

Microsoft
WinHEC
2005

***Pacifica* – Next Generation Architecture for Efficient Virtual Machines**

Steve McDowell
Division Marketing Manager
Computation Products Group
AMD
steven.mcdowell@amd.com

Geoffrey Strongin
Platform Security Architect
Computation Products Group
AMD
geoffrey.strongin@amd.com



Session Outline

- Driving Towards Virtualization
 - Solving the IT Department's Utilization Dilemma
 - Virtual Machine Approaches
 - System Architecture Matters
 - X86 Needs Help
- Pacifica Architecture
 - Core Architecture
 - Access Control
 - Interrupts
 - Secure System Management Mode
 - Device Protection

Session Goals

- Attendees should leave this session with the following:
 - A better understanding of virtualization use cases
 - Understanding of hardware assist for virtualization and AMD's Pacifica Technology
 - Knowledge of where to find resources for learning more about AMD and virtualization

Virtualization

Virtualization

is the pooling and abstraction of
resources

in a way that masks the physical nature
and boundaries of those resources

from the resource users

Problems With “Physical Boundaries”

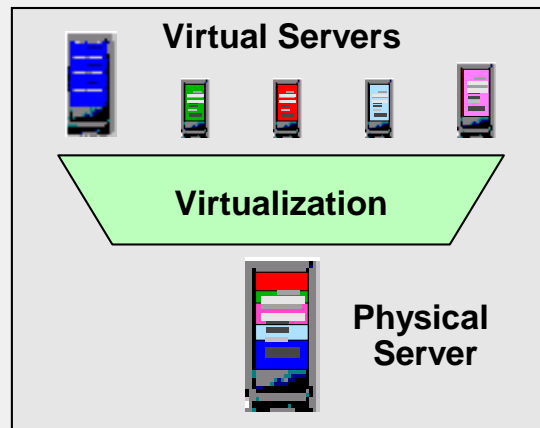
- Today, IT Departments often have lots of pools of excess capacity and no way to share them
 - Most applications are small
 - 93% of x86 server are 1 or 2-way
 - Small applications don't consume servers
 - Applications typically have dynamic workloads
 - Currently, x86 Servers run at 10–20% utilization
 - Mainframes typically run at 75-85% utilization
- The costs add up for running lots of under-utilized servers

Virtualization in Servers

Server Virtualization Example

Roles:

- Consolidations
- Dynamic provisioning/hosting
- Workload management
- Workload isolation
- Software release migration
- Mixed production and test
- Mixed OS types/releases
- Reconfigurable clusters
- Low-cost backup servers



Benefits:

- Higher resource utilization
- Greater usage flexibility
- Improved workload QoS
- Higher availability / security
- Lower cost of availability
- Lower management costs
- Improved interoperability
- Legacy compatibility
- Investment protection

Benefits

- Reduced Hardware Cost
 - Higher physical resource utilization
 - Smaller footprint (power, space, cooling, etc.)
- Improved flexibility and responsiveness
 - Resources can be adjusted dynamically
 - Enables On Demand and Adaptive Enterprise operating environments

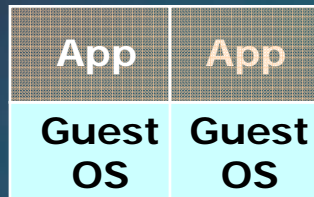
Virtualization in Clients

- Used for legacy support for enterprises who need to support applications on older OS's side-by-side with new technology
- Test & Development
 - Isolate development environments from production work
- Emerging use cases for management partitions, reducing IT support costs
- Heart of next generation security – allow trusted & untrusted partitions to co-exist,
 - Having partitions with different levels of security. The environment allows security policies to be reinforced.

