



AMD IOMMU Support on ESX

Wei Huang | September 02, 2009



Outline

- Why I/O virtualization
- AMD IOMMU design
- Demo: ATI graphics passthru
- Performance (graphics and 10G NIC)
- Summary



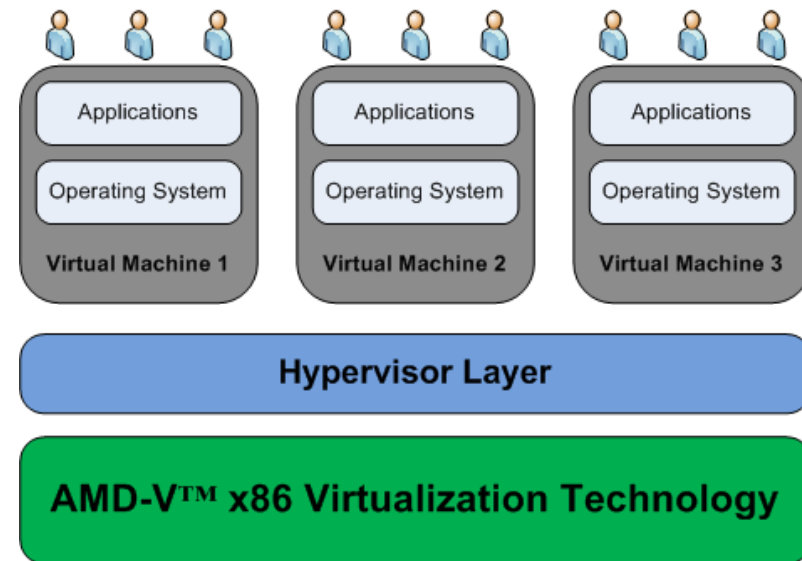
Outline

- Why I/O virtualization
- AMD IOMMU design
- Demo: ATI graphics passthru
- Performance (graphics and 10G NIC)
- Summary

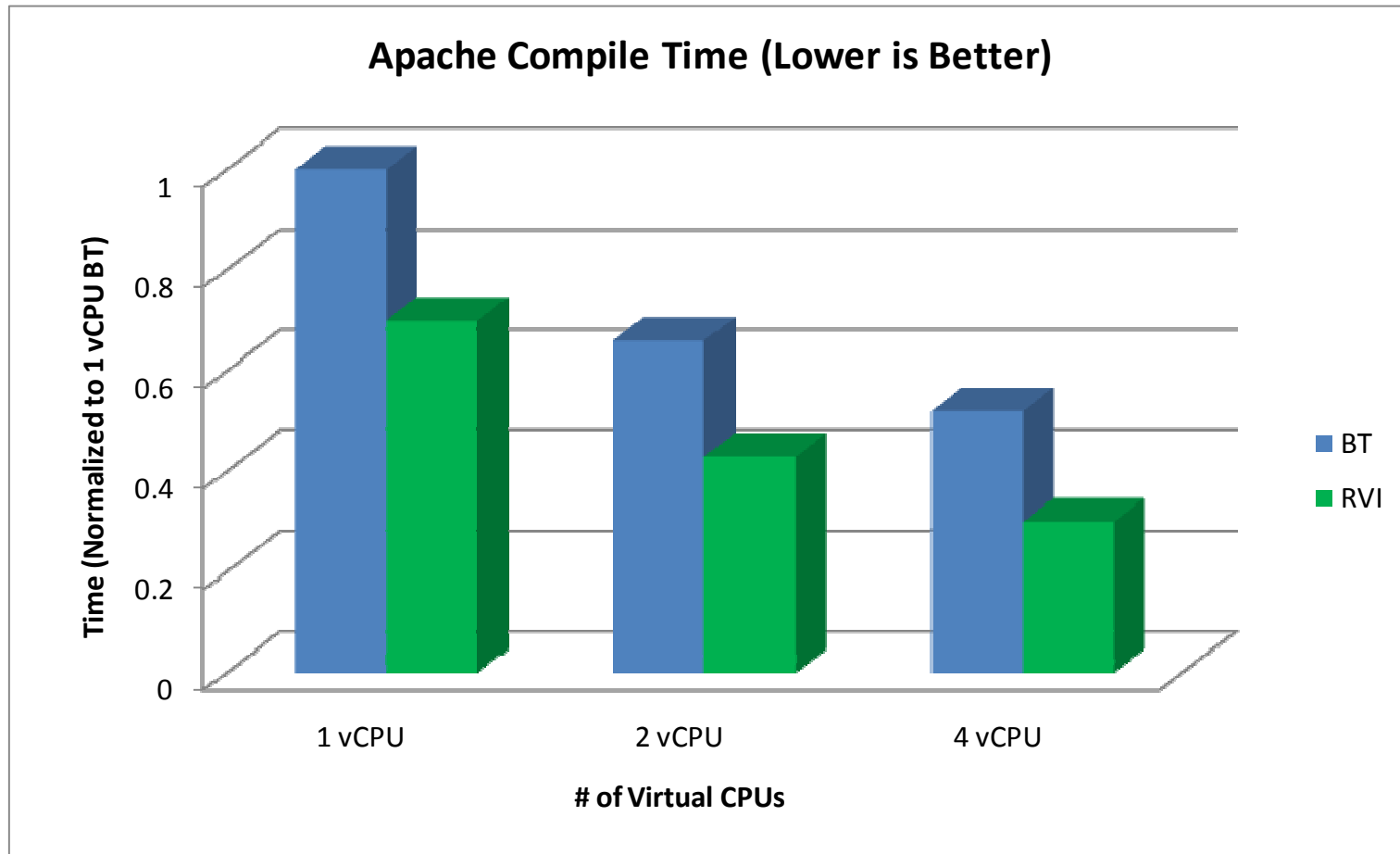


x86 CPU Virtualization

- x86 virtualization has improved dramatically recently.
- AMD-V™ offers innovative x86 virtualization solutions.
 - fast world switch
 - device exclusion vector
 - rapid virtualization indexing
 - page real mode
 - tagged TLB (ASID)



