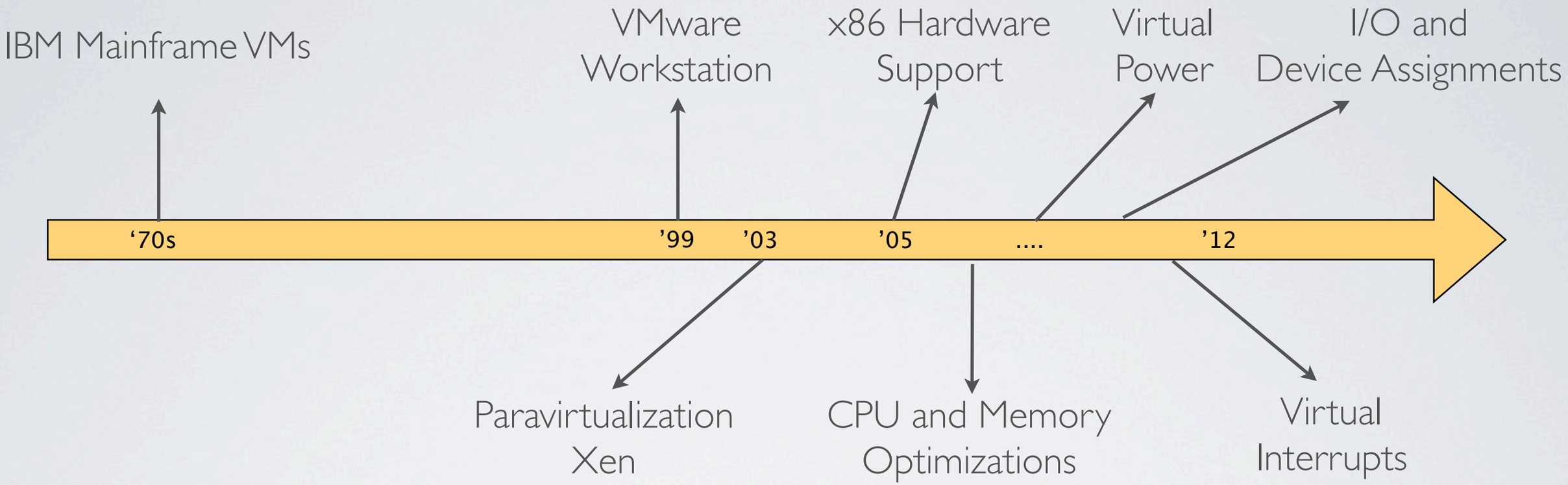
Virtualization

History and Future Trends

Christoffer Dall - Candidacy Exam - January 2013

Columbia University - Computer Science Department



Historical Overview

Taxonomy

Virtualization Software

[Sugerman 01]

[Kivity 09]

[Bugnion 12]

[Barham 03]

[Popek 74]

[Colp 11]

[Bugnion 97]

[Ben-Yehuda 10]

[LeVasseur 04]

Hardware Support

[Gum 83]

[Adams 06]

[Agesen 12]

[Uhlig 05]

[Bhargava 08]

[Wang 11]

Optimizations

[Waldspurger 02]

[Gordon 12]

[Liu 06]

[Willmann 08]

[Amit 11]

[Russell 08]

[Ahmad 11]

Power

[Stoess 07]

[Hwang 12]

[Nathuji 07]

[Ye 10]

[Krishnan 11]

[Tian 10]

[Andrus 11]

Virtualization

Virtualization Software

[Sugerman 01]

[Kivity 09]

[Bugnion 12]

[Barham 03]

[Popek 74]

[Colp 11]

[Bugnion 97]

[Ben-Yehuda 10]

[LeVasseur 04]

Hardware Support

[Gum 83]

[Adams 06]

[Uhlig 05]

[Bhargava 08]

[Wang 11]

Optimizations

[Waldspurger 02]

[Gordon 12]

[Liu 06]

[Willmann 08]

[Amit 11]

[Russell 08]

[Ahmad 11]

Power

[Stoess 07]

[Hwang 12]

[Nathuji 07]

[Ye 10]

[Krishnan 11]

[Tian 10]

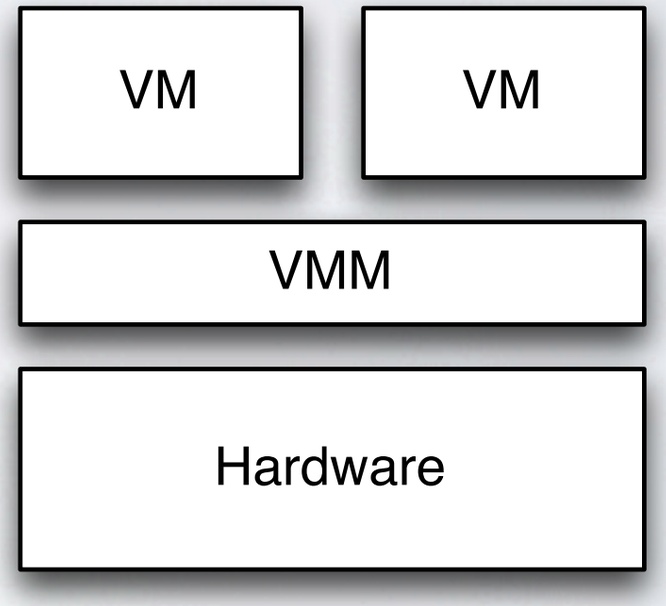
[Andrus 11]

Virtualization [Popek 74]

- The mechanism through which we facilitate Virtual Machines
- Defined intuitively by Popek and Goldberg [Popek 74] to be:
  *"...an **efficient**, **isolated duplicate** of the real machine."*

VMMs [Popek 74]

- Virtual Machine Monitor (VMM)
- Three properties
  1. Efficiency
  2. Resource control
  3. Equivalence



Virtualizable [Popek 74]