Virtual Machine Approaches

Carve a Server into Many Virtual Machines

Hosted Virtualization



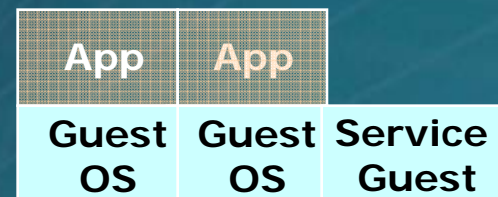
Virtualization Software

Host Operating System

X86 Hardware

- Virtualization software manages resources between Host and Guest OS's
- Application can suffer decreased performance due to added overhead

Hypervisor-based Virtualization



Hypervisor

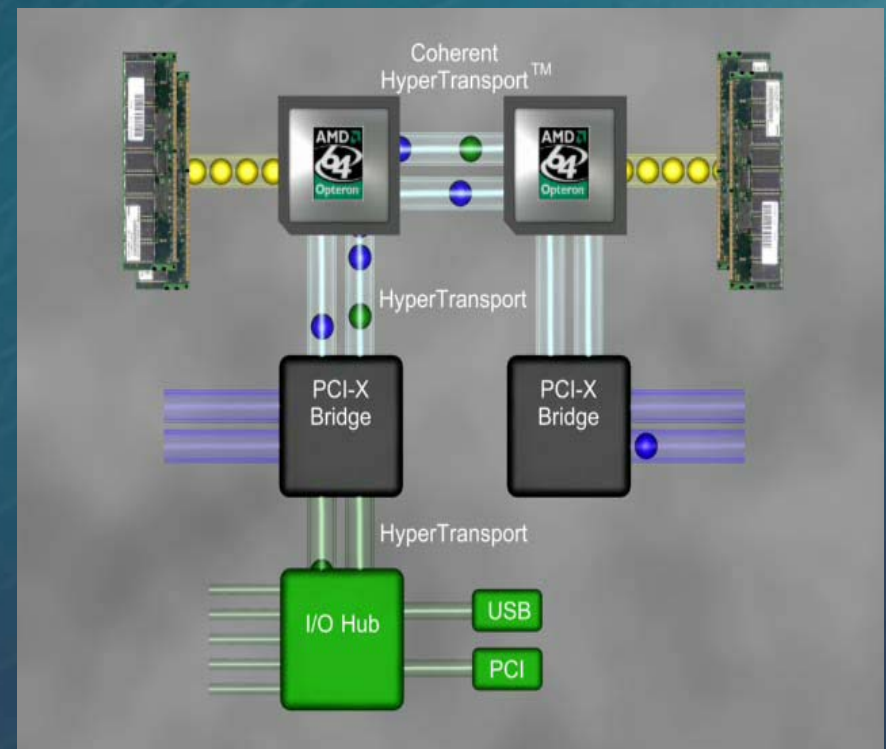
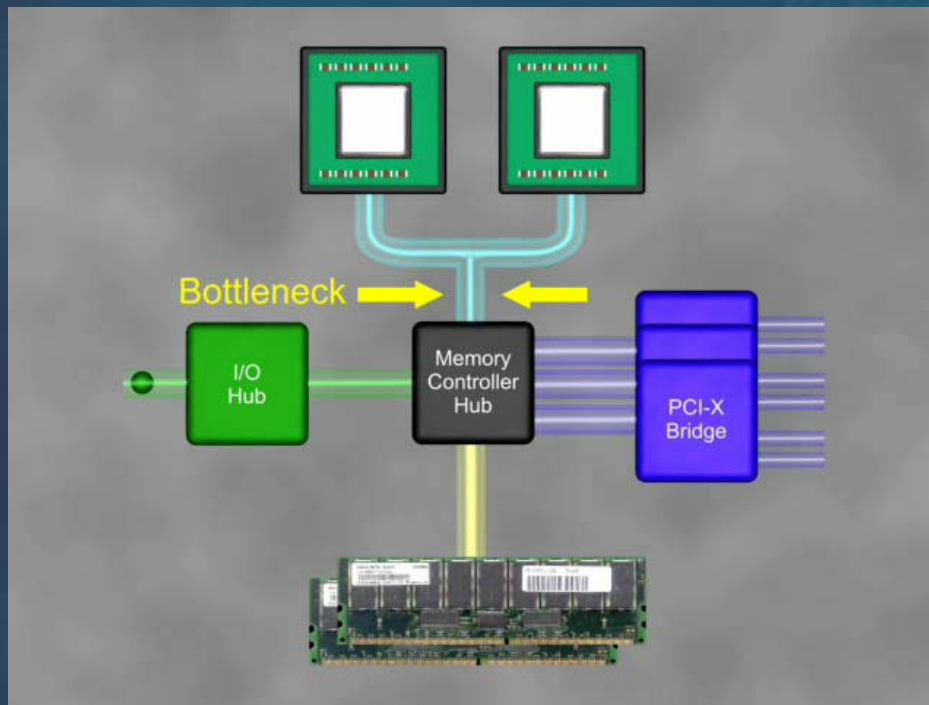
AMD64 w/ Pacifica

