Virtualization and Cloud System

Tien-Fu Chen

Dept. of Computer Science
National Chiao Tung Univ.

Operating System Virtualization

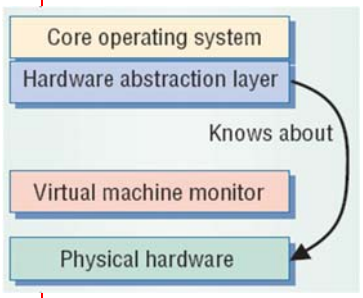
- A **virtual machine (VM)** is the virtual environment that emulates a physical computer's hardware and BIOS. **A guest OS** is the operating system installed on a VM.
- A **host computer** is the physical computer on which the VM software is installed
- Virtualization software creates and manages VMs and creates the virtual environment in which a guest OS is installed
- **Hypervisor** creates and monitors the virtual hardware environment, which allows multiple VMs to share physical hardware resources

Operating System Virtualization

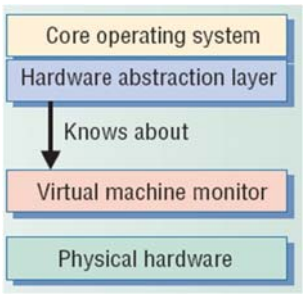
- Type 1 hypervisor runs directly on the host computer's hardware and controls and monitors guest OSs
- Type 2 hypervisor is installed in a general-purpose host OS and the host OS accesses host hardware on behalf of the guest OS
- A virtual disk consists of files residing on the host computer that represent a virtual machine's hard drive
- A virtual network is a network configuration created by virtualization
- A snapshot is a partial copy of a VM made at a particular moment

Source: Guide to Networking Essentials, 6th Edition

System-level Design Approaches

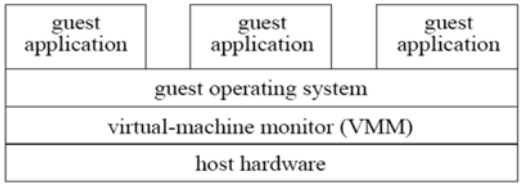


- **Full virtualization (direct execution)**
  - Exact hardware exposed to OS
  - Efficient execution
  - OS runs unchanged
  - Requires a "virtualizable" architecture
  - Example: VMWare

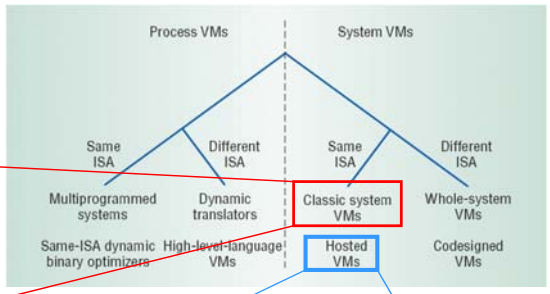


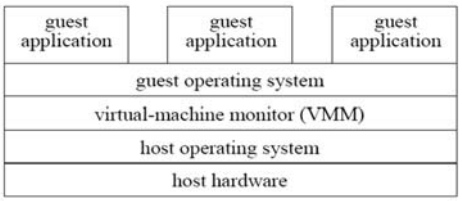
- **Paravirtualization**
  - OS modified to execute under VMM
  - Requires porting OS code
  - Execution overhead
  - Necessary for some (popular) architectures (e.g., x86)
  - Examples: Xen, Denali

System VMMs



Type 1

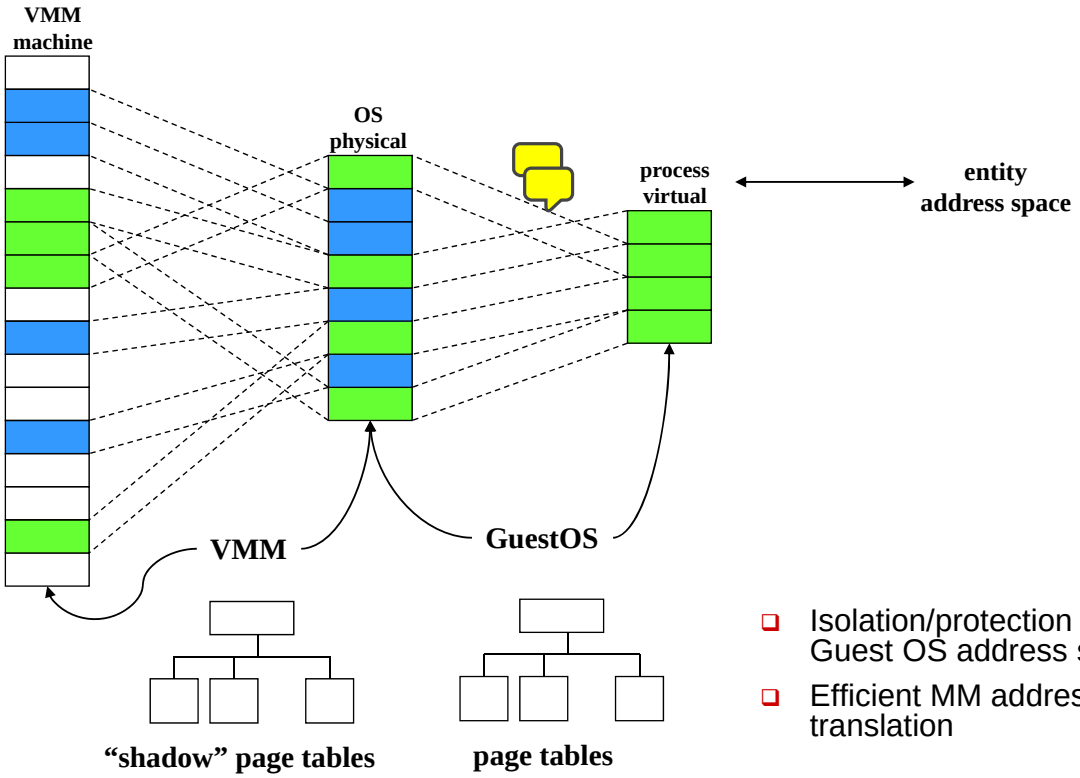




Type 2

- **Structure**
  - **Type 1: runs directly on host hardware**
  - **Type 2: runs on HostOS**
- **Primary goals**
  - **Type 1: High performance**
  - **Type 2: Ease of construction/installation/acceptability**
- **Examples**
  - **Type 1: VMWare ESX Server, Xen, OS/370**
  - **Type 2: User-mode Linux**

Memory Management



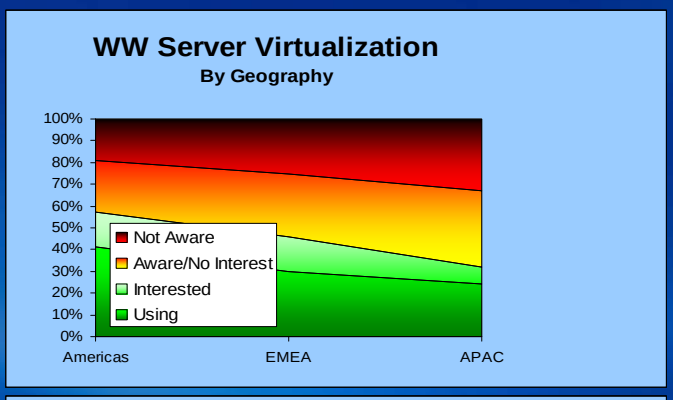
- Isolation/protection of Guest OS address spaces
- Efficient MM address translation

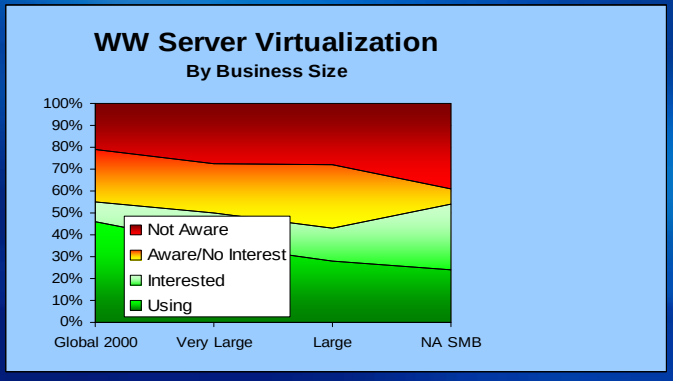
Hosted Virtualization Products

	VMware Workstation	VMware Player	Windows Virtual PC	VirtualBox
Price	$189 or free with Academic Program membership	Free	Free	Free
Host OS support	Windows, Linux, Mac OS X (with VMware Fusion)	Windows, Linux	Windows	Windows, Linux, Mac OS X, Solaris
Guest OS support	Windows, several Linux distributions, NetWare, Solaris, DOS	Same as Workstation	Windows XP and later	Windows, several Linux distributions, Solaris, Mac OS X Server, DOS, OS/2, others
Snapshots	Unlimited	None	One (with Disk Undo enabled)	Unlimited
Virtual network options	Bridged, NAT, host-only, custom	Bridged, NAT, host-only	Bridged, NAT, internal (guest-to-guest only)	Bridged, NAT, host-only, internal
Host integration tools	VMware Tools, Unity	VMware Tools, Unity	Integration Services, XP mode	Guest additions, seamless mode
Other features	Virtual teams, screen capture and screen movie capture, physical-to-VM conversion, developer tools			Command-line management interface, built-in remote desktop, developer programming interface, open-source edition

