

Cloud & Architectures

Partie 1



Partie 1 : Virtualisation

- Présentation
- Virtualisation : quelques définitions
- Exemple d'hyperviseur : VMware vSphere
- Autres fonctionnalités très utiles pour nos architecture
- Concrètement ?
- SLA : ce qui guide les choix d'architecture
- Virtualisation au sens large
- Quelles réponses aux besoins usuels d'une infra ?

Virtualisation : quelques définitions

...

Technologie de virtualisation

- Ensemble de techniques et d'outils permettant de faire tourner plusieurs systèmes d'exploitation sur un serveur
- Partage de ressources
- En respectant deux principes fondamentaux :
 - Le cloisonnement : chaque système d'exploitation à un fonctionnement indépendant sans aucune interférence mutuelle
 - La transparence : le fonctionnement en mode virtualisé ne modifie pas le fonctionnement du système ni des applications

Technologie de virtualisation

■ Intérêts :

- Economique : mutualisation du matériel, bénéfice en terme de coût d'acquisition, de possession (rack, électricité, climatisation, réseau) et d'exploitation
- Facilité d'administration : installation, déploiement et migration aisées des machines virtuelles entre serveurs physiques, simulation d'environnements de qualification ou de pré-production, création de plateforme de tests ou de développements réutilisables à volonté
- Sécurisation : séparation des systèmes virtuels et hôtes (invisibles), répartition des utilisateurs, allocation dynamique des ressources, dimensionnement des serveurs facilités

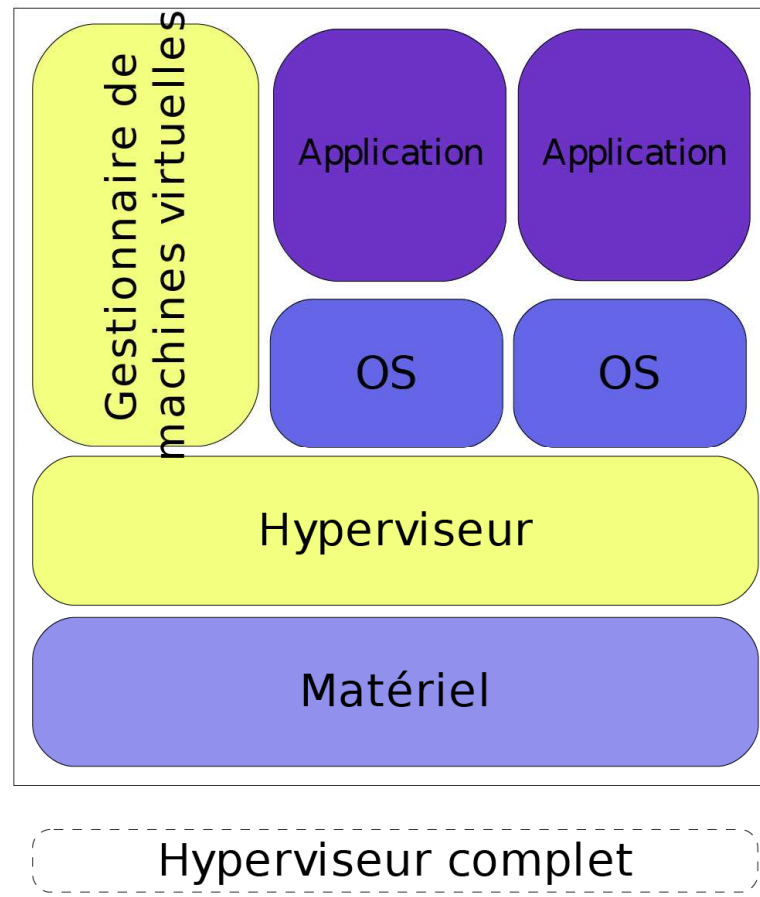
Technologie de virtualisation

- Différentes techniques :
 - **Type 1 / Hyperviseur complet / bare-metal** : Utilisation d'un noyau hôte léger permettant de faire tourner des systèmes d'exploitations natifs
 - **Type 2 / Hosted** : Utilisation d'un logiciel. Emulation partielle ou totale d'une machine
 - **Paravirtualiseur** : Utilisation d'un noyau hôte allégé permettant de faire tourner des systèmes d'exploitations invités, adaptés et optimisés
 - **Isolation** : Séparation forte entre différents contextes logiciels sur un même noyau de systèmes d'exploitation

Technologie de virtualisation

■ Hyperviseur complet :

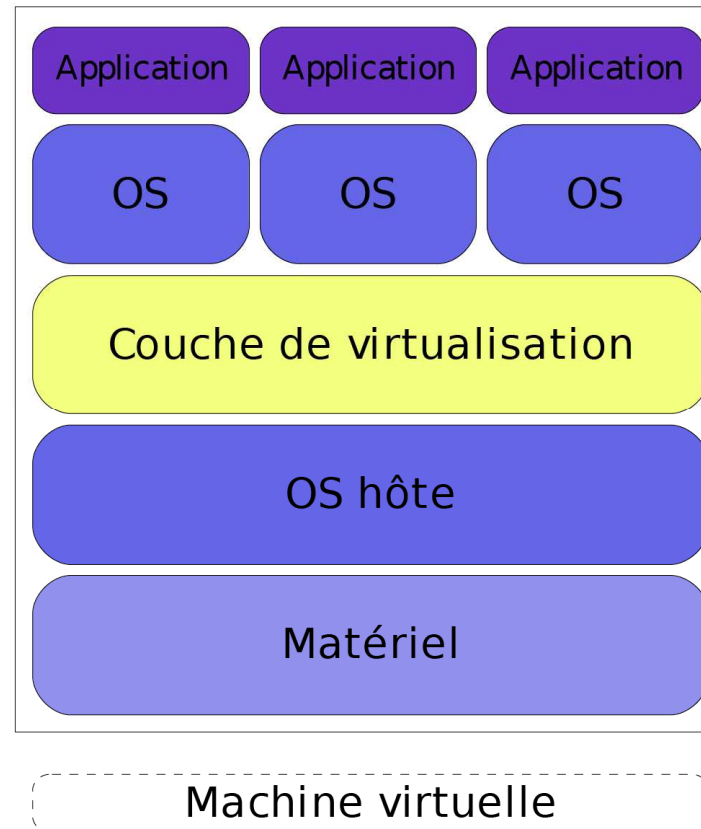
- ☐ Utilisation d'un micro-noyau
- ☐ Outils de supervision
- ☐ Emulation des I/O
- ☐ Instructions spécifiques
- ☐ Exemples :
 - QEMU
 - KVM
 - VMWare ESXi
 - XenServer



Technologie de virtualisation

■ Type 2 / Hosted :

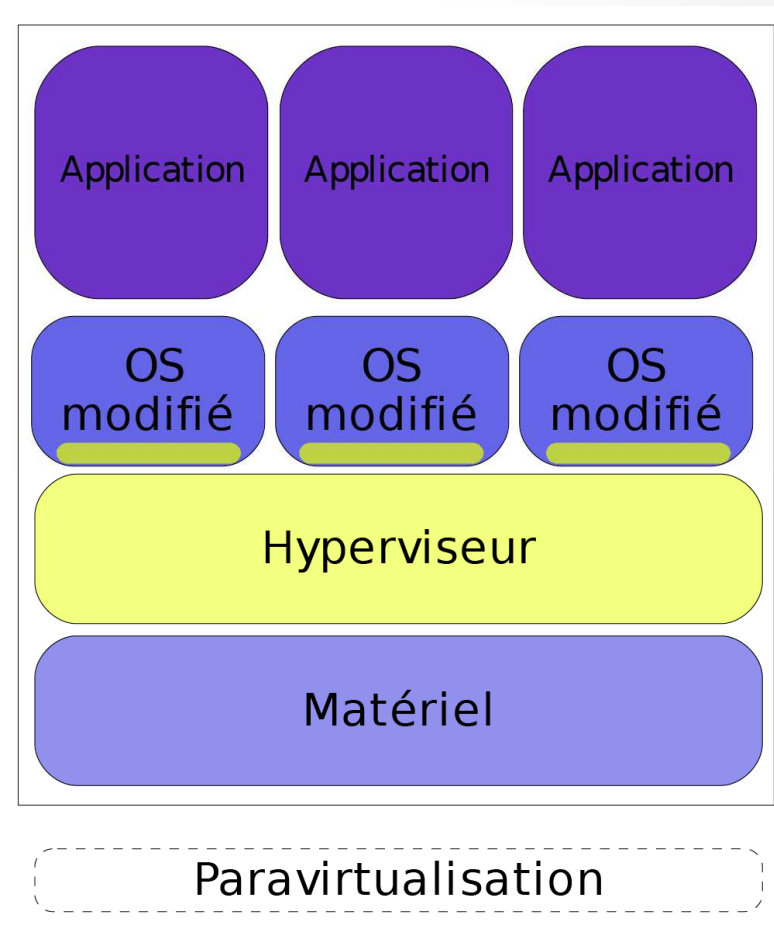
- Emulation logicielle
- Bonne isolation
- Coût en performance
- Exemples :
 - Qemu
 - VMWare
 - VirtualPC
 - VirtualBox



Technologie de virtualisation

■ Paravirtualiseur :

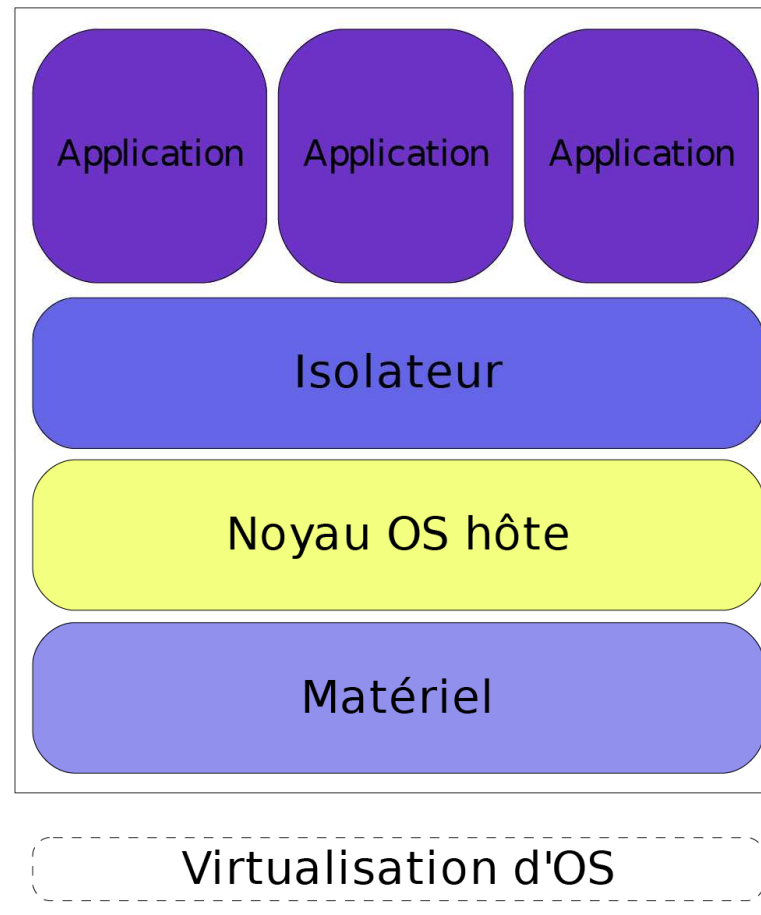
- Micro-noyau hôte optimisé
- OS invités adaptés et optimisés
- Sans instructions spécifiques
- Exemples :
 - XEN
 - KVM (avec Virtio)
 - VMWare ESXi (drivers paravirt)
 - Microsoft Hyper-V Server
 - Oracle VM