- Virtualization Software (Hypervisor) is the host environment.
- Enables better software performance by eliminating some of the associated overhead
- If Hardware is available, the Hypervisor can be designed to take advantage of it

System Architecture Makes a Difference

- Legacy Architectures based around front-side bus aren't scalable for today's virtualization needs

- AMD's Direct Connect Architecture removes the bottlenecks, enabling efficient partitioning



Examples: Today's Server Architectures

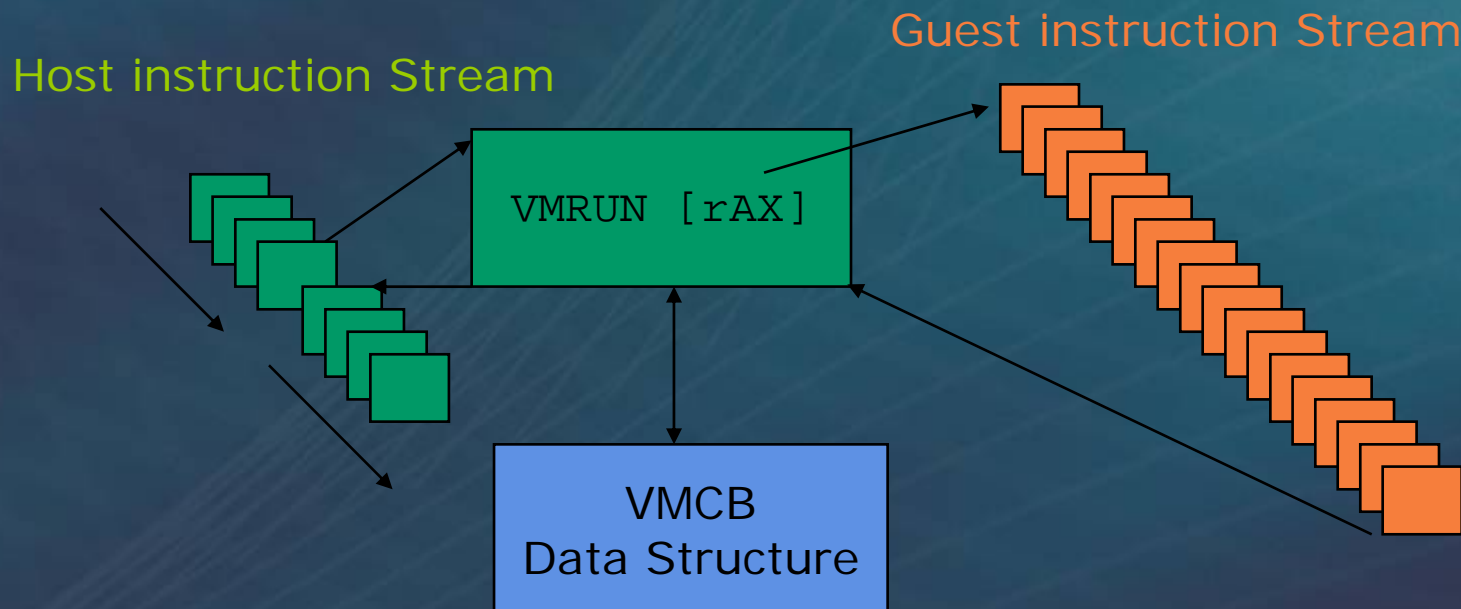
Efficiencies Needed on X86 for Virtualization

- Virtualization on the existing X86 architecture requires “unnatural acts” to achieve objectives
 - This level of emulation and code rewriting is not required on other architectures
- Existing approaches add performance overhead and undue complexity, leave security holes at the most physical levels
- AMD’s *Pacifica* technology takes the complexity out of the hypervisor, putting it into the CPU for higher performance, higher security, and less complexity
- Pacifica brings the X86 into the 21st century
 - On to the Pacifica Architecture...

