T.J. Watson Research Center

Hardware Virtualization Trends

Leendert van Doorn

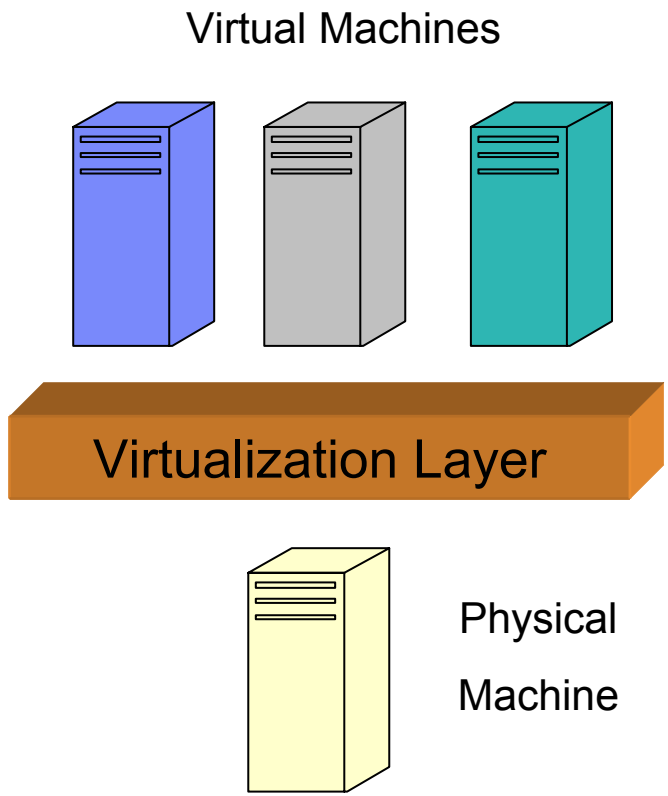
Outline

- **Virtualization 101**
- **The world is changing**
- **Processor virtualization (Intel VT-x, VT-x2, AMD SVM)**
- **Security enhancements: LT & Presidio**
- **Paravirtualization (software isolation approach)**
- **I/O Virtualization (AMD, Intel VT-d)**
- **Hypervisor Landscape**

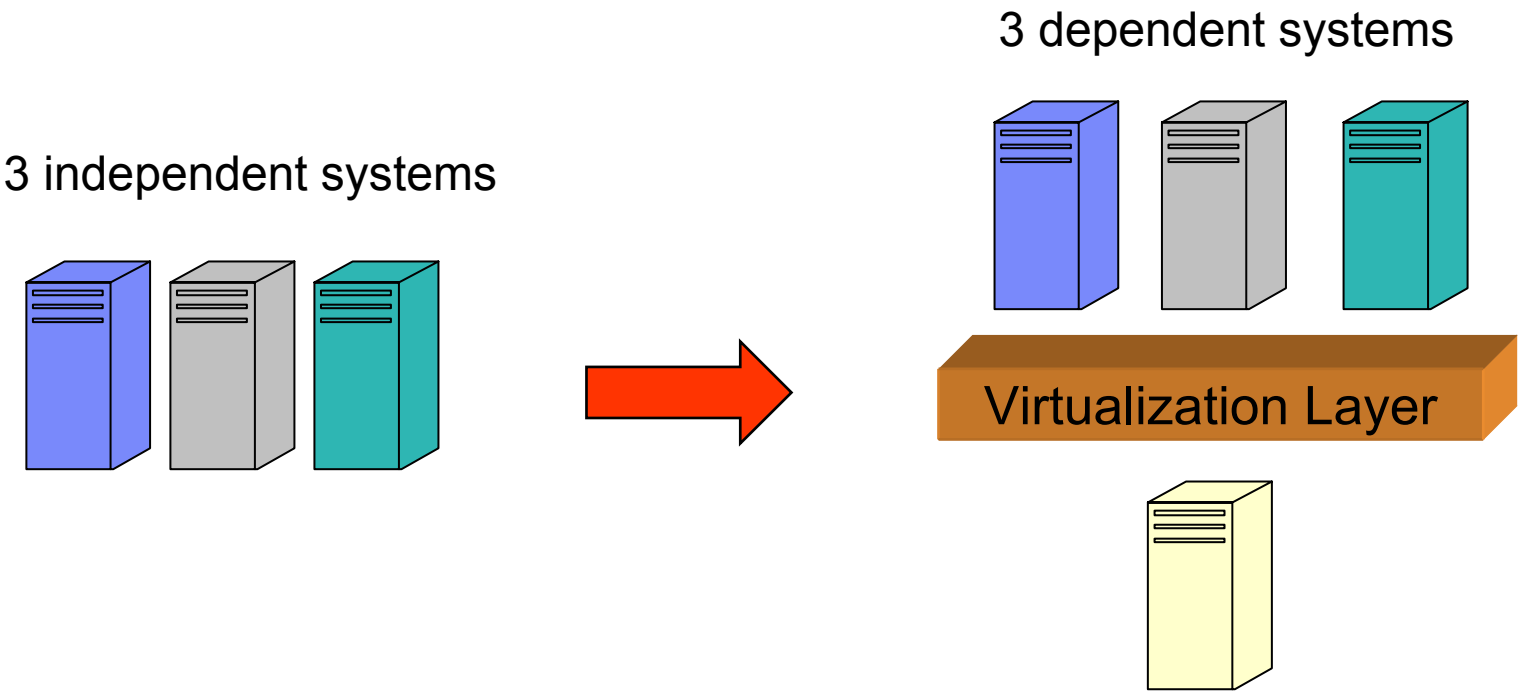
Talk is based on publicly available information

Virtualization In Servers

- **Reduce total cost of ownership (TCO)**
  - Increased systems utilization (current servers have less than 10% utilization)
  - Reduce hardware (25% of the TCO)
  - Space, electricity, cooling (50% of the operating cost of a data center)
- **Increase server utilization**
- **Management simplification**
  - Dynamic provisioning
  - Workload management/isolation
  - Virtual machine migration
  - Reconfiguration
- **Better security**
- **Legacy compatibility**
- **Virtualization protects IT investment**



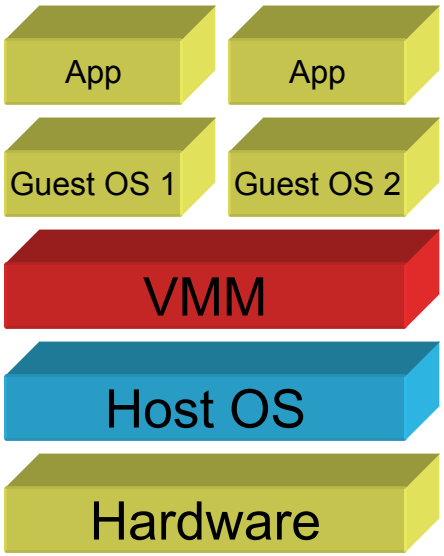
Virtualization is not a Panacea



- **Increasing utilization through consolidation decreases the reliability**
  - Need better hardware reliability (increased MTBF), error reporting, and fault tolerance
  - Need better software fault isolation