Example: AMD RVI on ESX 3.5

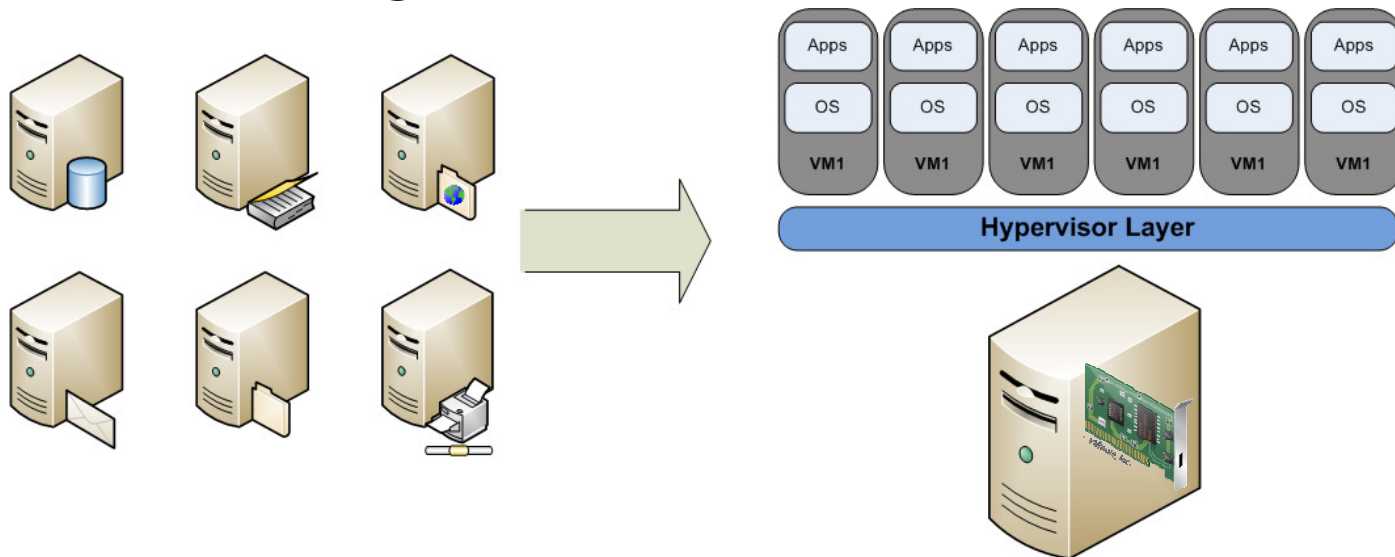


* Source: http://www.vmware.com/pdf/RVI_performance.pdf



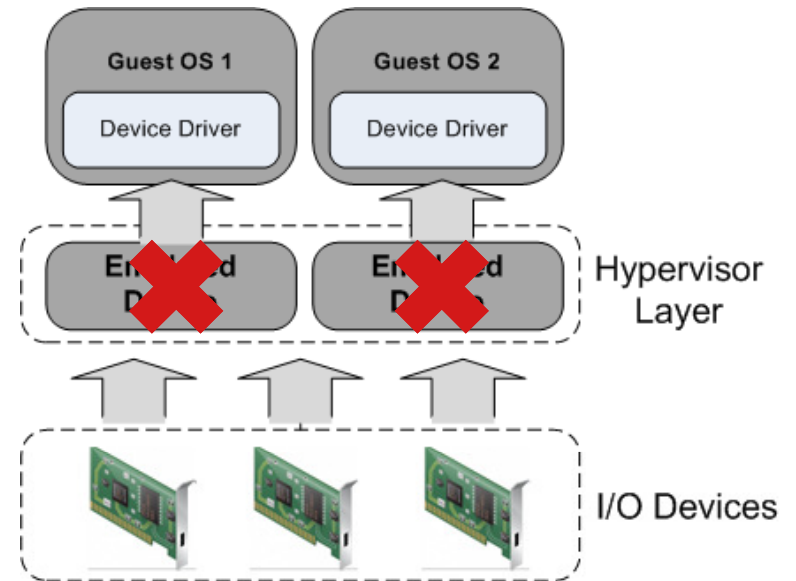
Trend of I/O Devices

- High ratio of server consolidation
 - ~5x guests run on the same server
 - Stress on I/O devices
- I/O device throughput is increasing
 - 10G NICs, storage I/O devices



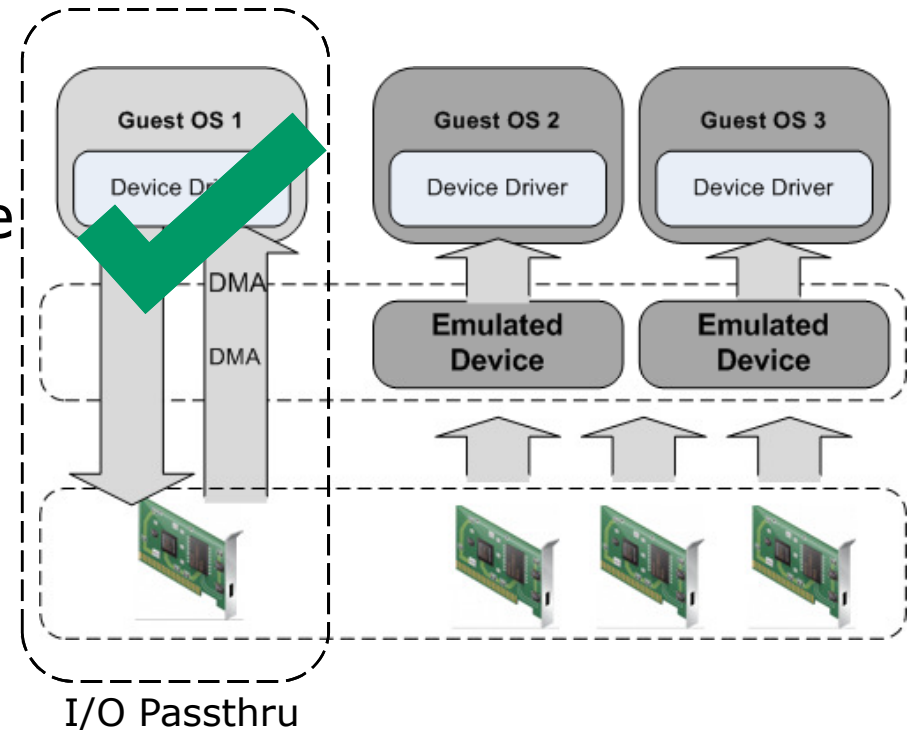
Overhead of Emulated I/O Devices

- DMA overhead
 - Intercept DMA accesses from I/O devices
- Interrupt delivery
 - Interrupt proxy receives and re-directs interrupt to guests



