x86 Virtualization

Corentin Derbois Marc Angel

corentin@lse.epita.fr null@lse.epita.fr
http://lse.epita.fr/

July 17, 2013

x86 Virtualization

Corentin Derbois, Marc Angel

Virtualization 101

Hardware/Software Techniques

Host/Guest Communication

1 Virtualization 101

2 Hardware/Software Techniques

3 Host/Guest Communication

What?

x86 Virtualization

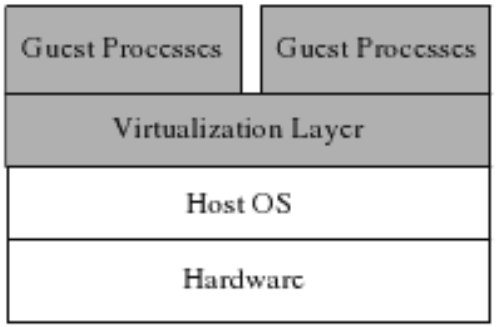
Corentin Derbois, Marc Angel

Virtualization 101
What
Why
How

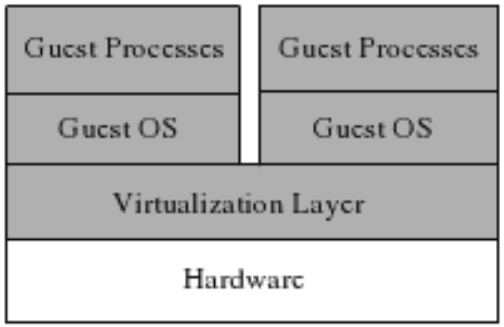
Hardware/Software Techniques

Host/Guest Communication

- Single computer, multiple OSs
- Hardware-level virtualization
  - As opposed to OS-level virtualization
    - LXC, OpenVZ, FreeBSD jails...



OS-level Virtualization



Hardware-level Virtualization

Why?

- Kernel Debugging
- Money
- Flexibility
- ...

How?

- Popek and Goldberg requirements
  - Fidelity
  - Safety
  - Performance
- Binary Translation
  - VMware, VirtualBox, KQEMU
- Paravirtualization
  - Xen
- Full Virtualization
  - KVM, VMware, VirtualBox, Xen...

Instruction Set Virtualization

- Run the VMM at a higher level of privilege
- trap-and-emulate
  - Sensitive instructions yield control to ring 0
  - The VMM emulates them
- Some instructions do not trap (popf, sidt...)
  - 17 of those

Software: Binary Translation

x86 Virtualization

Corentin Derbois, Marc Angel

Virtualization 101

Hardware/Software Techniques

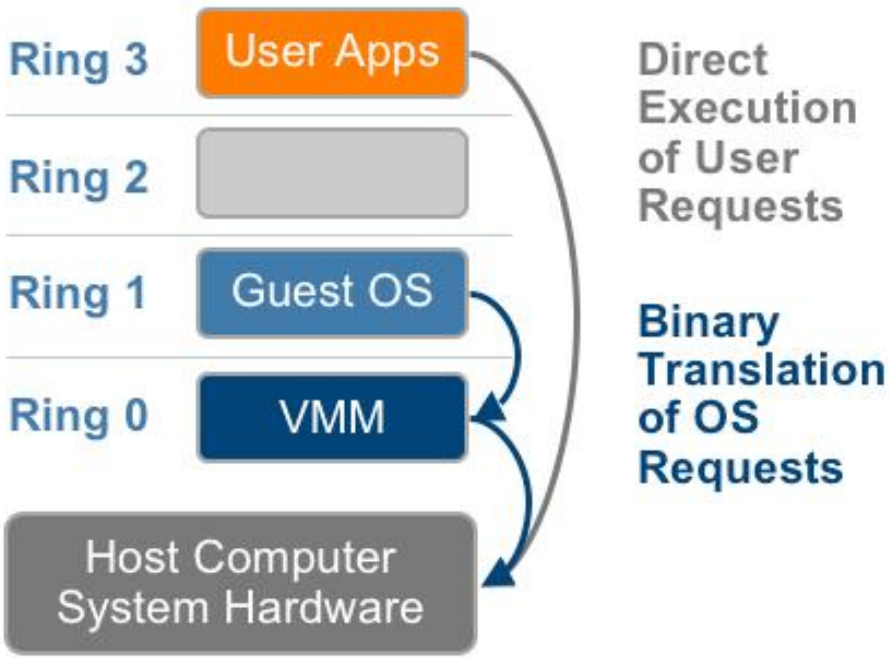
Instruction Set Virt.