Technologie de virtualisation

■ Isolateur :

- ☐ Séparation en contextes
- ☐ Régi par l'OS hôte
- ☐ Mais cloisonnés
- ☐ Un seul noyau
- ☐ N espaces utilisateurs
- ☐ Solution très légère
- ☐ Exemples
 - Linux-VServer
 - BSD Jail
 - OpenVZ
 - LXC



Exemple d'hyperviseur : VMware vSphere

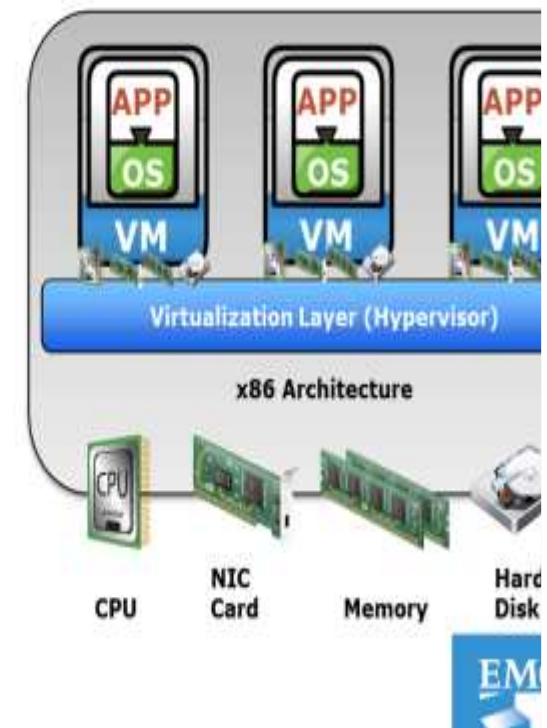
...

Compute Virtualization

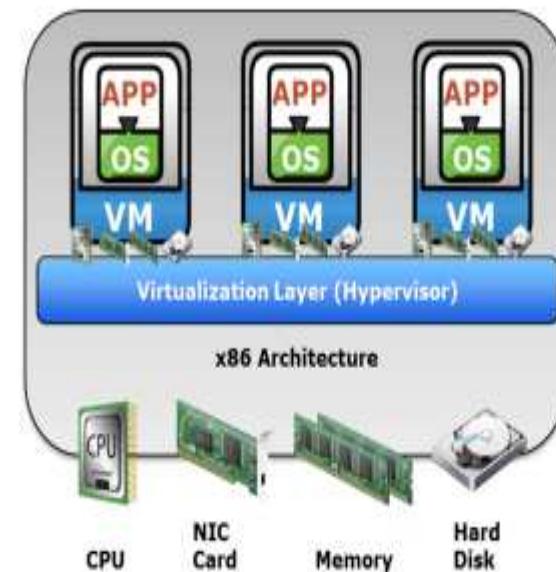
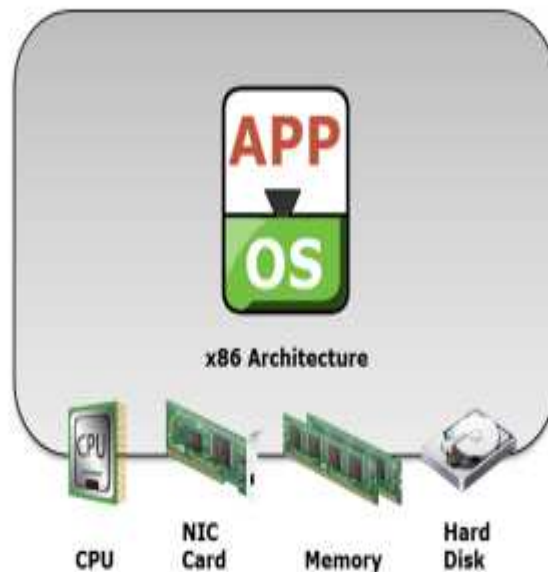
Compute Virtualization

It is a technique of masking or abstracting the physical compute hardware and enabling multiple operating systems (OSs) to run concurrently on a single or clustered physical machine(s).

- Enables creation of multiple virtual machines (VMs), each running an OS and application
 - VM is a logical entity that looks and behaves like physical machine
- Virtualization layer resides between hardware and VMs
 - Also known as hypervisor
- VMs are provided with standardized hardware resources



Need for Compute Virtualization



Before Virtualization	After Virtualization
<ul style="list-style-type: none">• Runs single OS per machine at a time	<ul style="list-style-type: none">• Runs multiple OSs per physical machine concurrently
<ul style="list-style-type: none">• Couples s/w and h/w tightly	<ul style="list-style-type: none">• Makes OS and applications h/w independent
<ul style="list-style-type: none">• May create conflicts when multiple applications run on the same machine	<ul style="list-style-type: none">• Isolates VM from each other, hence, no conflict
<ul style="list-style-type: none">• Underutilizes resources	<ul style="list-style-type: none">• Improves resource utilization
<ul style="list-style-type: none">• Is inflexible and expensive	<ul style="list-style-type: none">• Offers flexible infrastructure at low cost

Serveur physique vs Virtualisation

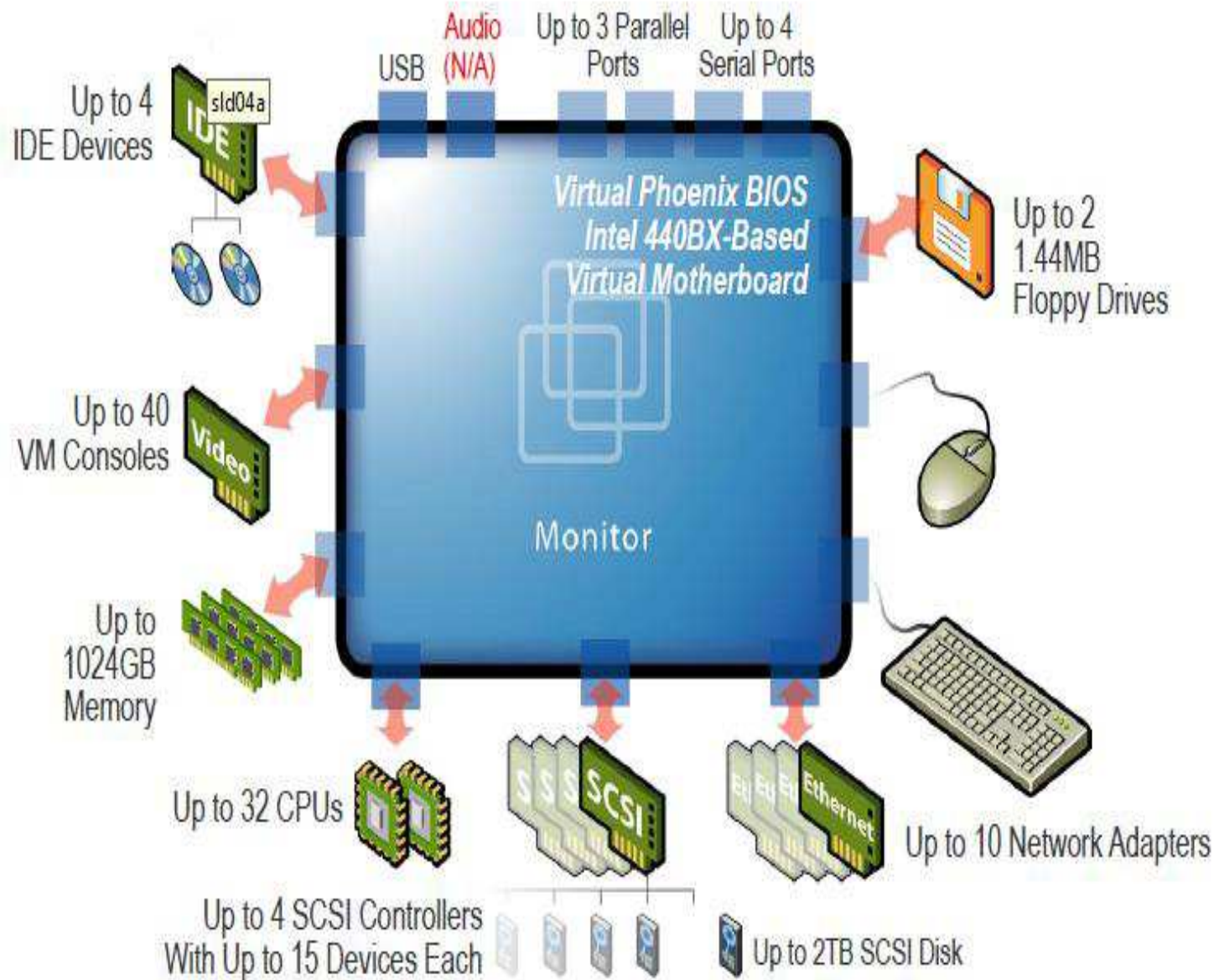


Traditional Architecture



Virtual Architecture

Virtual Motherboard of a VM



Key VMware vSphere Features

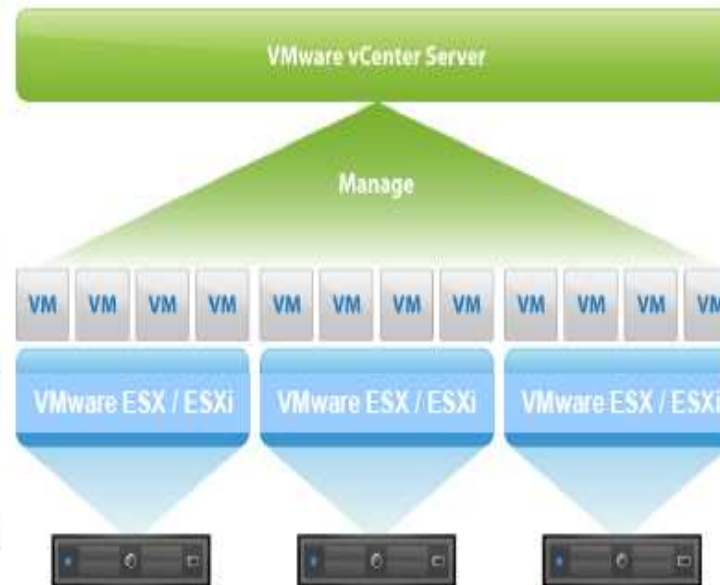
Hot Add Virtual Devices

- Hot add
 - CPU
 - Memory
- Hot add or remove
 - Storage devices
 - Network devices