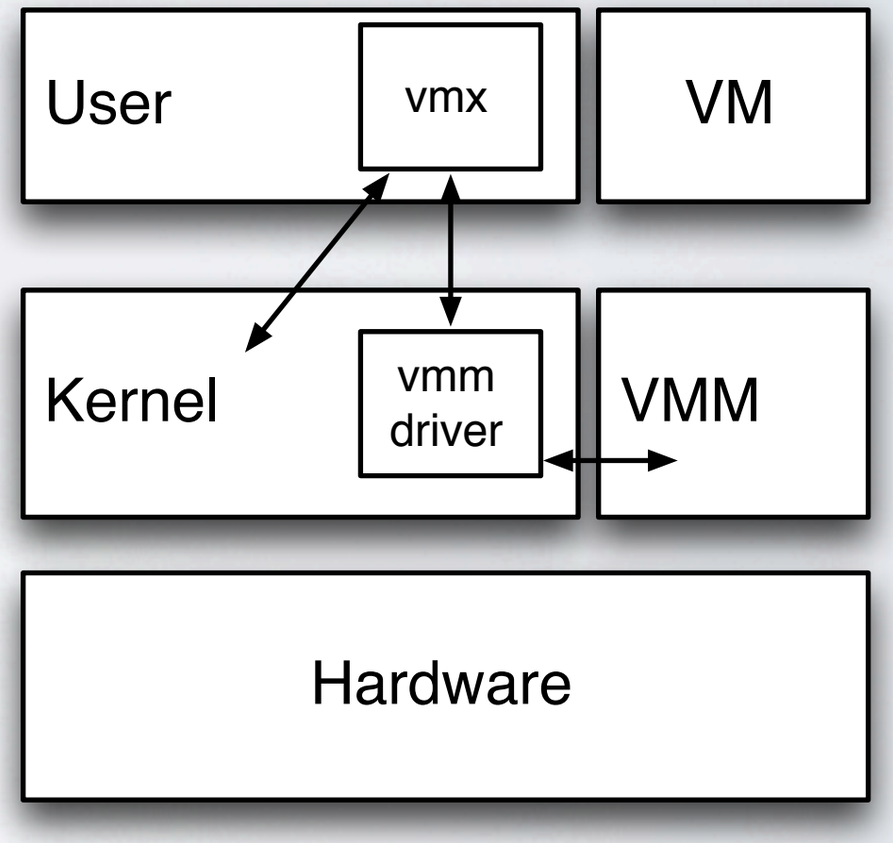
- Theorem:
  "For any conventional third generation computer, a virtual machine monitor may be constructed if the set of sensitive instructions for that computer is a subset of the set of privileged instructions"

- All sensitive instructions trap to the VMM

- Allows for trap-and-emulate

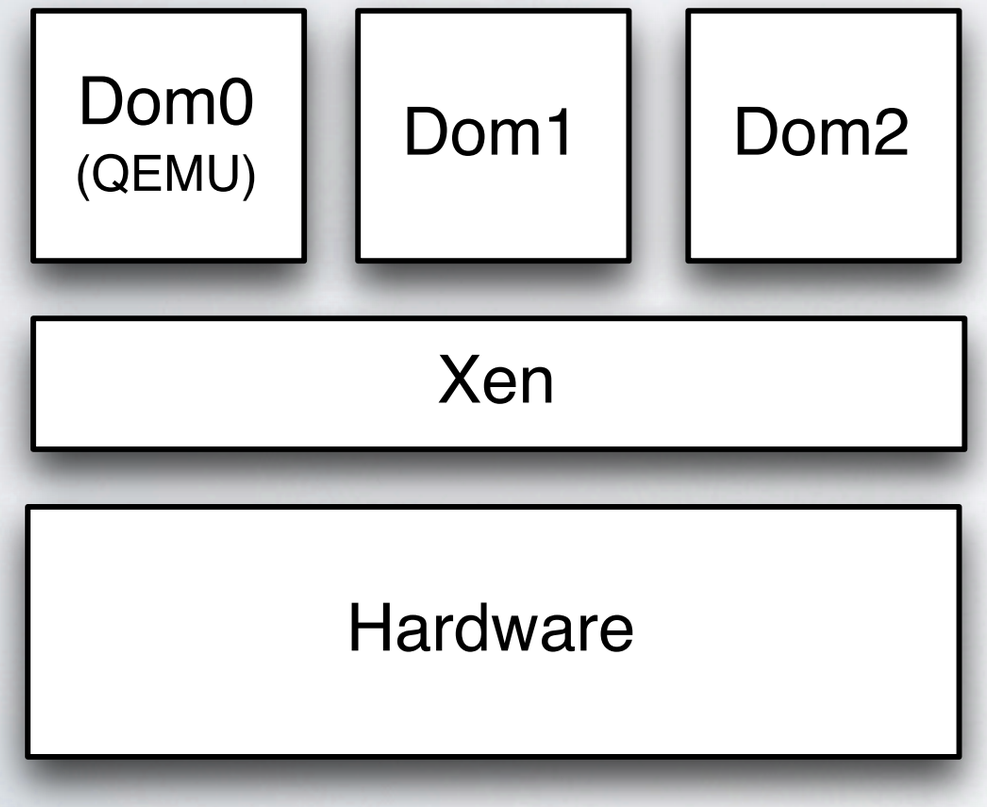
Revitalization in the nineties

- VMware workstation in '99 [Sugerman 01; Bugnion 12]