Virtualization Awareness Today*

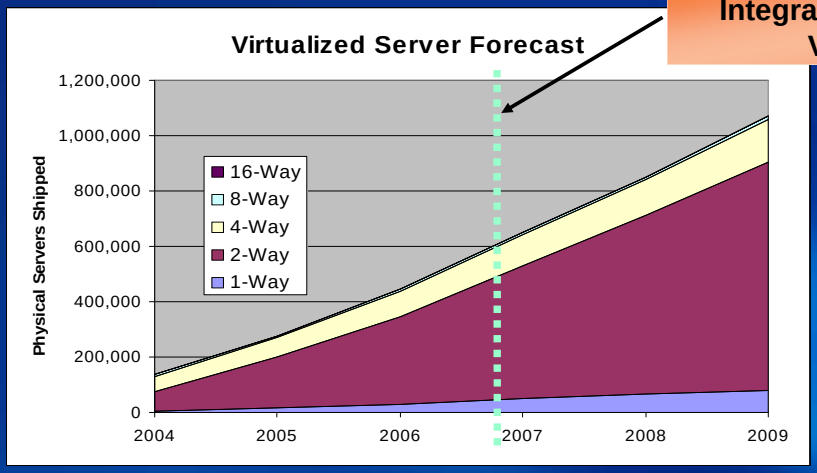
- 75% of enterprises aware of virtualization
- 34% Implementing virtualization by mid 2006
  - Very large biz at 46%; SMB at 25%!
  - North America leading; Other GEO's right behind!
- 60% increasing virtualization in next 12 months!





* Forrester 2-22-06 Server Virtualization Goes Mainstream; 1221 end user quant study

Virtualized x86 Server Market Overview*



- 80% of customers using virtualization do so for consolidation
- Virtualized server market growing from 4.5% today to >12% of all servers in 2009
  - Growing from 276K in 2005 to 1.1M units in 2009 (51% CAGR)
  - Feedback from the market: Aggressive projections for 2005; conservative for 2009

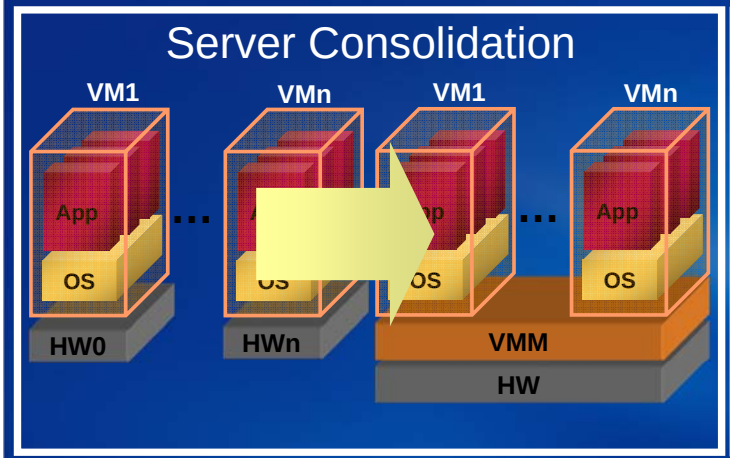
Virtualization: Significant growth due to compelling value

*Source: IDC WW Virtualization Forecast Aug-2005

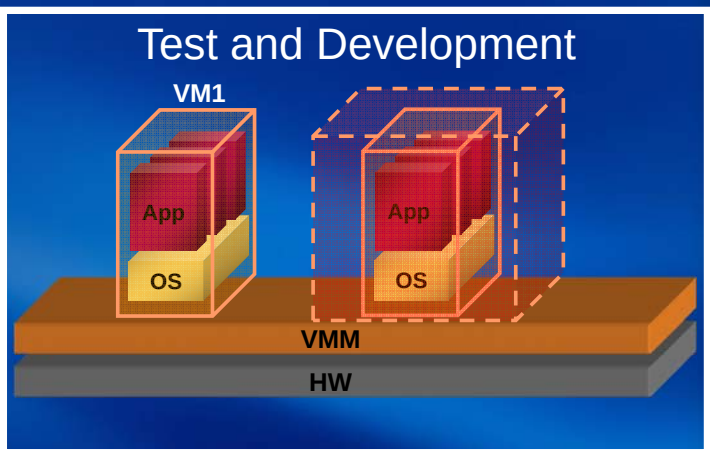
WinHEC 2006

Today's Uses

Virtualization addresses today's IT concerns



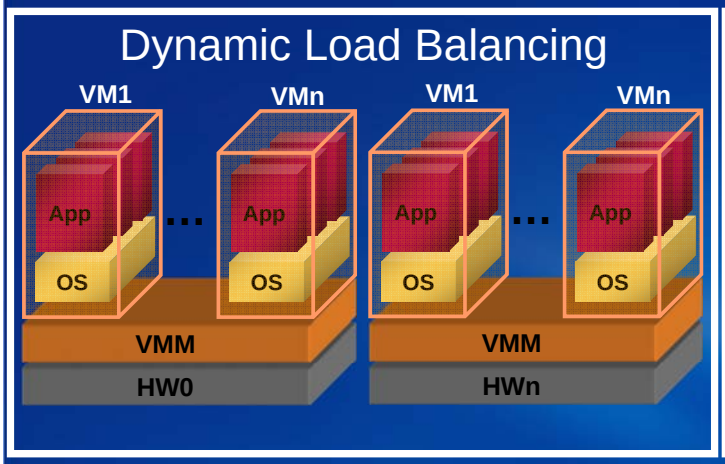
10:1 in many cases

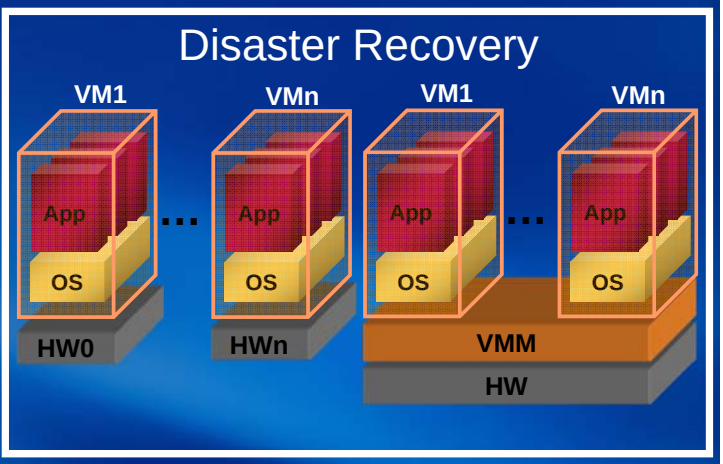


Enables rapid deployment

Microsoft WinHEC 2006

Emerging Usage Models





Goal: True "Lights Out" Datacenter

Instantaneous failover

Dynamic load balancing

Autonomics

Self healing

Microsoft WinHEC 2006

A More Reliable Server

Unique Intel x86 Reliability Features

Feature	Benefit	Description	Intel Xeon processor Based Servers	Other x86 Based Servers
Memory ECC	Data Integrity & Availability	Detects & corrects single-bit errors	☑	☑
Enhanced Memory ECC	Data Integrity & Availability	Retry double-bit errors vs. standard memory ECC that does single-bit errors only	☑	☐
Memory CRC (FBD)	Continued Operation & Availability	Address & command transmissions are automatically retried if a transient error occurs vs. the potential of silent data corruption	☑	☐
Memory Sparing	Data Availability	Predicts a "failing" DIMM & copies the data to a spare memory DIMM , maintaining server available & uptime	☑	☐
Memory Mirroring	Data Protection	Data is written to 2 locations in system memory so that if a DRAM device fails, mirrored memory enables continued operation and data availability	☑	☐
Symmetric Access to all CPUs	Server Continuity	Enables a system to restart and operate if the primary processor fails	☑	☐

A Better Business Foundation
Less Downtime, Higher Service Availability and Improved Confidence

Enabled by a combination of processor, chipset and platform memory technologies. Data as of March 6, 2006