Memory Virtualization

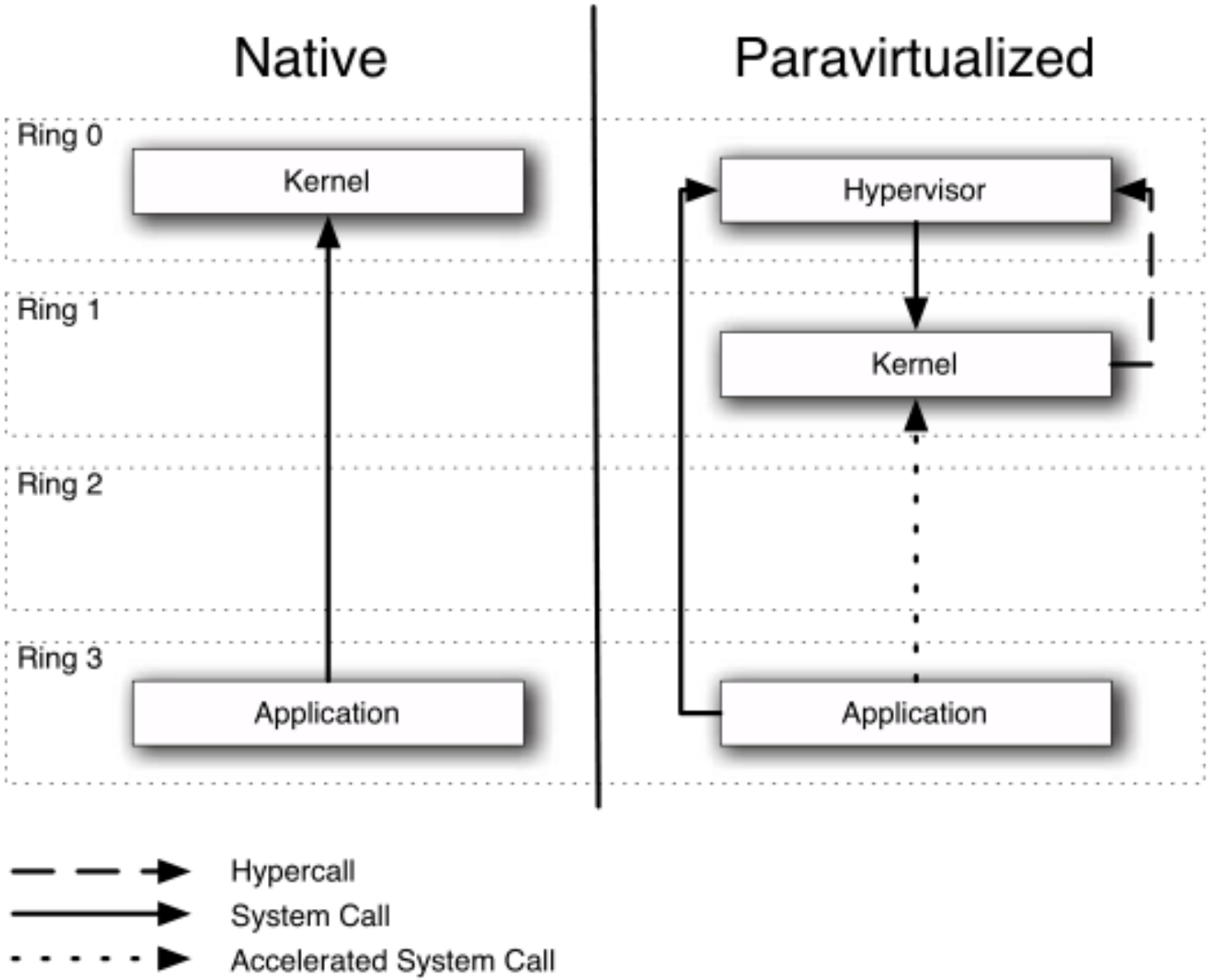
I/O Virtualization

Host/Guest Communication

- Replace critical instructions with traps
- Let the VMM emulate them
- Run userland code "as is"
- Need to emulate syscalls