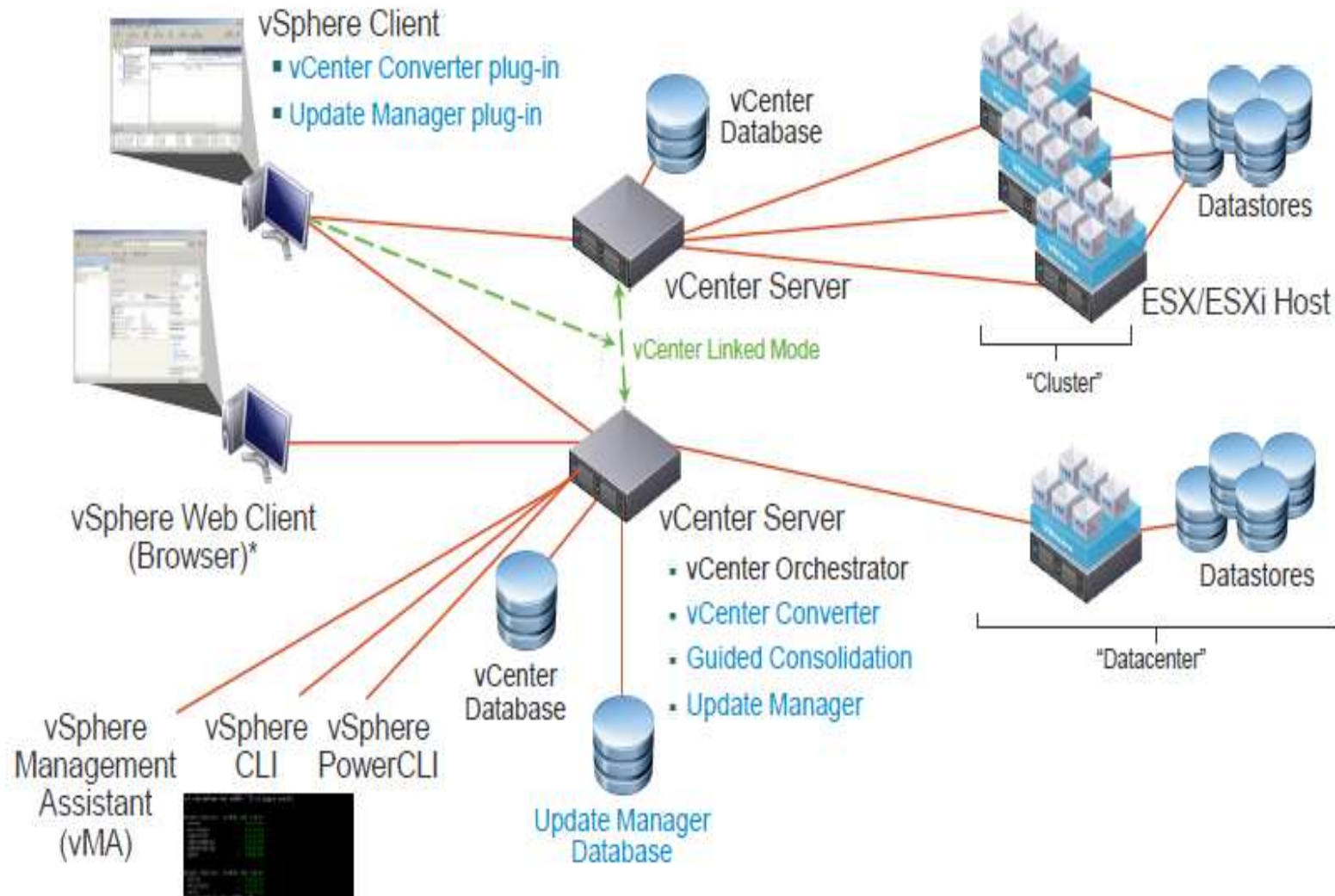
vSphere Components

- **vSphere** is a distributed software system with features enabled by a management server and hypervisor working together
- **VMware vCenter Server** is the management server and cluster coordinator (1 instance required)
- **VMware ESXi** is the hypervisor software running on each host (1 instance per server required)



vSphere editions are the licensed features enabled by vCenter Server working with ESXi

VMware vSphere Architecture

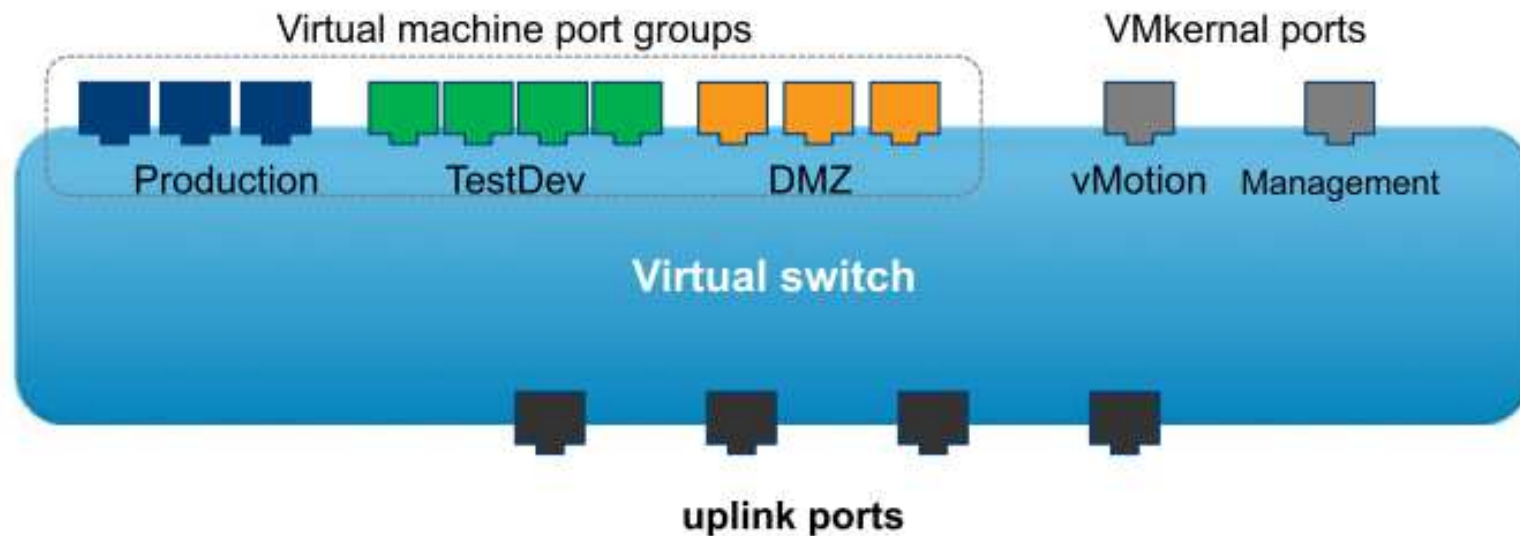


Types of Virtual Switch Connections

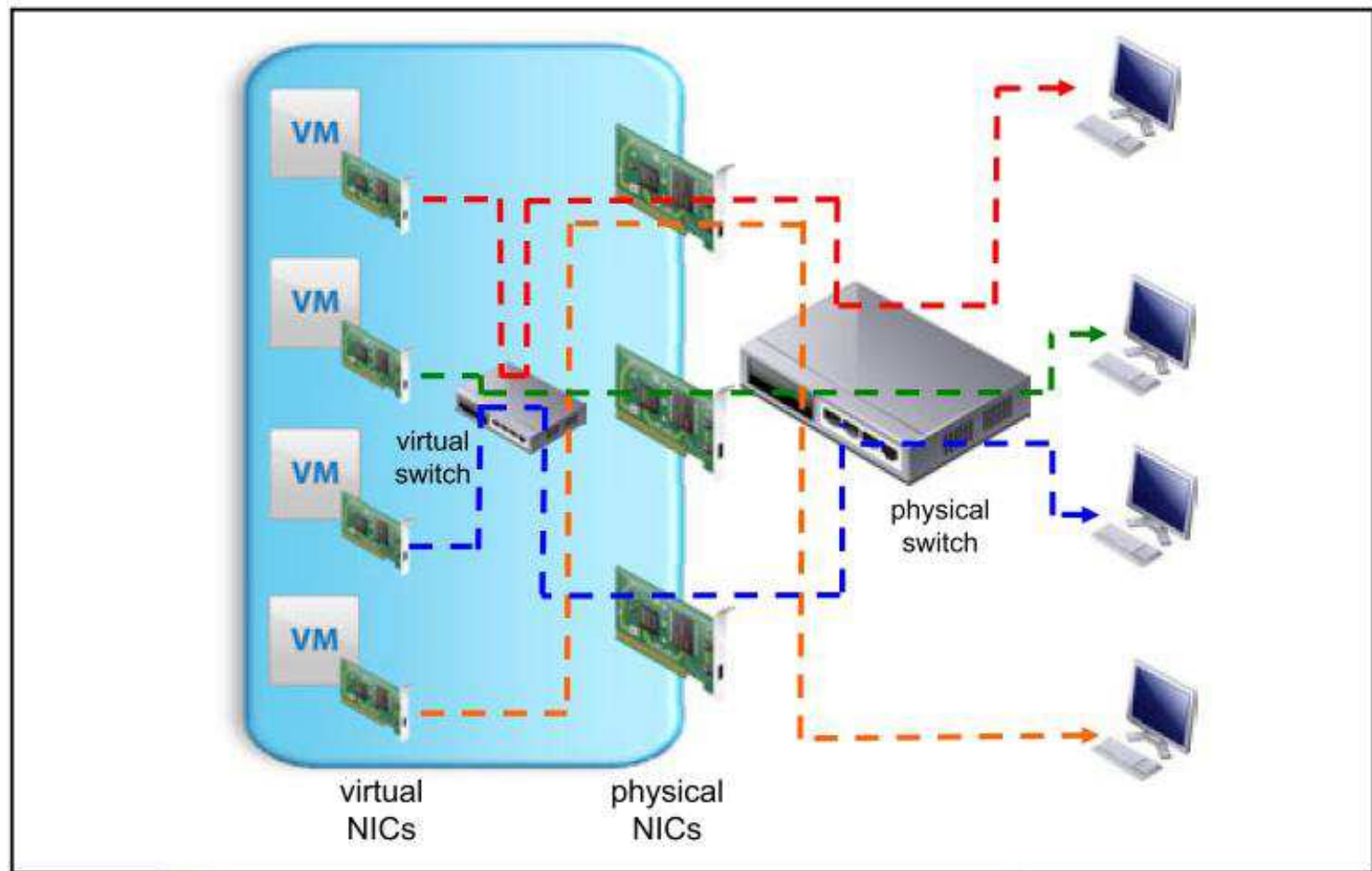
Slide 5-7

A virtual switch allows the following connection types:

- Virtual machine port groups
- VMkernel port:
 - For IP storage, vSphere vMotion migration, VMware vSphere® Fault Tolerance
 - For the ESXi management network

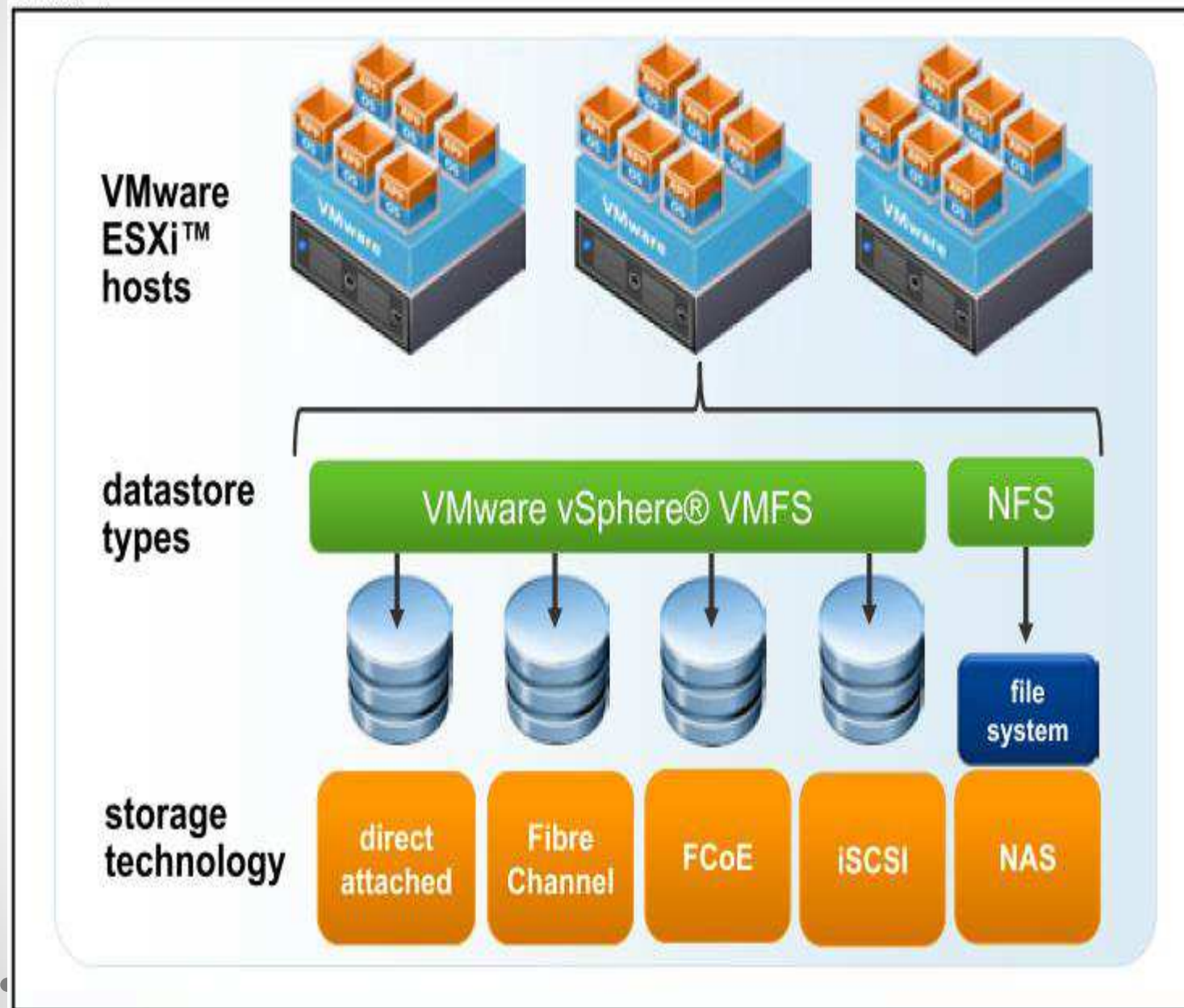


Virtual switch / physical switch



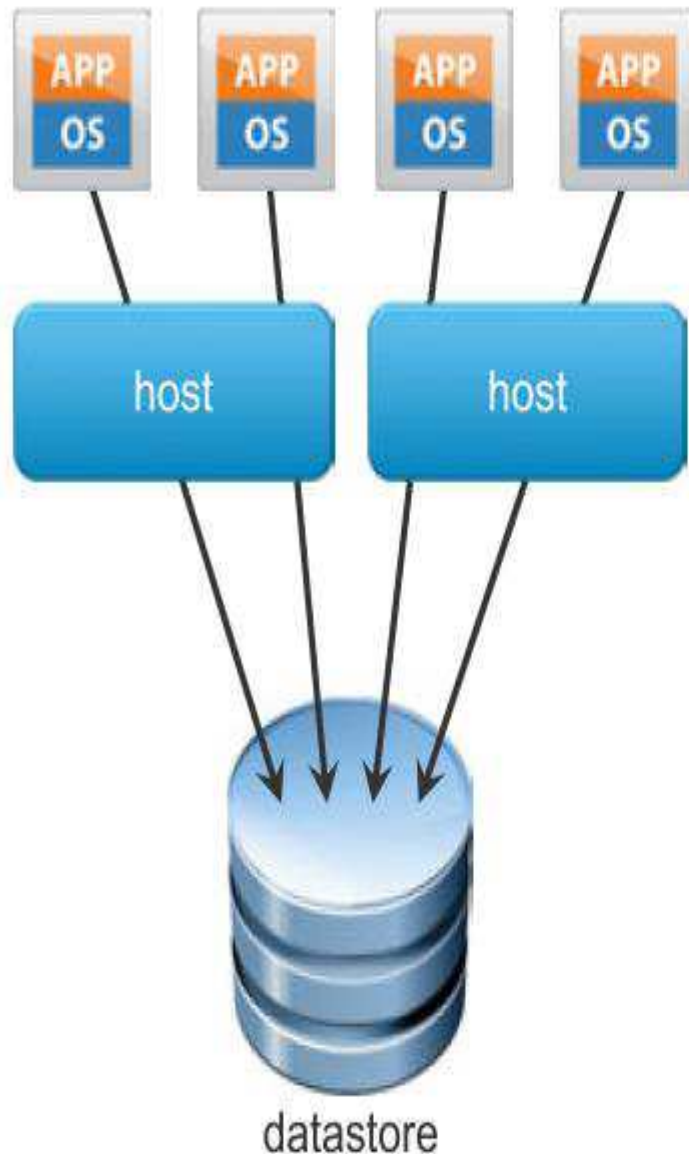
Storage Overview

Slide 6-7



Datastore

Slide 6-9



A datastore is a logical storage unit that can use disk space on one physical device or span several physical devices.

Types of datastores:

- VMFS
- NFS

Datastores are used to hold virtual machine files, templates, and ISO images.

Hard Disk Options

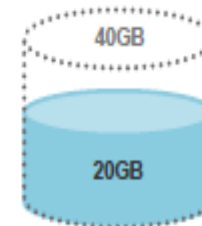
Thick

- Pre-allocated disk space
- Physical disk size = virtual disk size



Thin

- VM sees full logical disk at all times
- Physical disk size = used disk size
- Physical disk size grows as used

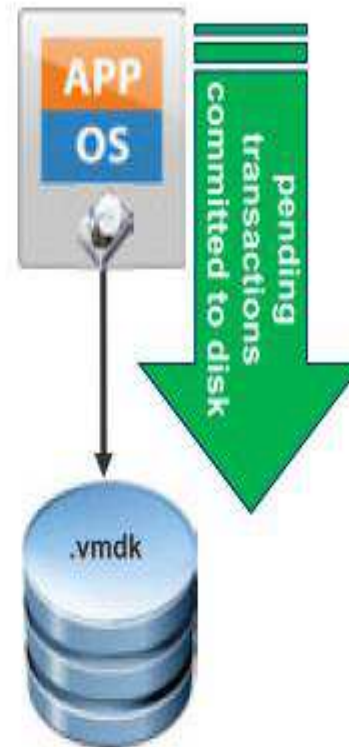


Autres fonctionnalités très
utiles pour nos architecture

...

Taking a Snapshot

Slide 7-53



You can take a snapshot while a virtual machine is powered on, powered off, or suspended.

A snapshot captures the state of the virtual machine:

- Memory state, settings state, and disk state

Snapshots are not backups.

High Availability

Slide 10-9

A highly available system is one that is continuously operational for an optimal length of time.

Which level of virtual machine availability is important to you?

Level of Availability	Downtime Per Year
99%	87 hours (3.5 days)
99.9%	8.76 hours
99.99%	52 minutes
99.999%	5 minutes

vMotion™

Description:

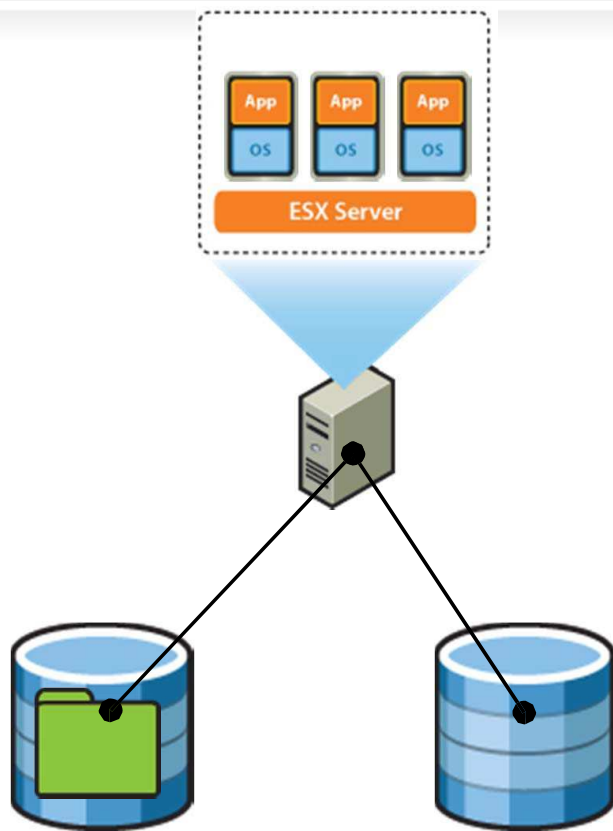
Enables the live migration of virtual machines
From one host to another with continuous
Service availability.

