Virtualization

EECE6029

Yizong Cheng

2/24/2016

Virtual Machines

- We want to run different servers on separate computers.
  - “sandboxing”, achieving isolation and fault tolerance
  - not cheap
- Solution: create the illusion of multiple machines on the same physical hardware– Consolidate servers
  - virtual machine monitor (VMM) or type 1 hypervisor on bare hardware
  - type 2 hypervisor making use of an underlying operating system

Hypervisors

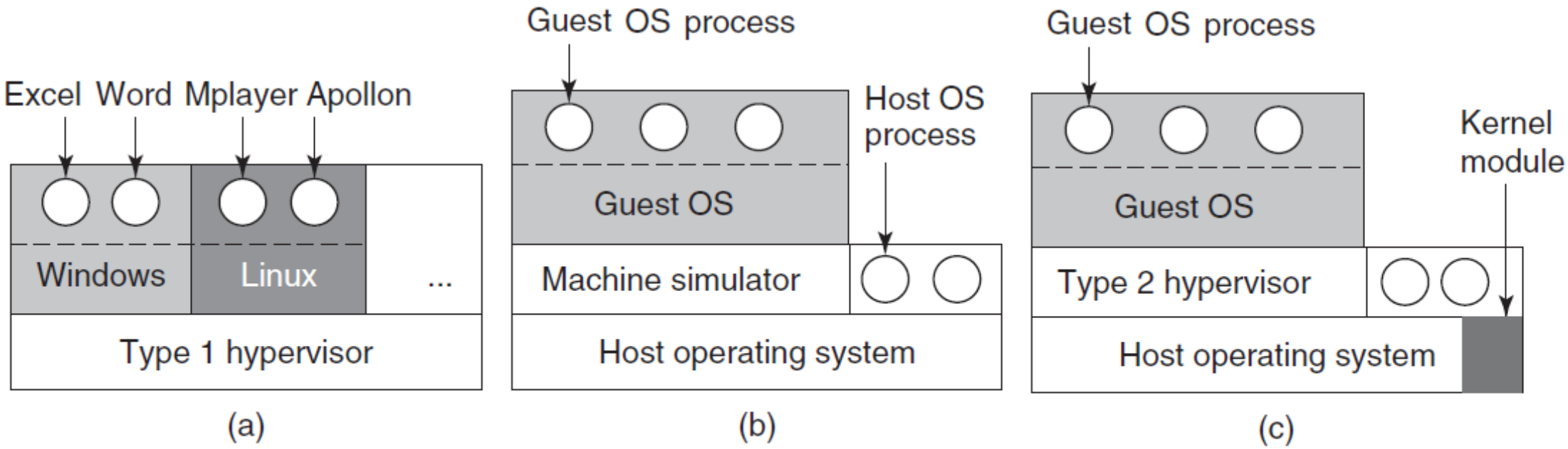

Figure 1-29. (a) A type 1 hypervisor. (b) A pure type 2 hypervisor. (c) A practical type 2 hypervisor.

History

- VM/370 (1979) virtual machine monitor (type 1 hypervisor) provides conversational monitor system (CMS) to remote users.
- Popek and Goldberg (1974), Formal requirements for virtualizable third generation architecture.
  - The requirement is that all sensitive instructions are privileged instructions.
  - x86 architecture fails the requirements. e.g. POPF is ignored when executed in user mode on Intel 386.
  - Intel started virtualizable CPU's (VT) and AMD with SVM after 2005 and then trap-and-emulate virtual machines are possible.
- Disco from Stanford (1990s) and then VMware, Xen, KVM, VirtualBox, Hyper-V, and Parallels.
  - binary translation is used.

Sensitive and Privileged Instructions

- Privileged instruction: one that causes a trap if executed in user mode.
  - In simpler language, if you try to do something in user mode that you should not be doing in user mode, the hardware should trap. -- Tanenbaum
  - A trap sends control to a trap handler, which is part of the operating system. – Knuth
- Sensitive instruction: one that behaves differently when executed in kernel mode than when executed in user mode.
- Popek and Goldberg (1974): A machine is virtualizable only if the sensitive instructions are a subset of the privileged instructions.