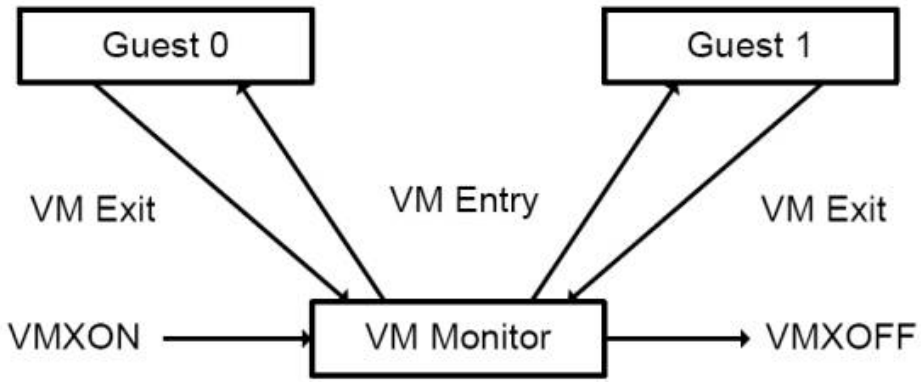
Software: Paravirtualization



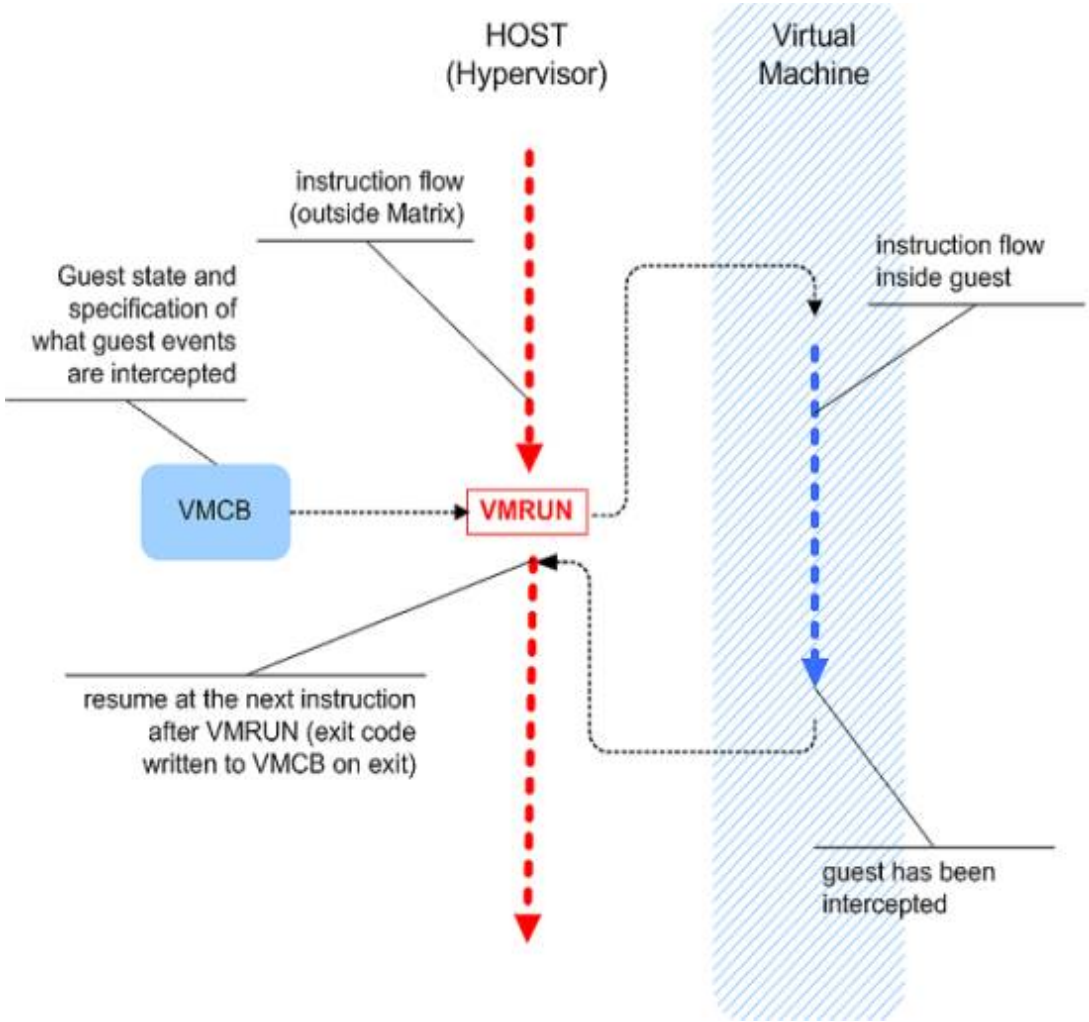
Intel & AMD Hardware Solution

- VT-x and AMD-v
- One ring to rule them all
  - new set of instructions at ring -1
- Guest OS goes back to ring 0