Core Pacifica Architecture

Core Pacifica Architecture: Virtual Machine Run

- Virtualization based on Virtual Machine Run (**VMRUN**) instruction
- VMRUN executed by host causes the guest to run
- Guest runs until it exits back to the host
- World-switch: host \rightarrow guest \rightarrow host
- Host resumes at the instruction following VMRUN



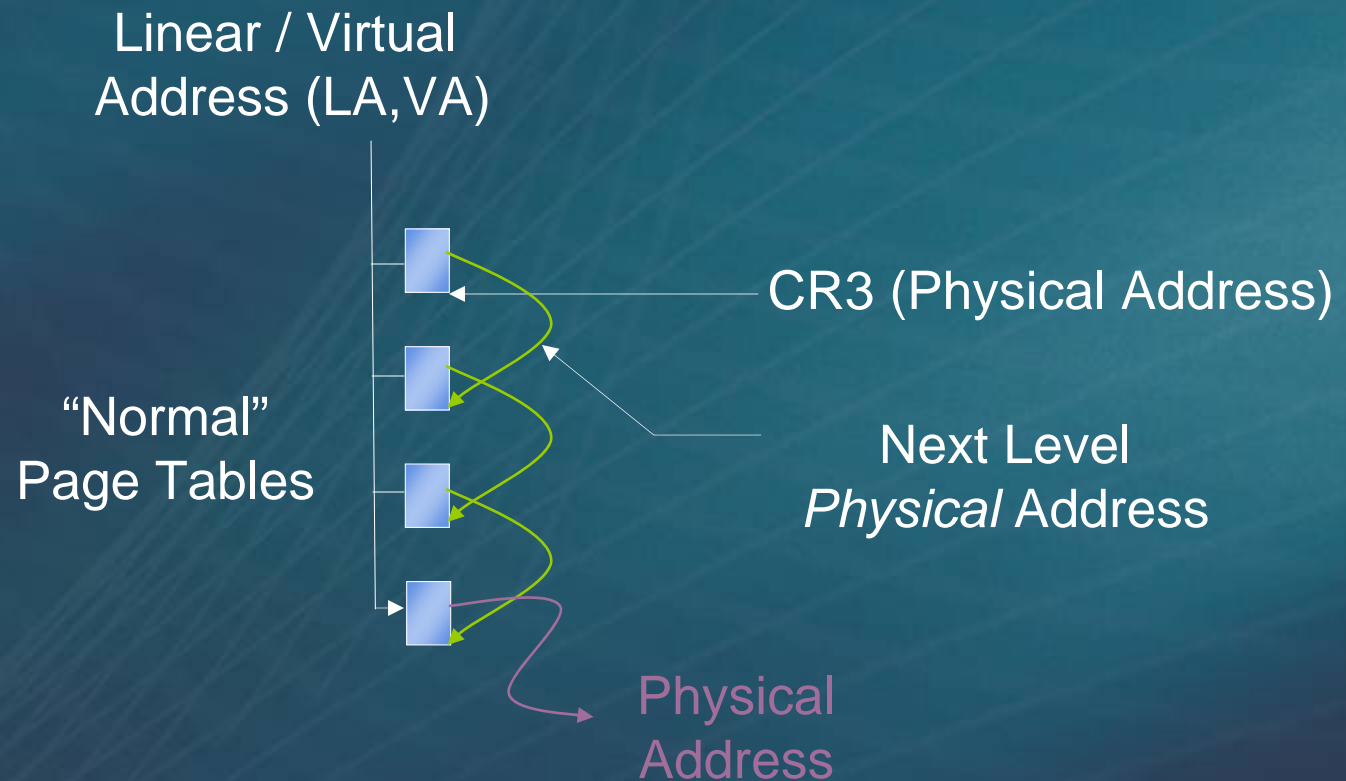
Core Pacifica Architecture: Intercepts

- Guest runs until:
 - It performs an action that causes an exit to the host
 - It explicitly executes the `VMMCALL` instruction
- The VMCB for a guest has settings that determine what actions cause the guest to exit to host
 - These **intercepts** can vary from guest to guest
 - Two kinds of intercepts
 - Exception & Interrupt Intercepts
 - Instruction Intercepts
 - Rich set of intercepts allow the host to set customize each guest's privileges
- Information about the intercepted event is put into the VMCB on exit