Type 1 and Type 2 Hypervisors

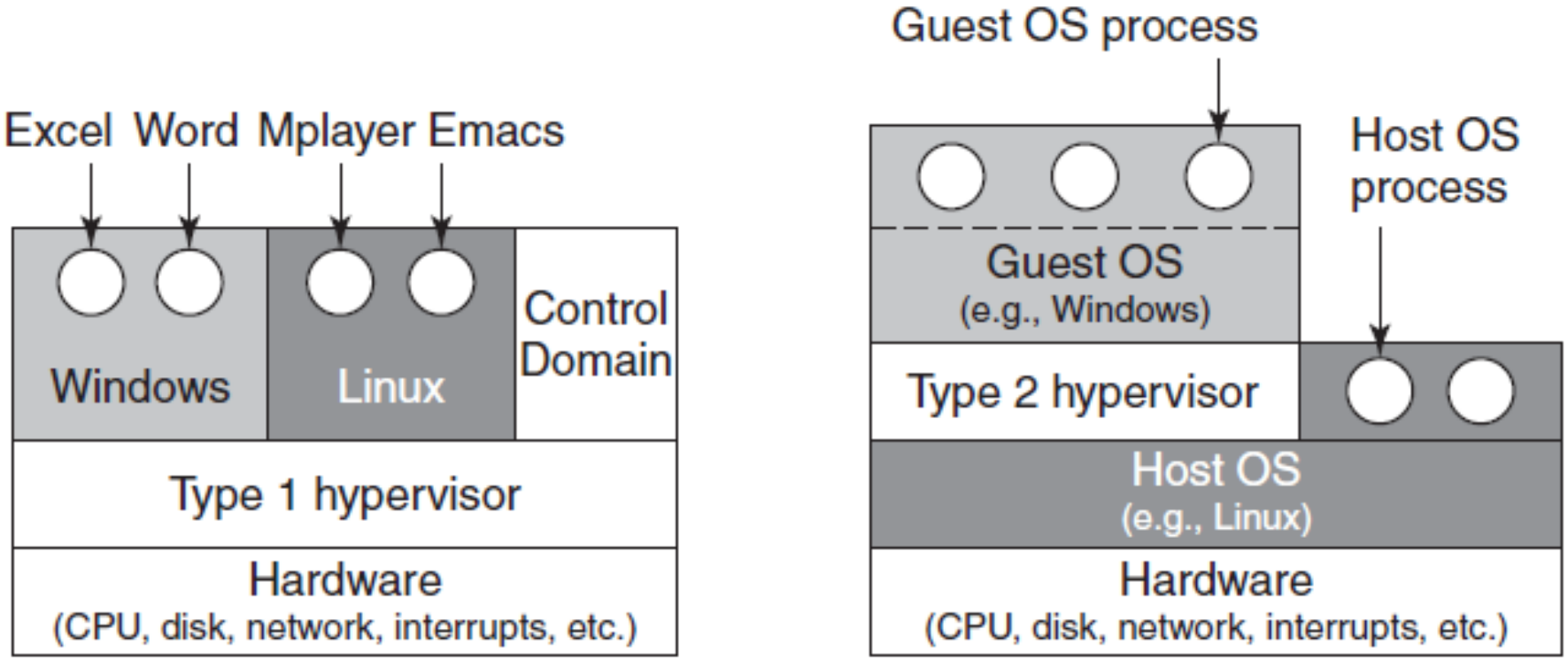

Figure 7-1. Location of type 1 and type 2 hypervisors.

Virtualization Methods

Virtualizaton method	Type 1 hypervisor	Type 2 hypervisor
Virtualization without HW support	ESX Server 1.0	VMware Workstation 1
Paravirtualization	Xen 1.0	
Virtualization with HW support	vSphere, Xen, Hyper-V	VMware Fusion, KVM, Parallels
Process virtualization		Wine

Figure 7-2. Examples of hypervisors. Type 1 hypervisors run on the bare metal whereas type 2 hypervisors use the services of an existing host operating system.

Virtual Kernel Mode

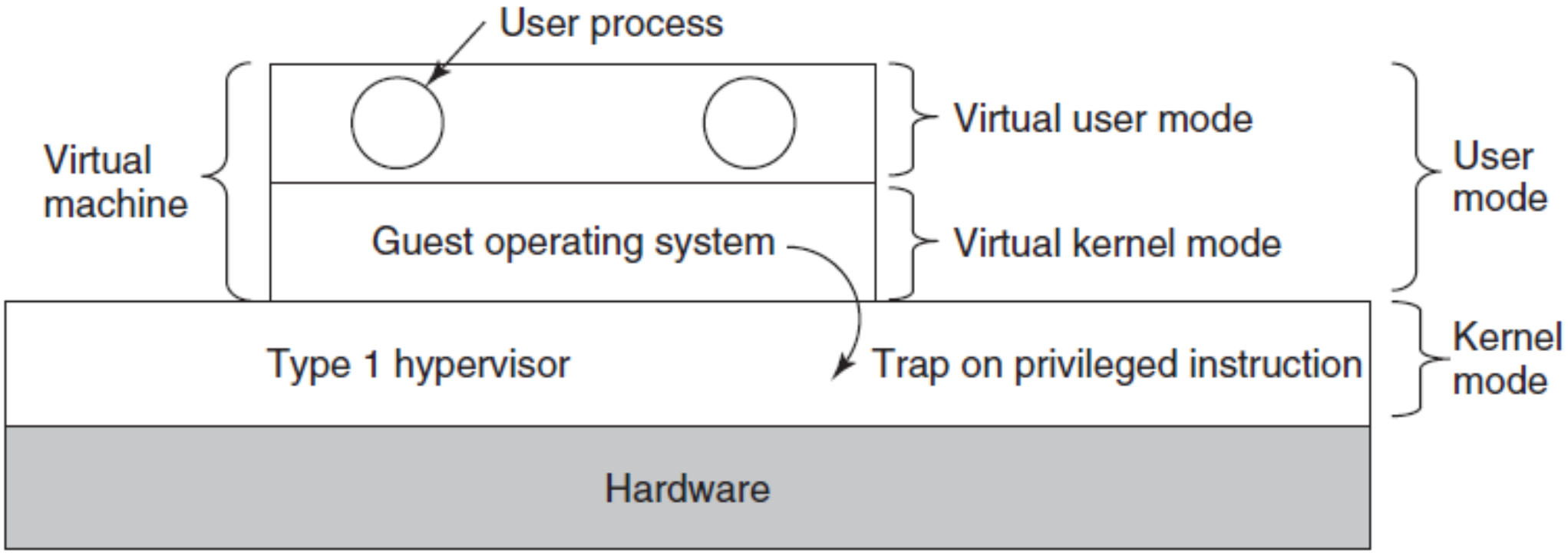

Figure 7-3. When the operating system in a virtual machine executes a kernel-only instruction, it traps to the hypervisor if virtualization technology is present.