Benefits:

- Revolutionary technology that is the basis for automated virtual machine movement
- Meets service level and performance goals



Storage vMotion – Live Migration Extend to storage



- > Live migration of VMs across Storage disks with no downtime
- > Minimizes planned downtime

Comparison of Migration Types

Slide 7-32

Migration Type	Virtual Machine Power State	Change Host or Datastore?	Across Virtual Data Centers?	Shared Storage Required?	CPU Compatibility
Cold	Off	Host or datastore or both	Yes	No	Different CPU families allowed
Suspended	Suspended	Host or datastore or both	Yes	No	Must meet CPU compatibility requirements
vMotion	On	Host	No	Yes	Must meet CPU compatibility requirements
Storage vMotion	On	Datastore	No	No	N/A
Enhanced vMotion	On	Both	No	No	Must meet CPU compatibility requirements

VMware High Availability

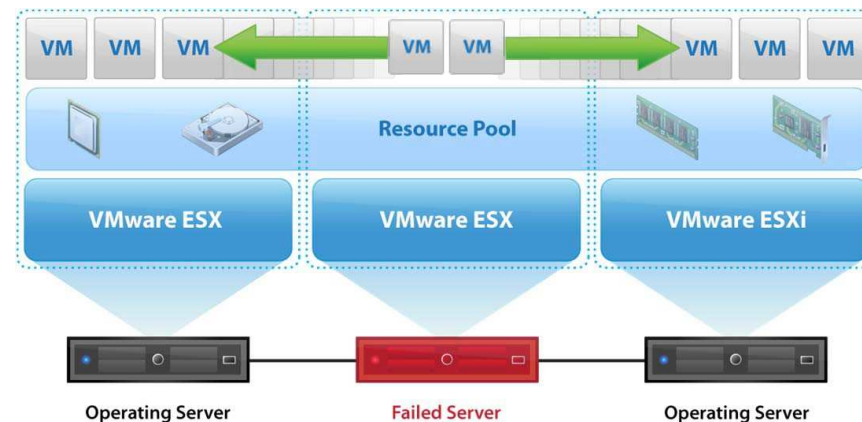
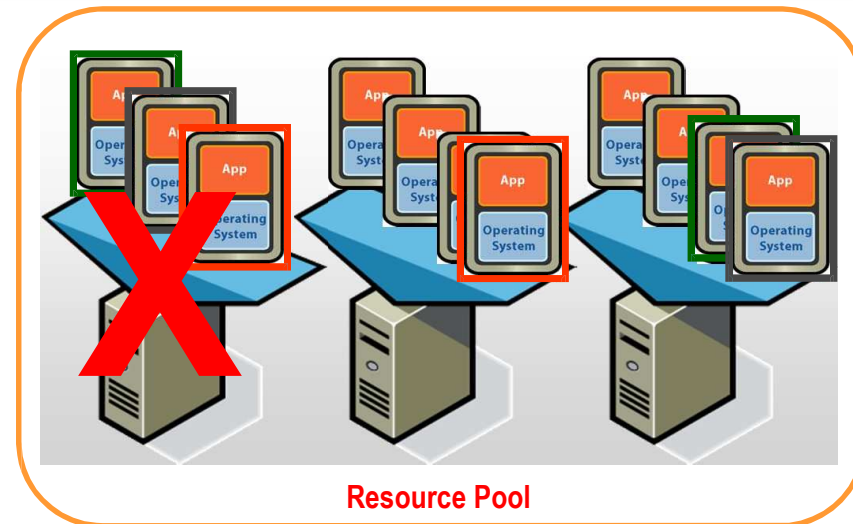
Description:

Enables the high availability of virtual machines by restarting them on a different vSphere host in the event of a failure.

*Automatic restart of Virtual Machine on host failure.

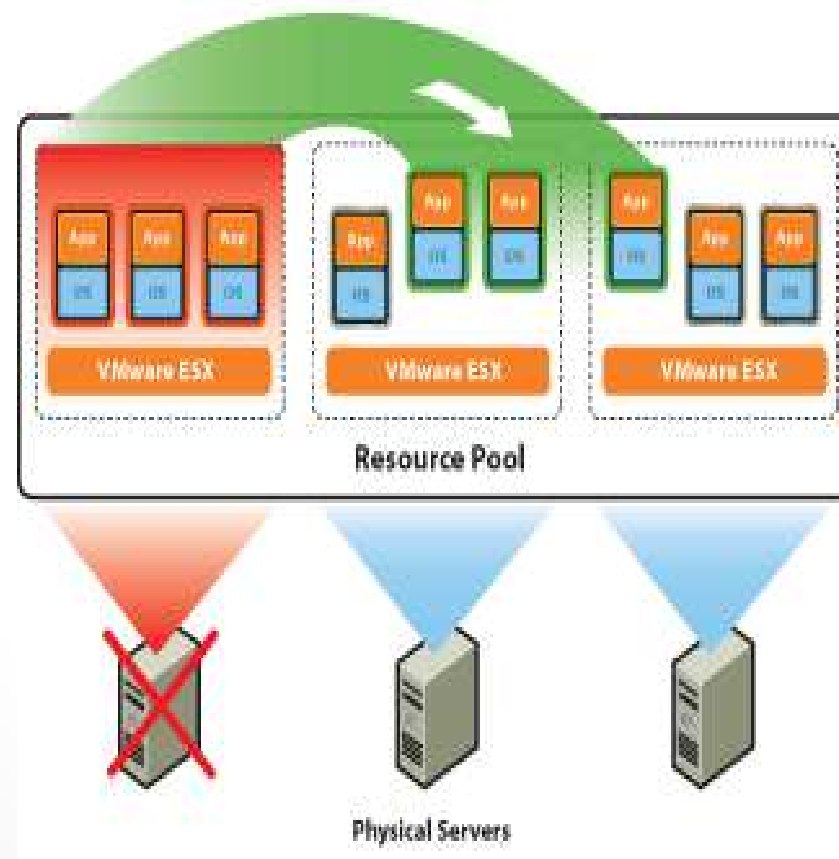
Benefits:

- Minimizes downtime and IT service disruption
- Reduce cost and complexity compared to traditional clustering



Vmware « HA »

protection contre les pannes hardware



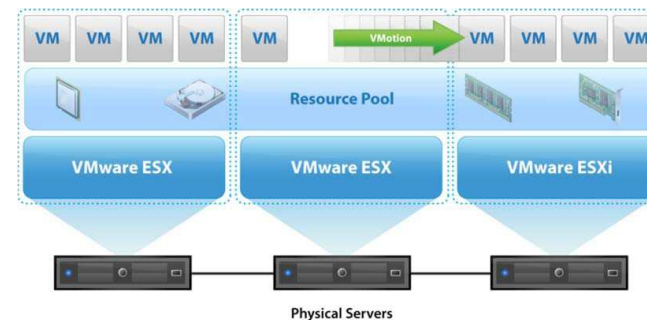
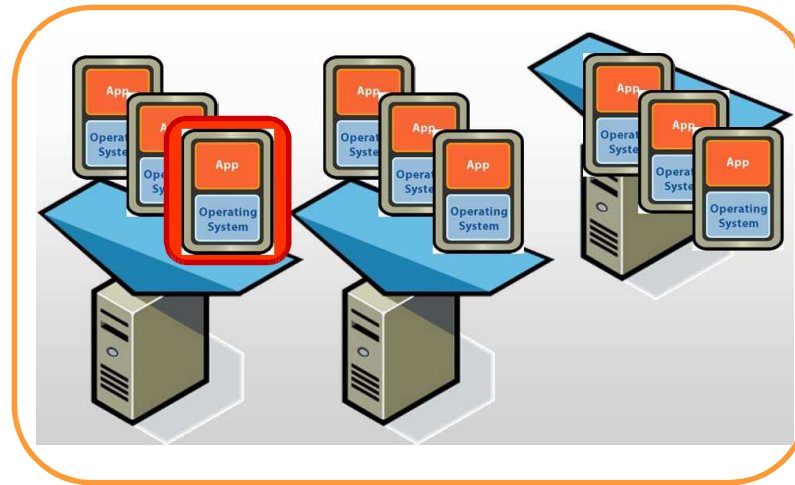
VMware Distributed Resource Scheduler (DRS)

Description:

Dynamically allocates and balances virtual machines to guarantee optimal access to resources

Benefits:

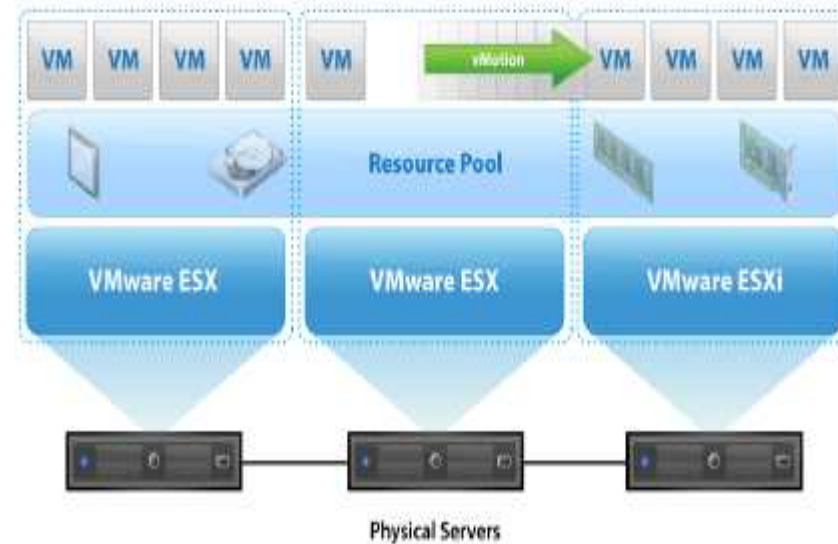
- Align resources to meet business goals
- Increase system administrator productivity
- Automate hardware maintenance
- Minimizes power consumption while guaranteeing service levels (DMP)