Virtual Machine Monitor Approaches

Type 2 VMM

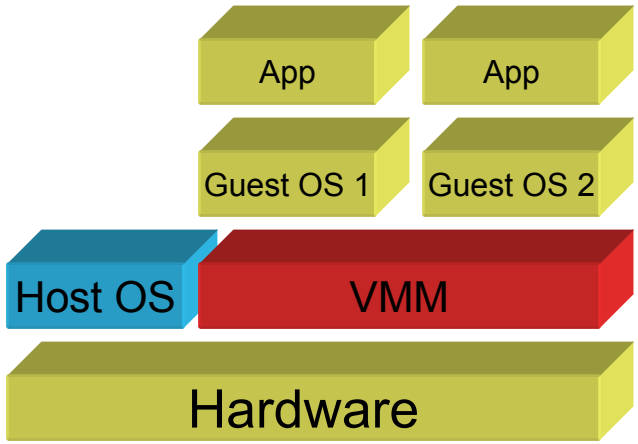


JVM

CLR

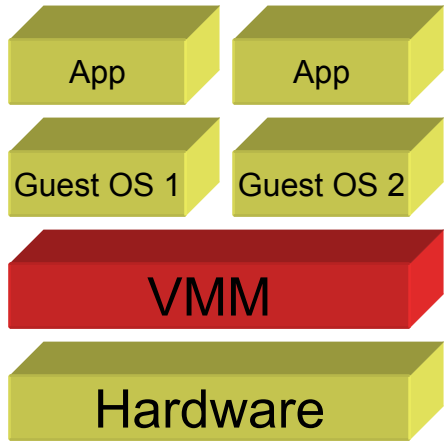
VMware Workstation

Hybrid VMM



MS Virtual Server

Type 1 VMM

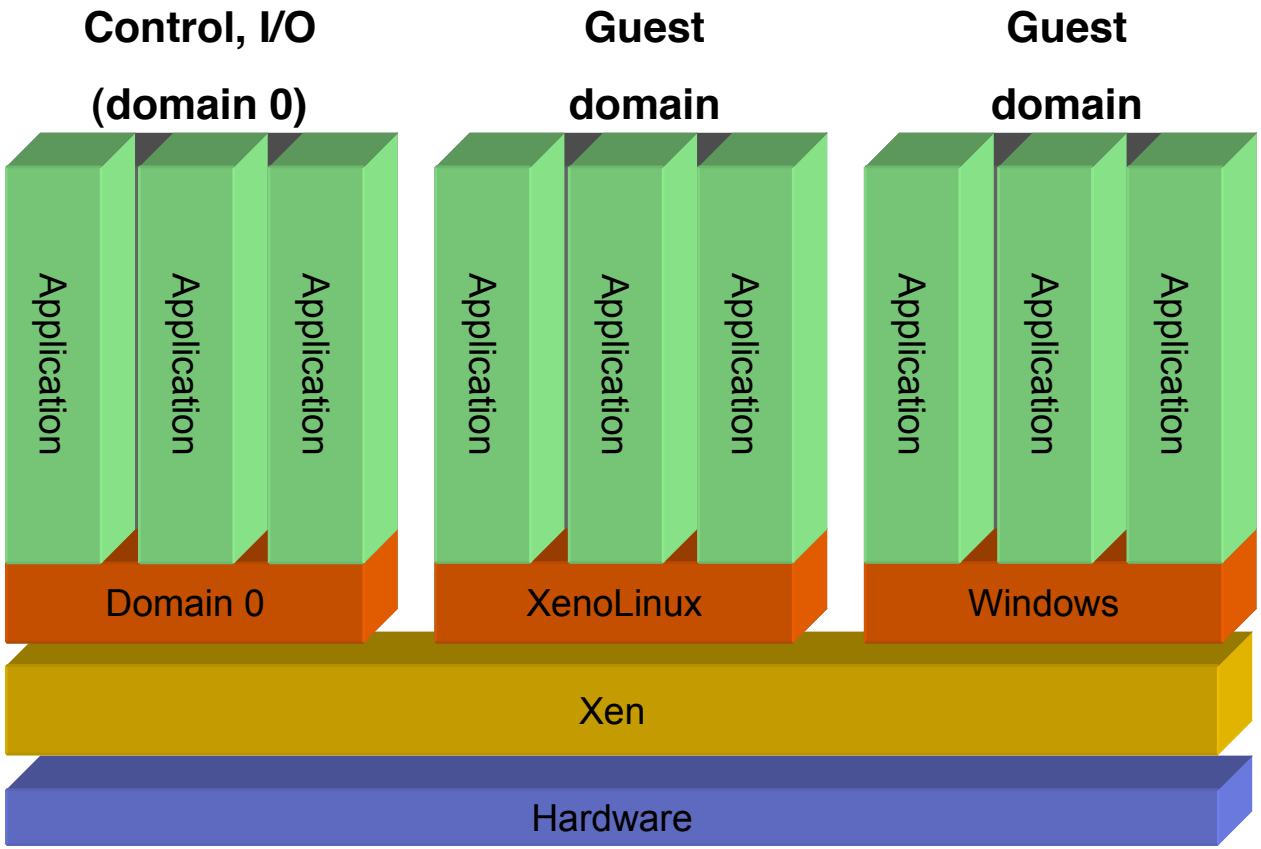


VMware ESX

Xen

MS Viridian

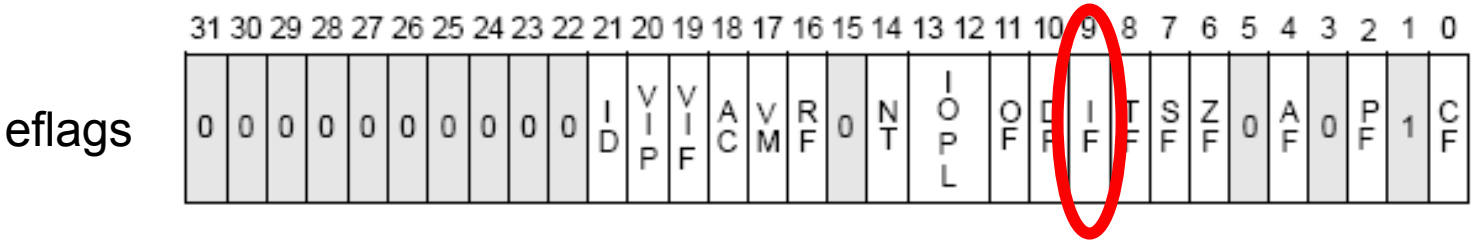
Example: Xen System Architecture



- **VMM and Microkernel are akin**

x86 Virtualization Problem

- **VMM needs to intercept the privileged instructions that are executed in the guest OS**
- **Traditional x86 architecture is not virtualizable**
  - POPF: pop value from stack, set flags



  - If in privileged mode: IF is set
  - If in non-privileged mode: IF is not set, no exception is raised

Virtualization Approaches

- **Full virtualization**
  - Binary rewriting
    - Inspect each basic block, rewrite privileged instructions
    - VMware, Virtual PC, qemu
  - Hardware assist (Intel VT-x, AMD SVM)
    - Conceptually, introduce a new CPU mode
    - Xen, VMware, MS Viridian
- **Paravirtualization**
  - Modify guest OS to cooperate with the VMM
  - Xen, VMware, L4, Denali

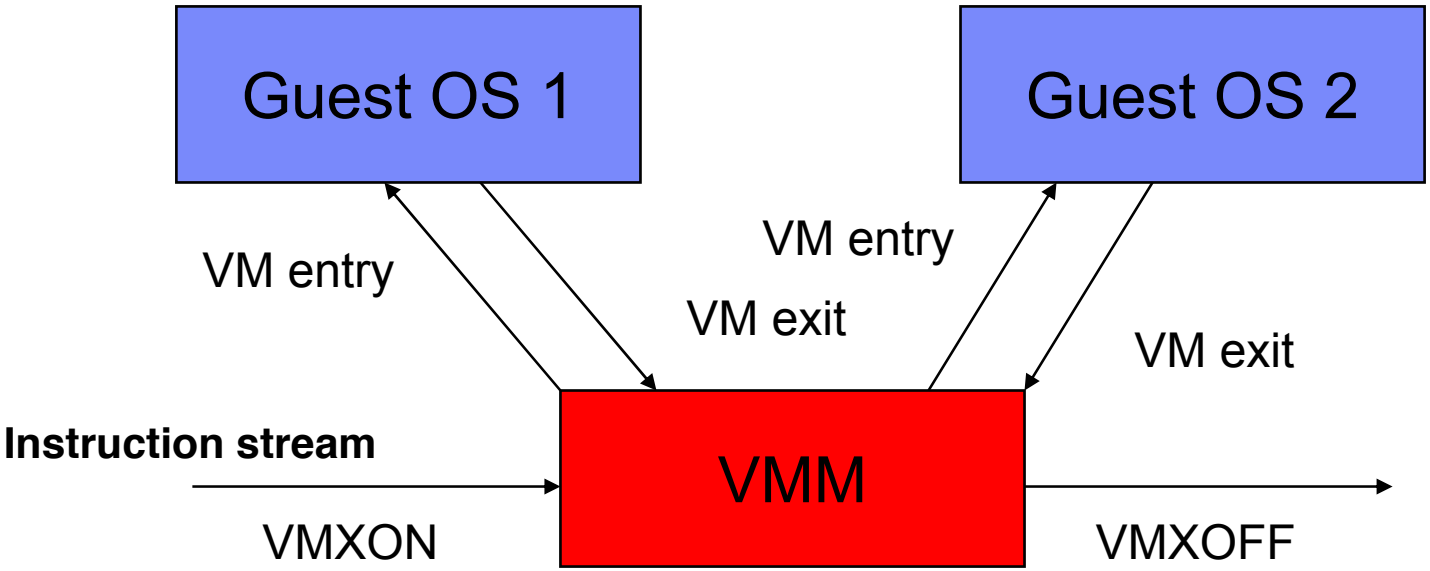
The World is Changing

- **Hardware virtualization support will become ubiquitous**
  - Every CPU will have (some) kind of virtualization support
  - Intel VT-x, AMD SVM (Pacifica), PowerPC, Cell
- **64-bit is here to stay**
- **Processor security features**
  - Dynamic root of trust (AMD SVM, Intel LT)
- **Massive multi/many core**
  - Computing cycles will become really cheap and the vendors cannot keep up
  - What do you do with 64 cores on a single socket? Reliability? Health monitoring?
- **I/O MMU**
  - Intel, AMD IOMMU, IBM Summit/Hurricane allow secure direct device access to partitions
  - PCI- Express will include virtualization support

Processor Virtualization Features

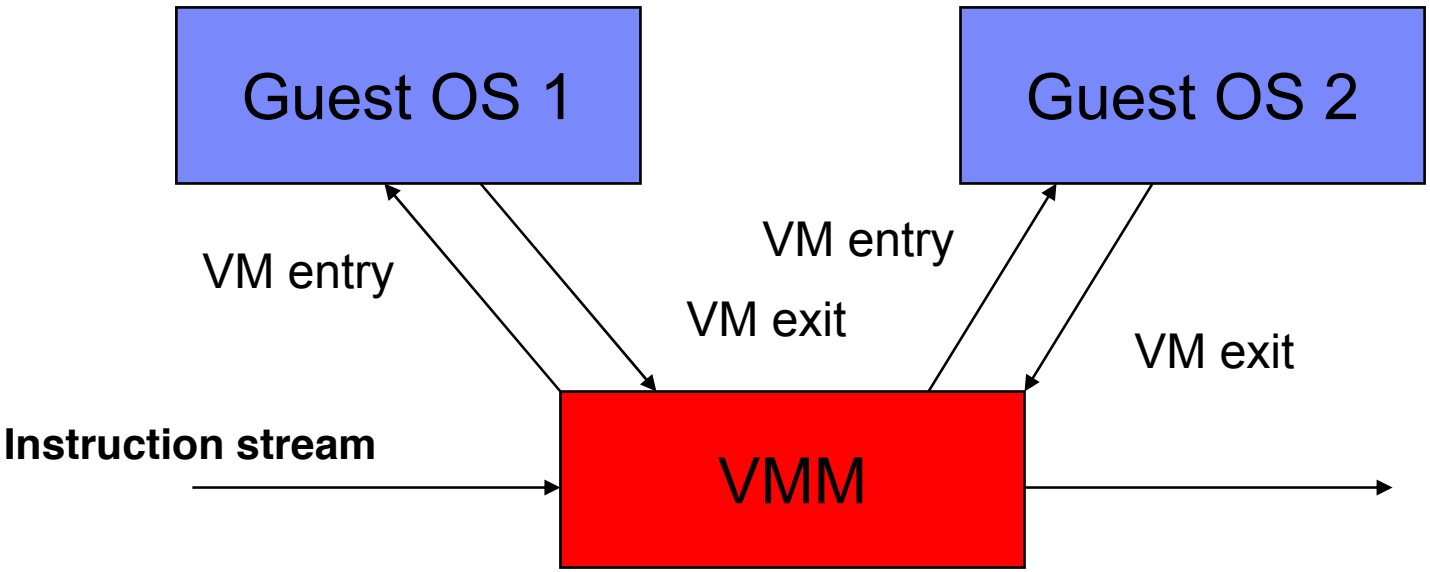
- **Both Intel and AMD defined processor extensions for their CPU architectures**
- **Intel: Vanderpool Technology (VT-x, VT-x2)**
- **AMD: Secure Virtual Machine (SVM, Pacifica), Rev F, Rev G, …**
- **From 10,000 ft. both look very similar:**
  - Container model (similar to mainframe SIE, *start interpretive execution*)