Binary Translation

- Hypervisors should provide safety, fidelity, and efficiency.
- The safest method is a interpreter but it is not efficient.
- A basic block is a straight-line sequence of instructions that ends with a branch.
- Before execution, the hypervisor replaces sensitive instructions in a basic block with calls to handlers.
- Translated blocks are cached.

Binary Translation

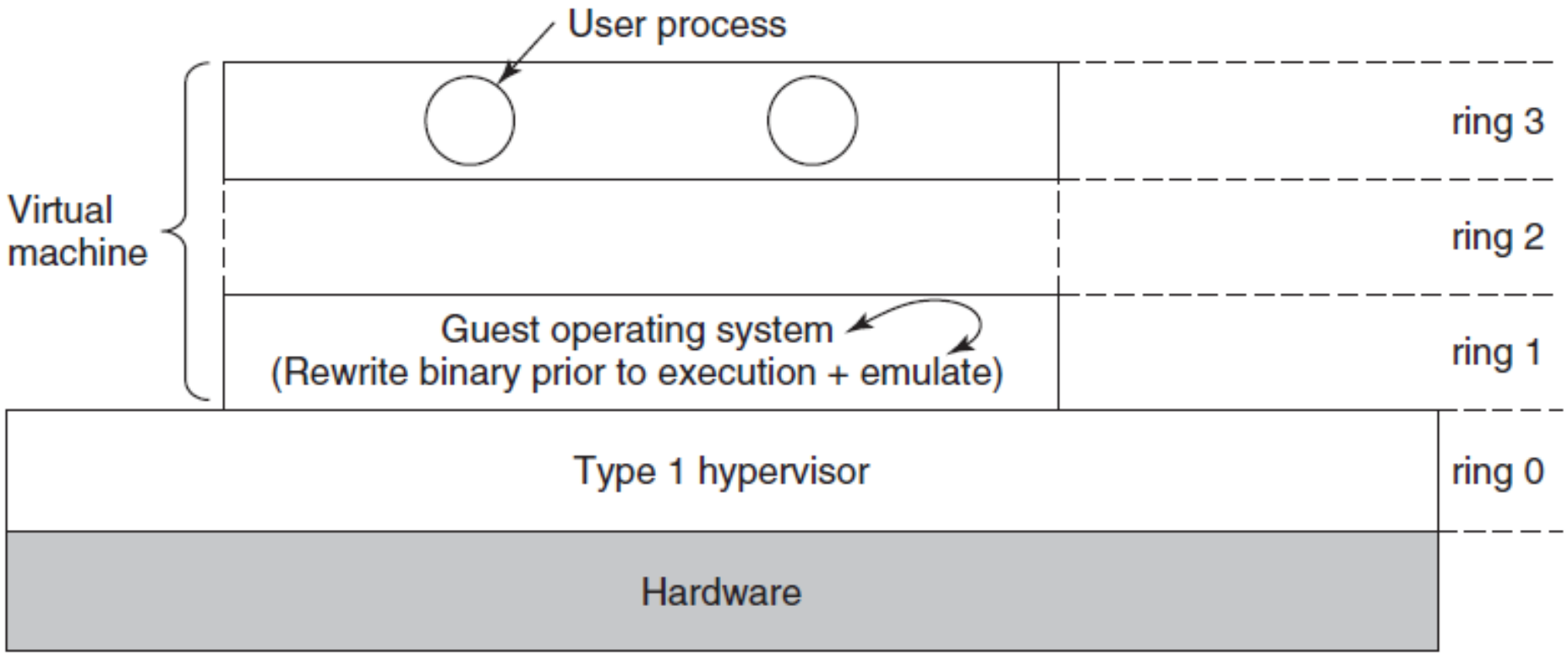

Figure 7-4. The binary translator rewrites the guest operating system running in ring 1, while the hypervisor runs in ring 0.