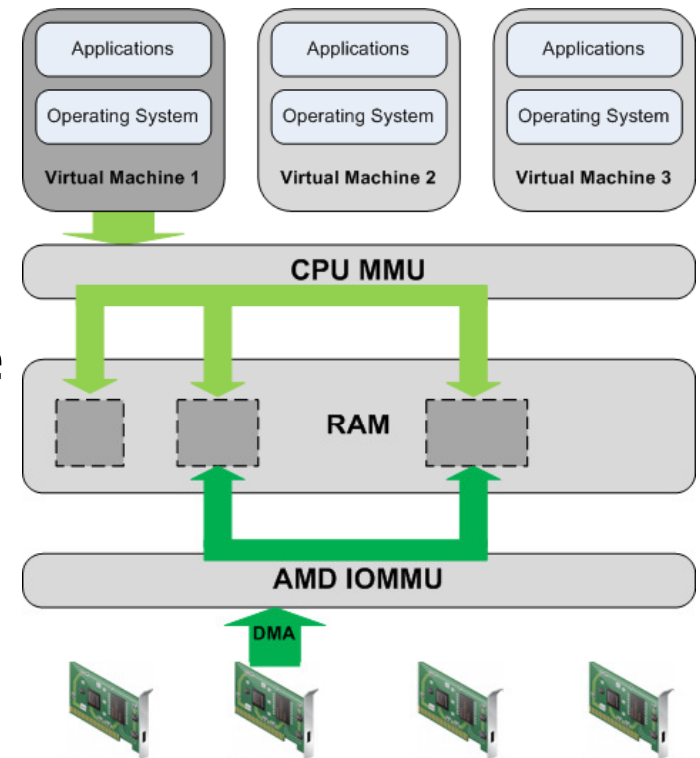
Solution: I/O Passthru

- Direct assignment of PCI devices to VMs
 - Server VMs
 - Big virtual appliances
- Guest driver owns the device



How IOMMU Translation Works

- Requirements
 - Hardware IOMMU in chipset
- Translates DMA accesses
 - An I/O page table for each device
 - Table walker walks I/O page table