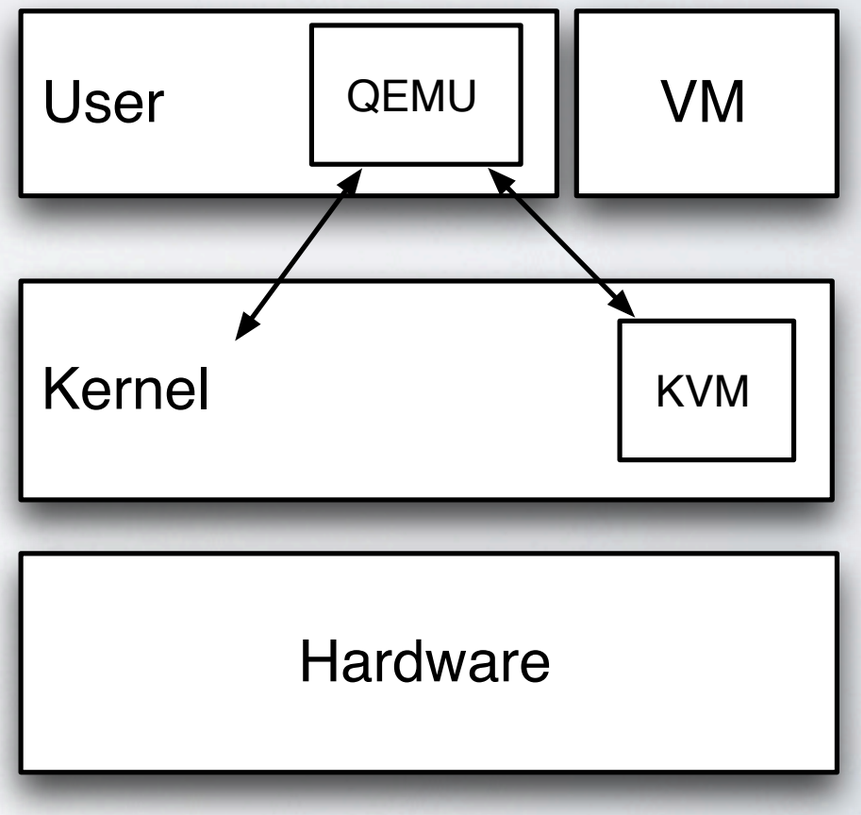


New Directions

Xen and the art of Virtualization [Barham 03]



Kernel-Based VM (KVM) [Kivity 09]



Hardware Support

Virtualization Software

[Sugerman 01]

[Kivity 09]

[Bugnion 12]

[Barham 03]

[Popek 74]

[Colp 11]

[Bugnion 97]

[Ben-Yehuda 10]

[LeVasseur 04]

Hardware Support

[Gum 83]

[Adams 06]

[Agesen 12]

[Uhlig 05]

[Bhargava 08]

[Wang 11]

Optimizations

[Waldspurger 02]

[Gordon 12]

[Liu 06]

[Willmann 08]

[Amit 11]

[Russell 08]

[Ahmad 11]

Power

[Stoess 07]

[Hwang 12]

[Nathuji 07]

[Ye 10]

[Krishnan 11]

[Tian 10]

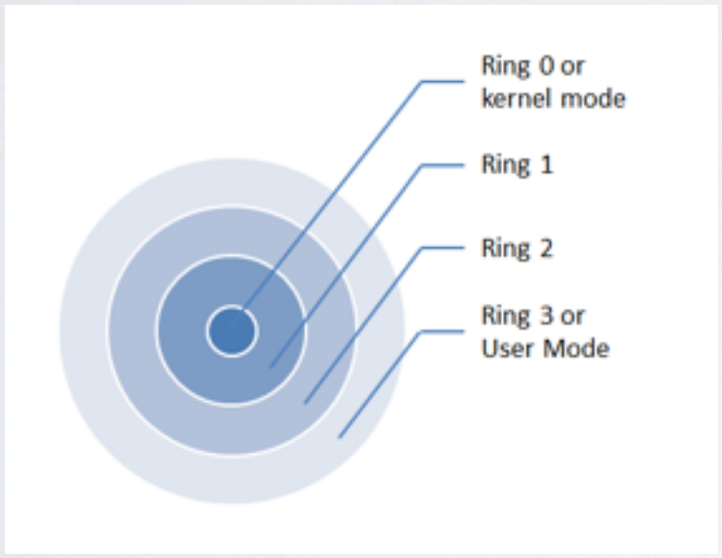
[Andrus 11]