Key VMware vSphere Features

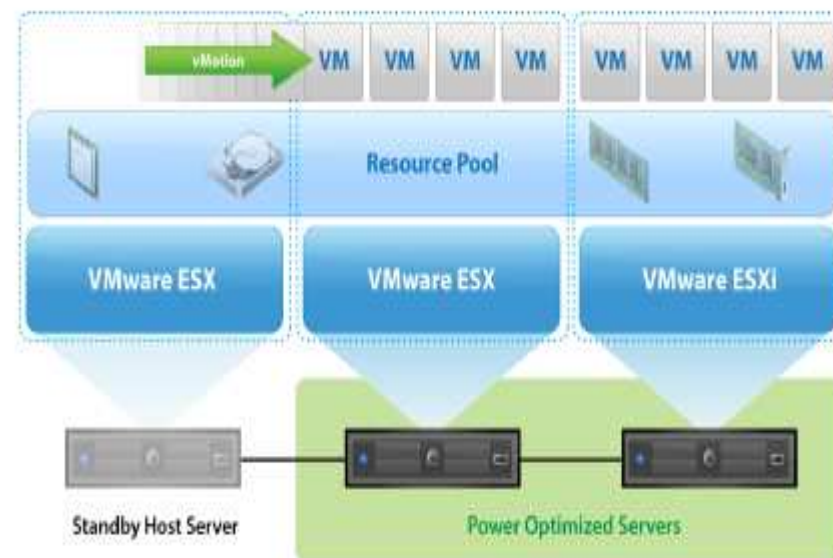
DRS

- Automated load balancing



DPM

- Optionally consolidate VMs onto fewer hosts and power off/on hosts as needed

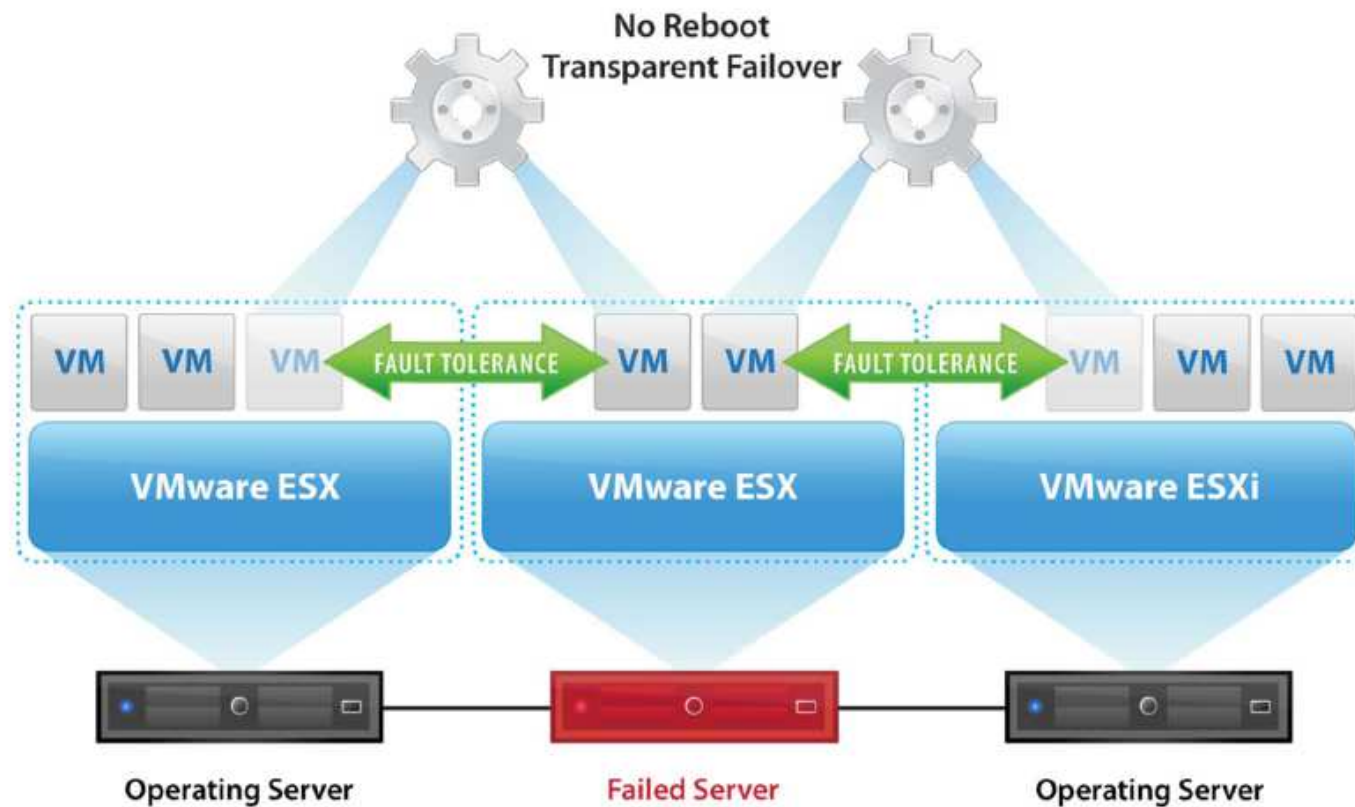


Affinité et anti-affinité

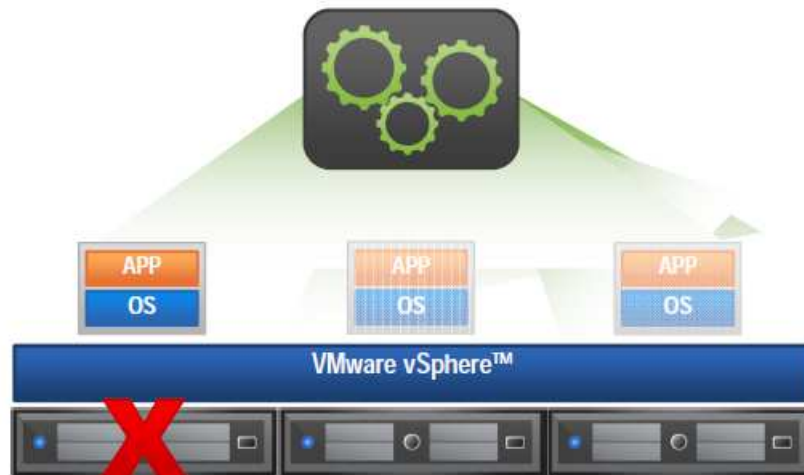
- Au niveau de l'exécution des VMs, possibilité de les regrouper sur un même ESX, ou au contraire de les dispatcher sur des ESX différents.
- Au niveau du stockage, possibilité de les regrouper sur un même datastore, ou au contraire de les dispatcher sur des datastores différents.

VMware Fault Tolerance

Near bumpless transfer of control on host failure



VMware Fault Tolerance [FT]



- ☐ Single identical VM's running in lockstep on separate hosts
- ☐ Zero data loss failover for all virtual machines in case of hardware failures
- ☐ Zero downtime, zero dataloss
- ☐ No complex clustering or specialized hardware required
- ☐ Single common mechanism for all applications and OS-es

Inherent Characteristics of Virtual Machines



Partitioning

- Run multiple operating systems on one physical machine
- Divide system resources between virtual machines



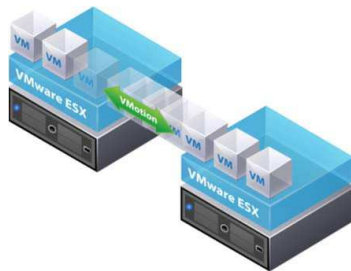
Isolation

- Fault and security isolation at the hardware level
- Advanced resource controls preserve performance



Encapsulation

- Entire state of the virtual machine can be saved to files
- Move and copy virtual machines as easily as files



Hardware Independence

- Provision or migrate any virtual machine to any similar or different physical server

Concrètement ?

...

Exemple d'Infra Vmware

en quelque chiffres

- Nombre de VMs = 1434
- Nombre de tenants = 313
- Nombre de vCPU = 2115
- Mémoire allouée (par les VMs) = 4829 Go
- Mémoire physique totale = 4096 Go
- Stockage alloué (par les VMs) = 94.5 To
- Stockage physique total = 55 To
- 8 serveurs Dell R630 de 512 Go de RAM + 2 NAS NetApp 2552 + réseaux Cisco en 10G
- Des étonnements ?

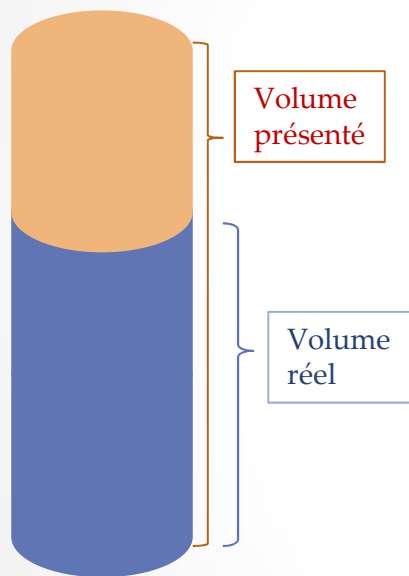
Rationalisation & optimisations

- Taux de consolidation :
nombre de VM par hyperviseur
- Over-provisioning :
 - Processeur : vCPU/pCPU = 4 (classiquement)
 - Mémoire : vRAM/pRAM peut être supérieur à 1
 - Stockage :

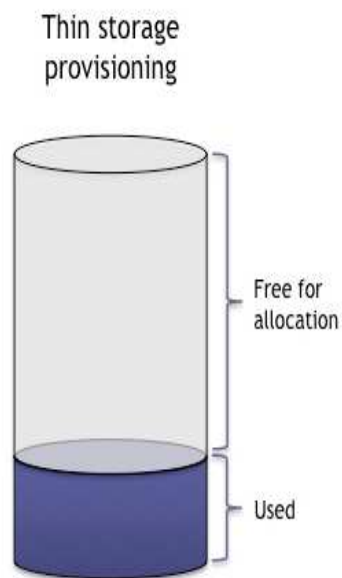
Attention :

- Tous les serveurs physiques ne peuvent pas supporter l'exécution de centaines de VMs !
- L'Over-provisionning doit être vraiment maîtrisé, sinon tout peut s'effondrer !

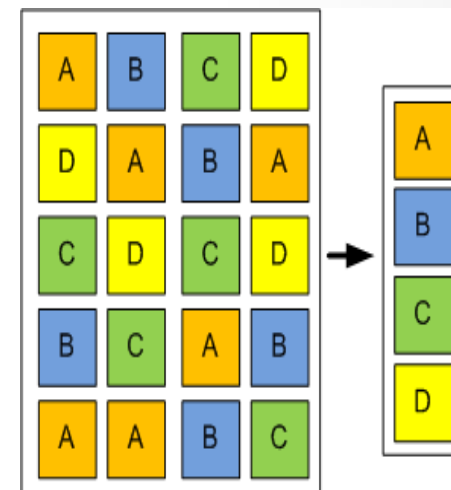
Mécanismes d'optimisation de la Virtualisation



Overprovisioning



Thin provisioning

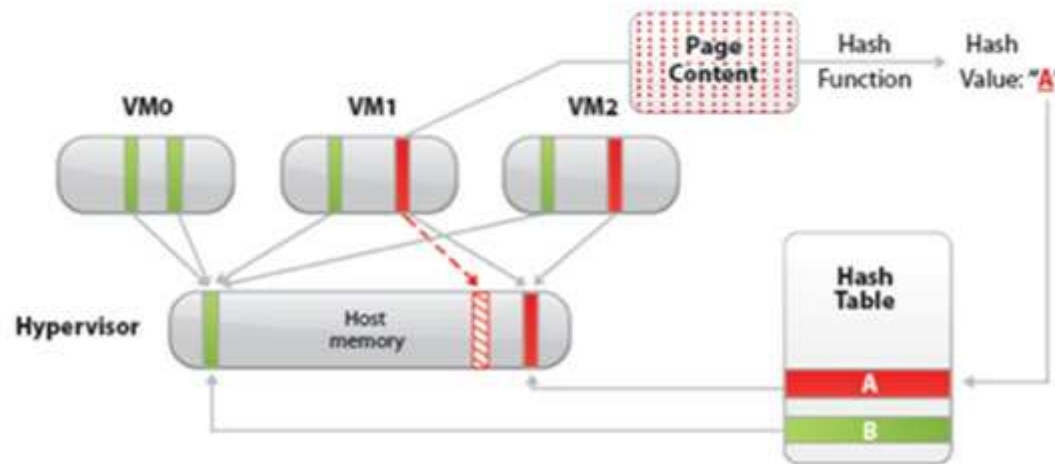


Deduplication

Déduplication de la mémoire (hyperviseur)

Transparent Page Sharing (TPS)

- An exclusive VMware memory reclamation technique where vSphere "reclaims memory by removing redundant memory pages with identical content"



D'autres systèmes similaires existent dans les autres hyperviseurs, comme KSM (Kernel Samepage Merging) pour KVM (Linux).