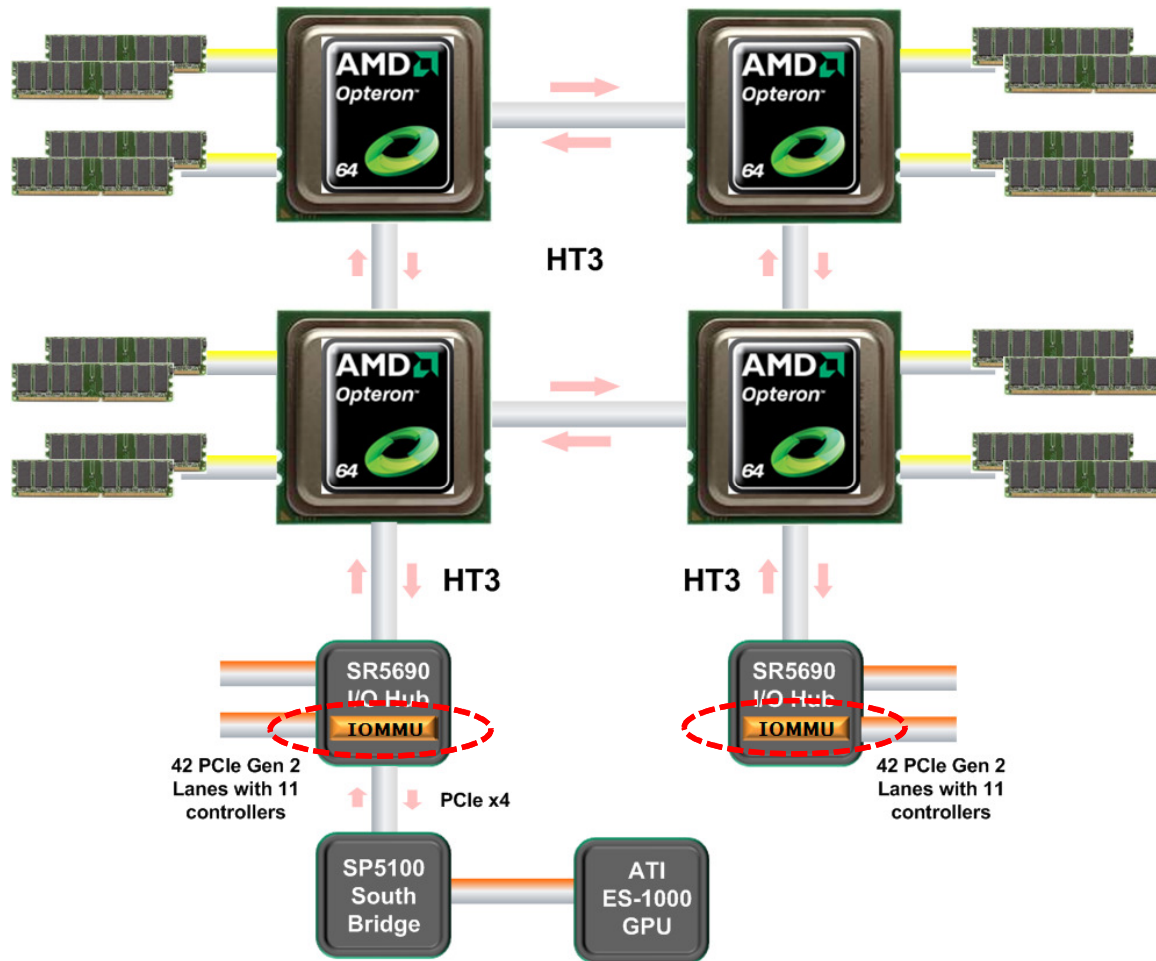


Outline

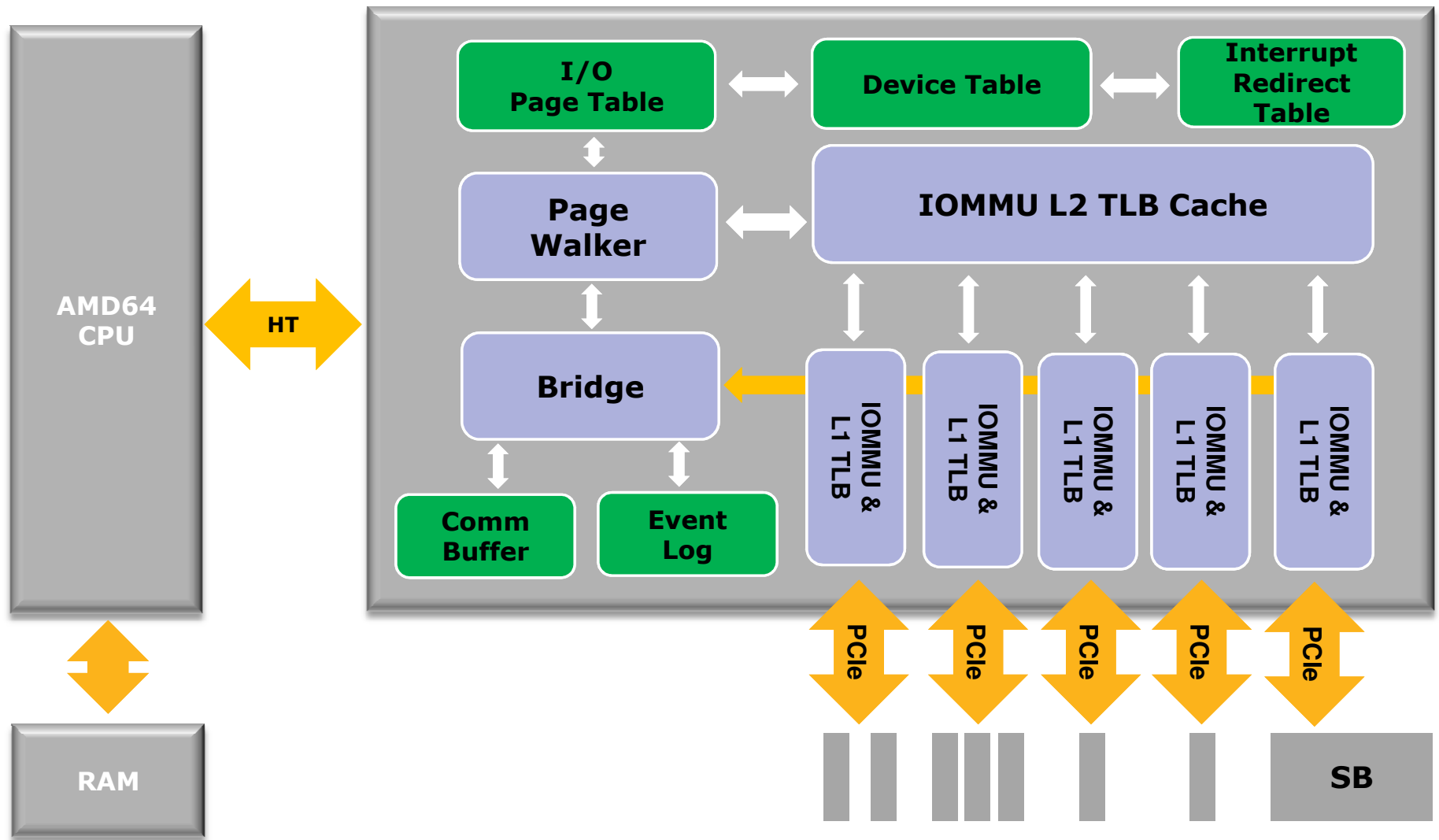
- Why I/O virtualization
- AMD IOMMU design
- Demo: ATI graphics passthru
- Performance (graphics and 10G NIC)
- Summary



Block Diagram of AMD's Fiorano Platform



A Simple IOMMU Diagram



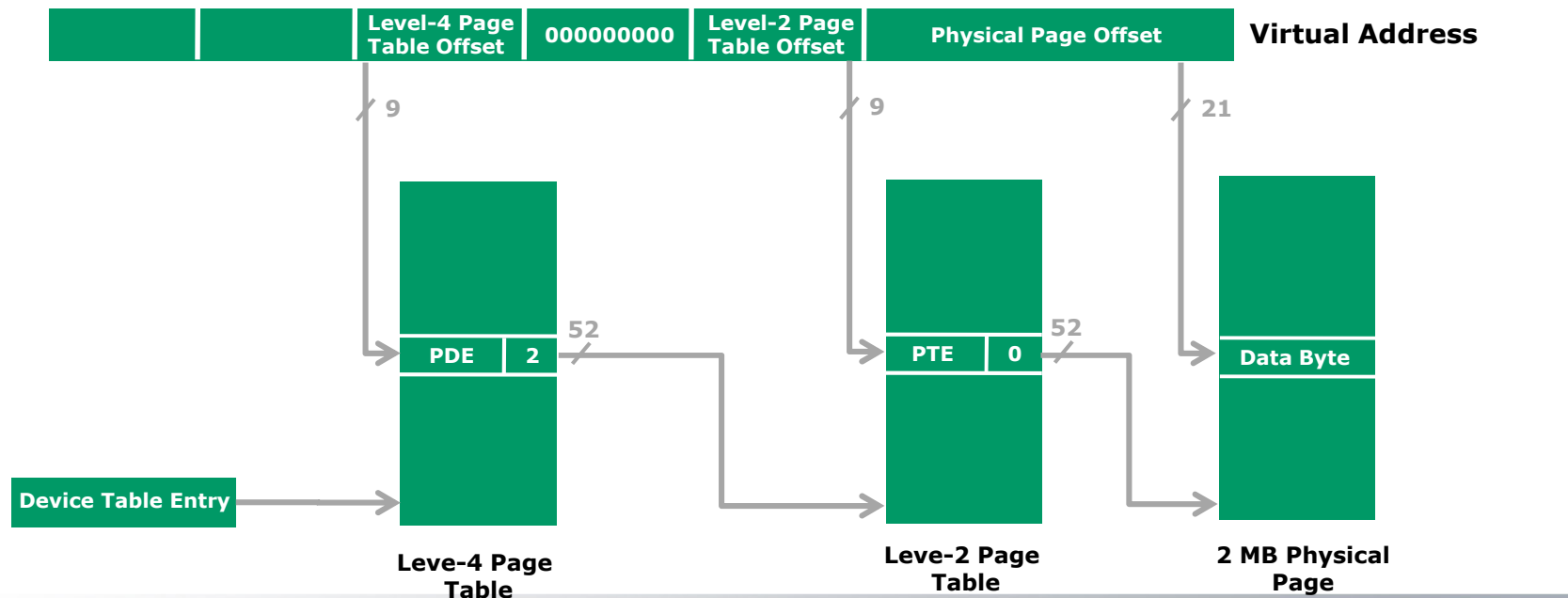
IOMMU Behavior

- Device table points to page table and interrupt remapping table for each device
- Software writes instructions to command buffer which the IOMMU executes asynchronously
- Errors are logged by the hardware to the event log