Paravirtualization

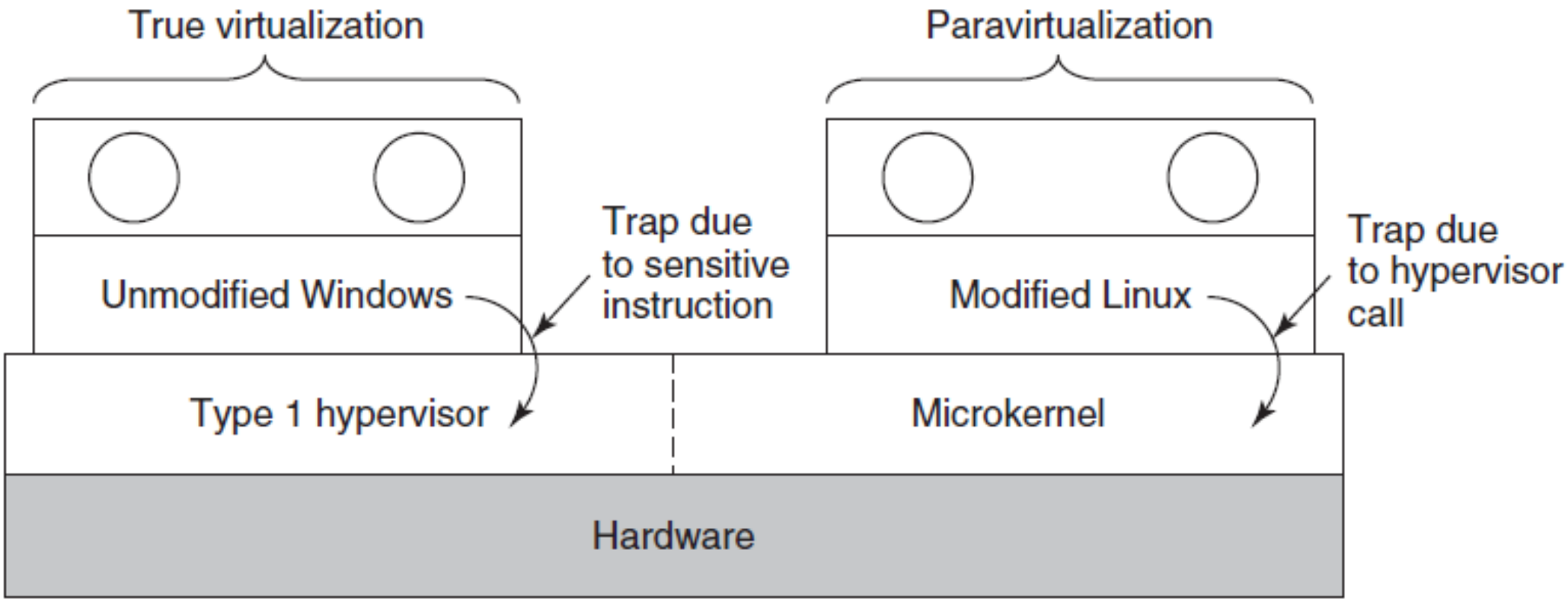

Figure 7-5. True virtualization and paravirtualization

Virtual Machine Interface (VMI)