Intel: VMX


Guest 0
Guest 1
VM Exit
VM Entry
VM Exit
VMXON
VM Monitor
VMXOFF

AMD: SVM


HOST
(Hypervisor)
Virtual
Machine
instruction flow
(outside Matrix)
Guest state and
specification of
what guest events
are intercepted
instruction flow
inside guest
VMCB
VMRUN
resume at the next instruction
after VMRUN (exit code
written to VMCB on exit)
guest has been
intercepted

Intel & AMD Hardware Solution

- Add protection to specific instructions
  - CPUID
  - LGDT
  - ...
- Two ways to handle critical instructions
  - Trigger VMEXIT
  - Let the processor handle them directly

Hardware: VMEXIT & native

- Processor data are stored in specific data structures
  - AMD: VMCB
  - Intel: VMCS
- Store to CRx, GDT, selectors...

Hardware: VMEXIT & native

x86 Virtualization

Corentin Derbois, Marc Angel

Virtualization 101

Hardware/Software Techniques

Instruction Set Virt.

Memory Virtualization

I/O Virtualization

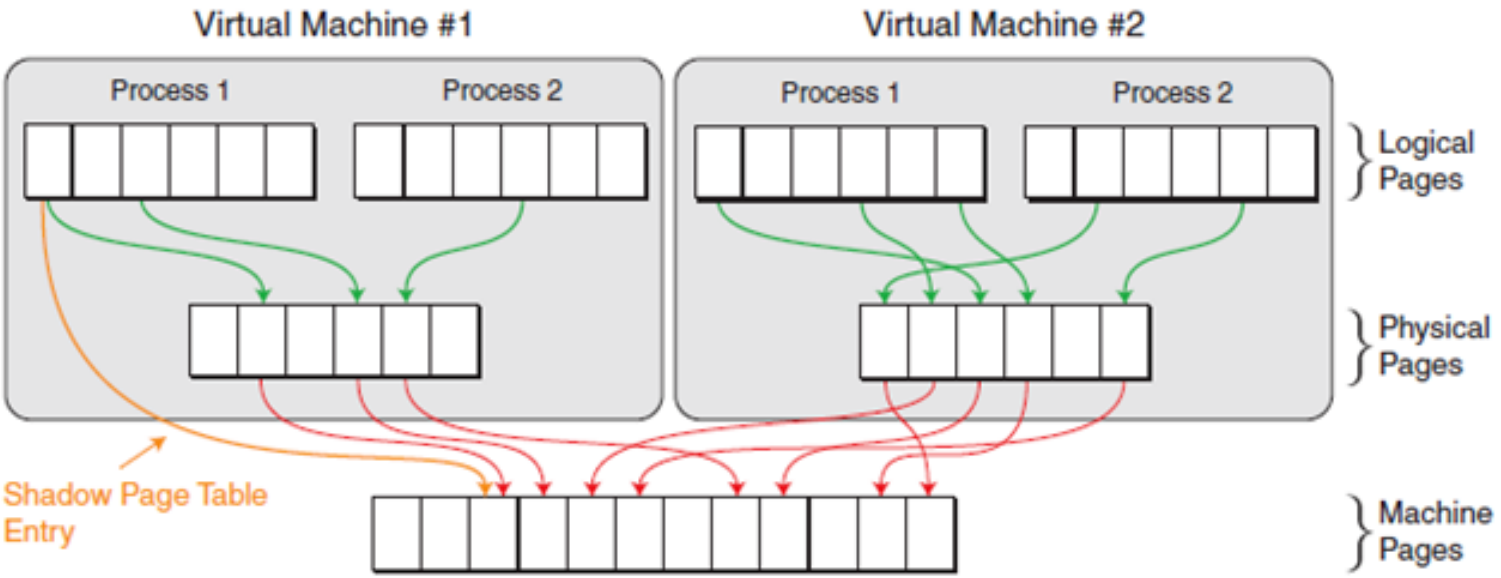
Host/Guest Communication

- Some behaviors can't be automatically handled by the CPU
  - I/O
  - CPUID
  - PageFault
- In this case, a VMEXIT is triggered to ask the host OS to emulate them