Special Features of I/O Page Table

- Page table entry is compatible to long mode format
- Supports up to 7 levels with full 64bit address range
- Various page table size (from 4KB to 4GB)
- Level skipping capability



AMD IOMMU Benefits

- Enhanced security
 - Memory protection
 - Domain isolation
- Higher throughput
 - Minimum overhead
- Better compatibility
 - Support non-emulatable devices
 - Legacy support for 32bit devices



AMD IOMMU Support on VMware ESX

- AMD IOMMU is fully supported by VMware ESX 4.0
- Loaded as a module called AMDIommu
- Module detects AMD IOMMU via ACPI IVRS table
- Completely automatic when used with vSphere client



Create Passthru Devices on ESX 4

- Guest configure file

```
...
pciPassthru0.present="TRUE"
pciPassthru0.msiEnabled="TRUE"
pciPassthru0.vmmIntEnabled="TRUE"
pciPassthru0.systemId = "48623f33-9d40-6989-cab1-0015170d948c"
pciPassthru0.ID="2:0.0"
pciPassthru0.deviceId = "4364"
pciPassthru0.vendorId = "11ab"
...
```

- Guest then finds native devices

```
unable to connect to host: No route to host (113)
server:~/esx-build$ vncviewer 10.236.48.234:0
Free Edition 4.1
C) 2002-2005 Realtek
www.realtek.com
16:26:08 2008
connected to
Server support
Using RFB protocol
Using default
Using pixel format
Using ZRLE encoding
Throughput 20
Throughput 20
Using pixel format
Using hex tile
mytesttesttt:~ # lspci
00:00.0 Host bridge: Intel Corporation 440BX/ZX/DX - 82443BX/ZX/DX Host bridge (rev 01)
00:01.0 PCI bridge: Intel Corporation 440BX/ZX/DX - 82443BX/ZX/DX AGP bridge (rev 01)
00:07.0 ISA bridge: Intel Corporation 82371AB/EB/MB PIIX4 ISA (rev 08)
00:07.1 IDE interface: Intel Corporation 82371AB/EB/MB PIIX4 IDE (rev 01)
00:07.3 Bridge: Intel Corporation 82371AB/EB/MB PIIX4 ACPI (rev 08)
00:0f.0 VGA compatible controller: VMware Inc [VMware SVGA III] PCI Display Adapter
00:10.0 SCSI storage controller: LSI Logic / Symbios Logic 53c1030 PCI-X Fusion-MPT Dual Ultra320 SCSI (rev 01)
00:11.0 PCI bridge: VMware Inc Unknown device 0790 (rev 02)
00:15.0 PCI bridge: VMware Inc Unknown device 07a0 (rev 01)
03:00.0 Ethernet controller: Marvell Technology Group Ltd. 88E8056 PCI-E Gigabit Ethernet Controller (rev 12)
mytesttesttt:~ #
```



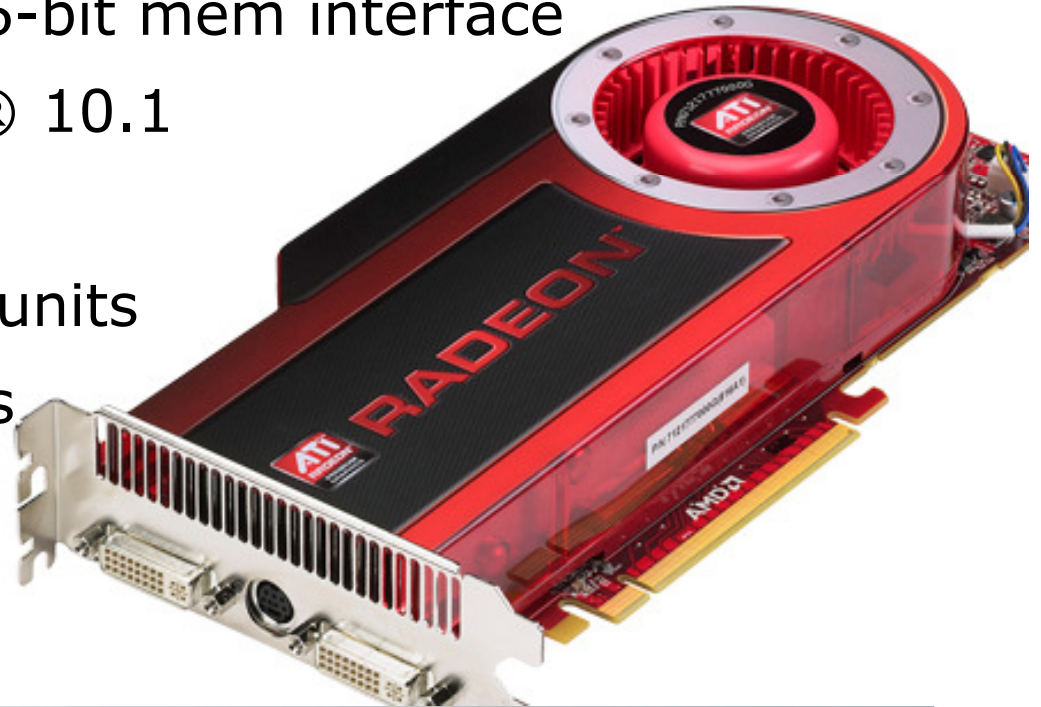
Outline

- Why I/O virtualization
- AMD IOMMU design
- Demo: ATI graphics passthru
- Performance (graphics and 10G NIC)
- Summary