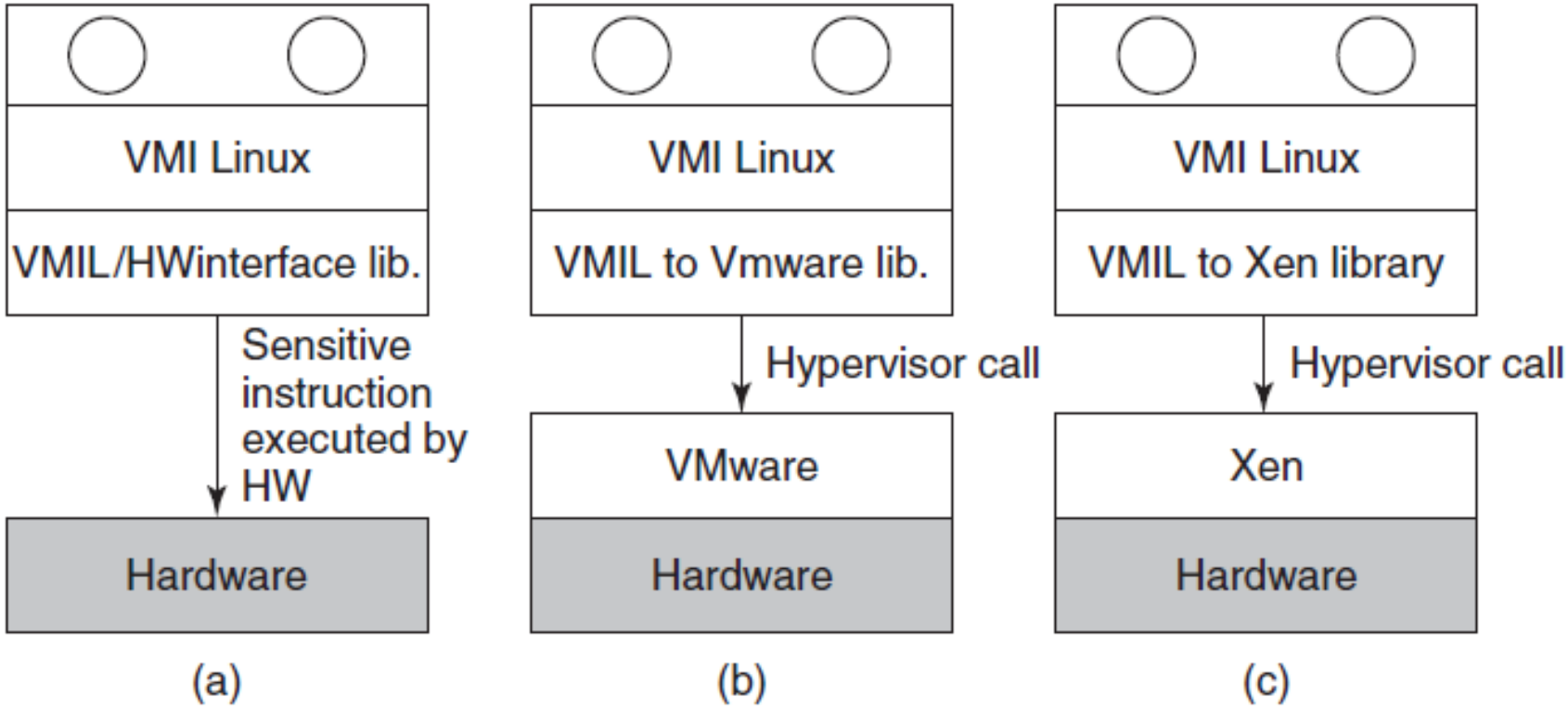

Figure 7-6. VMI Linux running on (a) the bare hardware, (b) VMware, (c) Xen

Nested Page Table

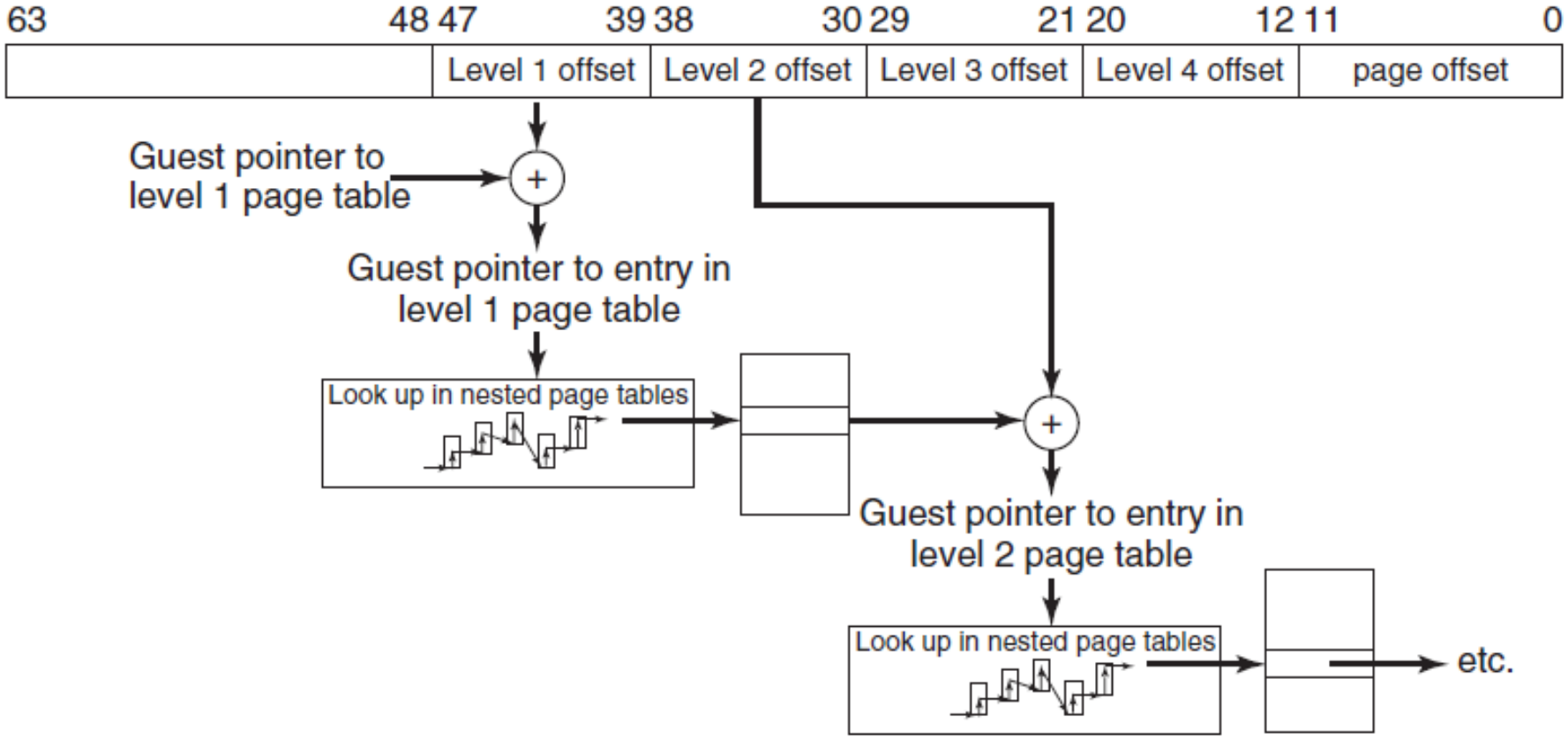

Figure 7-7. Extended/nested page tables are walked every time a guest physical address is accessed—including the accesses for each level of the guest's page tables.