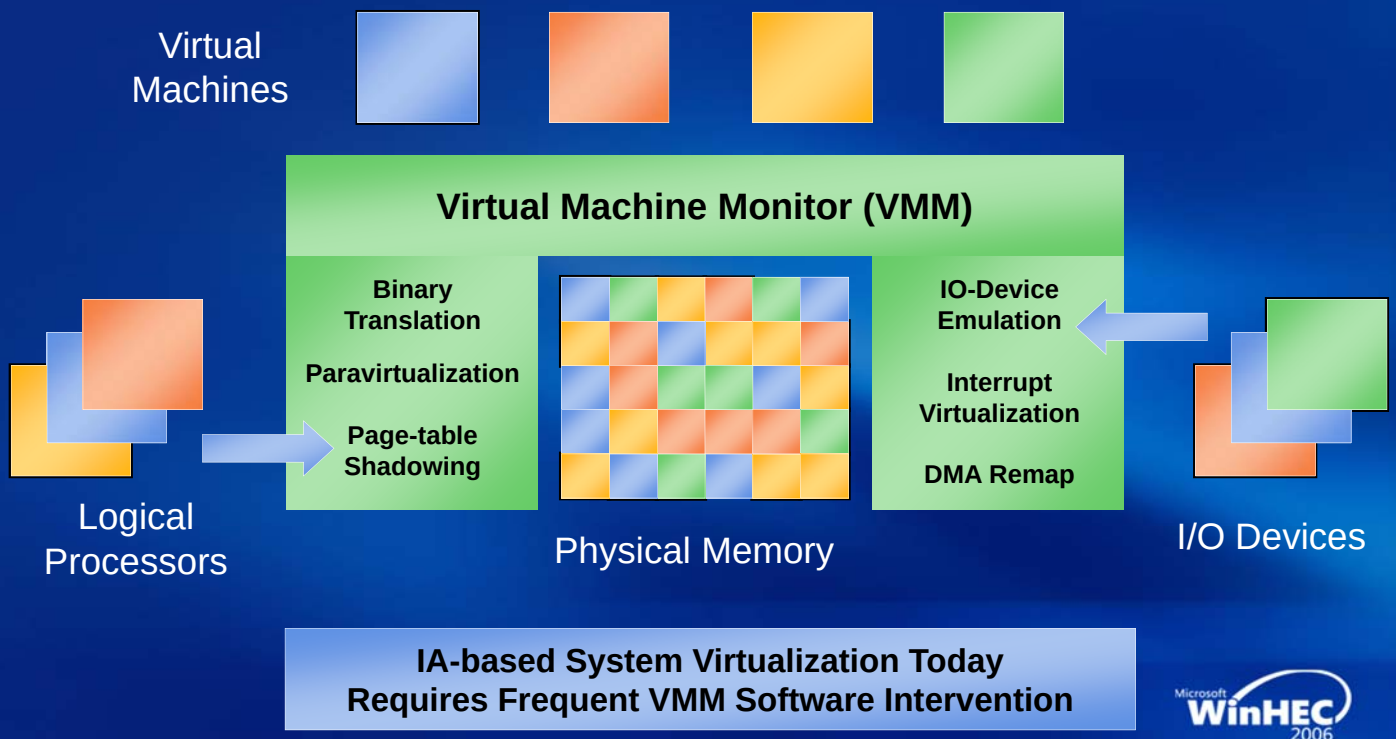
Intel Virtualization Technology (VT)

Provides silicon-based functionality that works together with compatible VMM software to provide new capabilities

- Enables richer software capabilities
  - 64-bit guest OS support in virtualized environment
  - Support for unmodified, heterogeneous guest operating systems to run on new VMM's
  - Intel is working with the industry
- Common virtualization standards from client to servers
- Broad availability of both client and server platforms since November 2005 for accelerated software development
  - Endorsements and beta SW available from multiple vendors
  - Support for VT in Microsoft Virtual Server 2005 R2 SP1

Intel VT Roadmap

IA System Virtualization Today

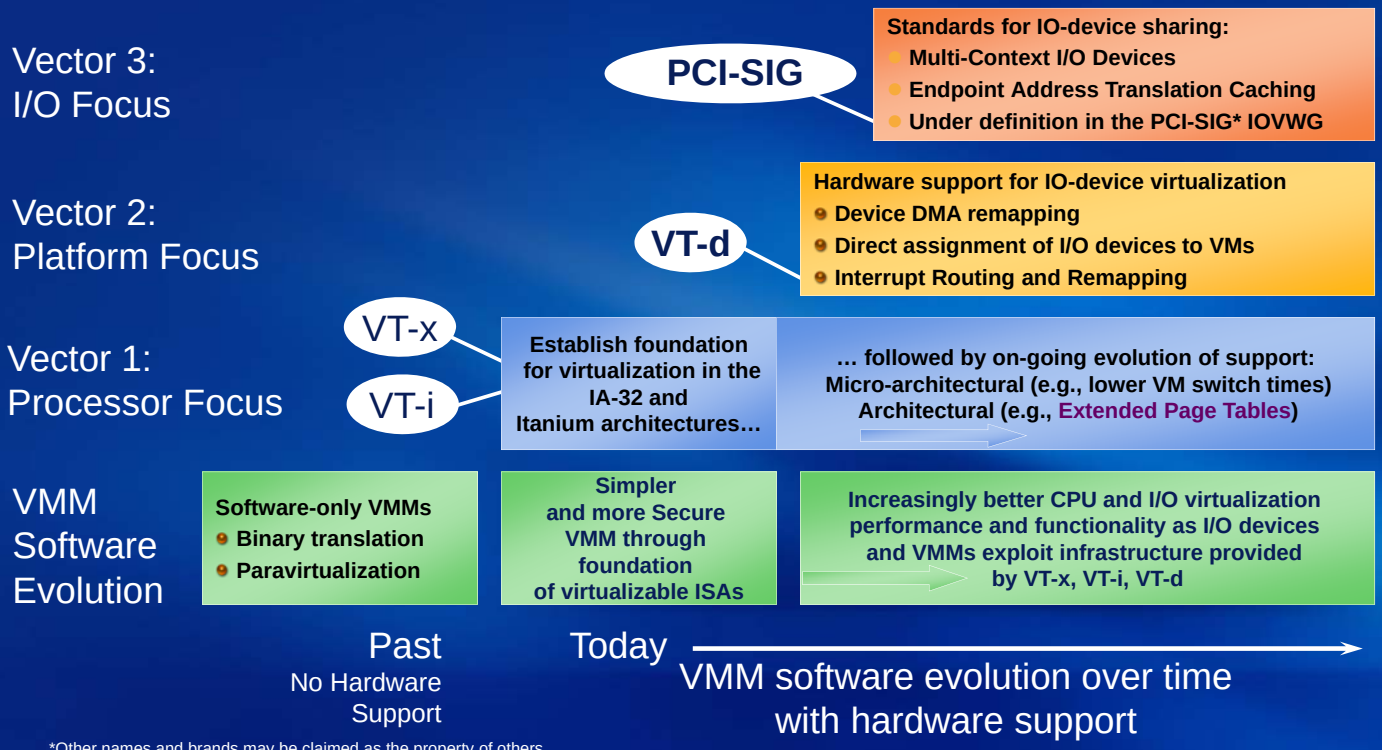


IA-based System Virtualization Today Requires Frequent VMM Software Intervention

IA Virtualization Today
Summary Of Challenges

- Complexity
  - CPU virtualization requires binary translation or paravirtualization
  - Must emulate I/O devices in software
- Functionality
  - Paravirtualization may limit supported guest OSes
  - Guest OSes “see” only simulated platform and I/O devices
- Reliability and Security
  - I/O device drivers run as part of host OS or hypervisor
  - No protection from errant DMA that can corrupt memory
- Performance
  - Overheads of address translation in software
  - Extra memory required (e.g., translated code, shadow tables)

Intel Virtualization Technology Evolution

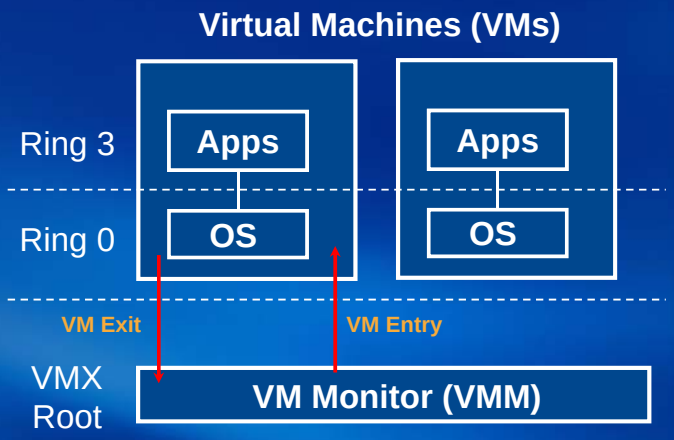


*Other names and brands may be claimed as the property of others

VT-x Overview: Intel Virtualization Technology For IA-32 Processors

CPU Virtualization With VT-x

- Two new VT-x operating modes
  - Less-privileged mode (VMX non-root) for guest OSes
  - More-privileged mode (VMX root) for VMM
- Two new transitions
  - VM entry to non-root operation
  - VM exit to root operation