Intel Vanderpool Technology (VT)



- **VT-x adds new instructions such as VMXON, VMXOFF, VMLAUNCH, VMRESUME, VMCALL, …**
- **VM entry is caused by a VMLAUNCH or a VMRESUME**
- **Each guest has a VMCS (VM control segment) for its state**

AMD Secure Virtual Machine (SVM) Technology



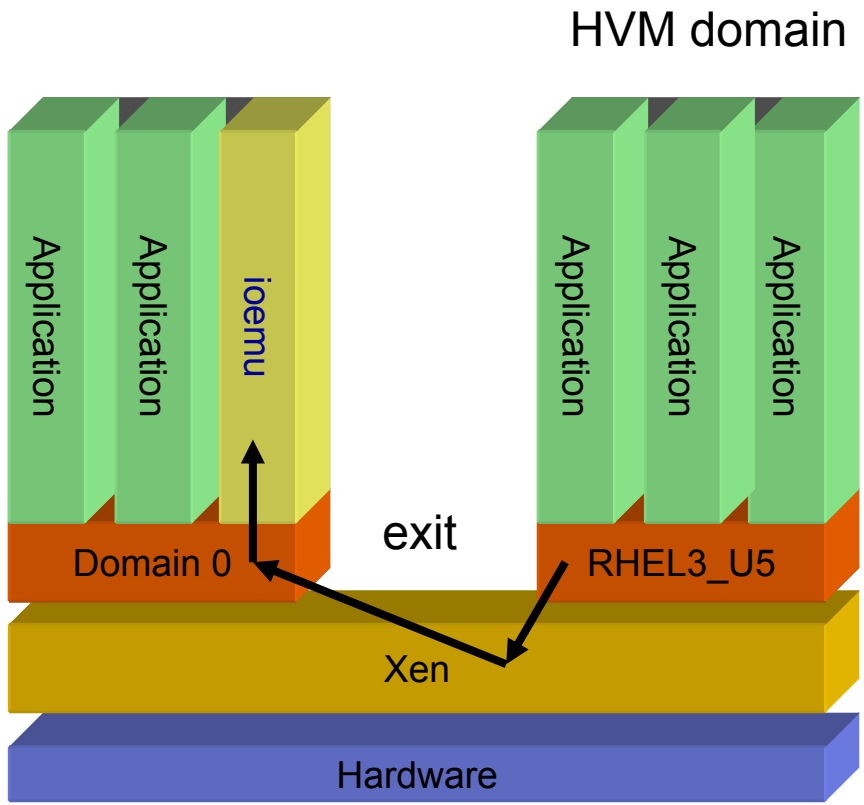
- **SVM adds the following new instructions: VMRUN, VMMCALL, …**
- **VM entry is caused by a VMRUN [rax]**
- **Each guest has a VMCB (VM control block) for its state**
- **Also known as Pacifica**

Intercepts and Exits

- **A guest runs until**
  - it performs an action that causes an exit
  - it executes a VMCALL/VMMCALL
- **Exit conditions are specified per guest:**
  - Exceptions (e.g., page faults) and interrupts
  - Instruction intercepts (CLTS, HLT, IN, OUT, INVLPG, MONITOR,MOV CR/DR, MWAIT, PAUSE, RDTSC …)
- **Intel VT-x has shadow registers**

Example: Full Virtualization Support for Xen

- **Most device emulation is implemented in ioemu (PCI, VGA, IDE, NE2100, …)**
- **High performance drivers, such as ioapic, lapic, vpit are implemented in Xen**
- **Developed by Intel, AMD and IBM**

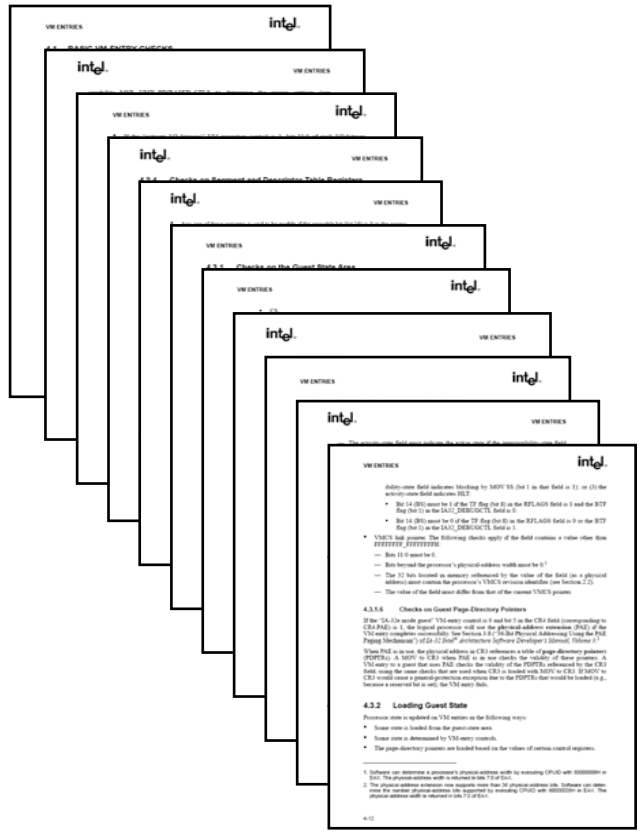


Example: Xen Implementation Statistics

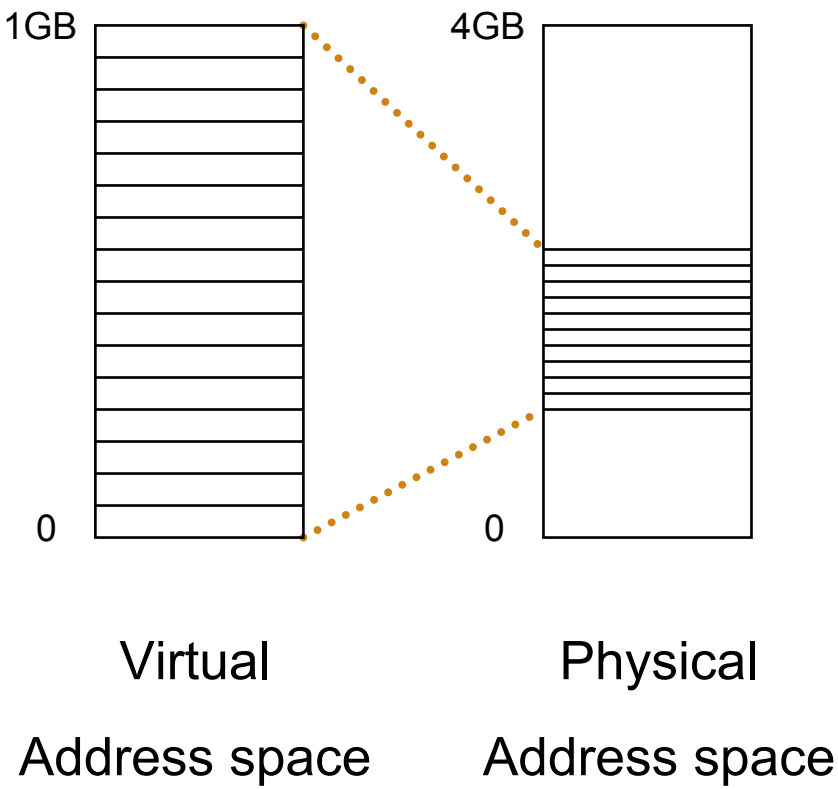
- **Lines of Code (C, assembly, headers):**
  - Xen: Intel VT-x specific code: 3718 (3.7%)
  - Xen: AMD SVM specific code: 5721 (5.6%)
  - Xen: Common HVM code: 5794 (5.7%)
  - Tools: Common HVM code: 86313 (85%)
- **Xen 3.0.2 contains both Intel VT-x and AMD SVM support**

The Cost Of VM Entry And Exit

- **VM entry and exits are very heavy weight and expensive operations**
- **VT-x specification has 11 pages of conditions that need to be checked just on a single VM entry!**
- **It is paramount to reduce number of exits**
  - Shadow page tables
  - Direct device assignment