Hardware Support Outline

- Intel VT-x and AMD-V
- Hardware vs. Software
- Hardware Memory Virtualization

Intel VT-x

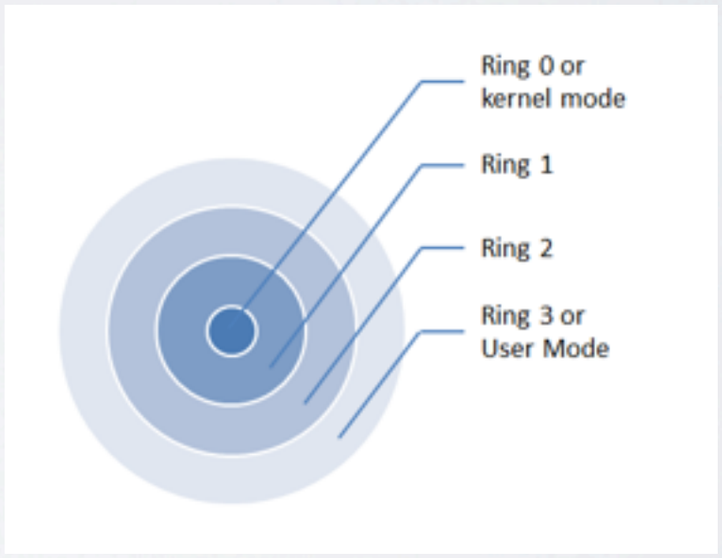
Non-Root traps sensitive instructions



Root

VM ENTRY

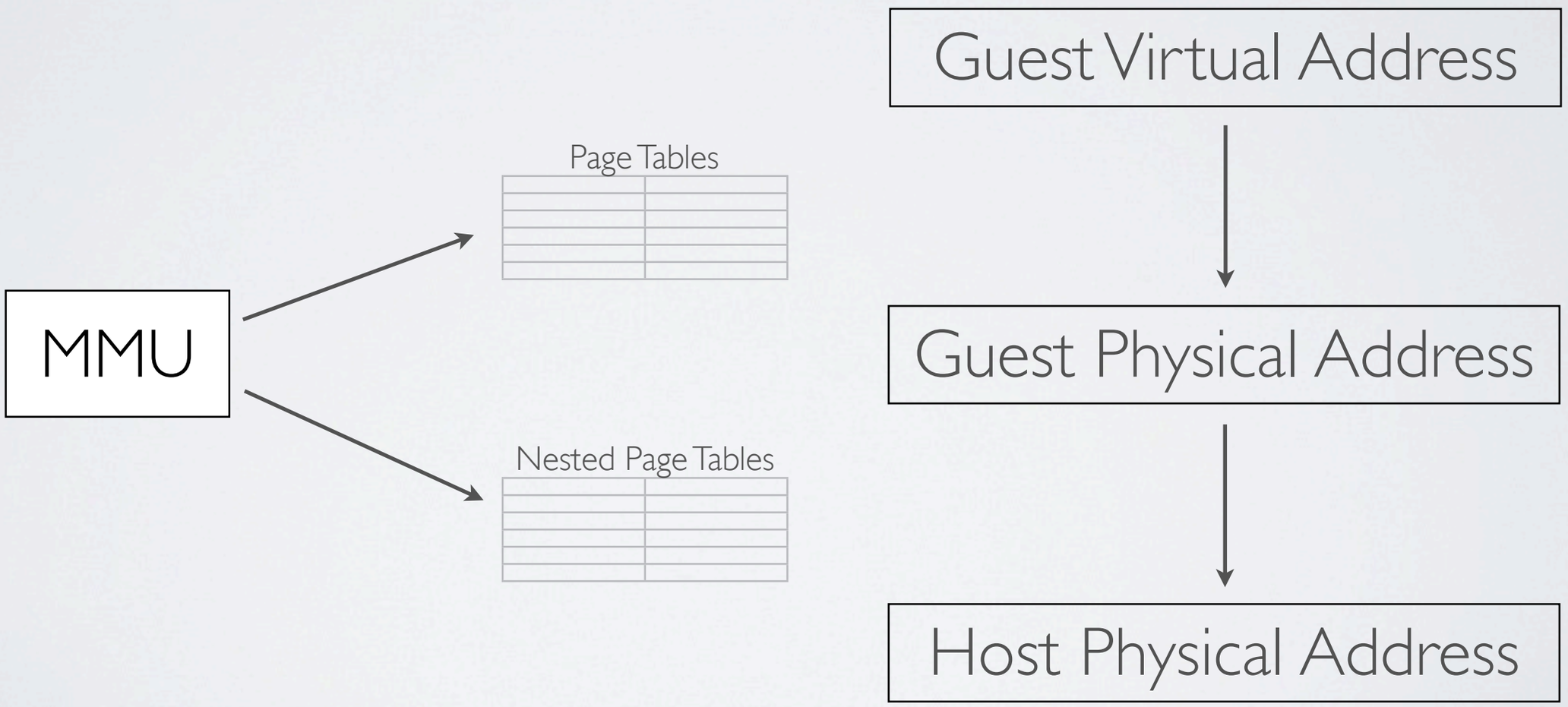
VM EXIT



Non-Root (VM)

Intel VT-x

Nested Page Tables (NPT)



Hardware Support

Not a new idea [Gum 83]

Hardware vs. Software

- Benefits not clear cut

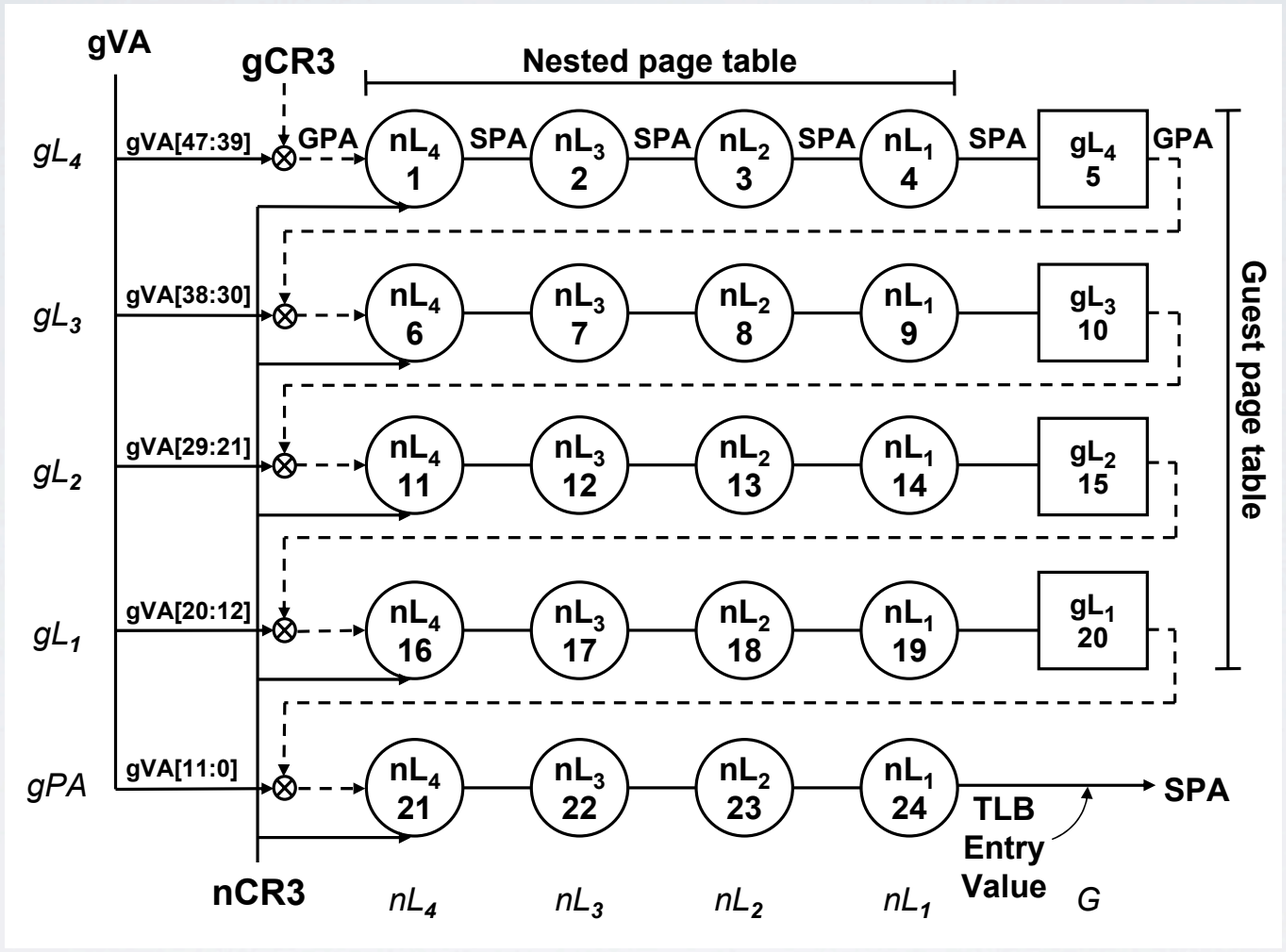
- Comparison with software [Adams 06] show that especially IO-bound workloads are actually faster.