Noisy Neighbors

parce que les infra virtualisées sont mutualisées



SLA :
ce qui guide les choix
d'architecture

...

Service-level agreement

La disponibilité s'exprime souvent en pourcentage :

Disponibilité en %	Indispo par année	Indispo par mois	Indispo par semaine
95 %	18,25 jours	36 heures	8,4 heures
98 %	7,30 jours	14,4 heures	3,36 heures
99 % (« deux neuf »)	3,65 jours	7,20 heures	1,68 heures
99,5 %	1,83 jours	3,60 heures	50,4 minutes
99,8 %	17,52 heures	86,23 minutes	20,16 minutes
99,9 % (« trois neuf »)	8,76 heures	43,2 minutes	10,1 minutes
99,95 %	4,38 heures	21,56 minutes	5,04 minutes
99,99 % (« quatre neuf »)	52,56 minutes	4,32 minutes	1,01 minutes
99,999 % (« cinq neuf »)	5,26 minutes	25,9 secondes	6,05 secondes

SLA : paramètres à prendre en compte

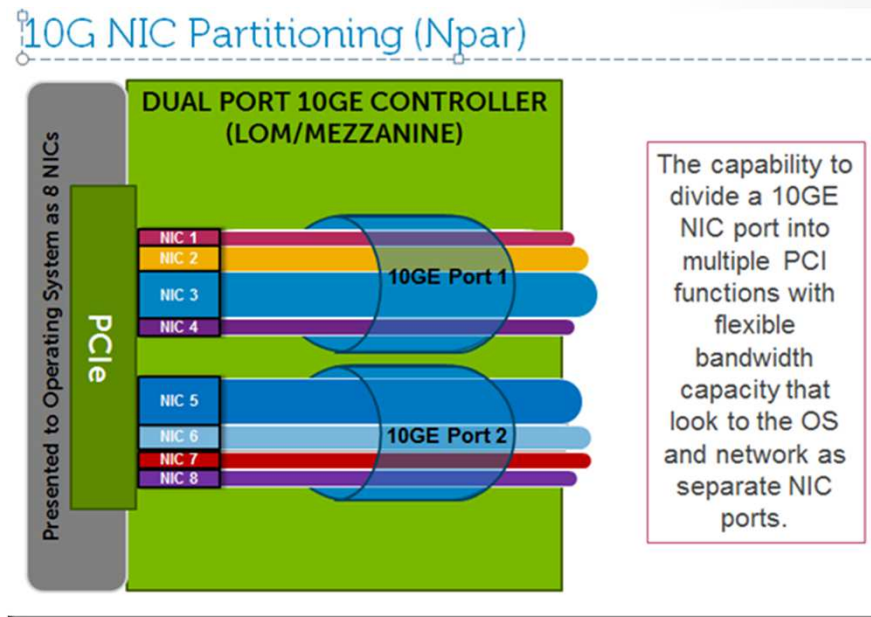
- GTD : Garantie de Temps de Disponibilité
- GTI : Garantie de Temps d'Intervention
- GTR : Garantie de Temps de Rétablissement
- RTO : Recovery Time Objective (retour du service)
- RPO : Recovery Point Objective (données perdues)

Virtualisation au sens large

...

Virtualisation du réseau

- VLAN 802.1 Q
- vSwitch
- vNIC
- Virtual Firewall
- 10G NIC partitioning



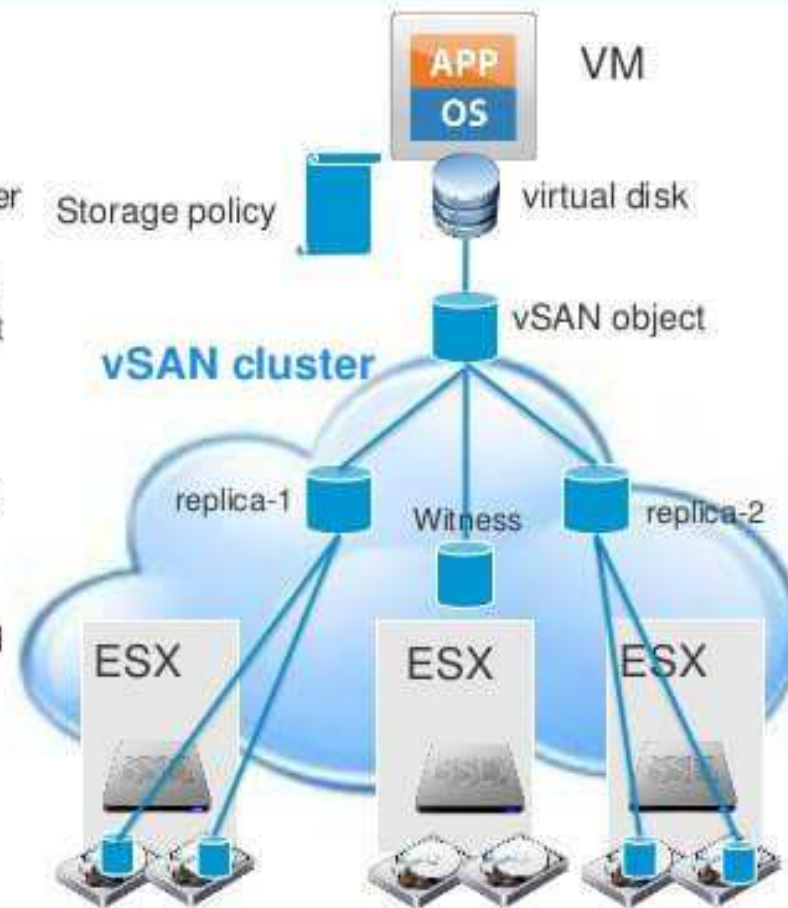
Stockage

- Types de baies de stockages
 - DAS - attachement direct
 - SAN – mode bloqué(FC, iSCSI)
 - NAS – mode fichier (NFS, CIFS)
- Types de réseaux de stockage
 - Fiber Channel (4, 8 ou 16 G)
 - Ethernet (1G, 10G,)
- Performance unitaire des disques en IOPS
 - SATA ~ 75
 - SAS 10k ~ 125
 - SAS 15k ~ 175
 - SSD ~ 10,000
- Fonction de snapshot possible au niveau des baies de stockage

Stockage distribué

Virtual SAN – Architecture

- Each ESX host contributes SSD and magnetic disk capacity
- Virtual SAN aggregates these resources into 1 global Datastore per vSphere cluster
- Each VM home directory and each virtual disk is now represented by a vSAN object
- Virtual machines run on the ESX hosts that belong to the cluster
- HA/DRS ensures the VM is restarted if a host crash
- Virtual SAN objects can be split into multiple components for performance and data protection. This is governed by the storage policies



Quelles réponses aux
besoins usuels d'une infra ?

...

Sauvegarde de VMs

- Sauvegarde classique par agents, pour les données
- Application de sauvegarde spécifiques aux infra virtualisées. Optimisations possibles :
 - Change Block Tracking (CBT) : identification des blocs modifiés depuis la dernière fois
 - Snapshot de VMs : figer les écritures sur un disque virtuel pendant la sauvegarde
- Modèle de sauvegarde en « disk2disk », parfois en « disk2disk2tape ».
- Baie de stockage spécialisé pour les sauvegardes (données froides) optimisé grâce à la déduplication.

PRA / DRS

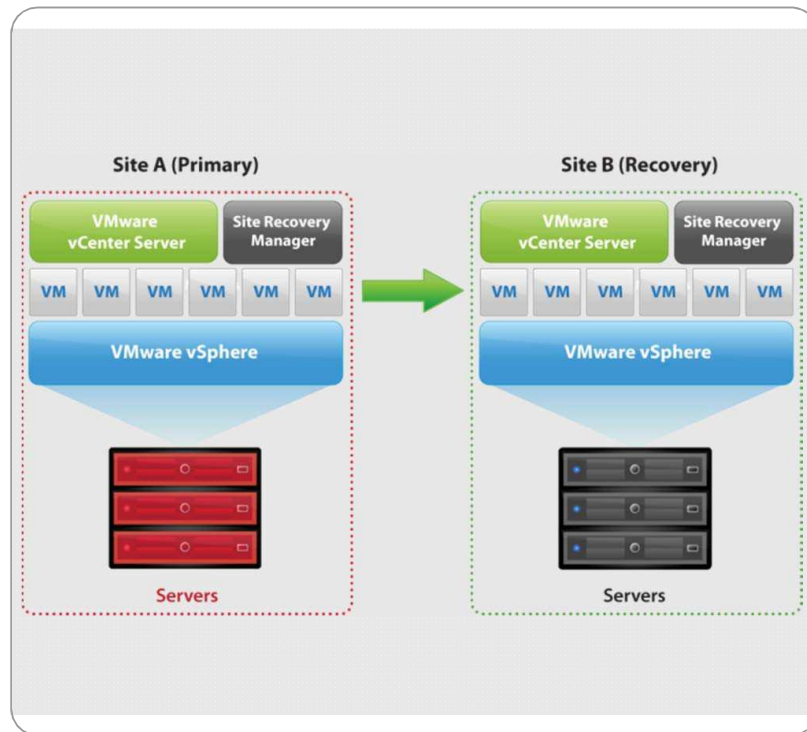
- Réplication entre les baies de stockages
- Réplication au niveau système
- Réplication au niveau applicatif
- **Attention** à la consistance des données quand l'application ne supporte pas la réplication bas niveau. (problème de synchronisation vis-à-vis du stockage)