Memory Virtualization

- Deduplication: share pages among virtual machines
- content-based page sharing in VMware

VMware

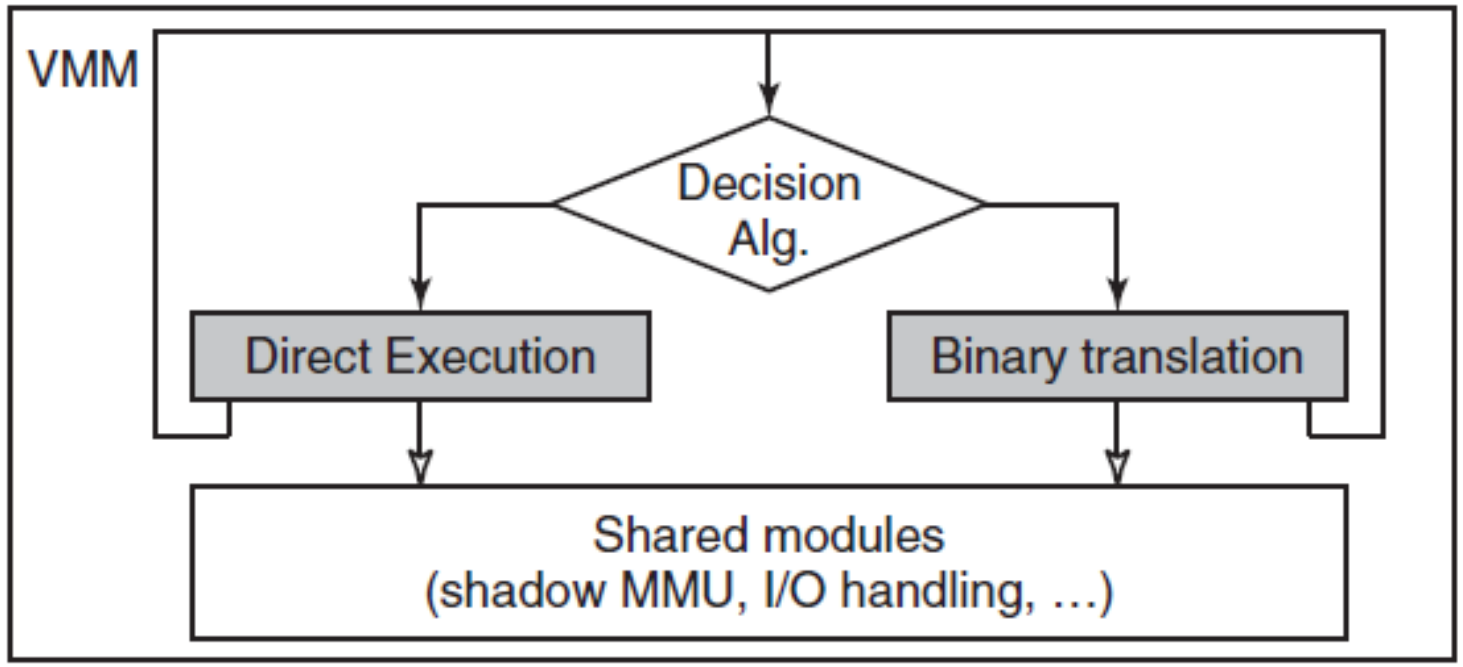

Figure 7-8. High-level components of the VMware virtual machine monitor (in the absence of hardware support).

	Virtual Hardware (front end)	*Back end*
Multiplexed	1 virtual x86 CPU, with the same instruction set extensions as the underlying hardware CUP	Scheduled by the host operating system on either a uniprocessor or multiprocessor host
	Up to 512 MB of contiguous DRAM	Allocated and managed by the host OS (page-by-page)

Emulated	PCI Bus	Fully emulated compliant PCI bus
	4x IDE disks 7x Buslogic SCSI Disks	Virtual disks (stored as files) or direct access to a given raw device
	1x IDE CD-ROM	ISO image or emulated access to the real CD-ROM
	2x 1.44 MB floppy drives	Physical floppy or floppy image
	1x VMware graphics card with VGA and SVGA support	Ran in a window and in full-screen mode. SVGA required VMware SVGA guest driver
	2x serial ports COM1 and COM2	Connect to host serial port or a file
	1x printer (LPT)	Can connect to host LPT port
	1x keyboard (104-key)	Fully emulated; keycode events are generated when they are received by the VMware application
	1x PS-2 mouse	Same as keyboard
	3x AMD Lance Ethernet cards	Bridge mode and host-only modes
	1x Soundblaster	Fully emulated