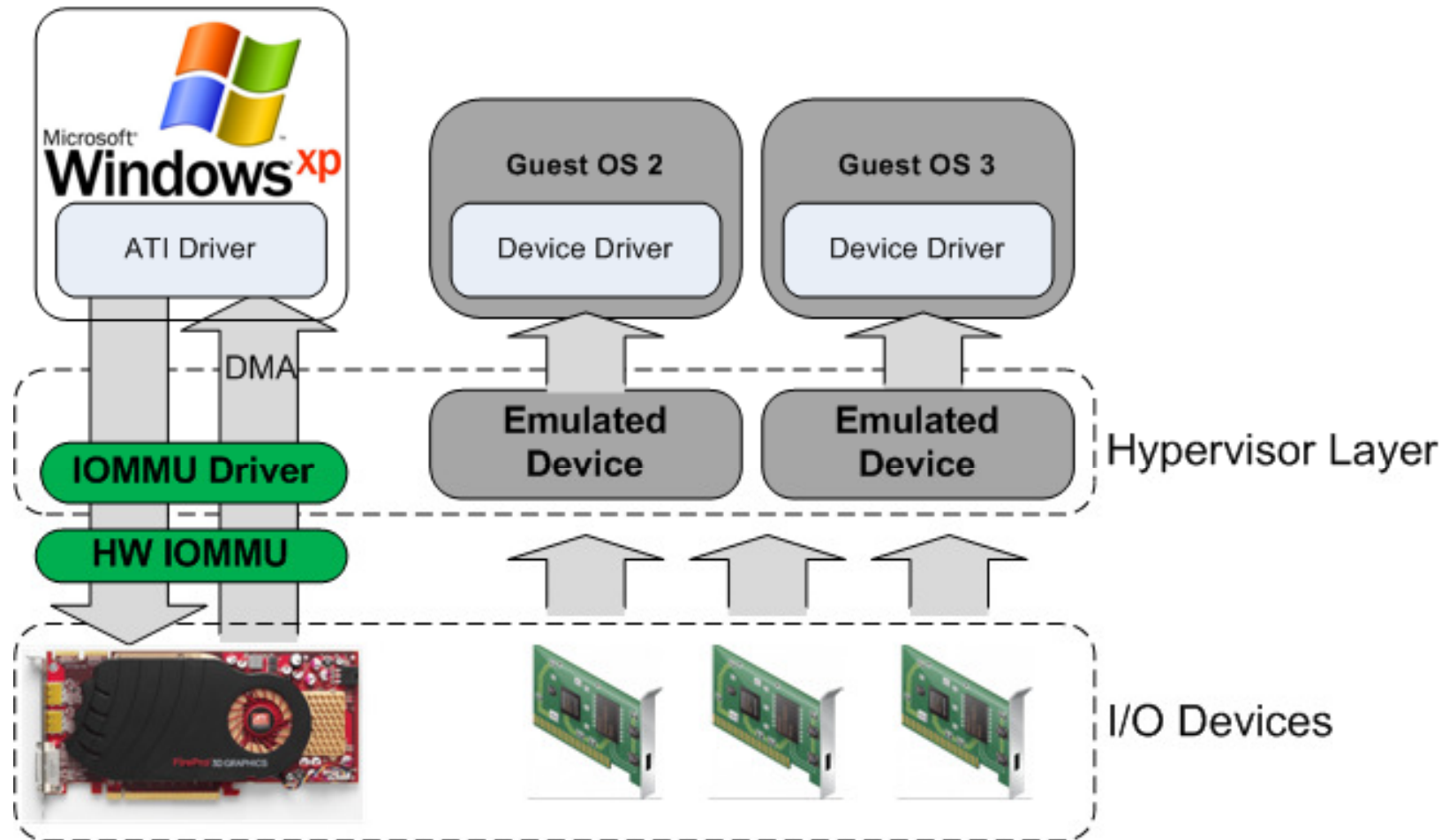


ATI Graphics Passthru on ESX 4

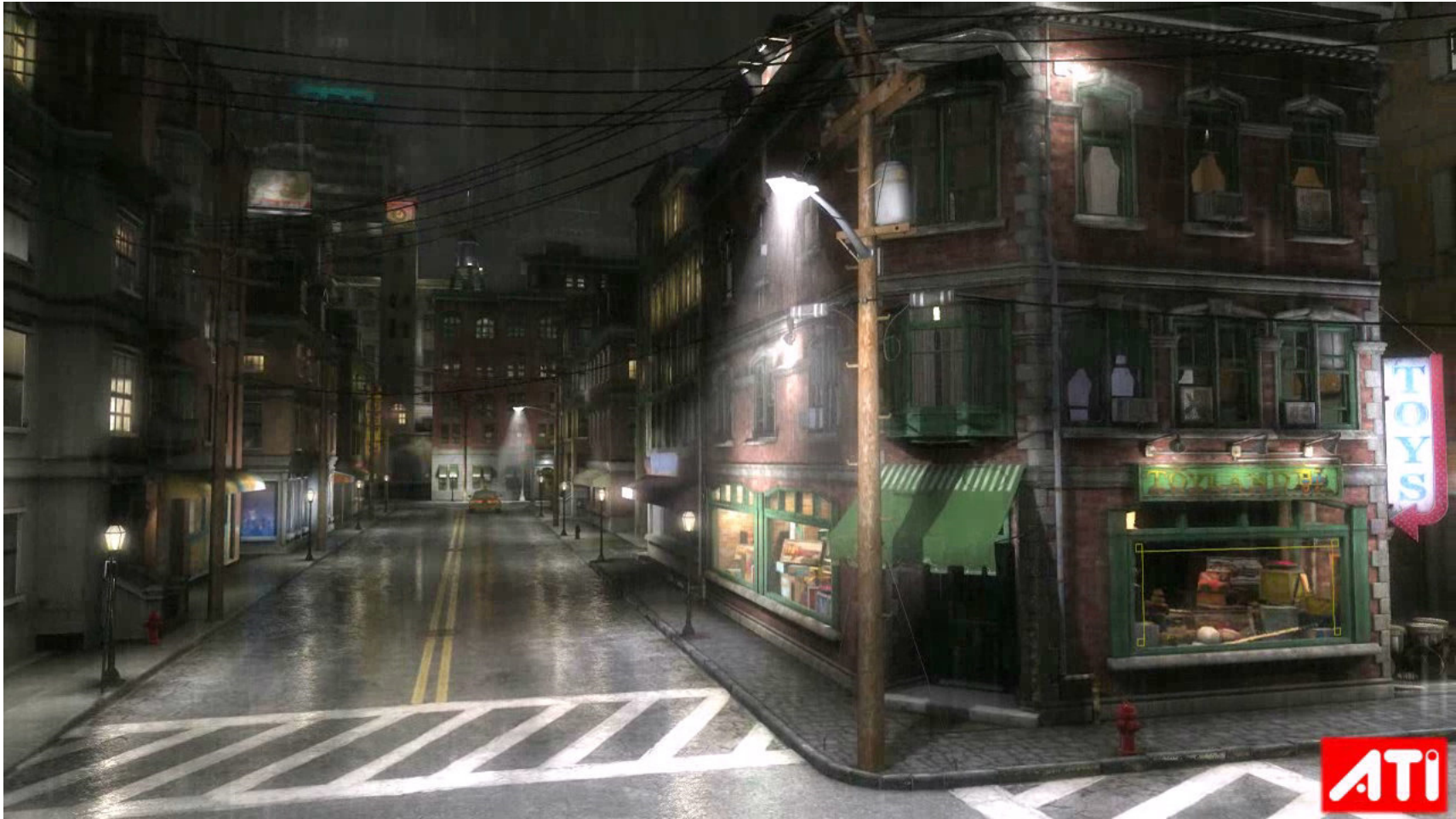
- Windows XP SP3 guest (2 vCPUs + 2GB mem)
- ATI Radeon 4870
 - 800 stream processing units (1.2 Tera FLOPs)
 - 512MB GDDR5 with 256-bit mem interface
 - PCI-E 2.0 with DirectX® 10.1
- ATI Firepro™ V5700
 - 320 stream processing units
 - 3D workstation graphics