PRA = Plan de Reprise d'Activité

DRS = Disaster Recovery System

VMware vCenter Site Recovery Manager

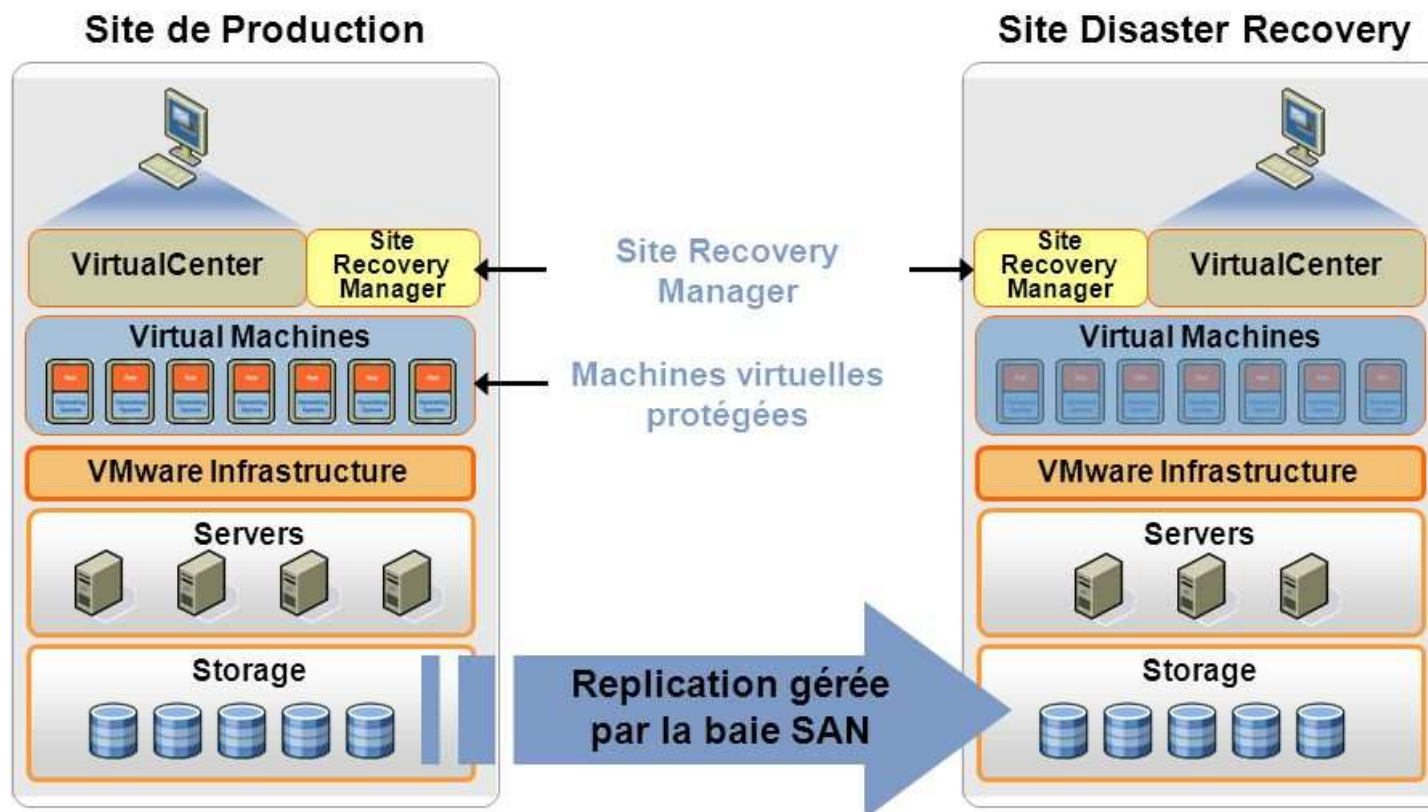
Site Recovery Manager leverages VMware vSphere to deliver advanced disaster recovery management and automation



- Simplifies and automates disaster recovery workflows:
Setup, Testing and Failover
- Turns manual recovery runbooks into automated recovery plans
- Provides central management of recovery plans from the VMware vSphere Client

Site Recovery Manager

VMware Site Recovery Manager - PRA



P2V, V2V

- Migration du physique vers le virtuel
- Conversion de formats de VMs (VMDK, AMI, VHD, QCOW2, RAW, ...)
- Format pivot d'import/export de VM : OVF/OVA
- Parfois, prise en charge des drivers (liens avec le matériel, ou le pseudo-matériel)
- La conversion de format n'est pas le plus compliqué ...

Le point critique, le réseau

- Déplacer une machine, physique ou virtuel, n'est pas vraiment compliqué.
- Par contre, continuer à communiquer avec le reste du système, c'est pas gagné
 - Configuration réseau de l'OS
 - Configuration réseau des applications (parfois « en dur » ou en base)
 - Dépendances avec des services transverse d'infrastructure (DNS, NTP, SMTP, proxy, ...)

vApp



- Regroupement de plusieurs VM dans une ensemble logique.
 - Possibilité de démarrer les VMs dans un ordre défini (back to front)
 - Possibilité d'arrêter les VMs dans le sens inverse (front to back)
 - Possibilité d'importer ou d'exporter toutes les VMs de la vApp, en OVF
 - Temporisation ou heartbeat via VM Tools
- Cas d'usage : service en multi-tiers

Que peut on virtualiser ?

- Techniquement, quasiment tout !
- Un frein persiste : licences de l'éditeur et modèle de facturation dissuasif ! (payer la licence au cpu, et pour tous les cpu du cluster de l'infra virtualisée)
- Cas particulier de cartes spécifiques, par exemple une carte HSM.

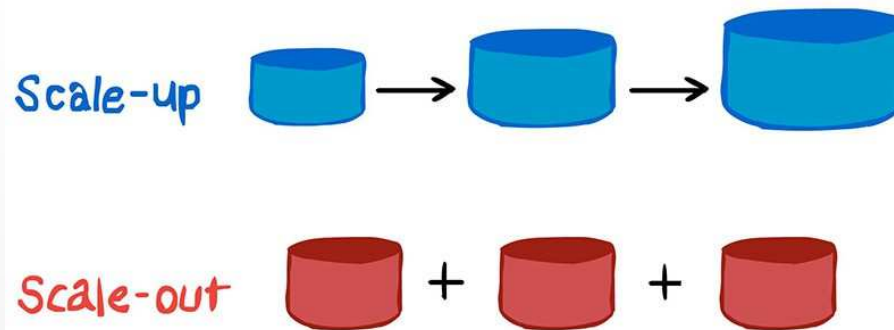
Vitesses de déploiement

CHANGEMENT DE VITESSE

Datacenter	Virtualisation	
		
Déploiement dans le mois	⇒ Déploiement dans la minute	
Pendant des années	Pendant des mois	
Développement en cascade	Agile	

Scalability (mise à l'échelle)

- Niveau infra de virtualisation
 - Ajout de serveurs hyperviseur
 - Ajout de stockages
 - Ajout de switch réseaux
 - Ajout de RAM sur les serveurs hyperviseur
- Niveau VM
 - Scalabilité verticale : ajout de vRAM et/ou de vCPU
 - Scalabilité horizontale : ajout de VM du même type (ex : serveur web)



Hyperconvergence

- **Les systèmes hyperconvergés**

Ils concentrent le stockage principal et les fonctions de calcul (“compute”) en une seule solution « hautement virtualisée » en s’appuyant sur une architecture hardware Intel X86 unique et extensible (“scale-out”). Le stockage est géré de façon virtuelle par logiciel.

- Plateforme logicielle d’hyperconvergence :

VMware (32,4%)

Nutanix (29,5 %)

Hardening Guides

Pour répondre à certain niveaux de sécurité, il est possible de sécuriser les hyperviseurs.

Pour le monde VMware, il existe des « hardening guides » pour les ESXi.

Suivant ses contraintes, on peut s'inspirer de

<https://www.vmware.com/security/hardening-guides.html> (VMware)

<https://nvd.nist.gov/> (DoD via le NIST)

<https://www.ssi.gouv.fr/> (ANSSI)

Bien dimensionner ses VM

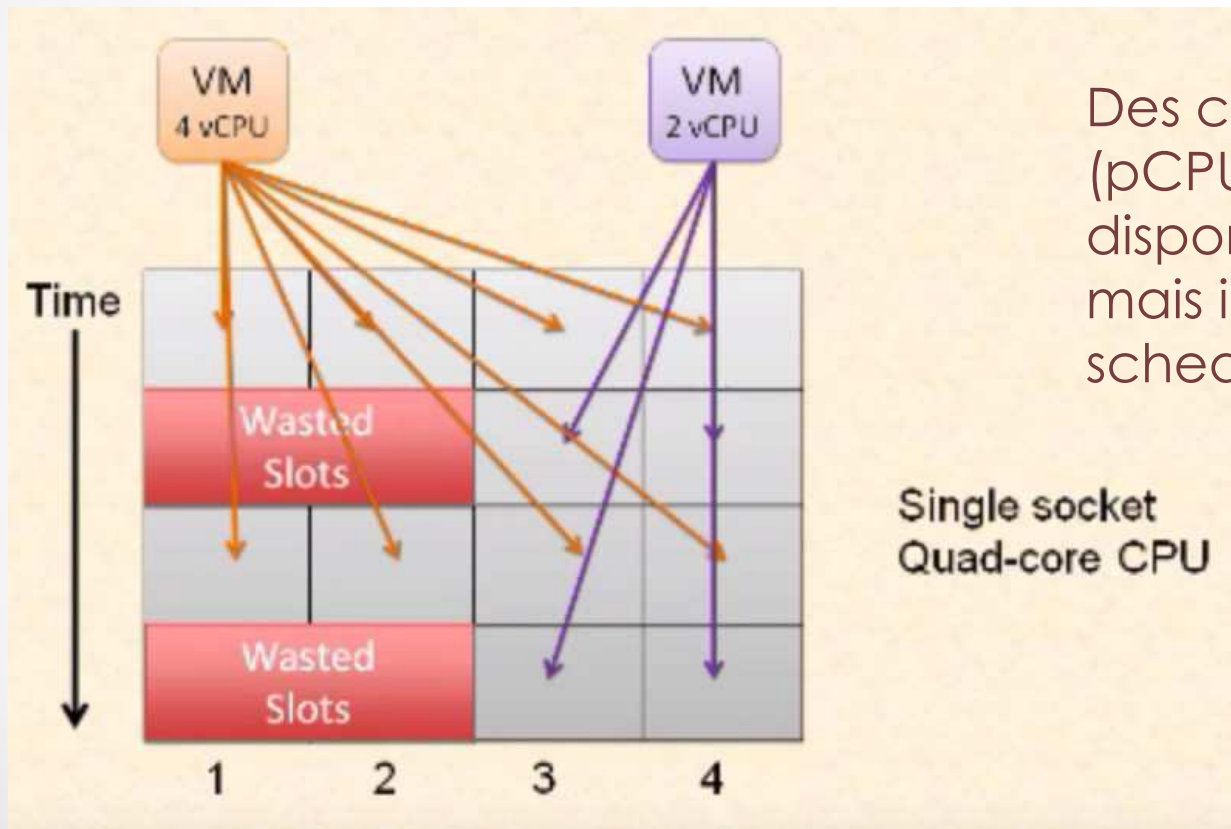
Il est important de « bien » dimensionner les VMs sur une infrastructure de virtualisation.

- **Mémoire** : les OS modernes utilisent la RAM disponible pour du cache. Si l'hyperviseur manque de RAM, il déclanchera le swap, la compression ou le ballooning
- **Processeur** : si trop de vCPU ont été attribués aux VMs, cela introduit de la latence « cpu ready time »

Les **impacts sur la performance arrivent vite !**

Scheduler vSphere

Cas de 2 VMs sur un processeur physique à 4 cores



Des cœurs physiques (pCPU) peuvent être disponibles (« ready ») mais inutilisés par le scheduler de l'hyperviseur.

Ajouter des vCPU aux VMs peut ralentir un hyperviseur !

Démos

...