  - example:
    in/out instruction sequences rewritten to single exit

Hardware vs. Software

- Hardware extensions have matured
- But, exits are still expensive...
- Still uses combination of hardware/software approach in commercial VMware products [Agesen 12].
- Software methods still useful for nested virtualization [Ben-Yehuda 10]

Nested Page Tables



Some concerning results [Bhargava 08]

Hardware Support

Virtualization Software

[Sugerman 01]

[Kivity 09]

[Bugnion 12]

[Barham 03]

[Popek 74]

[Colp 11]

[Bugnion 97]

[Ben-Yehuda 10]

[LeVasseur 04]

Hardware Support

[Gum 83]

[Adams 06]

[Uhlig 05]

[Bhargava 08]

[Wang 11]

Optimizations

[Waldspurger 02]

[Gordon 12]

[Liu 06]

[Willmann 08]

[Amit 11]

[Russell 08]

[Ahmad 11]

Power

[Stoess 07]

[Hwang 12]

[Nathuji 07]

[Ye 10]

[Krishnan 11]

[Tian 10]

[Andrus 11]

Optimizations

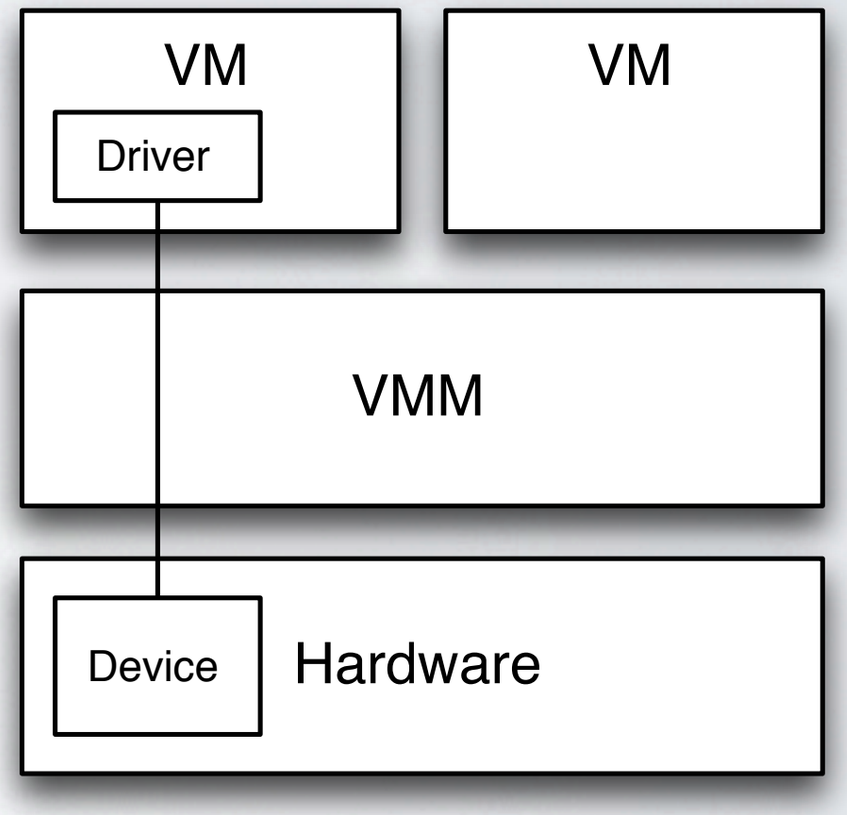
- Memory Consumption
- Paravirtualized drivers
- **Direct Device Assignment**
- **Virtual Interrupts**

Direct Device Assignment

- One major bottleneck is I/O
- Still far from bare metal
- Direct Device Assignment is a potential solution

Direct Device Assignment

- Liu et al. [Liu 06] propose to simply assign a device directly to a VM
- Requires device support
- Control messages trap to VMM
- IOMMUs can help [Willmann 08; Amit 11]