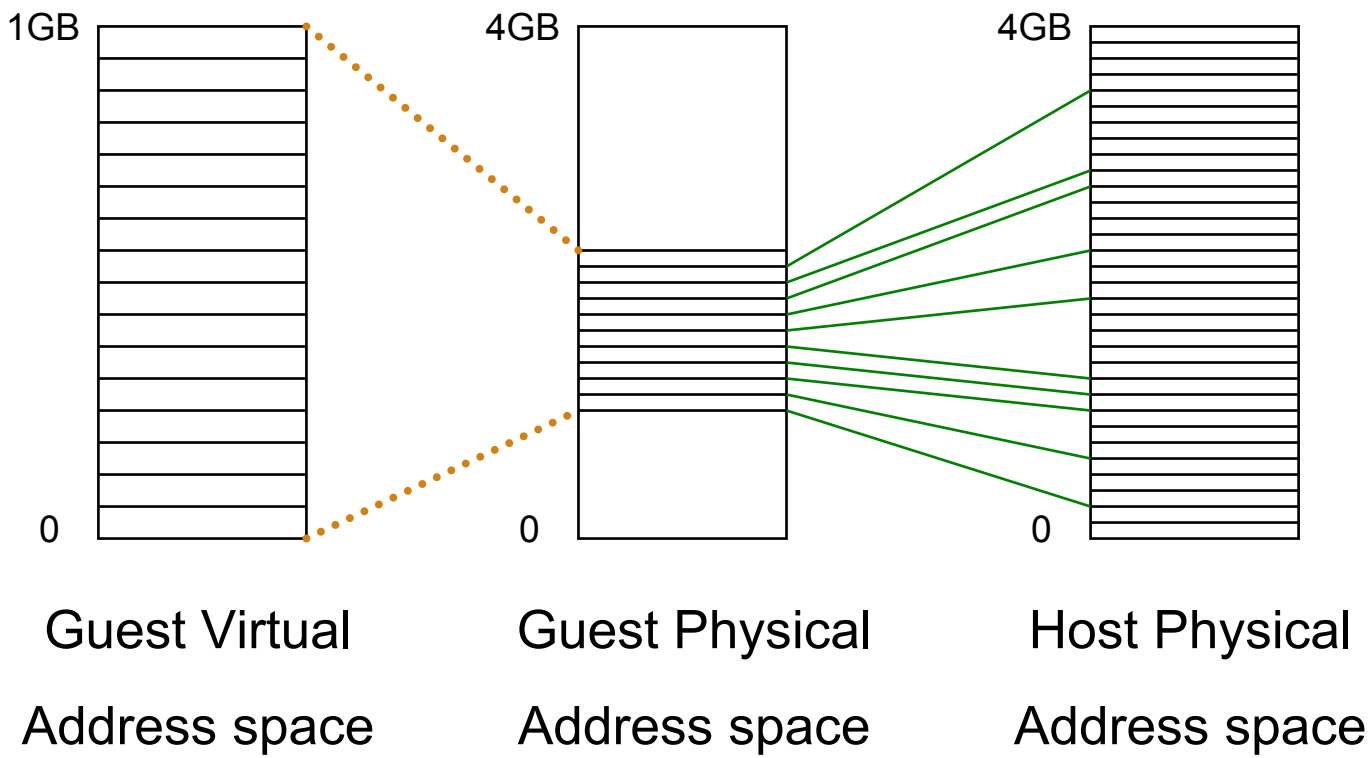


Traditional Virtual Memory Map

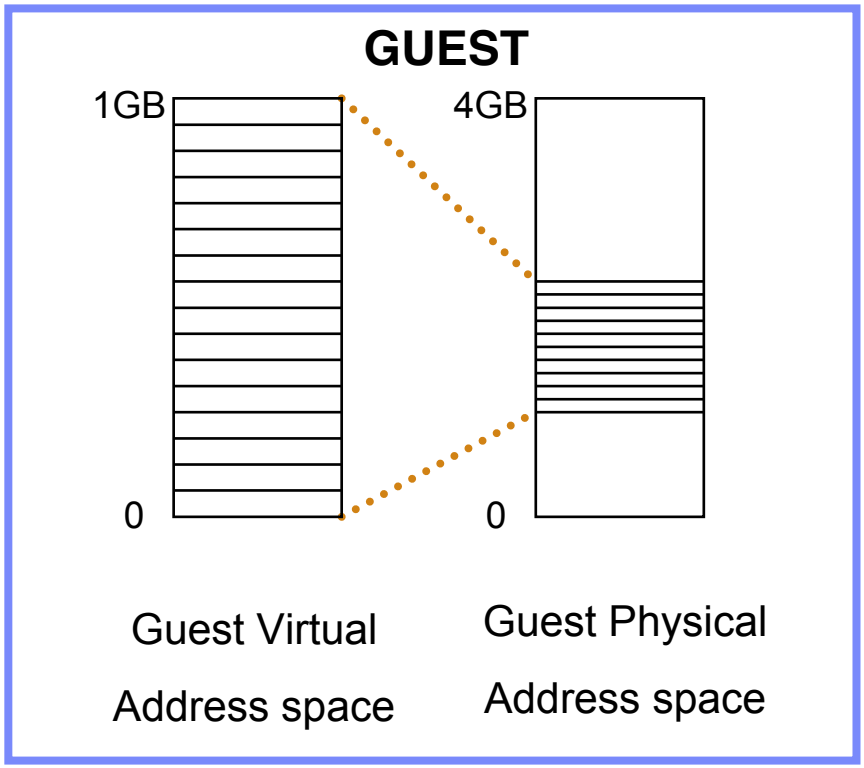


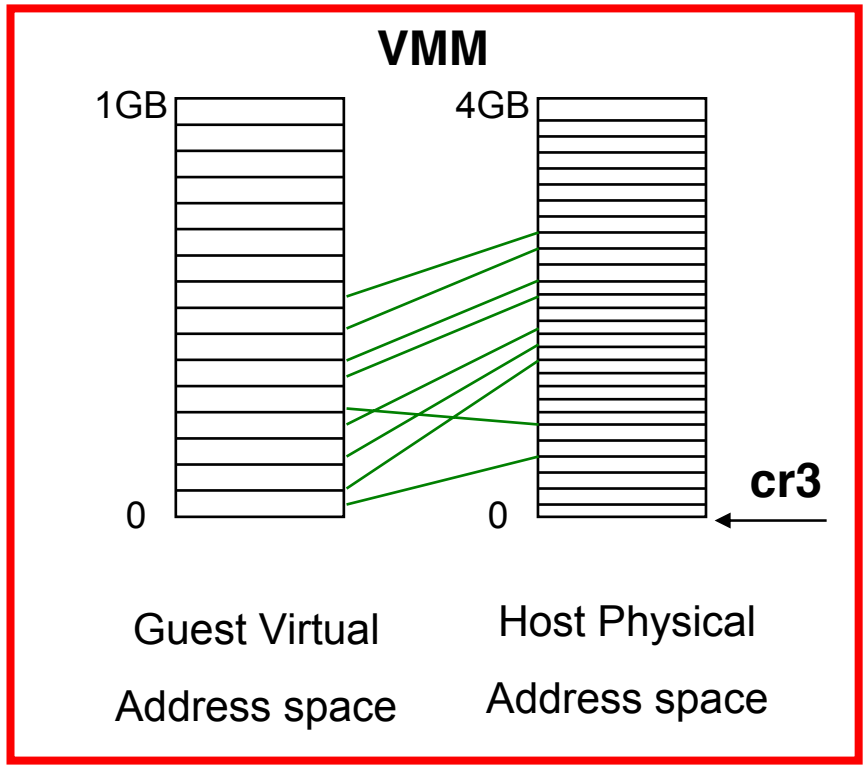
- **Virtual to physical translation (page table) is maintained by the OS**
- **The CPU walks the page tables automatically**
- **Page faults when page is not present or access violation**
- **CPU uses Translation-Lookaside-Buffer (TLB) to cache lookups**

