

Modèles supportés

- Symmetrix série 8000, DMX 800, 1000, 2000, 3000
- Avec PP: CLARiiON CX300, 400, 500, 600, 700
- HDS 9910, 9960, 9970, USP 100, 600, 1100
- HP XP48,512, 128, 1024

Virtual I/O Server : Hints and Tips (en vrac)

- **Le VIO serveur est une partition qui a besoin de ressources : cpu, mémoire, disque**
- **La virtualisation d'I/O intensives peut être très consommateur de ressources :**
  Privilégier les accès directs pour les grands débits I/O (Gigabit Ethernet, FC)
- **On peut doubler les VIOS pour protéger les accès disque et réseau**
- **L'installation peut se faire par CDROM ou en utilisant NIMOL (à partir de la console HMC)**
- **Utiliser les Redbook**
  sg247940: Advanced POWER Virtualization on IBM p5 Servers: Introduction and Basic Configuration

  Sg245768: Virtualization Performance Considération

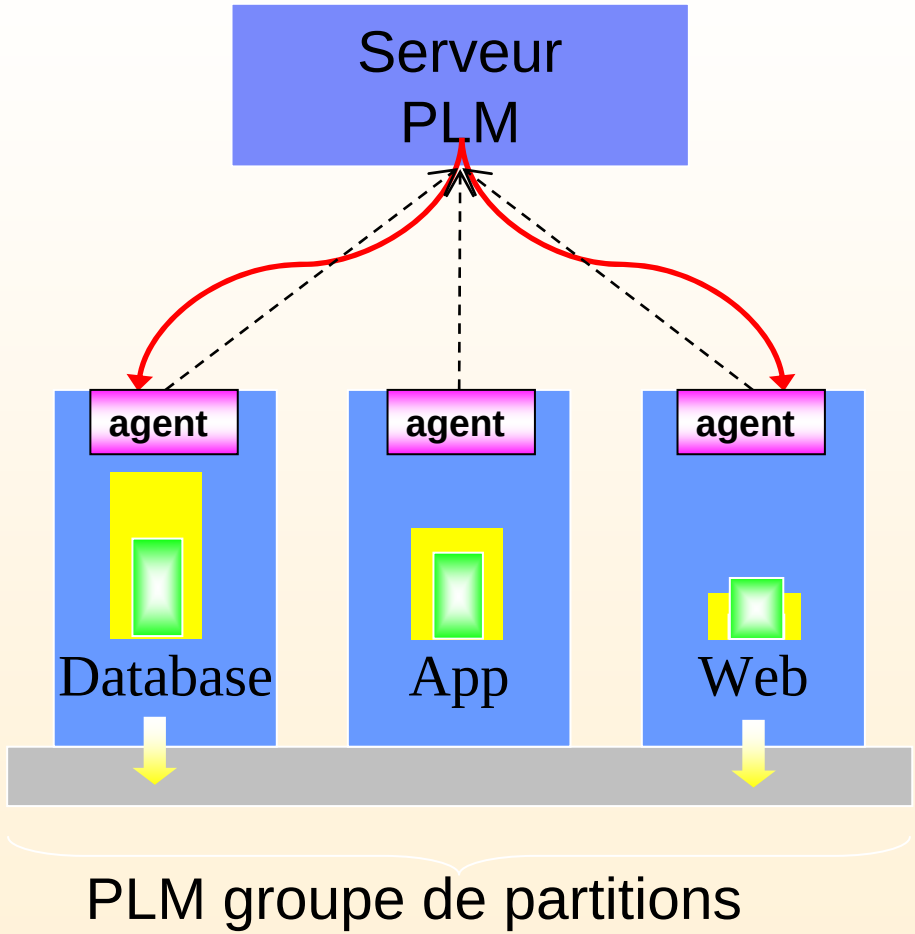
Fonctions complémentaires

PLM - Introduction

- Partition Load Manager (PLM) gère automatiquement la répartition des ressources mémoire et processeurs entre des partitions AIX.
- PLM fait partie de la fonction Advanced POWER Virtualization des serveurs Power5.
- PLM supporte les partitions à processeurs dédiés ou partagés.

  Tous les processeurs dans un groupe de partitions gérées par PLM doivent être de même type : dédiés ou partagés.

Partition Load Manager



Capacité d'Exécution

Utilisation

Note: dans l'exemple les partitions sont *bridées*

Advance Accounting

- Définir des classes de mesure
- Classifier les charges de travail
- Mesurer l'utilisation des ressources pour chacune de ces classes
- Enregistrer les utilisations de ressources
- Deux scénarios

  Basé sur les applications (projet)

  Basé sur les partitions (LPAR)

Facturation basée sur les applications

Pour une facturation basée sur un ensemble d'applications (projet)
Plusieurs classes de facturation dans une même instance d'OS

Classe A Programmes	Classe B Programmes	Classe C Programmes	Classe D Programmes

AIX
- Les programmes utilisateurs sont placés dans des classes
- Les règles de classification sont basées sur :
  - Le nom des applications et/ou
  - Le nom des utilisateurs et/ou
  - Le nom des groupes
- Le système mesure les consommations de ressources
- Ces mesures sont enregistrées dans des fichiers

Système de facturation transparent pour les applications

Facturation basée sur les partitions

Pour une facturation basée les partitions
Facturation par instance d'OS

Classe A	Classe B	Classe C	Classe D
LPAR A	LPAR B	LPAR C	LPAR D
AIX	AIX	AIX	AIX

- Le système mesure les consommations de ressources (cpu par exemple) dans chacune des partitions
- Ces mesures sont enregistrées dans des fichiers

Système de facturation transparent pour les applications