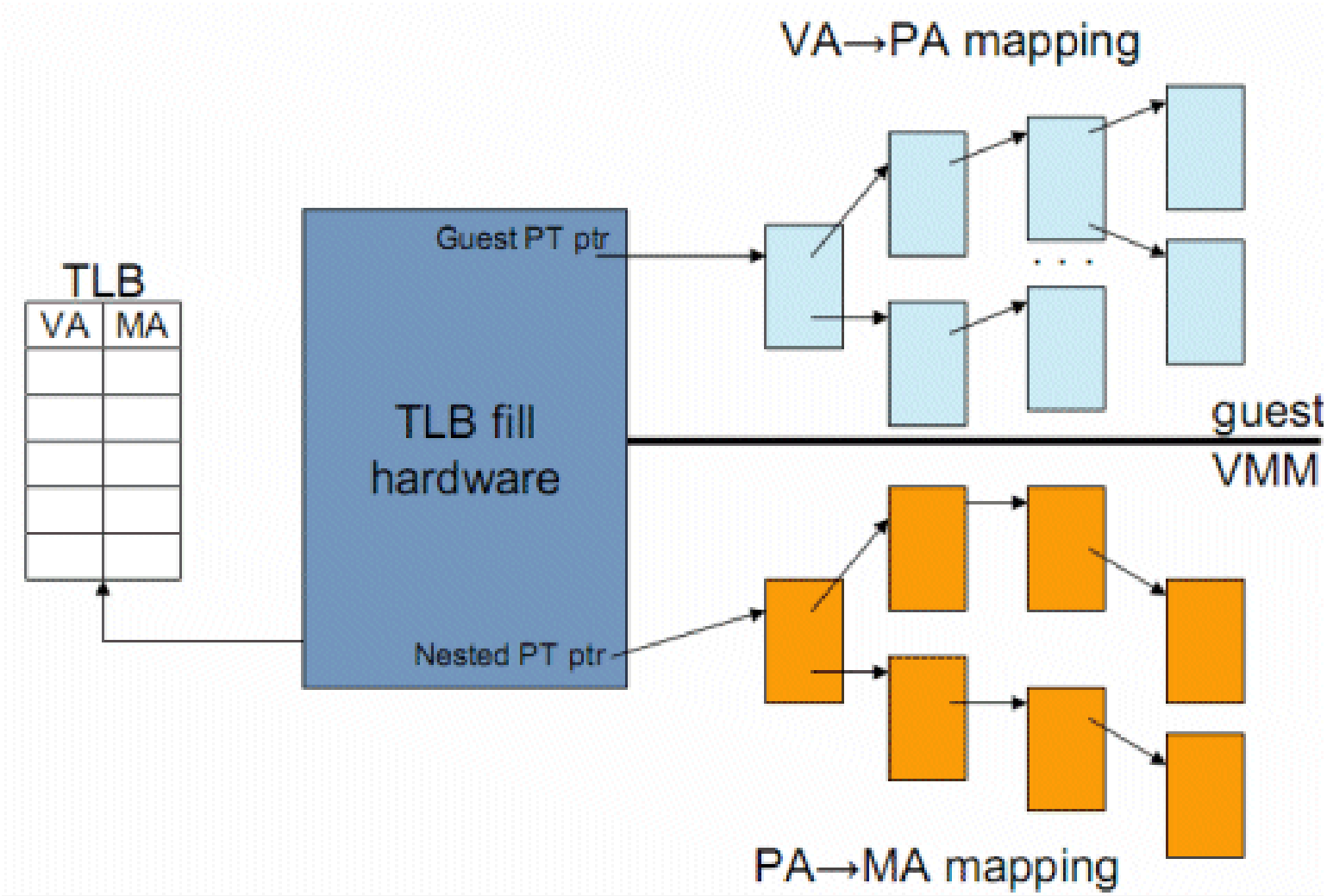
MMU Virtualization

- Three levels of memory
  - Guest virtual address space
  - Guest physical address space
  - VMM physical memory