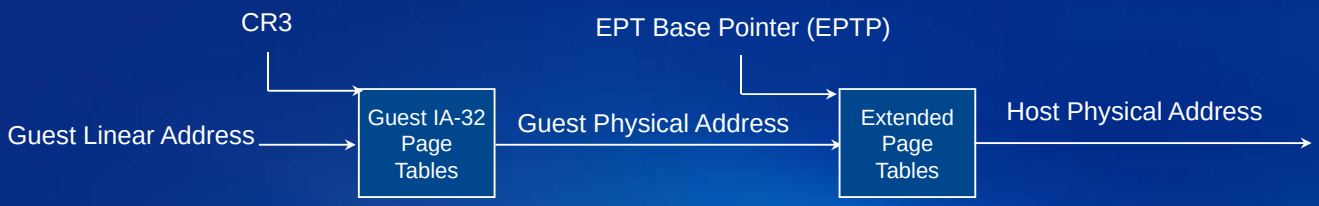
- Execution controls determine when exits occur
  - Access to privilege state, occurrence of exceptions, etc.
  - Flexibility provided to minimize unwanted exits
- VM Control Structure (VMCS) controls VT-x operation
  - Also holds guest and host state

Extended Page Tables (EPT)

- A VMM must protect host physical memory
  - Multiple guest operating systems share the same host physical memory
  - VMM typically implements protections through "page-table shadowing" in software
- Page-table shadowing accounts for a large portion of virtualization overheads
  - VM exits due to: #PF, INVLPG, MOV CR3

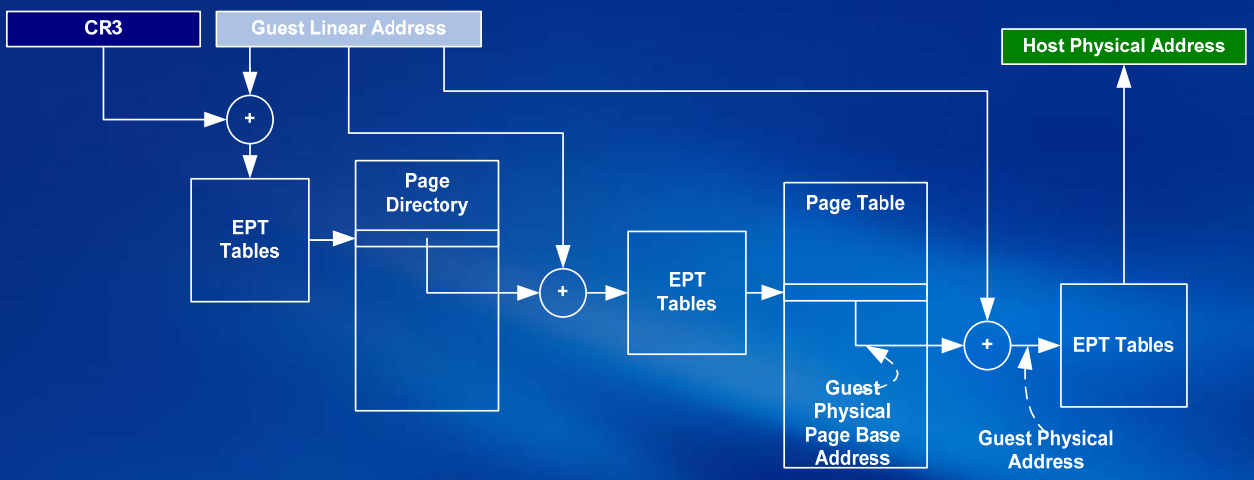
Goal of EPT is to reduce these overheads

What Is EPT?



- Extended Page Table
- A new page-table structure, under the control of the VMM
  - Defines mapping between guest- and host-physical addresses
  - EPT base pointer (new VMCS field) points to the EPT page tables
  - EPT (optionally) activated on VM entry, deactivated on VM exit
- Guest has full control over its own IA-32 page tables
  - No VM exits due to guest page faults, INVLPG, or CR3 changes

EPT Translation: Details



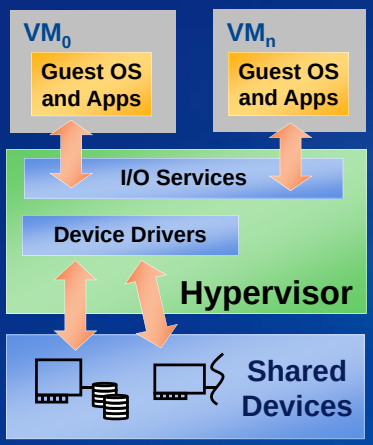
- All guest-physical memory addresses go through EPT tables
  - (CR3, PDE, PTE, etc.)
- Above example is for 2-level table for 32-bit address space
  - Translation possible for other page-table formats (e.g., PAE)

VT-d Overview: Intel Virtualization Technology For Directed I/O

Microsoft WinHEC 2006

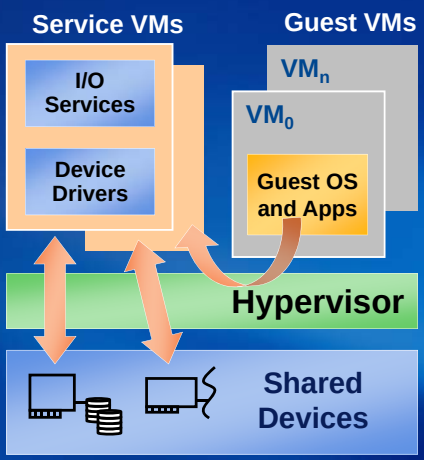
Options For I/O Virtualization

Monolithic Model



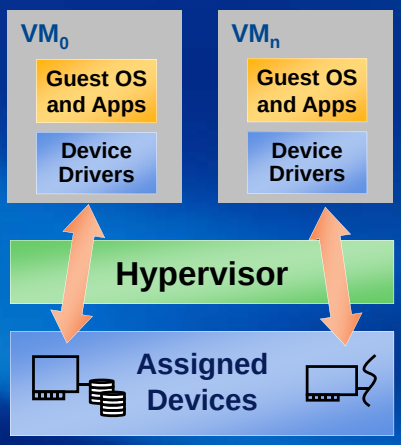
- Pro: Higher Performance
- Pro: I/O Device Sharing
- Pro: VM Migration
- Con: Larger Hypervisor

Service VM Model



- Pro: High Security
- Pro: I/O Device Sharing
- Pro: VM Migration
- Con: Lower Performance

Pass-through Model



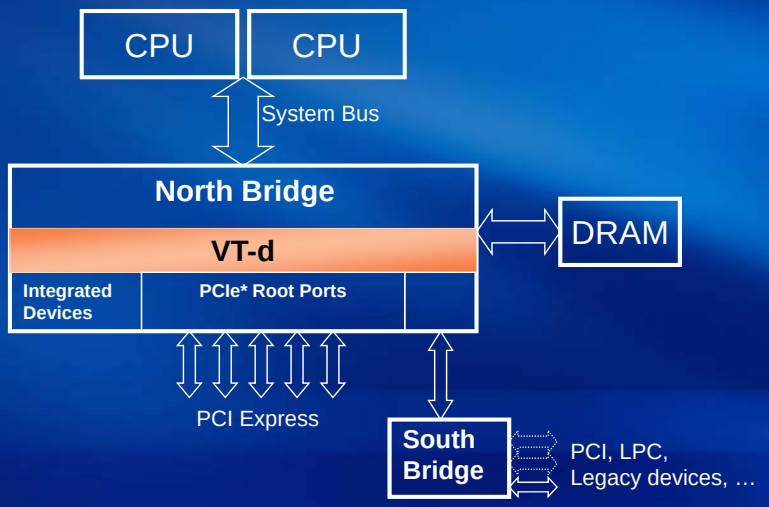
- Pro: Highest Performance
- Pro: Smaller Hypervisor
- Pro: Device assisted sharing
- Con: Migration Challenges

VT-d Goal: Support all Models

Microsoft WinHEC 2006

VT-d Overview

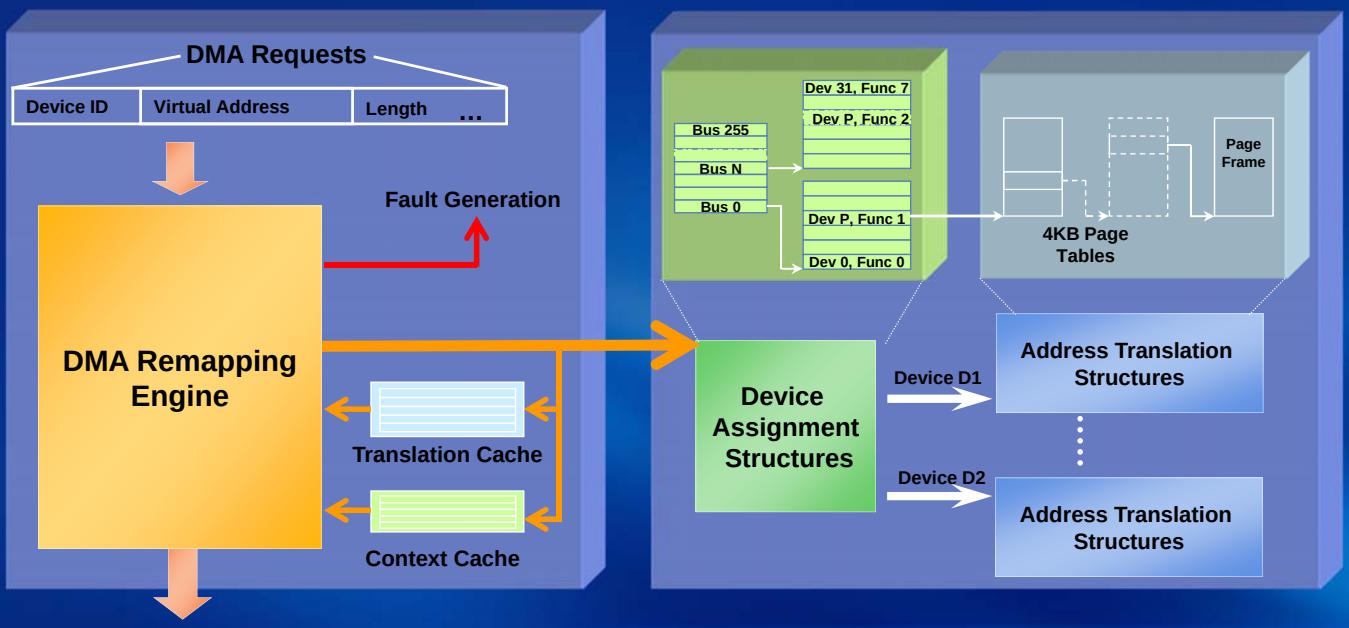
- VT-d is platform infrastructure for I/O virtualization
  - Defines architecture for DMA remapping
  - Implemented as part of platform core logic
  - Will be supported broadly in Intel server and client chipsets