Virtualized Memory Map



Shadow Page Tables



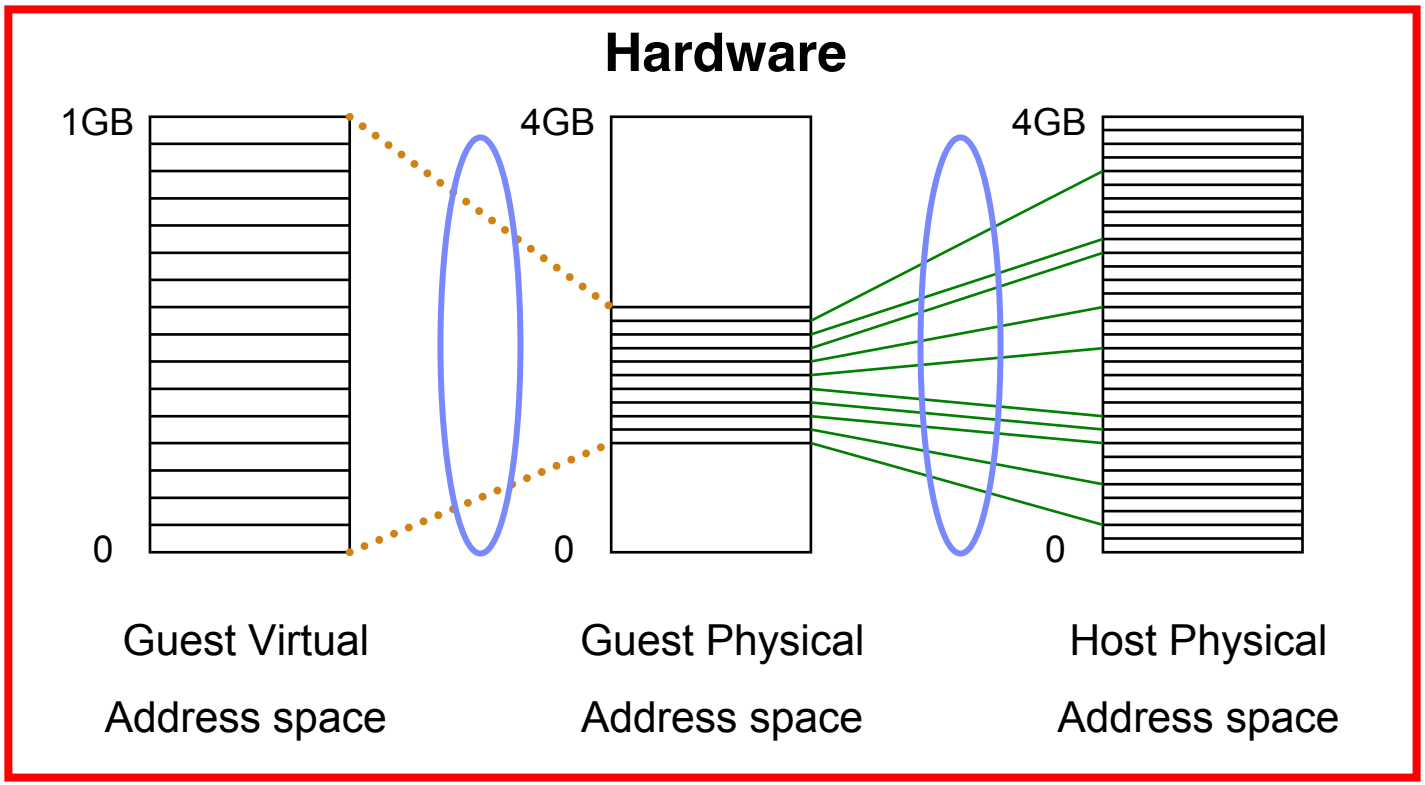


- **VMM maintains a shadow copy of the guest page table to translate from guest virtual to host physical**
- **Hardware only sees the shadow copy**

Shadow Page Table Issues

- **Managing the Shadow Page Table is expensive**
  - All page faults are handled by the VMM, it has to walk the guest page tables, and instantiate a shadow entry
  - VMM needs to propagate access and modify bits
    - A&M bits are used by the demand paging algorithms
    - The hardware modifies the shadow page table entry
    - VMM needs to emulate A&M behavior for the guest
    - May take up to 3 actual page faults per one guest page fault
- **Next generation virtualization extension will do this in hardware (Intel VT-x2, AMD SVM)**

Double (Page Table) Walker



- **AMD Nested Page Table support**
- **Intel Extended Page Table (EPT)**

Processor Security Features

- **Both Intel and AMD are working on hardware security constructs to enhance their virtualization offerings**
- **Primarily client focused and driven by Microsoft's NGSCB designs**
- **These enhancements include processor modifications to support**
  - Isolation (VMM)
  - Trusted computing (TCG)
  - Trusted keyboard/graphics I/O
- **Intel Lagrande Technology**
- **AMD Presidio Technology**

AMD SVM Dynamic Root of Trust

- **New instruction SKINIT that essentially reboots the CPU into a known state**
  - Start a 64KB secure loader
  - Interrupts disabled and other processors idled
  - Inhibit DMA to the secure loader memory area
  - Measurement of the secure loader is stored in the TCG trusted platform module (a secure passive storage device)
- **Measurement can be attested to by a remote party**

Intel VT-x / AMD SVM Comparison

Intel VT-x (2005)

- **VMENTER, VMRESUME, VMREAD, VMWRITE**
- **VMCS – VM control segment**

Intel VT-x2 (200?)

- **Extended Page Tables (EPT)**

Intel LT (200?)

- **SENTER (security)**
- **DMA exclusion vector (security)**

AMD SVM (2006)

- **VMRUN**
- **VMCB – VM control block**
- **ASID tagged TLB (performance)**
- **Paged realmode**
- **SKINIT (security)**
- **DMA exclusion vector (security)**

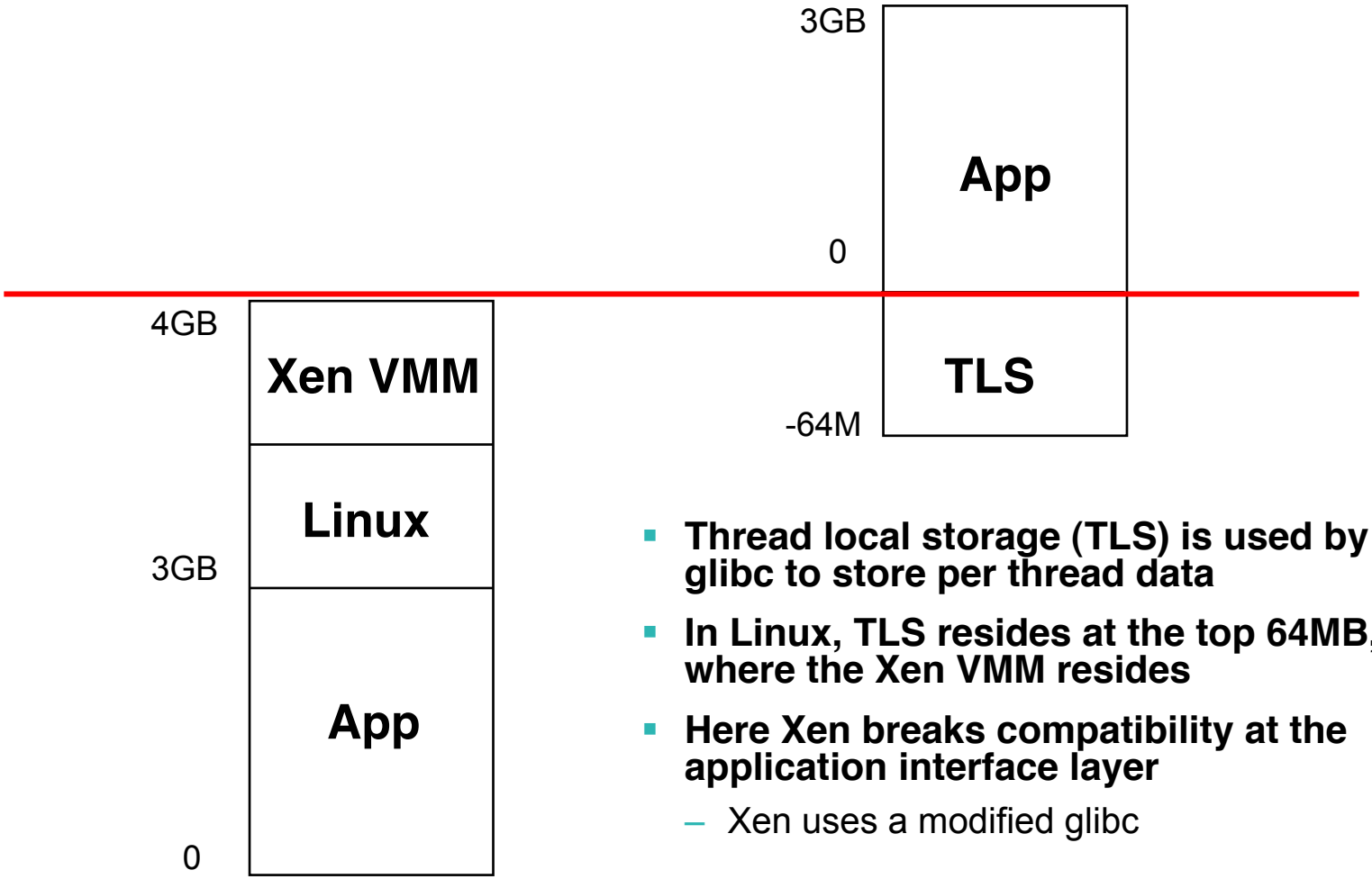
AMD SVM (2007)