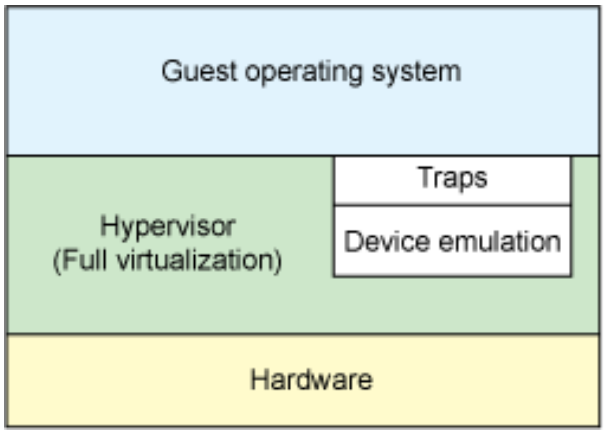
Software: Shadow Page Tables

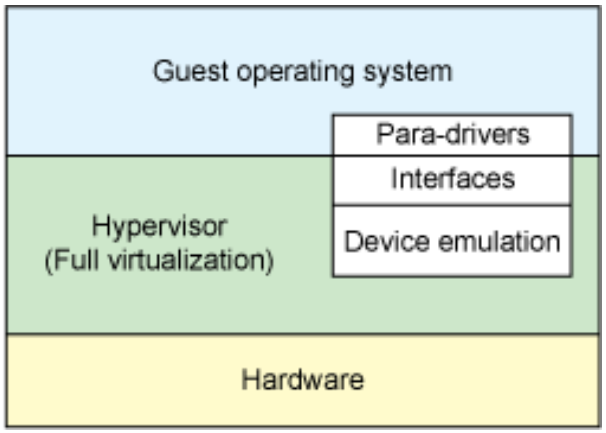


Hardware: Intel EPT, AMD RVI

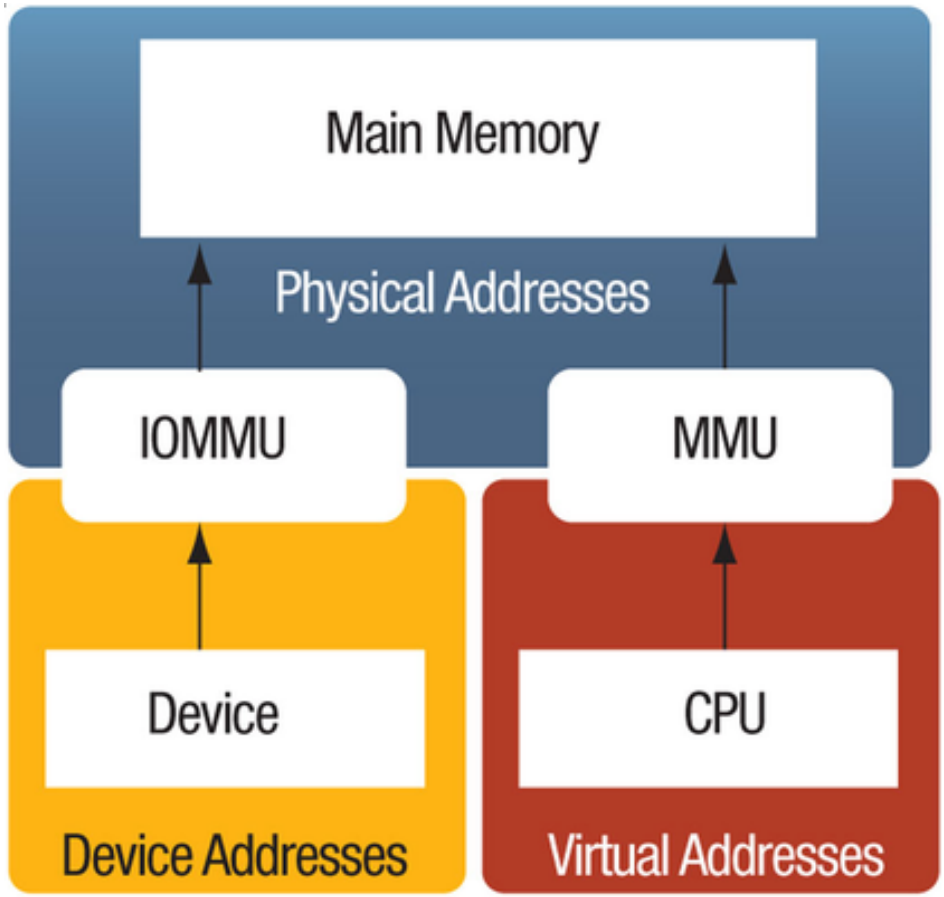


I/O Virtualization





IOMMU



x86 Virtualization

Corentin Derbois, Marc Angel

Virtualization 101

Hardware/Software Techniques

Instruction Set Virt.

Memory Virtualization

I/O Virtualization

Host/Guest Communication

CPUID

- Triggers VMEXIT
- Offers a decent interface for Question/Answer
- Static
- Xen
  - CPUID is overwritable in PVM
  - Can get specific value from Xen

I/O Ports

- Triggers VMEXIT
- Offers a large choice to make I/O requests
- Dynamic discussion at each VMEXIT
- VMware
  - Port: 0x5658
  - Can get lots of information:
    - Processor Speed
    - VMware version
    - Memory size
    - ...