VT-d Usage

- Basic infrastructure for I/O virtualization
  - Enable direct assignment of I/O devices to unmodified or paravirtualized VMs
- Improves system reliability
  - Contain and report errant DMA to software
- Enhances security
  - Support multiple protection domains under SW control
  - Provide foundation for building trusted I/O capabilities
- Other usages
  - Generic facility for DMA scatter/gather
  - Overcome addressability limitations on legacy devices

VT-d Architecture Detail



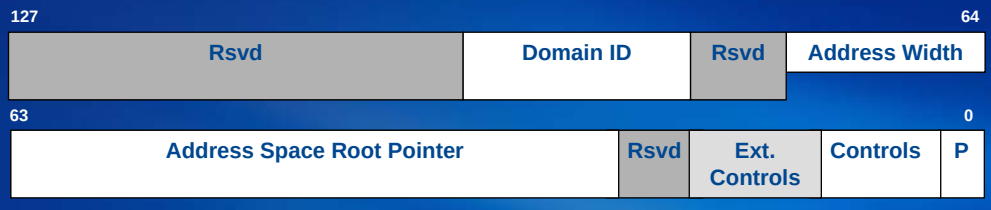
Memory Access with System Physical Address

Memory-resident Partitioning And Translation Structures

Microsoft WinHEC 2006

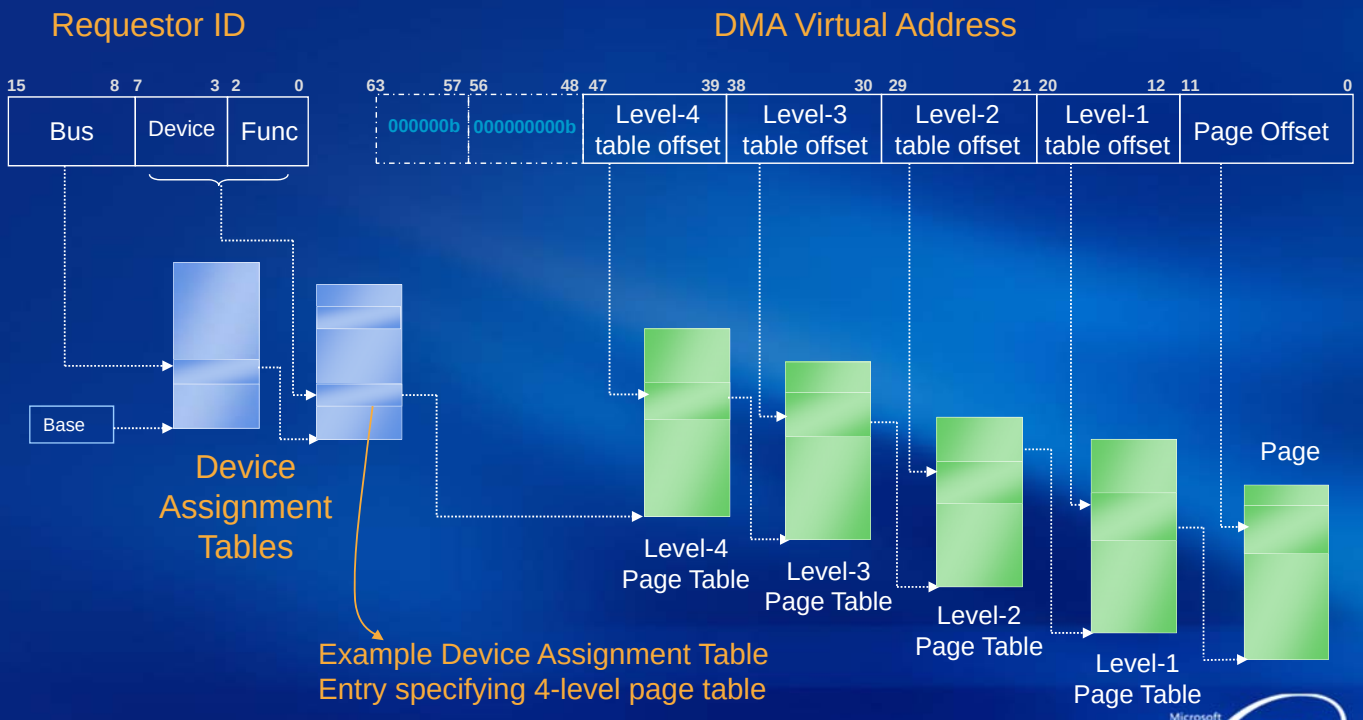
VT-d: Remapping Structures

- VT-d hardware selects page-table based on source of DMA request
  - Requestor ID (bus / device / function) in request identifies DMA source
- VT-d Device Assignment Entry



- VT-d supports hierarchical page tables for address translation
  - Page directories and page tables are 4 KB in size
  - 4KB base page size with support for larger page sizes
  - Support for DMA snoop control through page table entries
- VT-d Page Table Entry

VT-d: Hardware Page Walk



VT-d: Translation Caching

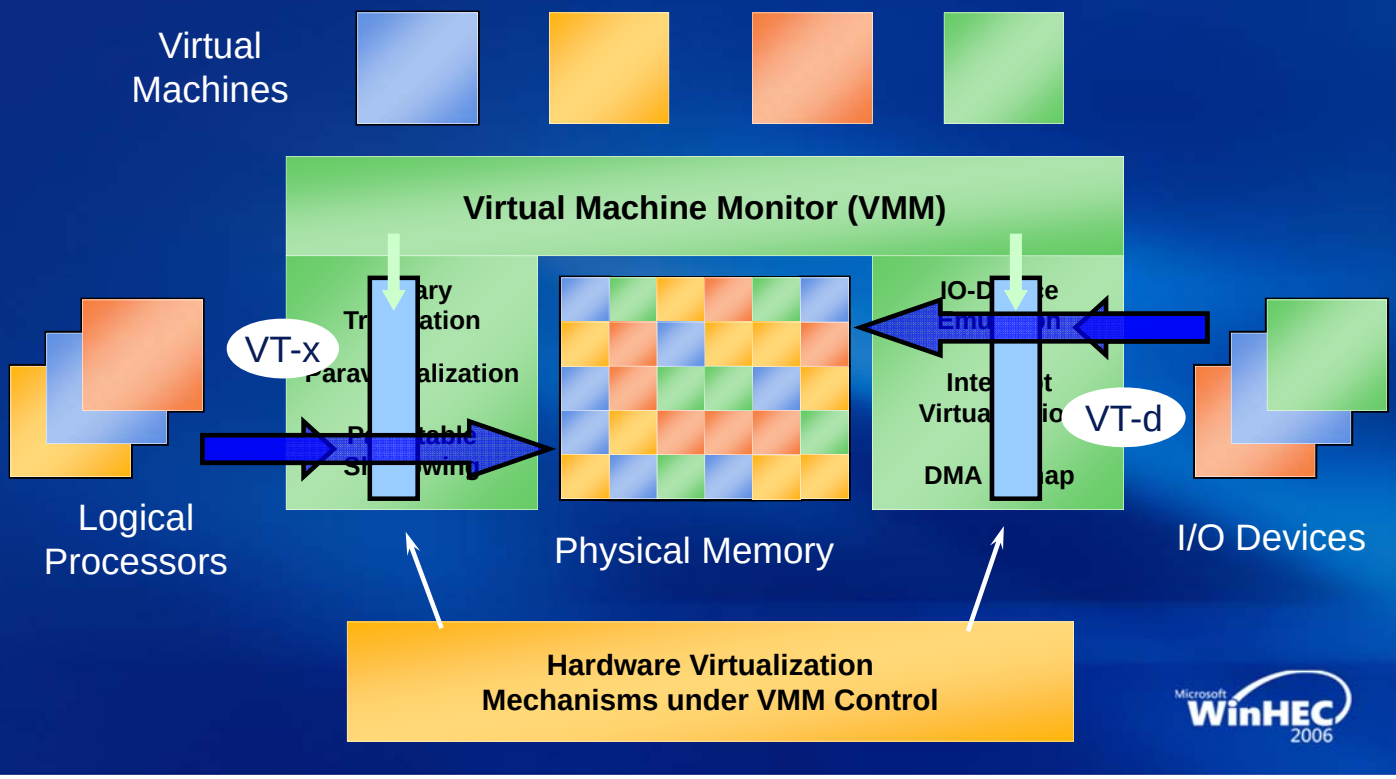
- Architecture supports caching of remapping structures
  - Context Cache: Caches frequently used device-assignment entries
  - IOTLB: Caches frequently used translations (results of page walk)
  - Non-leaf Cache: Caches frequently used page-directory entries
- When updating VT-d translation structures, software enforces consistency of these caches
  - Architecture supports global, domain-selective, and page-range invalidations of these caches
  - Primary invalidation interface through MMIO registers for synchronous invalidations
  - Extended invalidation interface for queued invalidations