Virtual Interrupts

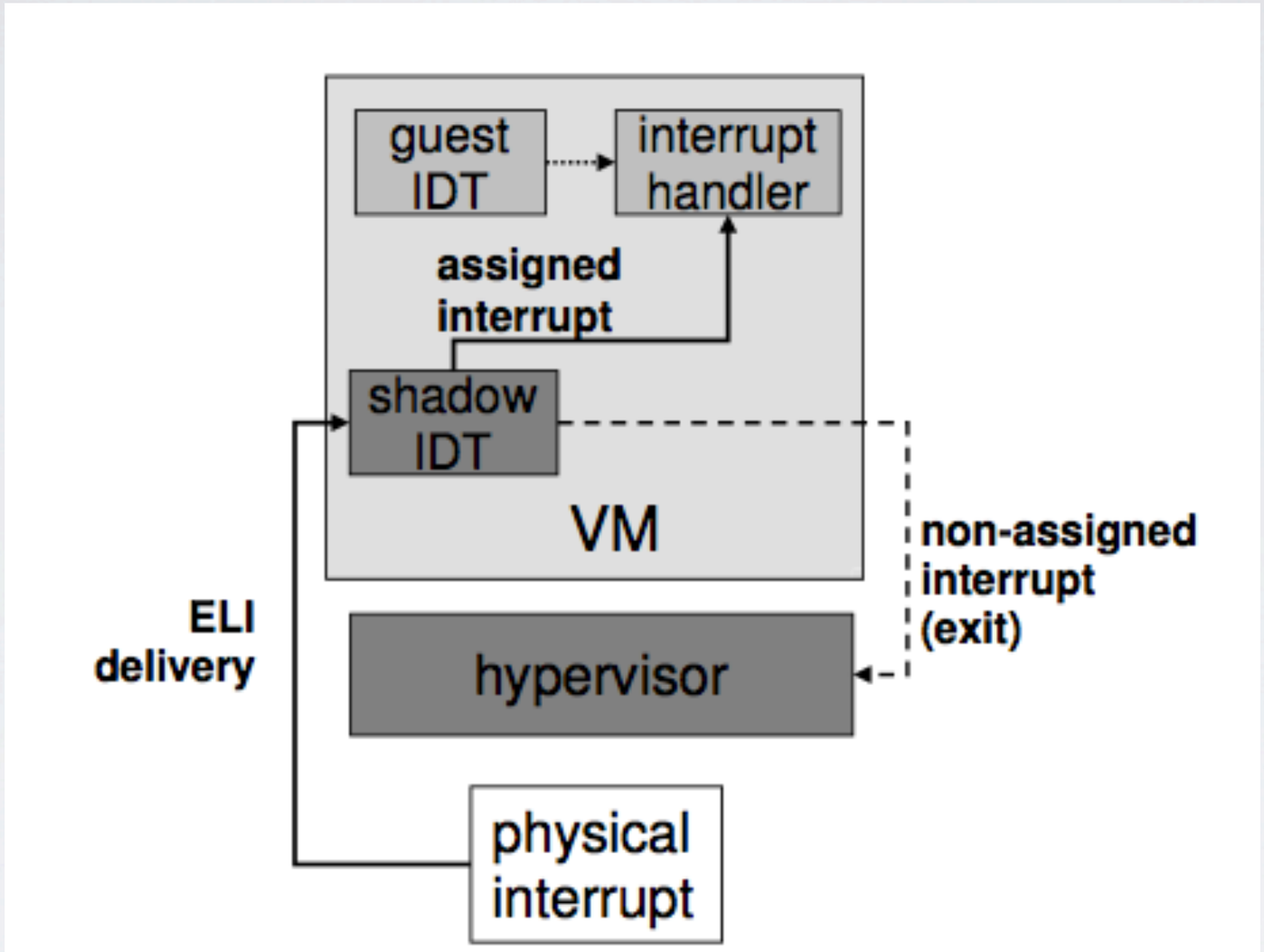
- Interrupts can limit CPU performance and I/O throughput
- Worse in VMs due to VMEXITs
- Especially bad on systems with high IRQ frequency

Virtual Interrupts

IRQ Device	NP	Handler
Host Device	1	
Host Device	1	
Host Device	1	
Guest Device	0	handle_disk_irq()
Host Device	1	
Guest Device	0	handle_nic_irq()

- ELI [Gordon 12]: VMs handle interrupts for certain devices
- ELI uses a shadow Interrupt Descriptor Table (IDT) to let guest handle interrupts directly, while remaining in control of host devices

Virtual Interrupts



Hardware Support

Virtualization Software

[Sugerman 01]

[Kivity 09]

[Bugnion 12]

[Barham 03]

[Popek 74]

[Colp 11]

[Bugnion 97]

[Ben-Yehuda 10]

[LeVasseur 04]

Hardware Support

[Gum 83]

[Adams 06]

[Uhlig 05]

[Bhargava 08]

[Wang 11]

Optimizations

[Waldspurger 02]

[Gordon 12]

[Liu 06]

[Willmann 08]

[Amit 11]

[Russell 08]

[Ahmad 11]

Power

[Stoess 07]

[Hwang 12]

[Nathuji 07]

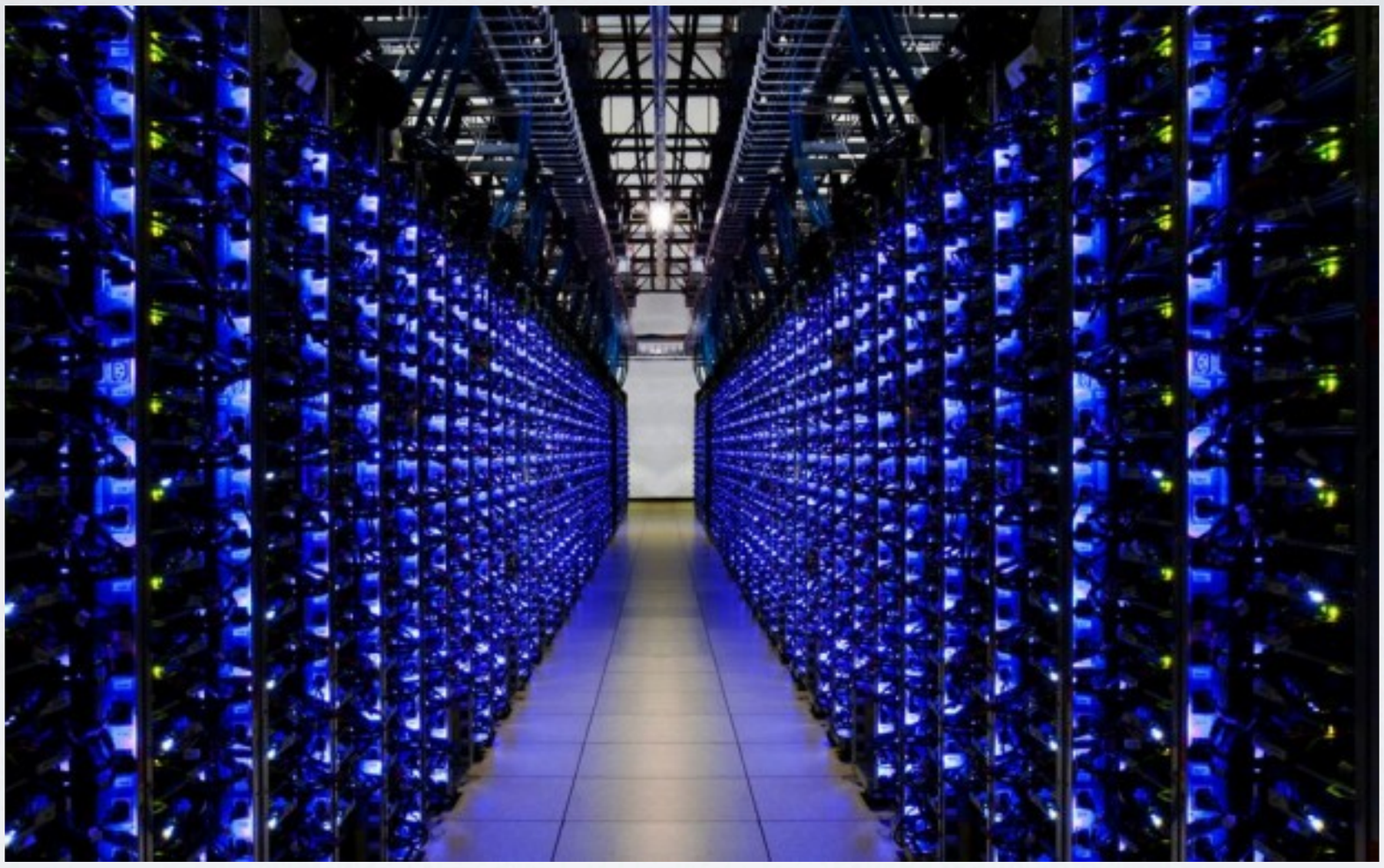
[Ye 10]