PCI

x86 Virtualization

Corentin Derbois, Marc Angel

Virtualization 101

Hardware/Software Techniques

Host/Guest Communication

CPUID

I/O Ports

PCI

Virtio

- PCI offers a decent interface to communicate
- Some HVM use it to make their video driver and do some communication
- Mainly for Desktop drivers
- VirtualBox
  - BEEF -> video driver
  - CAFE -> some other driver
- VMware
  - PCI driver for SVGA monitor

Virtio

x86 Virtualization

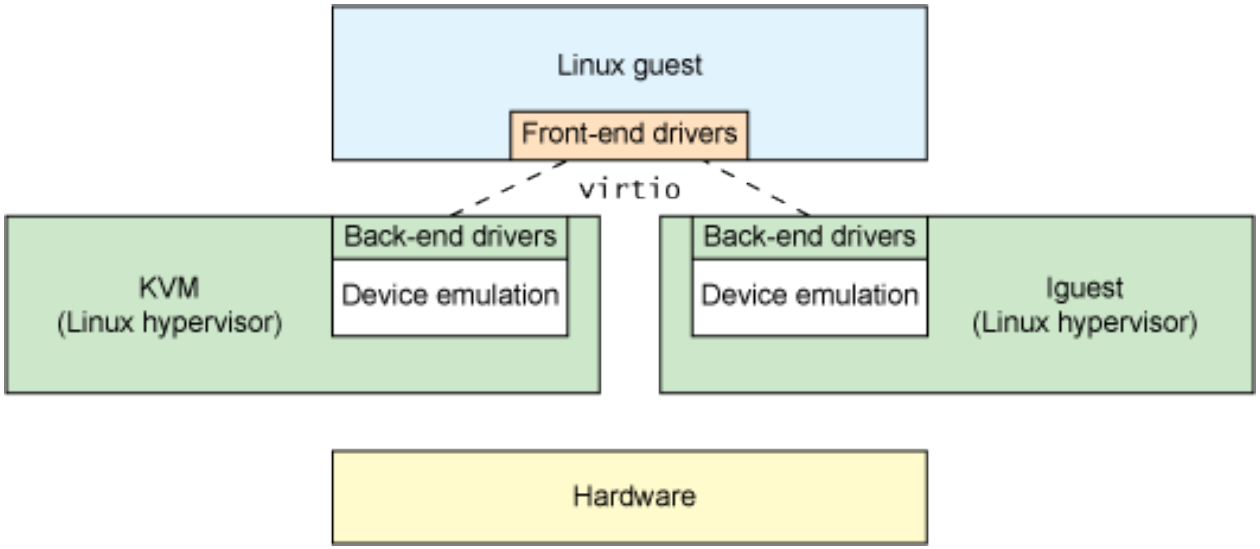
Corentin Derbois, Marc Angel

Virtualization 101

Hardware/Software Techniques

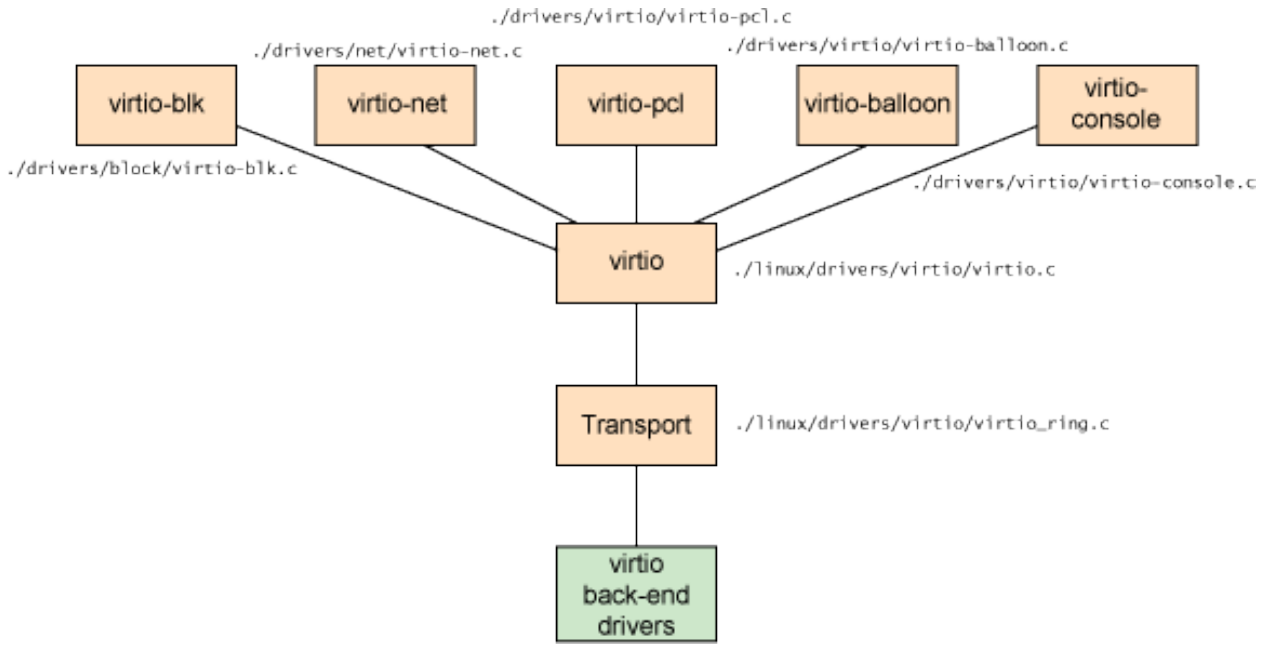
Host/Guest Communication

CPUID
I/O Ports
PCI
Virtio



- A common framework for I/O virtualization for hypervisors
- Main I/O virtualization platform in KVM
- High performance

Virtio Architecture



Supported Devices

- Network
- Block
- Console
- Entropy
- Balloon
- Rpmsg
- SCSI Host

Virtio Devices

- Presented by the host as a regular PCI device
  - Vendor ID: 0x1AF4 (Qumranet)
  - Device ID for each type of device