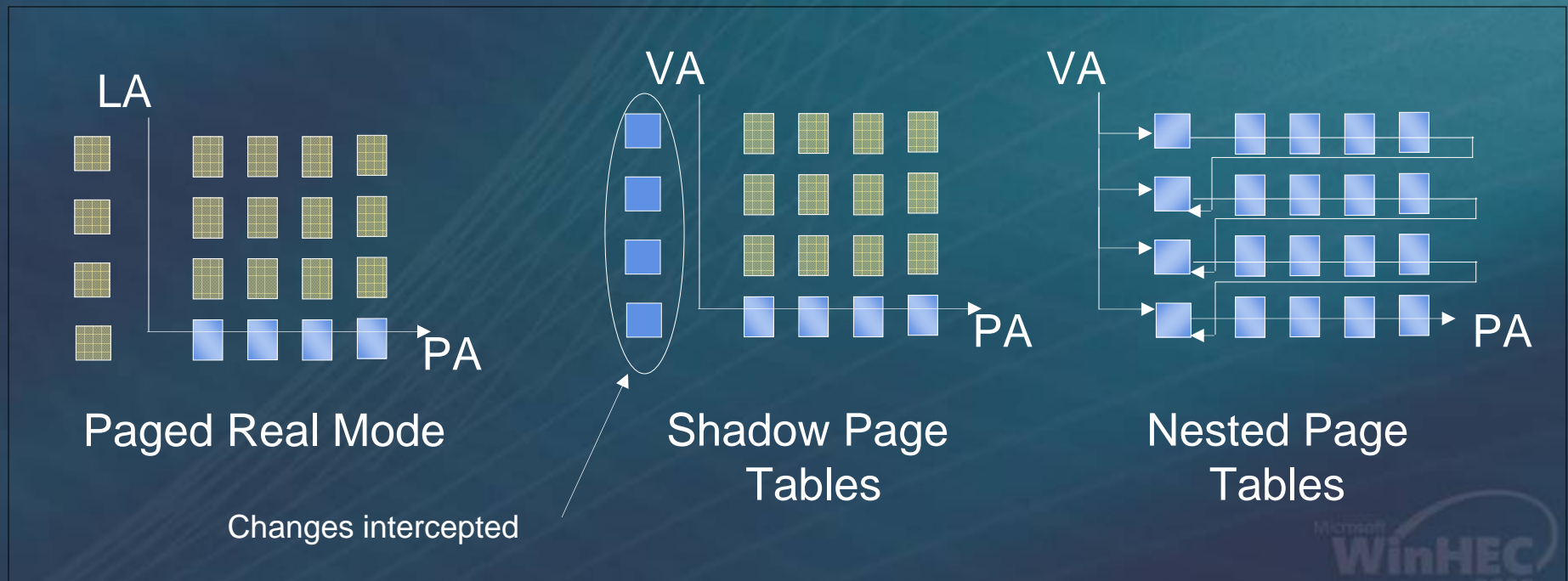
Core Pacifica Architecture: Virtual Machine Control Block

- All CPU state for guest is located in the Virtual Memory Control Block (**VMCB**) data structure
- VMRUN: Entry
 - Host state is saved to memory
 - Guest state loaded from VMCB
 - Guest runs
- VMRUN: Exit
 - Guest state is saved back to VMCB
 - Host state loaded from memory
- Host state saved using Model Specific Register (**MSR**): `vm_hsave_pa`

Address Translation: Page Tables



Address Translation: Modes w/Virtualization



Core Pacifica Architecture: Shadow Page Tables

- Memory Protection – CPU accesses
 - Shadow Page Tables (SPT)
 - Nested Page Tables
- SPT Constraints on host design
 - Host intercepts guest CR3 Reads/Writes
 - Host monitors guest edits to guest page tables
 - Guest page tables are marked “read only”
 - Host constructs and manages SPT in software
 - Software strategies for this are mature
- Guest never sees the “real” page tables or the real content of Control Register 3 (CR3)
- Address Space ID’s (ASID) implemented to improve Translation Look-aside Buffer (TLB) performance
 - VMRUN sets guest ASID

Core Pacifica Architecture: CPU Access Protection

- SPT sets guest access rights to physical address space
 - No guest access is possible unless a mapping is present in the SPT
 - Covers DRAM and Memory Mapped Input/Output (MMIO)
 - Minimum granularity 4k-bytes
- VMCB contains a pointer to an IO Permission Map (**IOPM**) that controls guest access rights to IO Ports
 - Granularity is to 1-byte port
- VMCB contains a pointer to an MSR permission map that control guest access to MSRs