[Krishnan 11]

[Tian 10]

[Andrus 11]

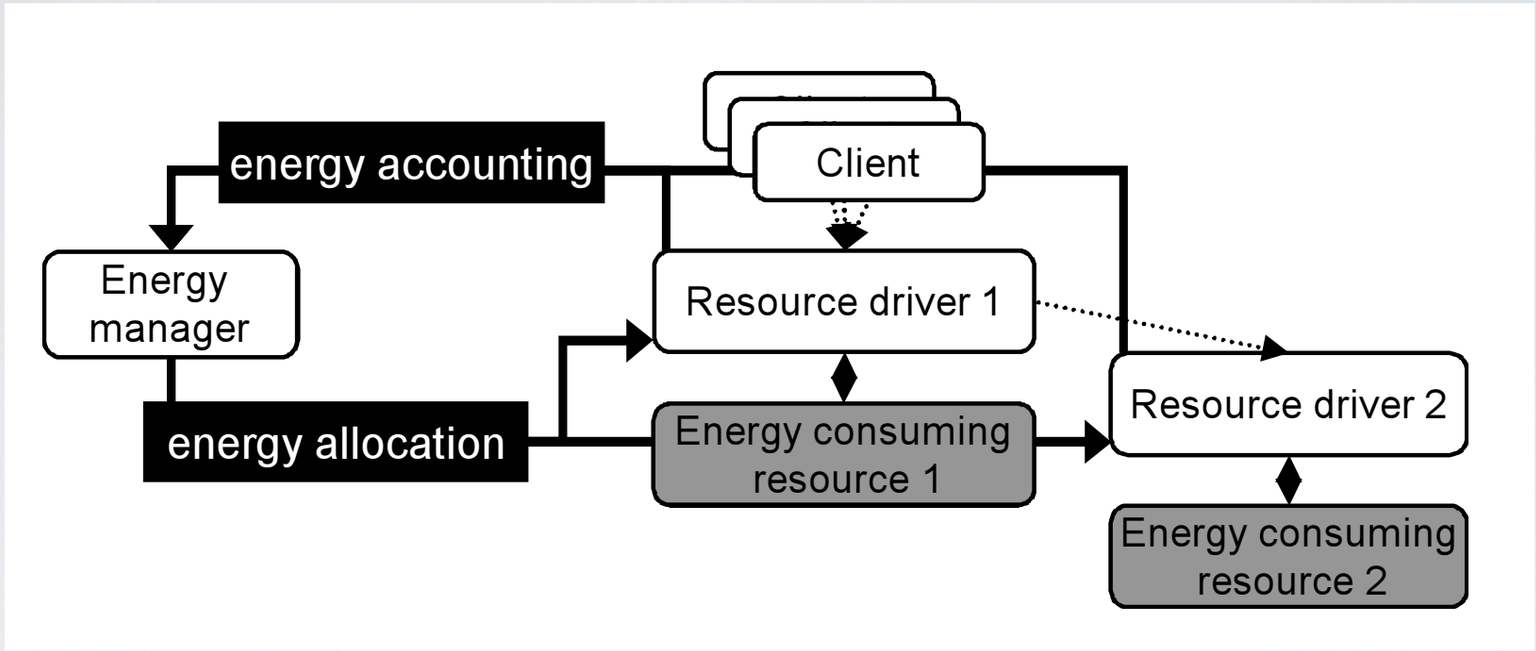
Data Centers



Amazon AWS hosts 46,000 servers

8 MW power consumption at $88 million / year

Power Architecture for VMs



- Accounting model [Stoess 07]
- Problem is that VMMs are uninformed about VMs and VMs don't know their real power consumption (VMEXITs)
- Accounting model suggests passing allocation hints up the stack, and statistics down the stack

VirtualPower [Nathuji 07]