VT-d: Extended Features

- PCI Express protocol extensions being defined by PCISIG for Address Translation Services (ATS)
  - Enables scaling of translation caches to devices
  - Devices may request translations from root complex and cache
  - Protocol extensions to invalidate translation caches on devices
- VT-d extended capabilities
  - Enables VMM software to control device participation in ATS
  - Returns translations for valid ATS translation requests
  - Supports ATS invalidations
  - Provides capability to isolate, remap and route interrupts to VMs
  - Support device-specific demand paging by ATS capable devices

VT-d Extended features utilize PCI Express enhancements being pursued within the PCI-SIG

VT-x & VT-d Working Together

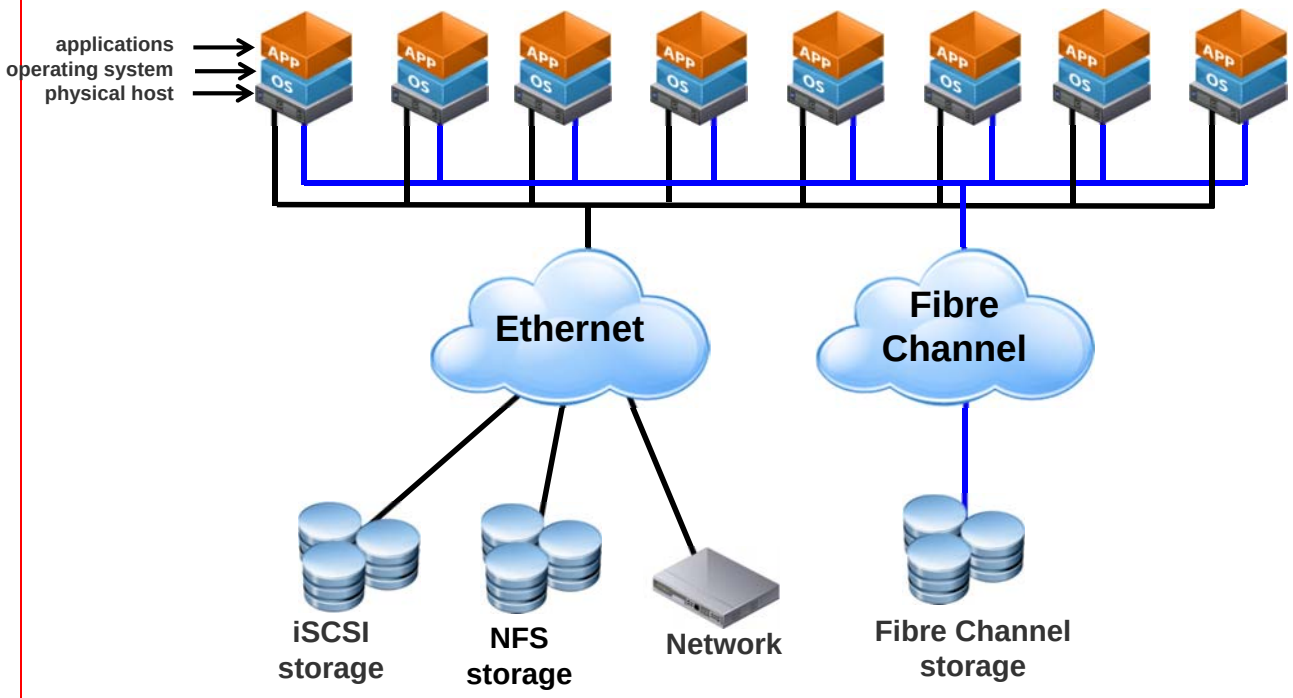


How Intel Virtualization Technology Address Virtualization Challenges

- Reduced Complexity
  - VT-x removes need for binary translation / paravirtualization
  - Can avoid I/O emulation for direct-mapped I/O devices
- Improved Functionality
  - 64-bit guest OS support, remove limitations of paravirtualization
  - Can grant Guest OS direct access to modern physical I/O devices
- Enhanced Reliability and Protection
  - Simplified VMM reduces “trusted computing base” (TCB)
  - DMA errors logged and reported to software
- Improved Performance
  - Hardware support reduces address-translation overheads
  - No need for shadow page tables (saves memory)