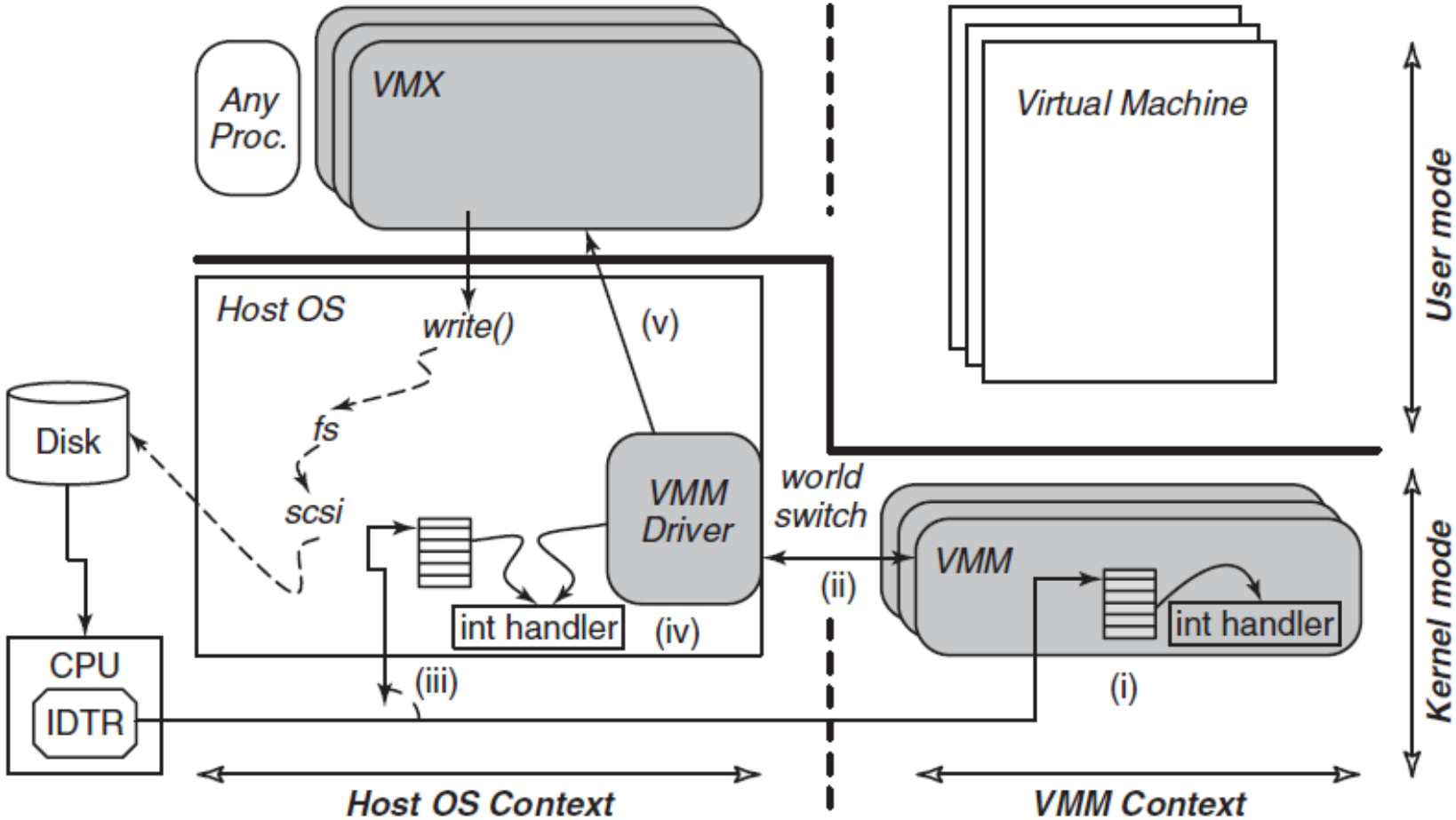

Figure 7-10. The VMware Hosted Architecture and its three components: VMX, VMM driver and VMM.