- **Nested paging (performance)**

Paravirtualization

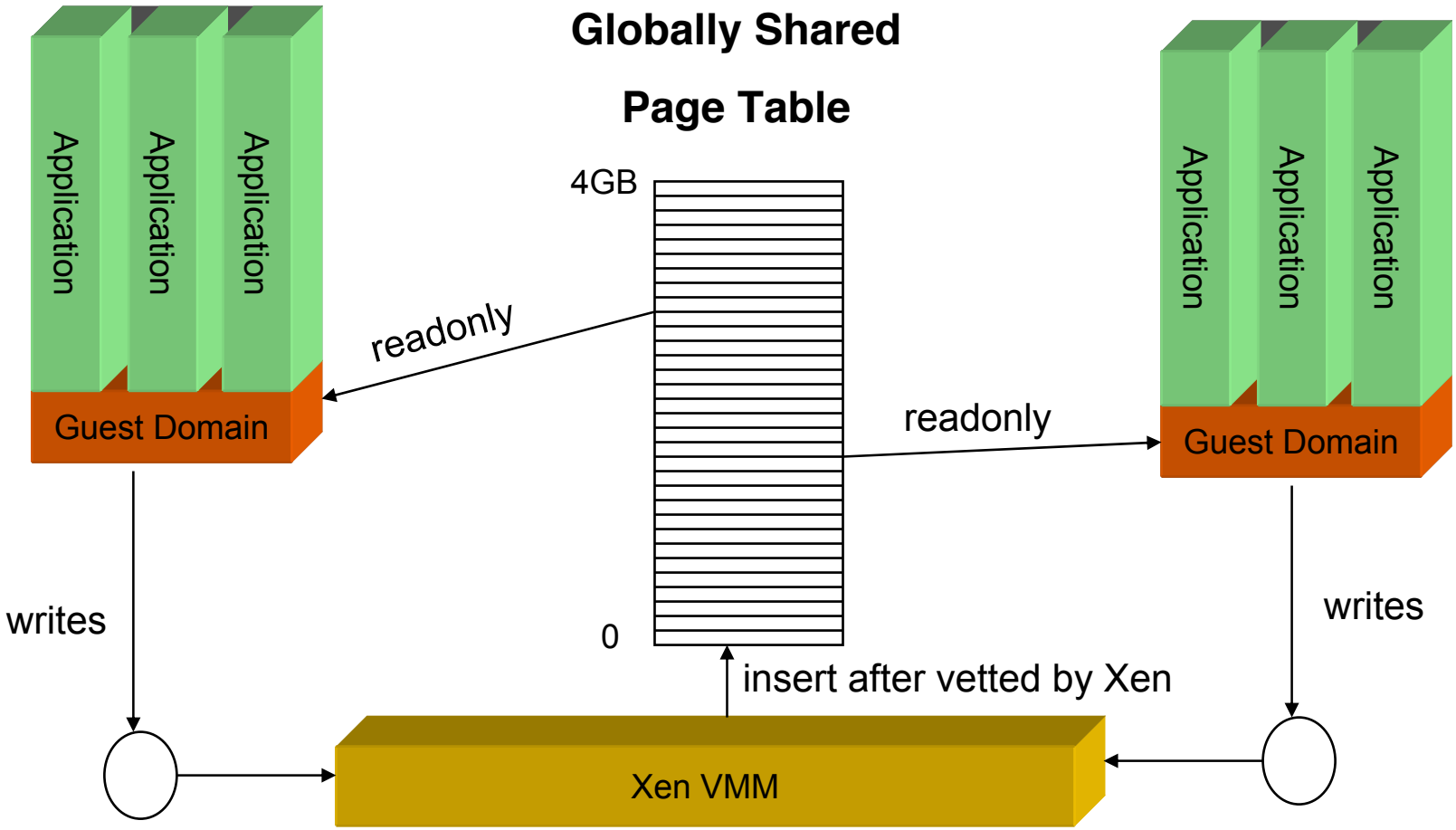
- **What do you do when you don't have hardware support?**
- **Modify guest kernel to cooperate with the hypervisor**
  - Introduce hypercalls
  - All privileged instructions become hypercalls
  - such as: mov-to-cr3, sti, cli, etc.
- **All guest applications & libraries are unmodified**
- **This is the approach IBM i/p-Series took, but also Xen, Microsoft's Viridian, L4, Denali and Virtual Iron**
  - VMware has now proposed a Virtual Machine Interface (VMI)
- **Paravirtualization has the potential to obviate the need for all the CPU virtualization extensions**
  - Not quite, certain features such as Thread Local Storage are hard to do in a paravirtualized environment

Example: Thread Local Storage



- **Thread local storage (TLS) is used by glibc to store per thread data**
- **In Linux, TLS resides at the top 64MB, where the Xen VMM resides**
- **Here Xen breaks compatibility at the application interface layer**
  - Xen uses a modified glibc

Example: Xen Writable Page Tables



VMware's Virtual Machine Interface

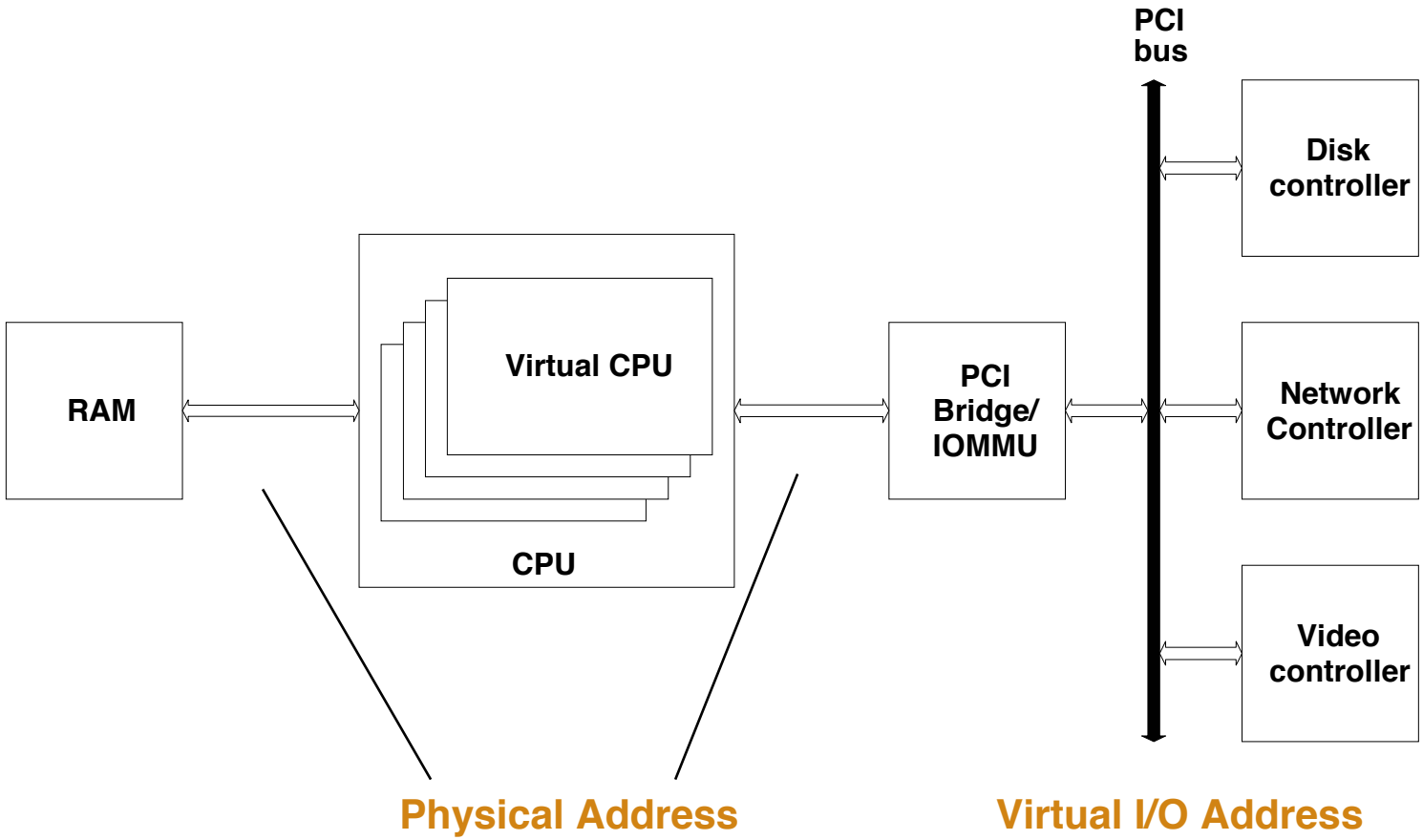
CPU state	**Interupt state, interrupt mask, halt, reboot, …**
CPU descriptors	**Gdt, idt, ldt, tr, …**
CPU control	**Read/write MSR, CR0..4, …**
CPU information	**Cpuid, read/write tsc, pmc, …**
Stack/privilege transitions	**Update kernel stack, iret, sysexit**
I/O	**In/out, delay, IOPL, invd**
APIC	**Read/write**
Timer	**Get wall clock, frequency, cycles, set alarm, …**
MMU	**Set linear mapping, flush TLB, invalidate, set PTE, swap PTE, …**

CPU Virtualization Techniques Comparison

	Performance	Legacy guest support	VMM complexity
Binary rewriting	**medium**	**yes**	**high**
paravirtualization	**high**	**no**	**medium**
Hardware assist (current gen)	**low**	**yes**	**medium-low**
Hardware assist (next gen)	**medium**	**yes**	**medium-low**