Core Pacifica Architecture: Interrupts

- Processor response to HW interrupts is setup in the VMCB
- Two Options:
 - Hardware interrupts while guest is running are intercepted causing exit to host
 - Host manages physical APIC
 - Host determines interrupt routing and distribution
 - Host injects virtual interrupts into guests as needed
 - Hardware support for virtual interrupts:
`v_irq, v_vector, v_prio, v_tpr, PHYS_IF`
 - Interrupts serviced directly in the guest
 - Guest manages physical APIC
 - Host can still inject virtual interrupts
- Global Interrupt Flag (**GIF**)
 - Protects host code critical-regions

Core Pacifica Architecture: System Management Mode

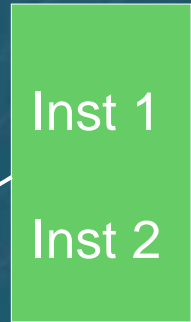
- Pacifica implements a flexible architecture for System Management Interrupt (SMI)/SMM
 - Full legacy support for SMI from within host or guest
 - SMI Intercepts:
 - Allow host to scrub state if needed followed by native SMI from host
 - Support for “containerized” SMM
 - SMM Mode control via **SMM_CTL_MSR**
 - Allow host to scrub state and dispatch the SMM handler from a VMCB

Pacifica: Containerized SMM Flow

Host

```
Top:  
...  
VMMRUN [rAX]  
...  
(Examine Exit Code)  
...  
If external SMM  
(Setup SMM save state)  
VMRUN [rAX]  
...  
Loop Top
```

Guest



Inst 1

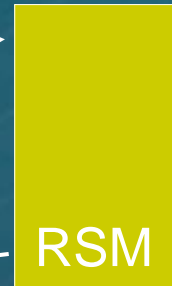
Inst 2

SMI

SMI Intercept

SMM Code

SMM Entry Point



RSM

RSM Intercept

SMM Save State

Pacifica: Paged Real Mode (New)

- SMM code is designed to start in real mode
- Memory protections rely on paging, guests *must* run with paging enabled
- Pacifica Solution: Paged Real Mode
 - Only available for guests
 - `cr0.pg=1, cr0.pe=0`
 - Host must intercept page faults
 - Real-mode address translation (segment+offset) = Linear address → translation via SPT → physical address
 - Correct composition of SPT's is host responsibility
 - Guest is assuming linear, 0-based mapping

Pacifica: DMA Protection

- Protection Domains
 - Mapping from bus/device ID to protection domain
- Device Exclusion Vector (DEV)
 - One DEV per protection domain
 - Permission-checks all upstream accesses
 - 1 bit per physical 4K page (0.003% tax; 128K / 4G) of the system address space
 - Protection for both DRAM and Memory Mapped IO space
 - *Contiguous* table in physical memory

Summary Slide

- Virtualization is being used in several Server scenarios today
- AMD expects that virtualization will prove valuable for PC clients too
- There are ways to modify the X86 architecture, so that virtualization is easier to accomplish, performs better, and provides more security
- AMD Pacifica Technology is being developed for future AMD64 CPUs for Servers and Clients
- Key technologies include adding new instructions, supporting different methods of handling page tables, handle host and guest interrupts (including SMI/SMM), and provide DMA protection

Call To Action

- Read the *Pacifica* specification to understand hardware assisted virtualization, available at www.amd.com
- Continue to ensure that your device and driver works with AMD64 on ALL 64-bit enabled Windows operating systems.
 - Pacifica Technology is for AMD64 CPUs
- Access Developer information at DevCentral <http://www.amd.com/devcentral>
- Sign up for AMD's development center at <http://devcenter.amd.com>

Additional Resources

- Web Resources:
 - Main Page: <http://www.amd.com>
 - Developer Center: <http://devcenter.amd.com>
 - DevCentral: <http://www.amd.com/devcentral>

questions

Microsoft[®]

Your potential. Our passion.[™]

© 2005 Microsoft Corporation. All rights reserved.

This presentation is for informational purposes only. Microsoft makes no warranties, express or implied, in this summary.