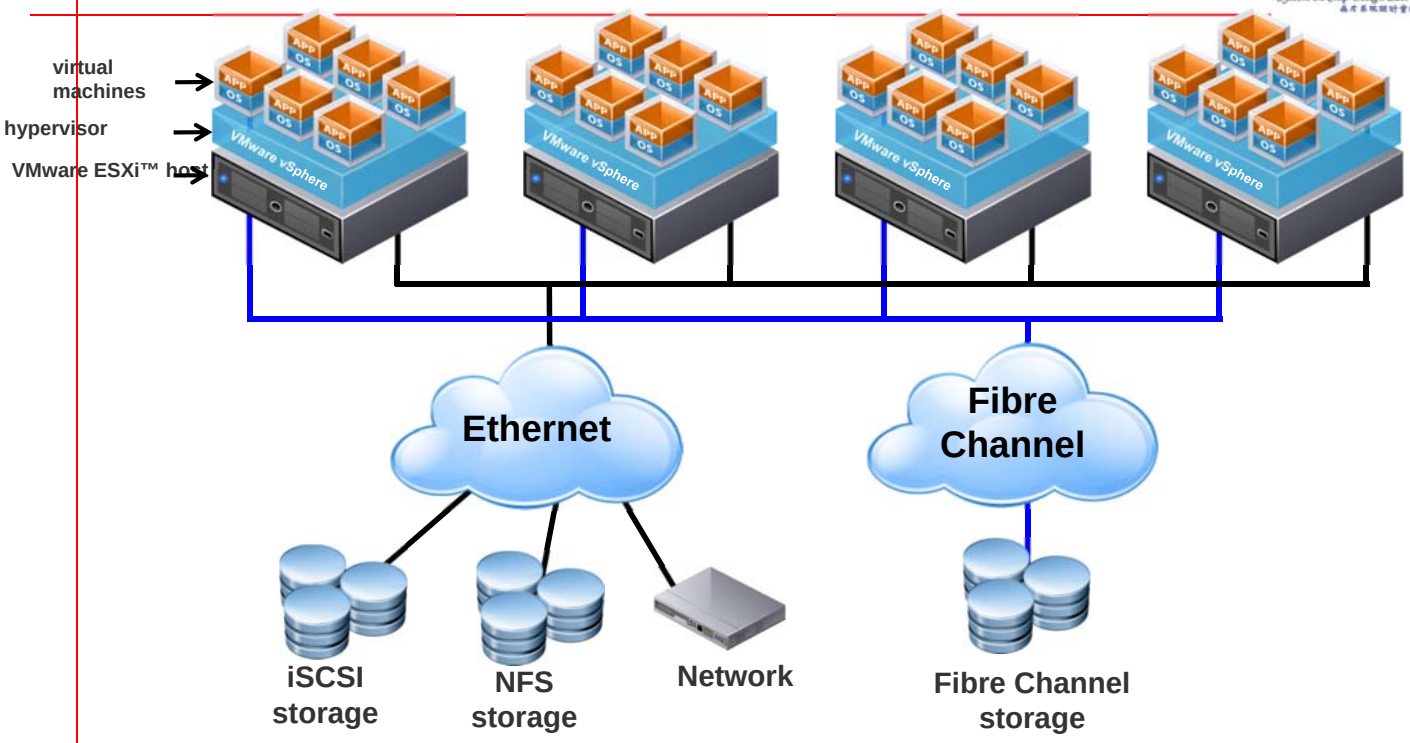
Microsoft WinHEC 2006

Physical Infrastructure



Source: VMware vSphere: Overview

Virtual Infrastructure

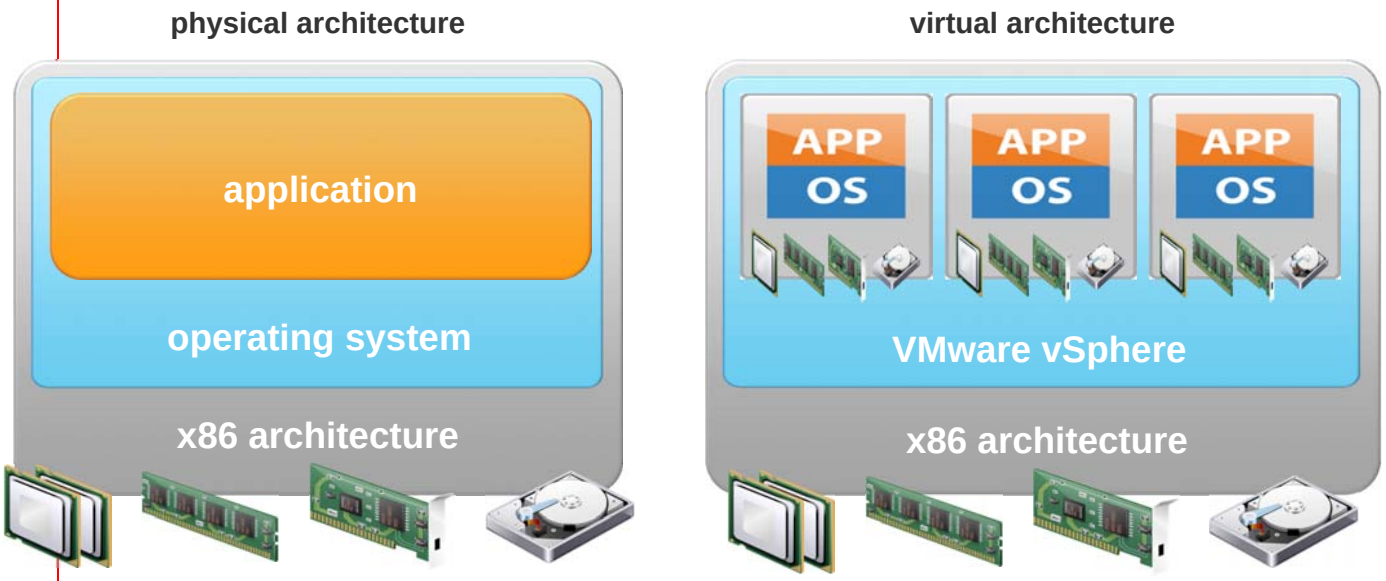


Source: VMware vSphere: Overview

Bare-Metal Virtualization

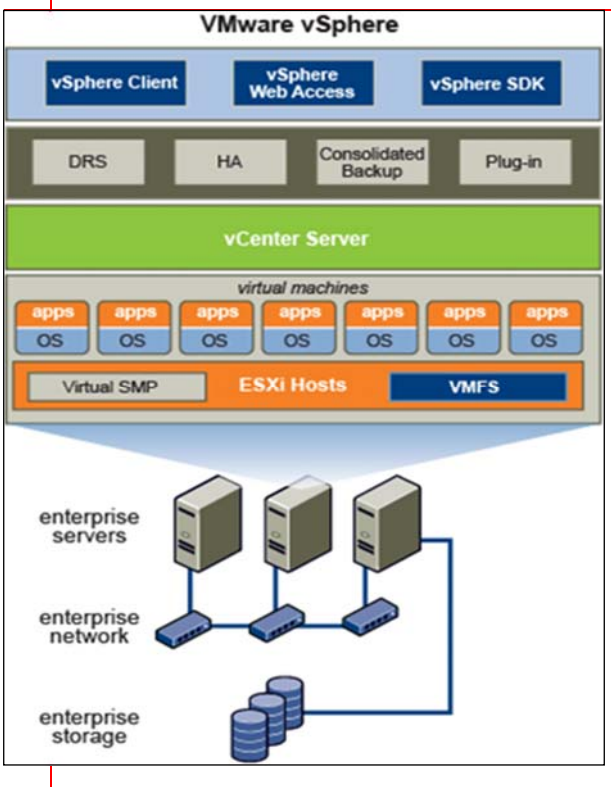
- Uses a Type 1 hypervisor
- Targeted mainly for production virtualization in data centers
- Installed directly on hardware and has more stringent host machine requirements
- Offers more features for managing VMs than hosted virtualization
  - Microsoft Hyper-V – introduced with Windows Server 2008 and can be installed as a server role
  - Citrix XenServer – Uses Linux as a management OS on the host
  - VMware vSphere – includes VMware ESX Server, which is installed directly on the physical server without a management OS

Physical Versus Virtual Architecture



Source: VMware vSphere: Overview

What Is VMware vSphere?



An infrastructure virtualization suite that provides virtualization, management, resource optimization, application availability, and operational automation capabilities

It consists of the following components:

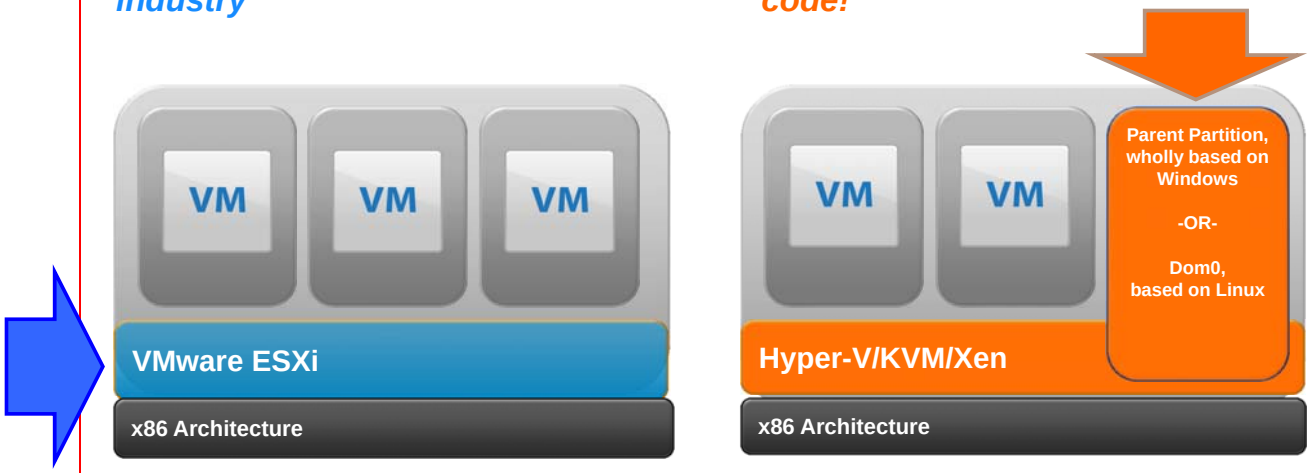
- VMware ESXi
- VMware vCenter Server™
- VMware vSphere® Client™
- VMware vSphere® VMFS
- VMware vSphere® Virtual Symmetric Multiprocessing

Source: VMware vSphere: Overview

VMware Differentiation by Hypervisor

First ultra-slim x86 hypervisor in industry

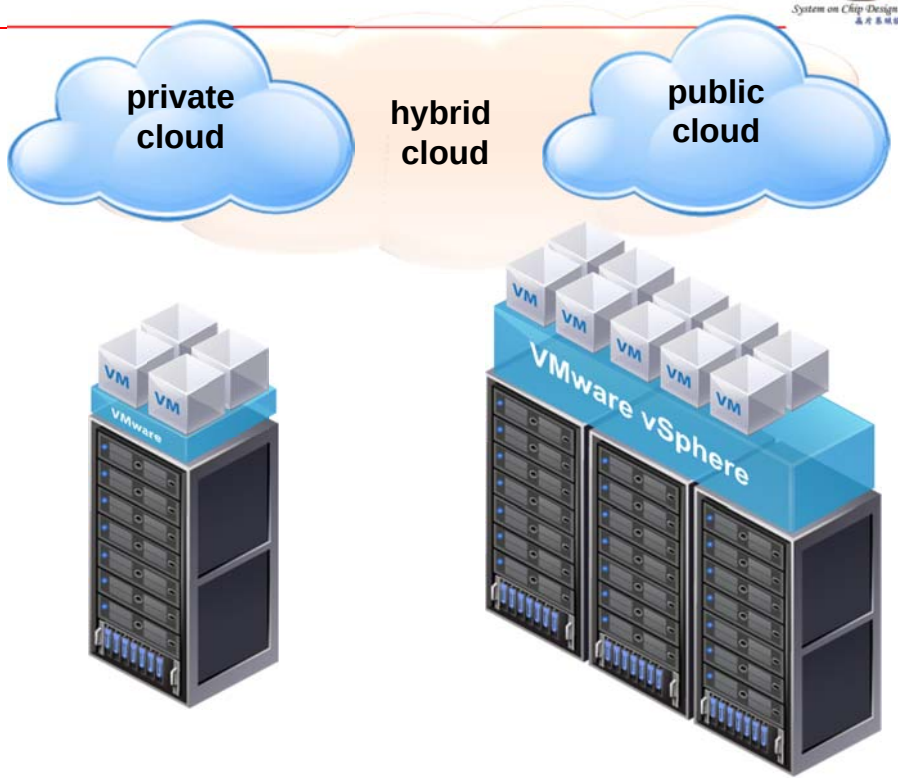
Lots of legacy operating system code!