ATI Graphics Passthru Diagram



Demos : ATI Toy Shop

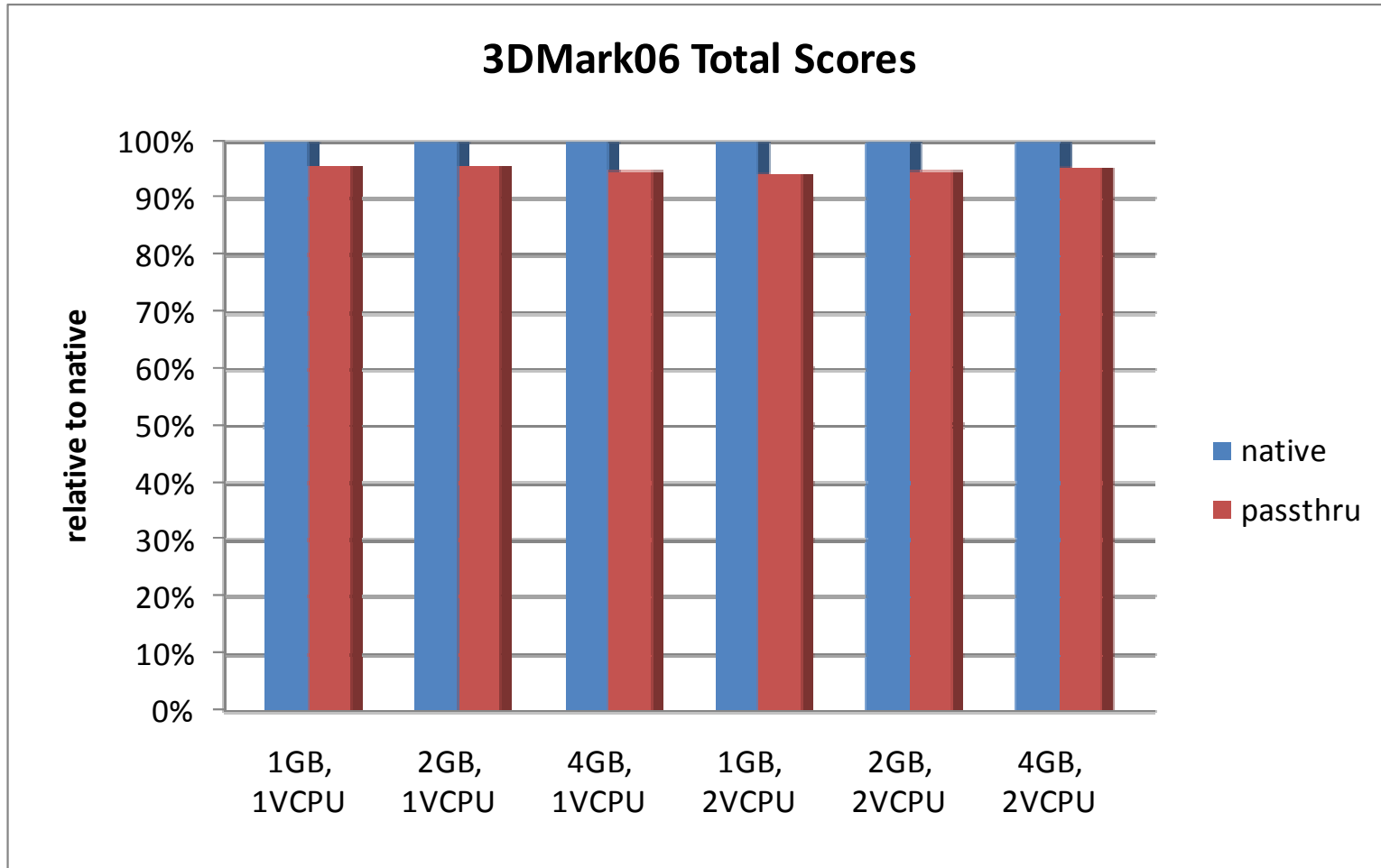


Outline

- Why I/O virtualization
- AMD IOMMU design
- Demo: ATI graphics passthru
- Performance (graphics and 10G NIC)
- Summary