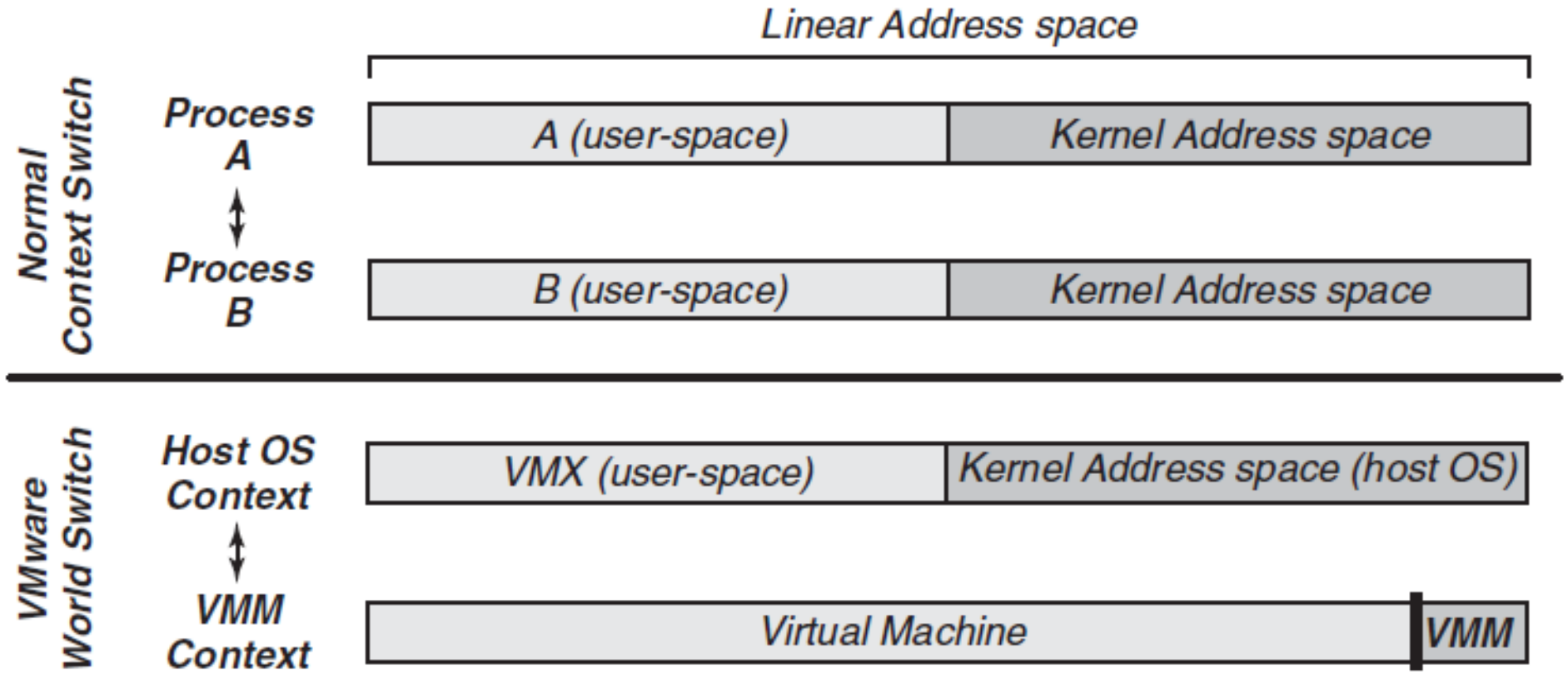

Figure 7-11. Difference between a normal context switch and a world switch.

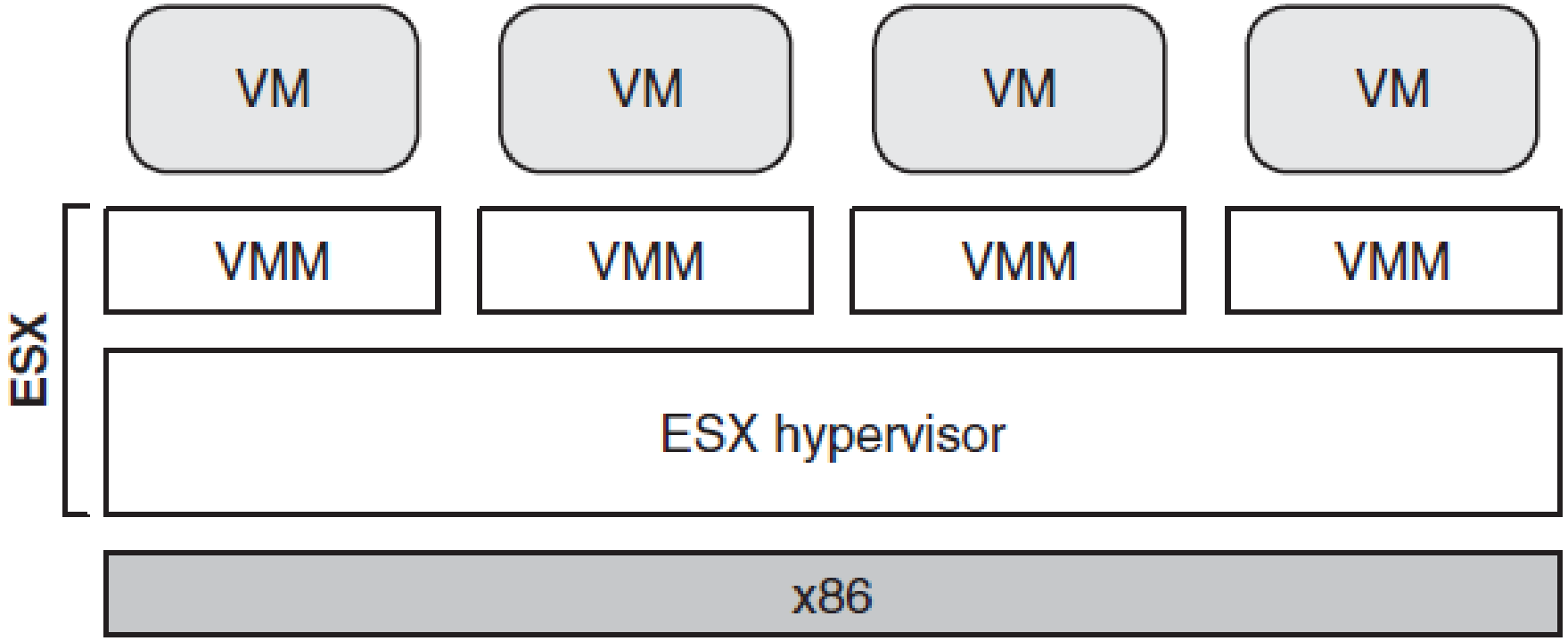

Figure 7-12. ESX Server: VMware's type 1 hypervisor.