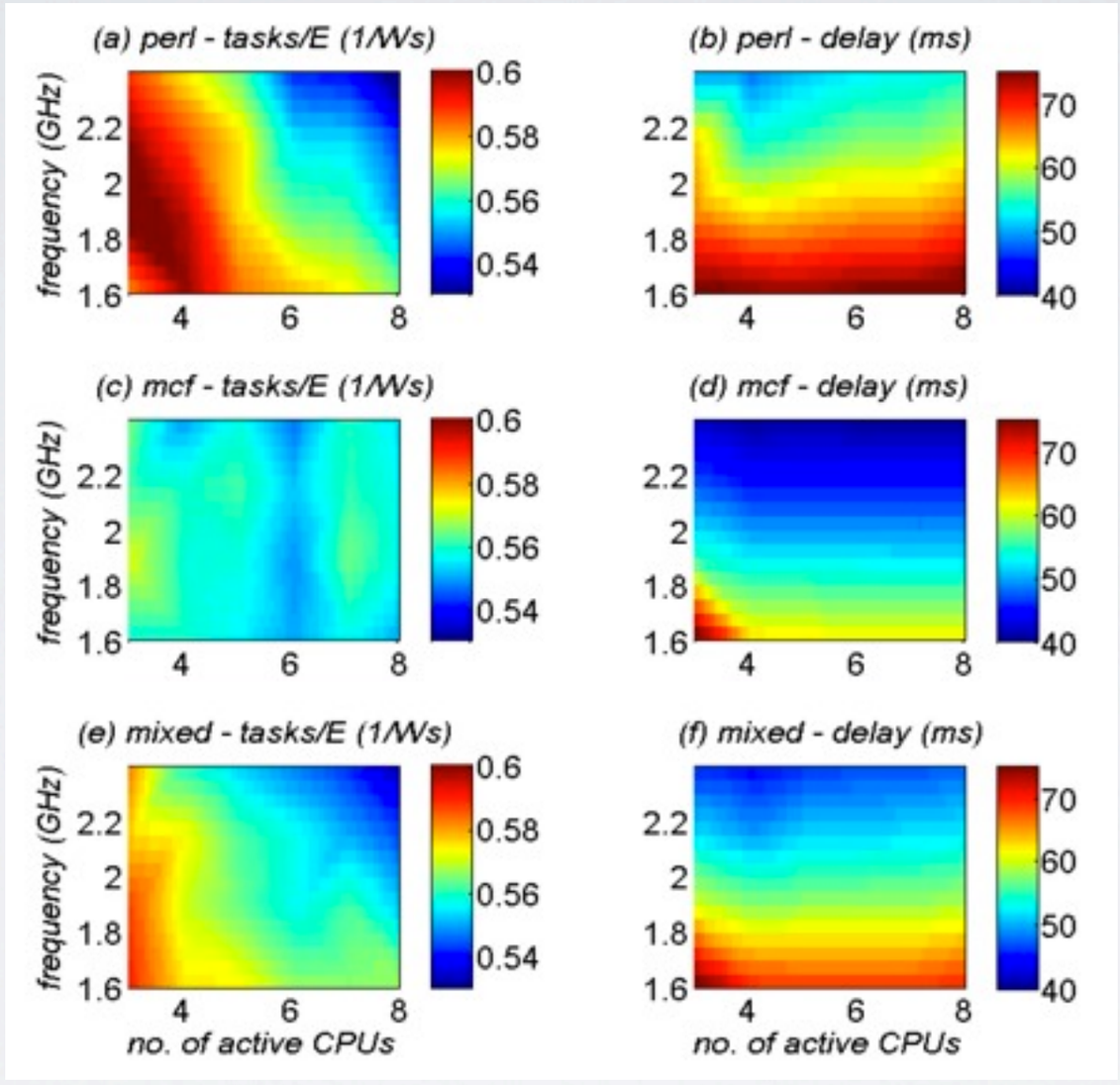
- Main idea: Present *virtual* ACPI states to VMs

- How does virtual ACPI states map to hardware states?

- Soft Scaling

- Information from ACPI can be aggregated in data centers with heterogenous hardware

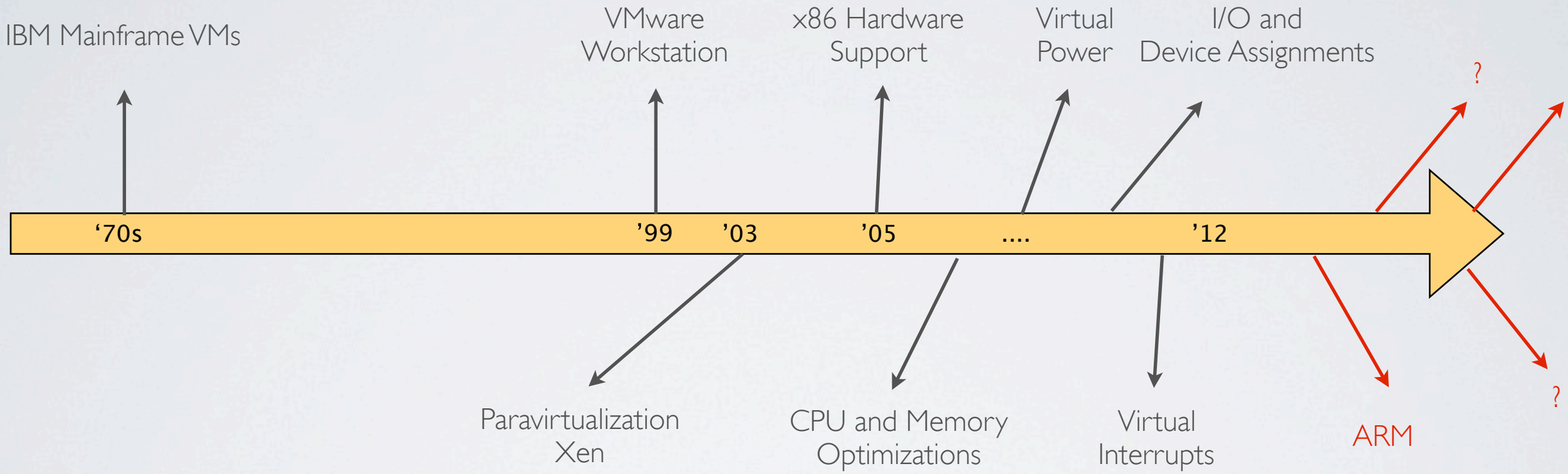
CPU Consolidation



The Effectiveness of CPU Consolidation [Hwang 12]

Interesting related observation about exclusive caches in [Krishnan 2011]

Conclusion


IBM Mainframe VMs
VMware Workstation
x86 Hardware Support
Virtual Power
I/O and Device Assignments
?
?
'70s
'99
'03
'05
....
'12
Paravirtualization Xen
CPU and Memory Optimizations
Virtual Interrupts
ARM
?

VM Power Metering

- Model: Upper and lower bounds on CPU and on Memory [Krishnan 11]
- Power consumption depends on how memory bound
- Depends on cache structures, L1 snooping wakes up other cores

I/O Power Savings

- Prolong sleep periods of mechanical disks [Ye10]
  - Early flushes, and buffering writes
- Expose virtual per-device power state agents to VMs [Tian 10]

vIC [Ahmad 11]

- Virtual Interrupt Coalescing
- Balance coalescing with latency
- Define a ratio R = Virtual IRQs / Hardware IRQs
- The lower R, the more coalescing
- Dynamically determine R based on Commands-In-Flight (CIF) and estimated IOPS