  - Configuration header at the start of the BAR
- Memory mapped header for embedded devices without PCI support

Virtio PCI Header

Bits	32	32	32	16	16	16	8	8
Read/Write	R	R+W	R+W	R	R+W	R+W	R+W	R
Purpose	Device Features bits 0:31	Guest Features bits 0:31	Queue Address	Queue Size	Queue Select	Queue Notify	Device Status	ISR Status

Can be followed by device specific headers:

- MAC addresses for network devices
- Other information for block devices (cylinder/head/sector counts...)

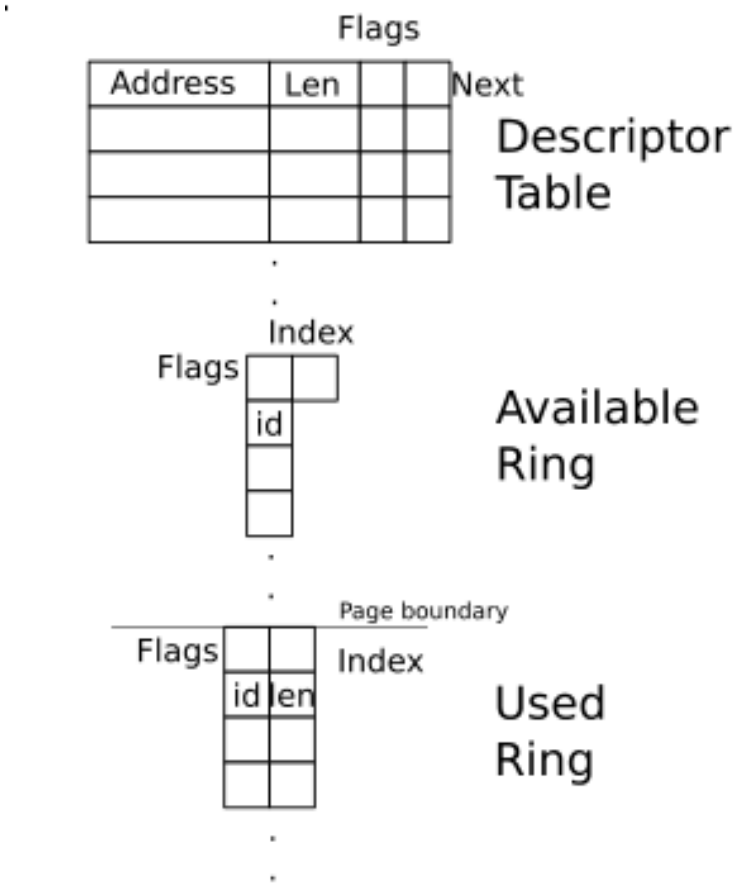
Virtio PCI Device Init

Bits	32	32	32	16	16	16	8	8
Read/Write	R	R+W	R+W	R	R+W	R+W	R+W	R
Purpose	Device Features bits 0:31	Guest Features bits 0:31	Queue Address	Queue Size	Queue Select	Queue Notify	Device Status	ISR Status

1. RESET
2. ACKNOWLEDGE
   - Valid virtio PCI device
3. DRIVER
   - We know how to use the device
4. DRIVER_OK
   - Virtqueue configuration
   - Feature exchange

Virtqueues

- 0 or more virtqueues per devices
- Spans 2 pages



Conclusion

Questions?

Thank you