ATI 4870 Passthru Performance

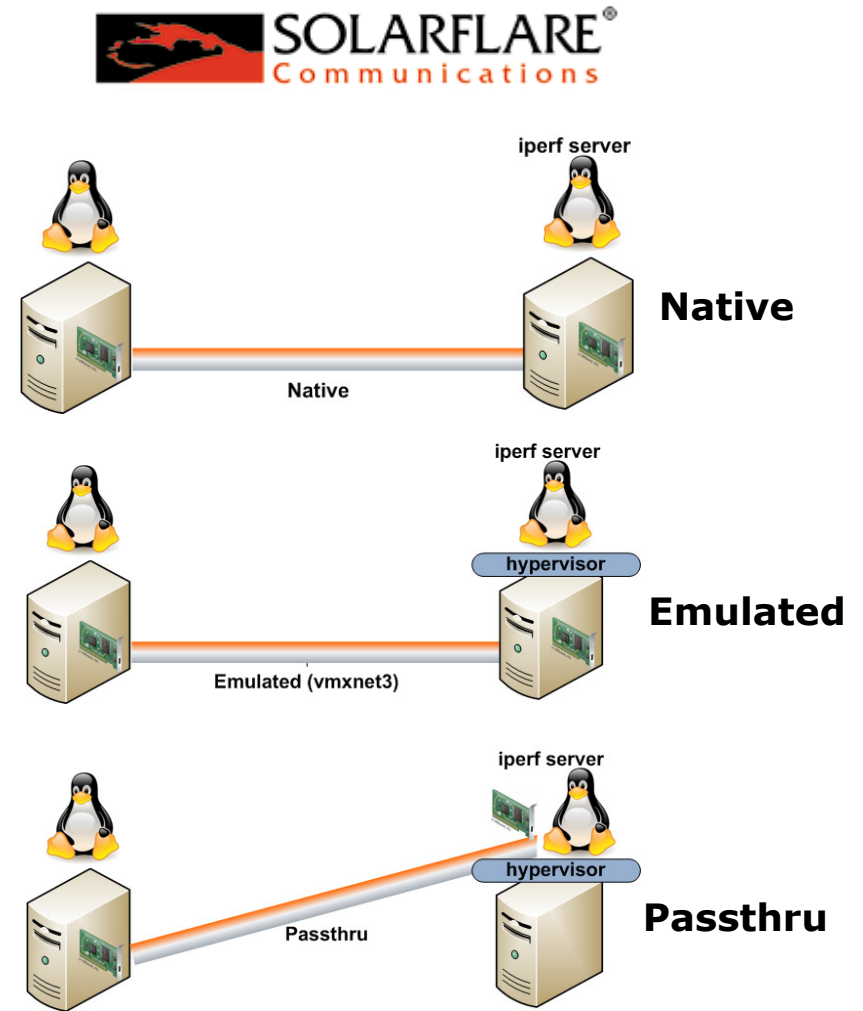


- (1) Performance test results may not be indicative of actual results when tested or deployed in user environments.
- (2) Test configuration not currently certified or supported by VMware not currently certified or supported by VMware

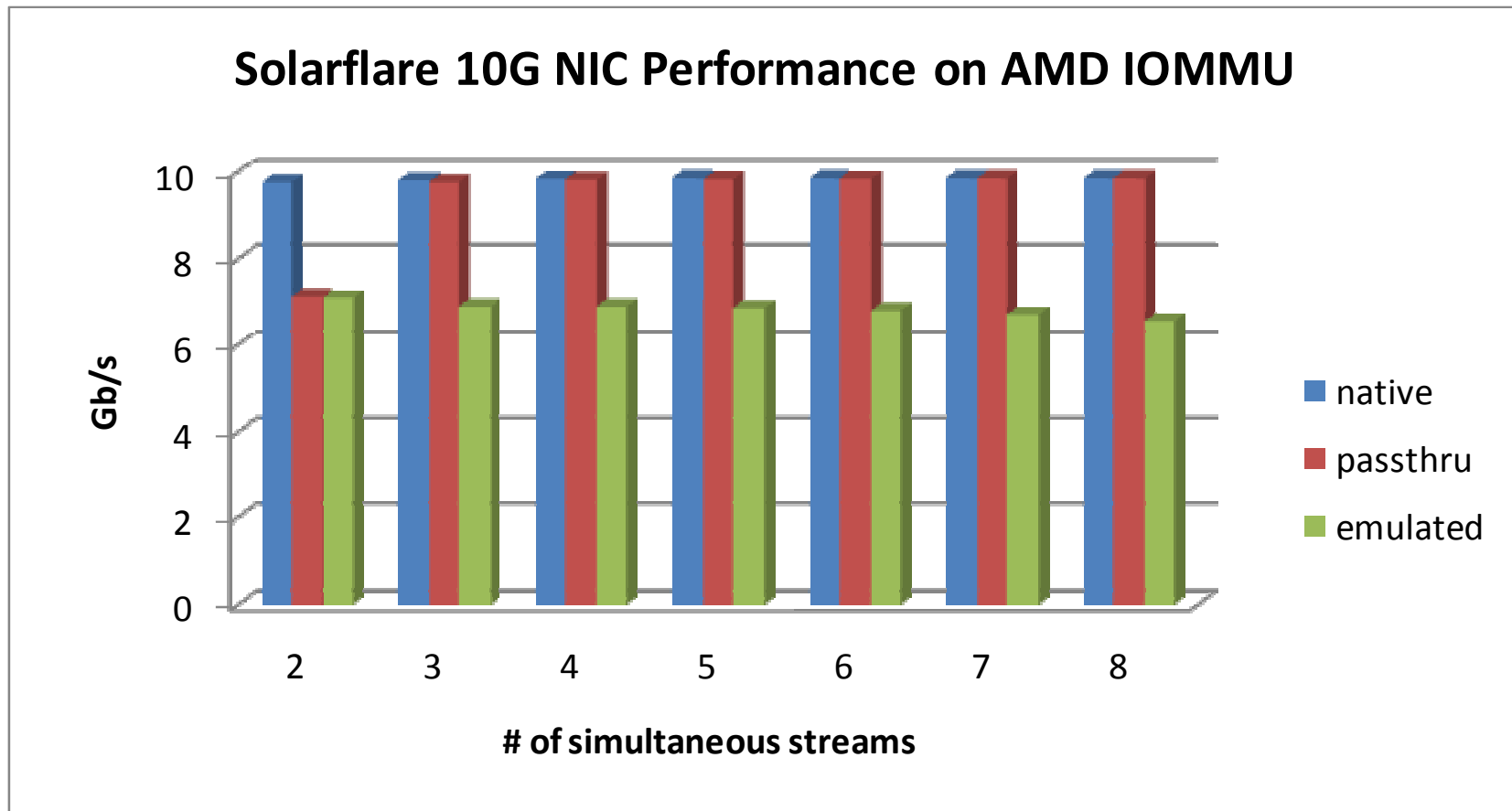


Solarflare 10G NIC Passthru Test

- Solarflare 10G NICs
 - Solarstorm SFC4000
- Tests
 - RHEL 5.3 64bit
 - MTU=9000
 - Guest VM as iperf server
- Three modes
 - Native
 - Emulated (vmxnet3)
 - Passthru



10G NIC Performance



- (1) Performance test results may not be indicative of actual results when tested or deployed in user environments.
- (2) Test configuration not currently certified or supported by VMware not currently certified or supported by VMware



Outline

- Why I/O virtualization
- AMD IOMMU design
- Software interface of AMD IOMMU
- Demo: ATI graphics passthru
- Performance
- **Summary**



Summary

- I/O virtualization is a key to achieve zero-overhead virtualization.
- AMD IOMMU offers rich features for I/O virtualization.
- AMD SR5690 chipset can meet the demands of high-performance PCIe devices (graphics, 10G NICs, etc.).
- VMware ESX 4.0 fully supports AMD IOMMU.



Q & A



Trademark Attribution

AMD, the AMD Arrow logo and combinations thereof are trademarks of Advanced Micro Devices, Inc. in the United States and/or other jurisdictions. Other names used in this presentation are for identification purposes only and may be trademarks of their respective owners.

©2009 Advanced Micro Devices, Inc. All rights reserved.



Backup: SR-IOV