Bolting virtualization to general purpose operating system increases risk and decreases reliability

How vSphere Fits into Cloud Computing

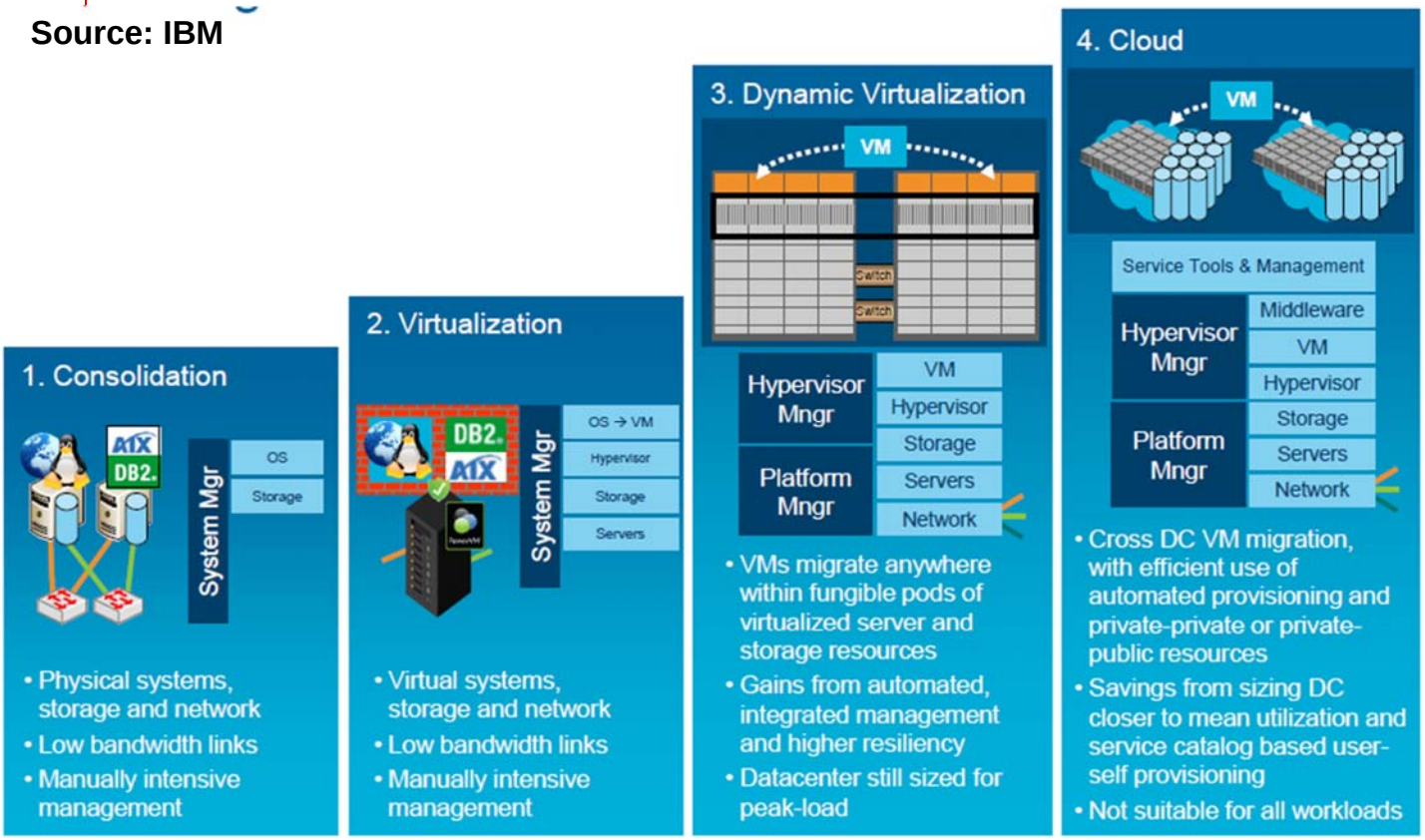
- Installing vSphere 5 creates a virtual infrastructure.
- Your virtual machines run in this virtual infrastructure.
- VMware vCloud Director™ enables you to create a cloud.
- Third-party providers can host public or private clouds.
- VMware® clouds empower you to run your virtual machines in a private, public, or hybrid cloud to fit your business needs.



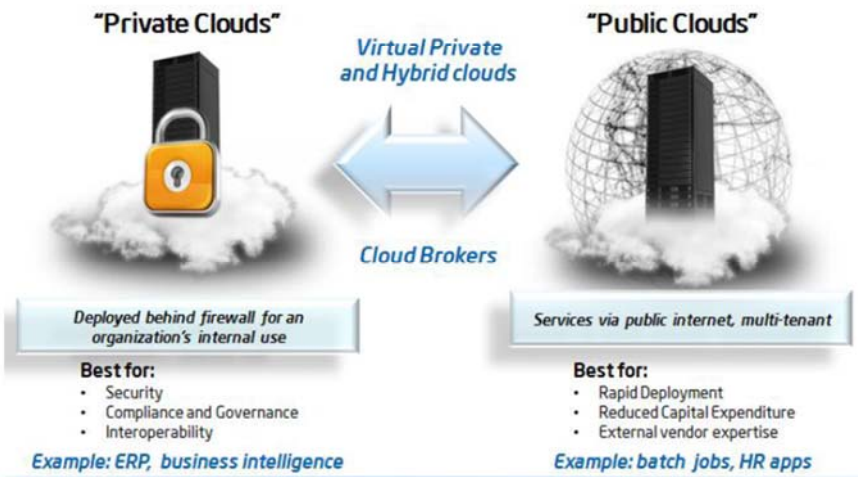
Source: VMware vSphere: Install, Configure, Manage

Data center transformations are driven by increasing levels of virtualization

Source: IBM

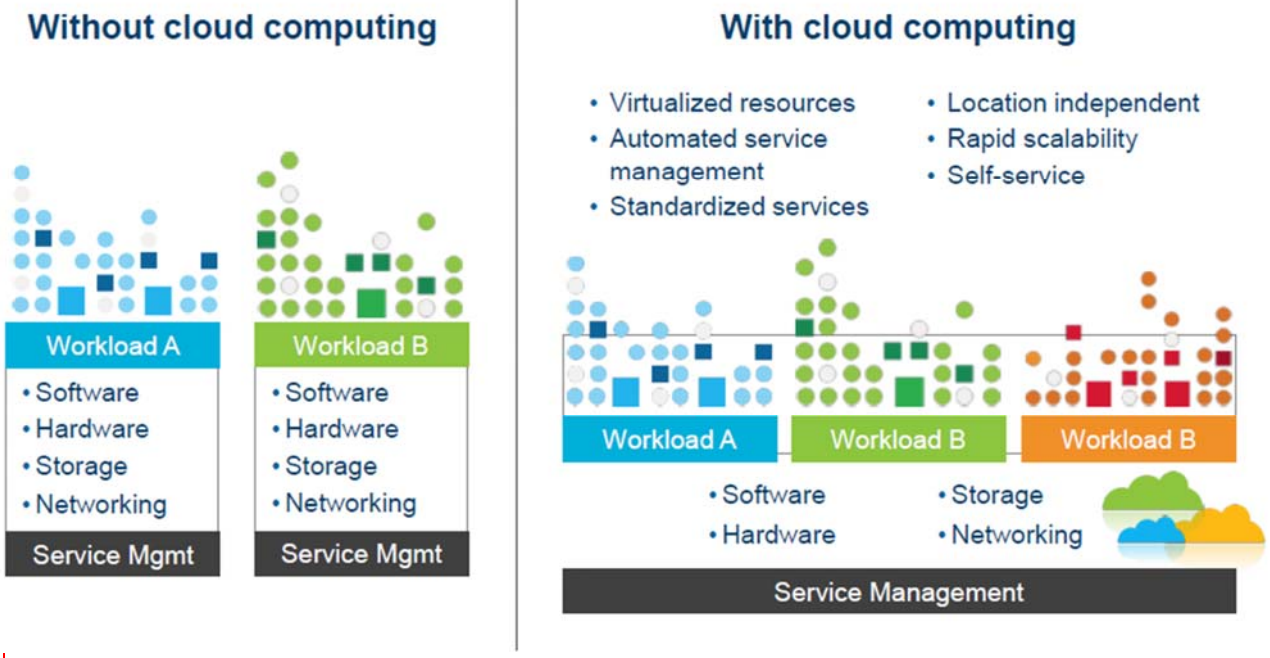


Control Issue has Driven Investment in Private Clouds



Virtualization support and Cloud system

What is different about cloud computing?



Source: Elements of cloud computing taken from NIST, Gartner, Forrester and IDC cloud computing definitions

Delivering the Cloud platform through a spectrum of delivery models