low medium high

Typical System Architecture

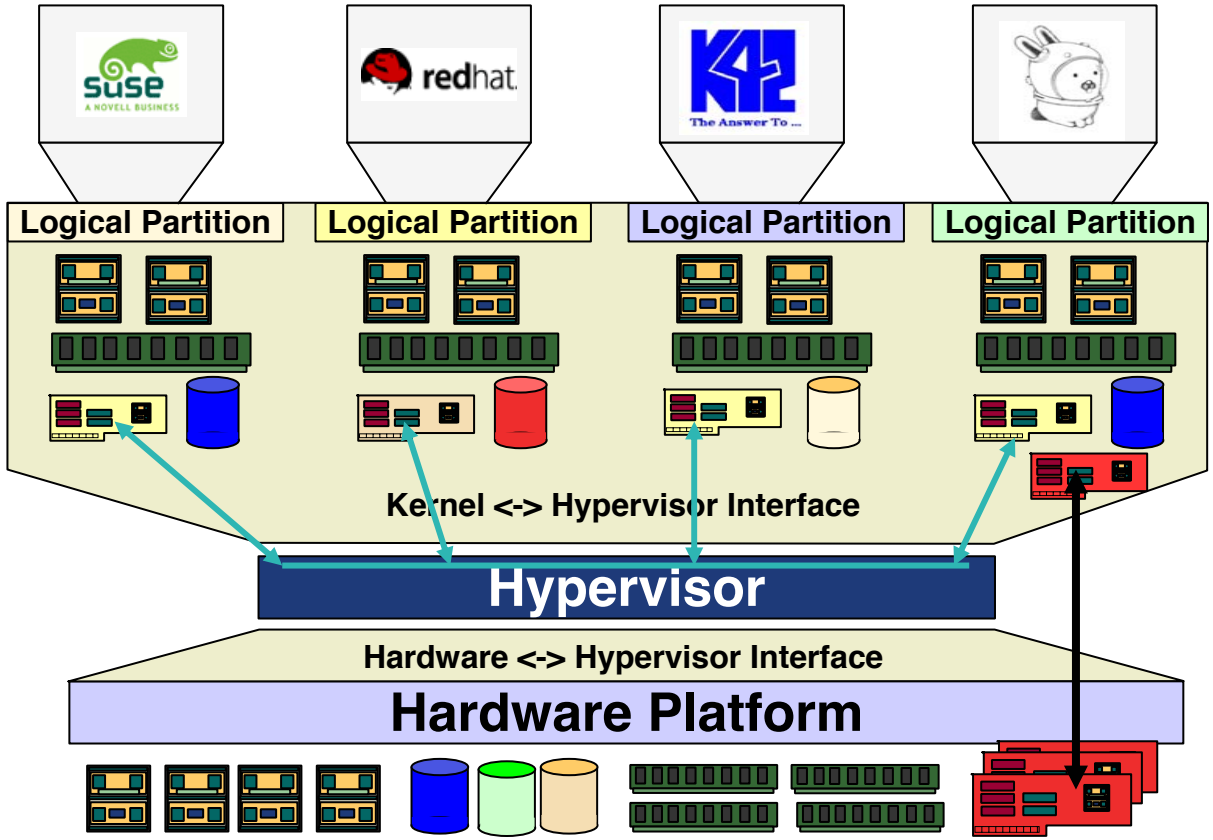


I/O Memory Management

- **I/O MMU translates I/O virtual addresses to host physical addresses**
- **Main functions**
  - Memory isolation
  - Address remapping (guest PA == I/O VA for devices assigned to guest)
  - Eliminate bounce buffers (32-bit devices into 64-bit PCI space)
- **Protect system against BUSMASTER devices**
- **AMD disclosed their design in January 2006, Intel followed in March 2006 (VT-d)**
  - Again, from a 10,000 ft. view, both are very similar
- **IBM has the same Summit IOMMU in p/i/xSeries**

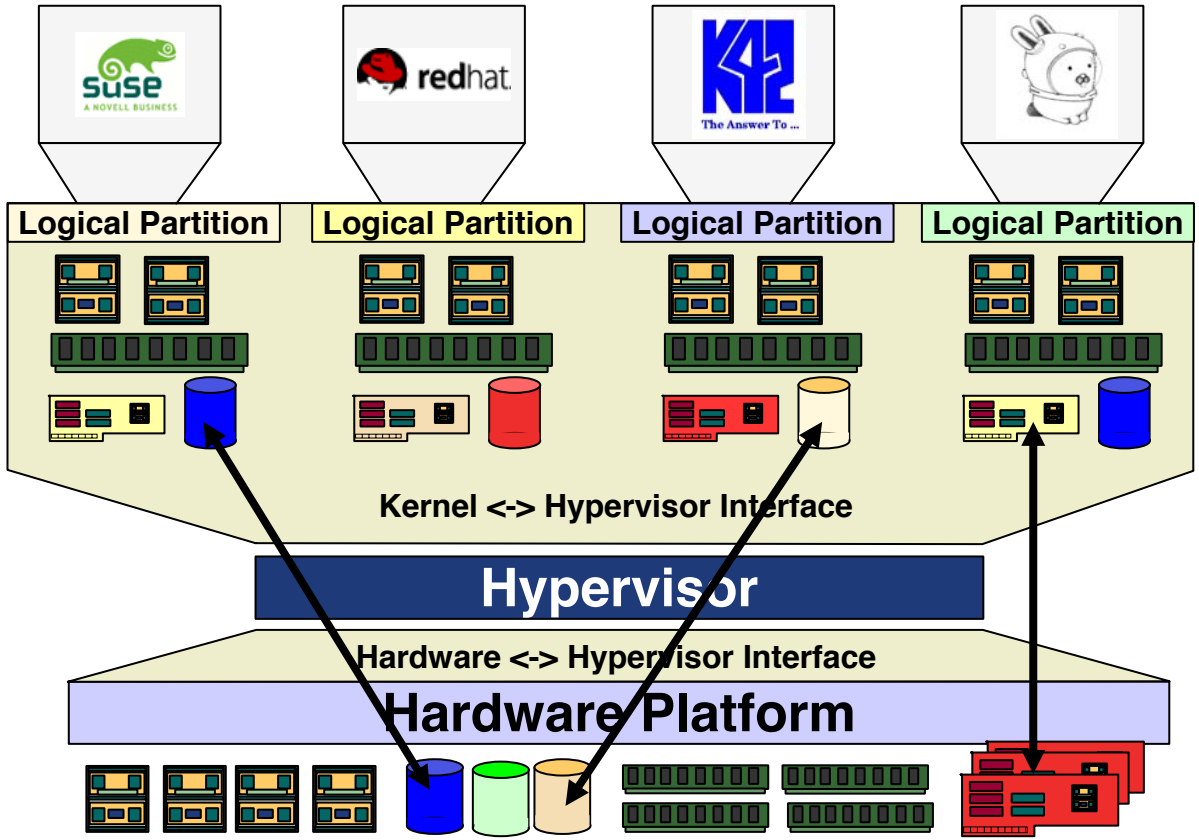
I/O Hosting Partition

- With I/O hosting domain/partition, all real drivers extracted from DomU.
- Can have multiple Device Domains to support different devices.
- Reasonable performance possible through batching, (page flipping)… (“Unmodified device driver reuse via virtual machines” OSDI04…)
- But performance is just not good enough to get rid of all native devices.



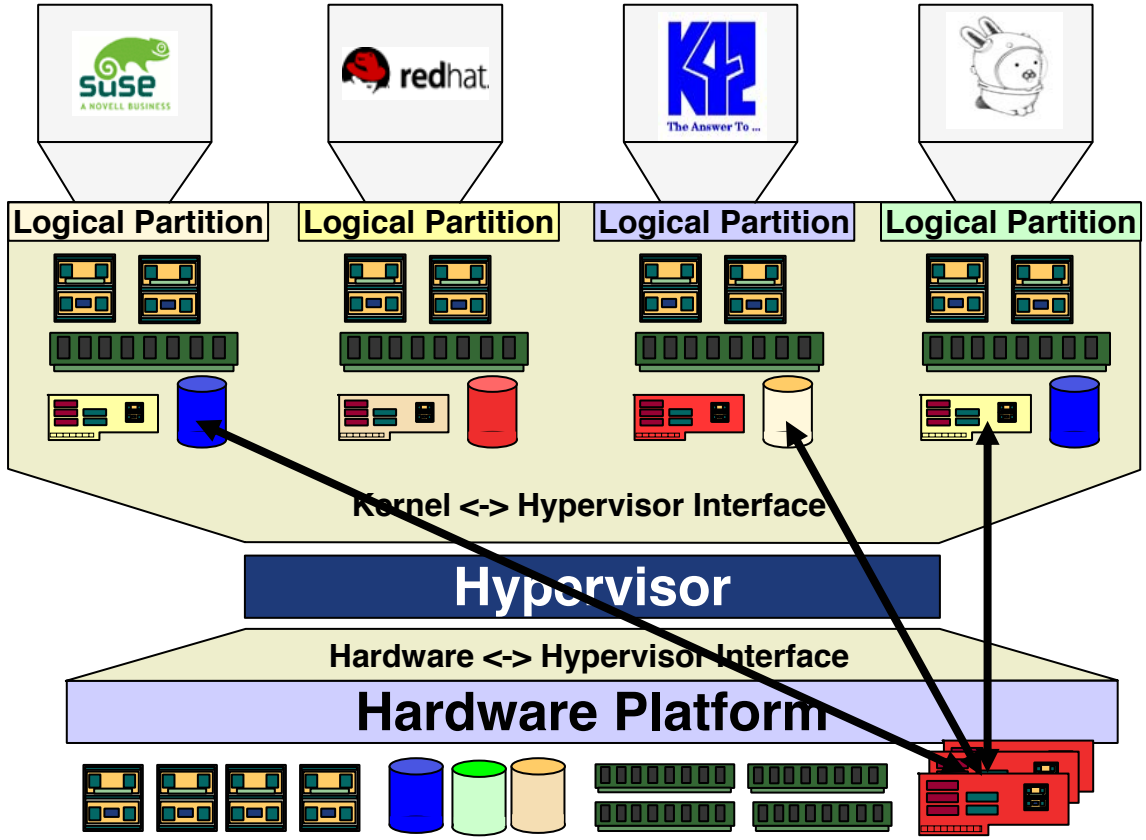
Direct Device Assignment

- With IOMMU can directly give partitions control over Bus/Dev/Func.
- Same HW results in improved reliability.
- With right HW can support fully virtualized OS (e.g., windows).
- Clean support for legacy OSes and for highest performance devices (majority probably still virtualized)
- Migration becomes impossible.
- DD in each OS

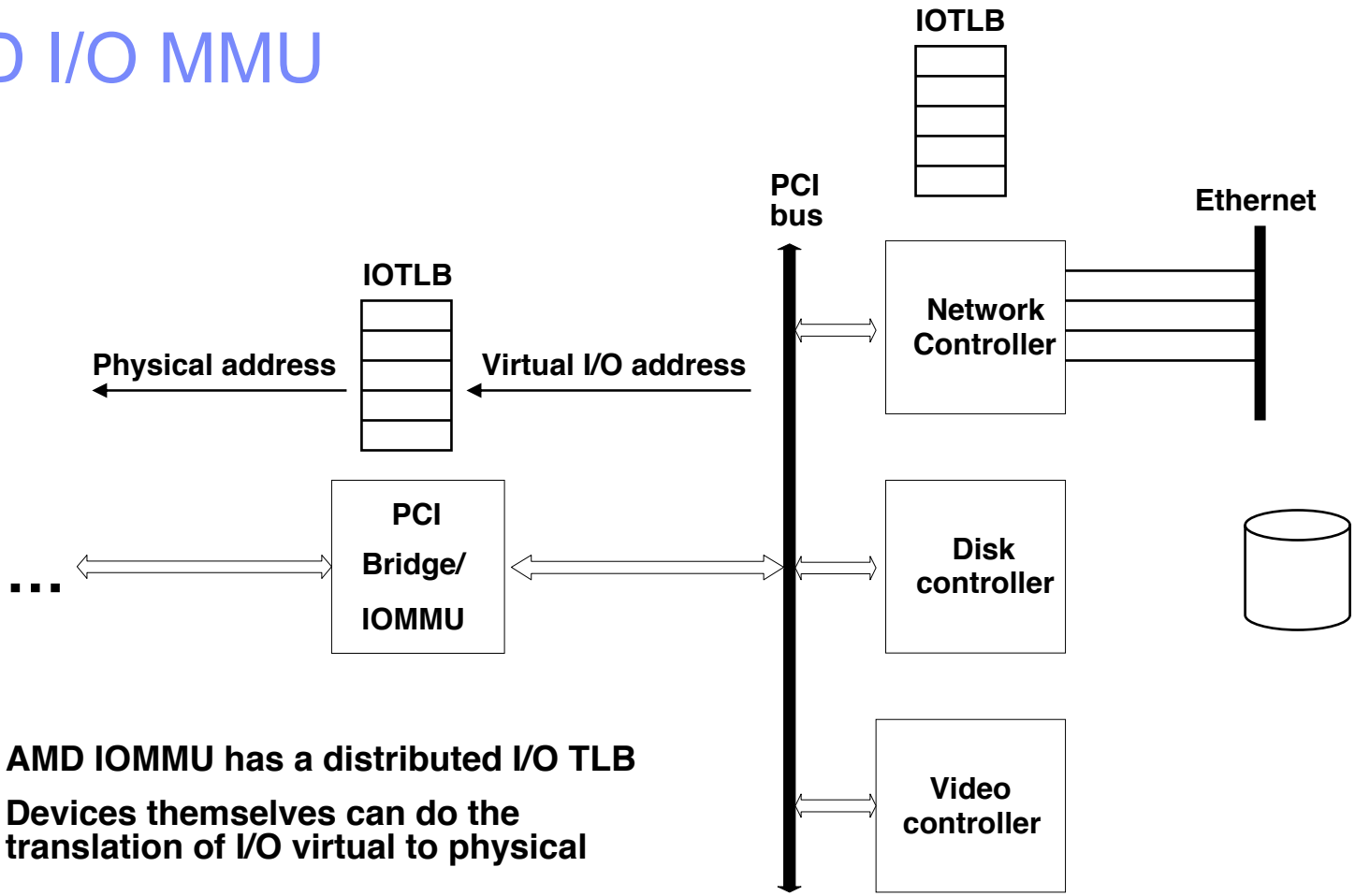


Self Virtualizing Devices

- Self virtualizing devices allow direct access by partition, e.g., infiniband.
- No overhead (throughput, latency, serialization) to context switch to device domain.
- Is exception for high performance devices… no migration.
- DD in each OS



AMD I/O MMU



- AMD IOMMU has a distributed I/O TLB
- Devices themselves can do the translation of I/O virtual to physical

Physical Address

Virtual I/O Address

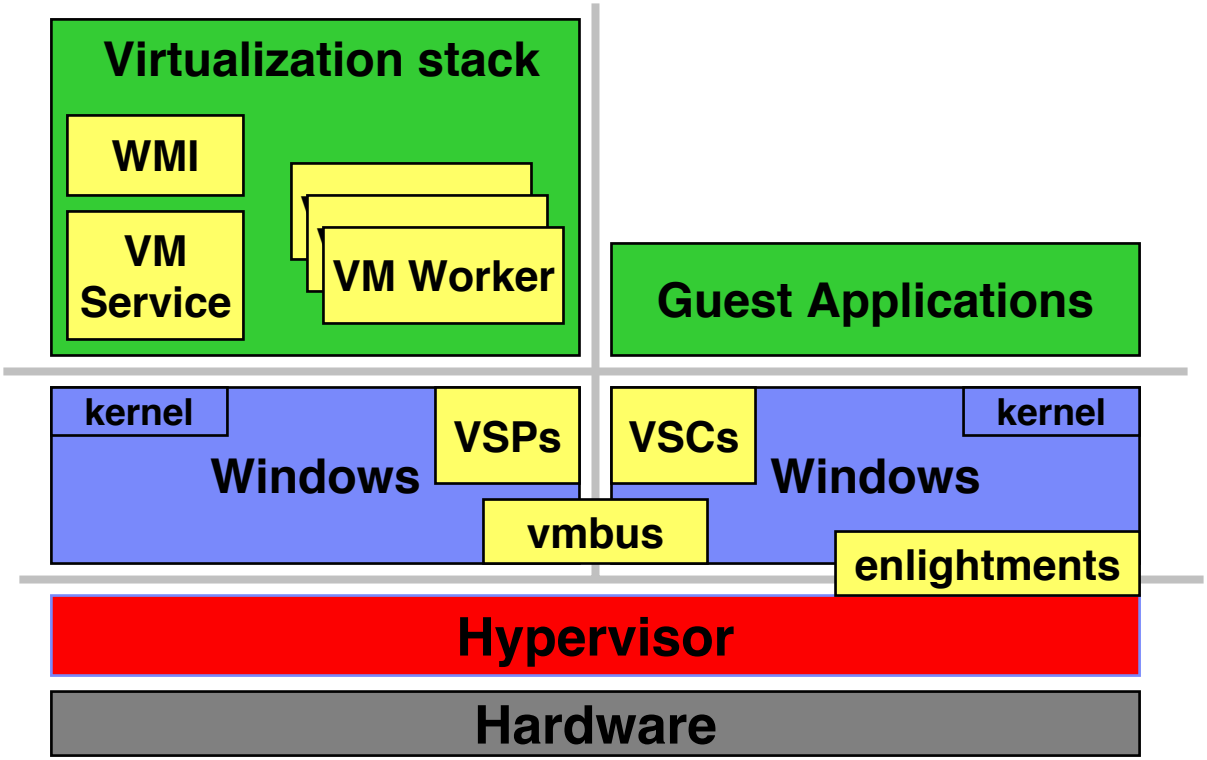
Partition Migration

- **IOMMU provides isolation and remapping for direct device assignment**
- **The current generation of IOMMU proposals do not handle partition migration**
  - Problem: PCI has no ability to restart a request and does not notify failed PCI requests in all cases
- **PCI SIG is working on adding support for this**

Hypervisor Landscape

- **VMware is the undisputed leader in the x86 virtualization space**
  - Its binary translation technology is superior
  - Only uses VT-x in x86-64 because unlike AMD, Intel does not provide segment limits
  - Very mature product
- **Xen is an open source hypervisor shipped as part of RedHat and Suse Linux**
  - Uses paravirtualization for Linux
  - VT-x/SVM for unmodified guest OS support
  - Immature product
- **And then there is the big unknown …**

Microsoft Viridian



- Will be released after LongHorn Server (end 2007)
- Will run Windows and Linux
- Uses Intel VT-x, AMD SVM, and paravirtualization (enlightments)

Summary

- **Technology Trends**
  - Hardware virtualization support will become ubiquitous
  - 64-bit is here to stay
  - Processor security features
  - Massive multi/many core
  - I/O MMU
- **Challenges**
  - First generation virtual-ization support is just a technology preview
  - Increase system reliability
    - Hardware fault tolerance
    - Software robustness
  - Management of virtual environments

Sources

- **AMD SVM specification,** http://www.amd.com/us-en/assets/content_type/white_papers_and_tech_docs/24594.pdf
- **AMD SVM availability of Nested Paging,** http://www.amd.com/us-en/assets/content_type/DownloadableAssets/JoeMenardAMDAnalystDayv2.pdf
- **AMD IOMMU specification,** http://www.amd.com/us-en/assets/content_type/white_papers_and_tech_docs/34434.pdf
- **Intel VT-x and VT-d specification,** http://www.intel.com/technology/computing/vptech/
- **Intel VT-d release date (2007),** http://www.virtualization.info/2006/03/intel-working-on-new-virtualization.html
- **Intel LT specification,** http://www.intel.com/technology/security/
- **Merom, Conroe details,** http://en.wikipedia.org/wiki/List_of_Intel_Core_2_microprocessors
- **Microsoft Viridian,** http://h40132.www4.hp.com/upload/ch/de/MicrosoftVS2005R2HPEvent.pdf, http://download.microsoft.com/download/4/1/e/41e56f34-8f90-405e-9daf-f8aeea249935/InTrack14dec_Presentatie_clean.ppt#35
- **VMWARE and CPU virtualization technology,** http://download3.vmware.com/vmworld/2005/pac346.pdf
- **VMWARE Virtual machine Interface,** http://www.vmware.com/pdf/vmi_specs.pdf

BACKUP

Code Word Soup

Intel VT-x, VT-x2	**Vanderpool (processor virtualization) Technology for x86 architecture**
Intel VT-i	**VT for IPF (Itanium)**
Intel VT-d	**Intel's IOMMU implementation**
Intel LT	**Intel's Lagrande (security) Technology**
AMD SVM (Pacifica)	**AMD's Secure Virtual Machine CPU extensions**
AMD Presidio	**AMD's Lagrande equivalent**

TCG Static Root of Trust