Introduction to Virtualization Technology

Argentina Software Development Center

Software and Solutions Group

Gisela Giusti

October 11, 2007

asdc

Argentina Software Development Center

intel Software

Software @ Intel

- 50+ R&D centers in 20+ countries
- 10.000+ software engineers

Western United States

Arizona
Folsom, CA
Santa Clara, CA
Southern CA
Colorado
New Mexico
Portland, OR
Utah
Washington

Eastern / Midwestern United States

Illinois
Massachusetts
New Hampshire
Texas
Virginia

Latin America

Argentina

Israel / Western Europe

Koln, Germany
Munich, Germany
Ulm, Germany
Israel
Stockholm, Sweden
Winnersh, UK

Russia

Moscow
Nizhniy Novgorod
Novosibirsk
Sarov
St. Petersburg

China

Beijing
Hong Kong
Shanghai
Shenzhen
Xi'An Zizhu

Asia

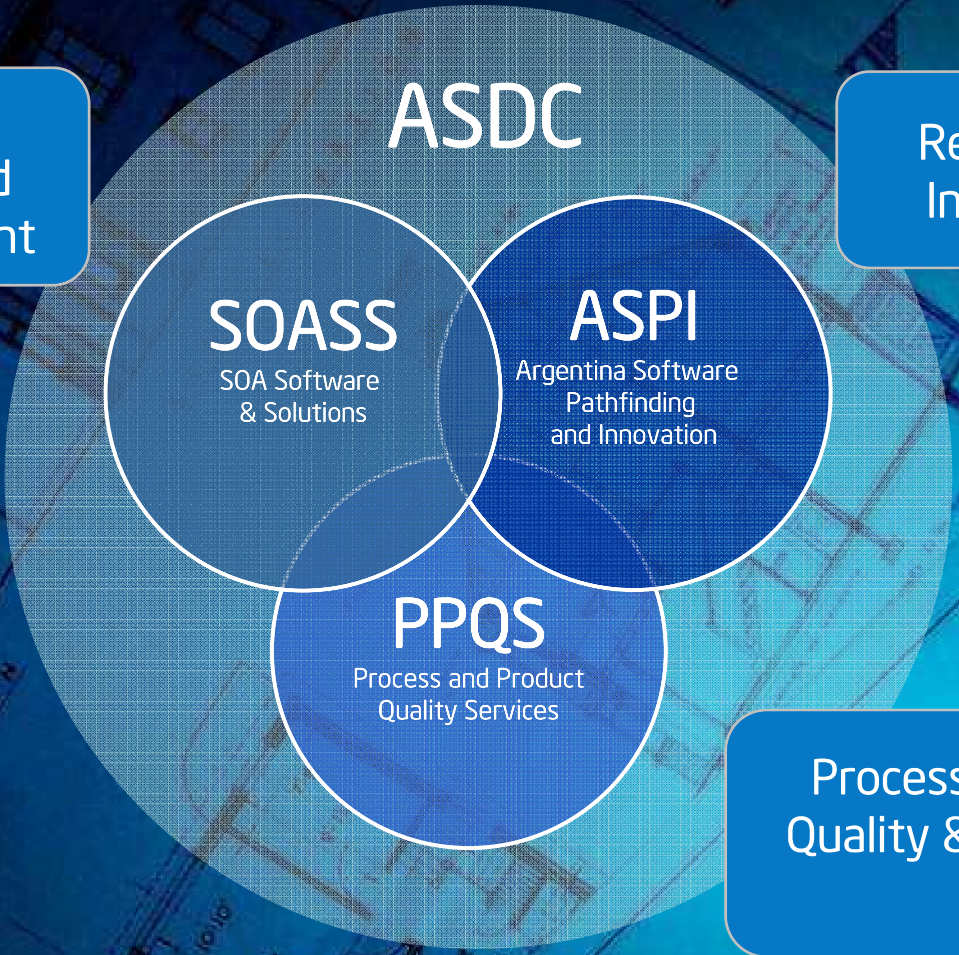
Sydney, Australia
Bangalore, India
Mumbai, India
Japan

Argentina Software Development Center (ASDC)

- Mission
  - To be a Software Center of Excellence for Intel in the region
- ASDC is part of Intel's Software and Solutions Group (SSG)
- Initiated activities in May 2006
- We work together with engineering groups all over the world
- Creating new products is part of our job
- We keep growing
  - Currently, 60 people
  - The center anticipates growing to 400+ engineers by 2011
  - Avg seniority is 6 years, 30% of the workforce has over 10 years experience
  - 16% PhDs, 23% MS
  - Publications and university teaching

ASDC Engineering Groups

Software Design and Development



ASDC

SOASS
SOA Software & Solutions

ASPI
Argentina Software Pathfinding and Innovation

PPQS
Process and Product Quality Services

Research & Innovation

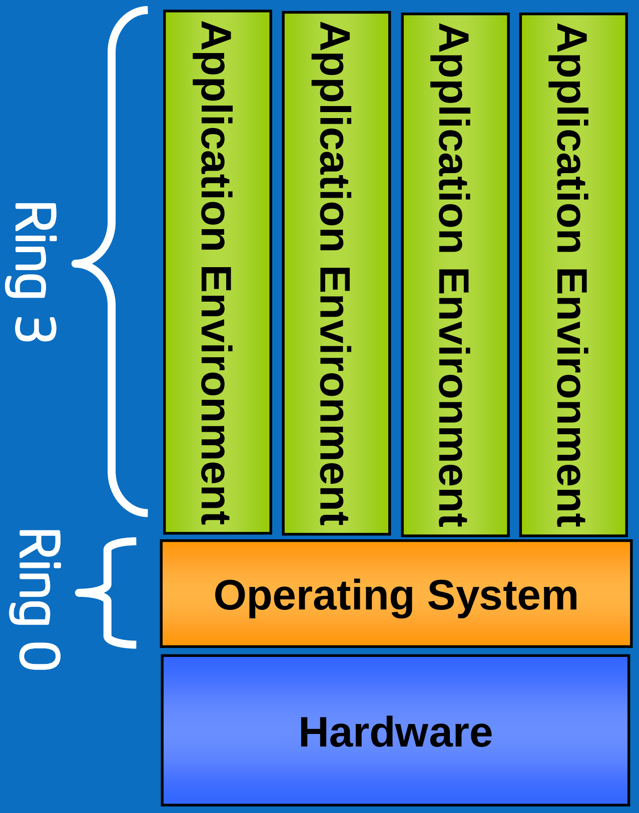
Process Engineering, Quality & Development Tools

Agenda

- Non virtualized environments and Virtualized Environments
- Virtualization Usage Models
- The Virtual Machine Monitor (VMM)
- Challenges of running a VMM
- SW Solution for IA-32 arch without Intel-VT
- Top Ring Deprivileging holes
- Software workarounds to support Ring Deprivileging
- The Intel ® Virtualization Technology (VT-x)
  - New operating modes
  - New transition mechanisms
  - Virtual Machine Control Structure (VMCS)
  - New instructions
- Conclusions

Non-virtualized Environment

- The OS controls access to the hardware resources.
- The instruction set is divided into privileged and non-privileged.
- The machine can be in two modes of operation: user and supervisor
- Only non-privileged instructions can be executed in user mode.
- Any instruction can be executed in supervisor mode.

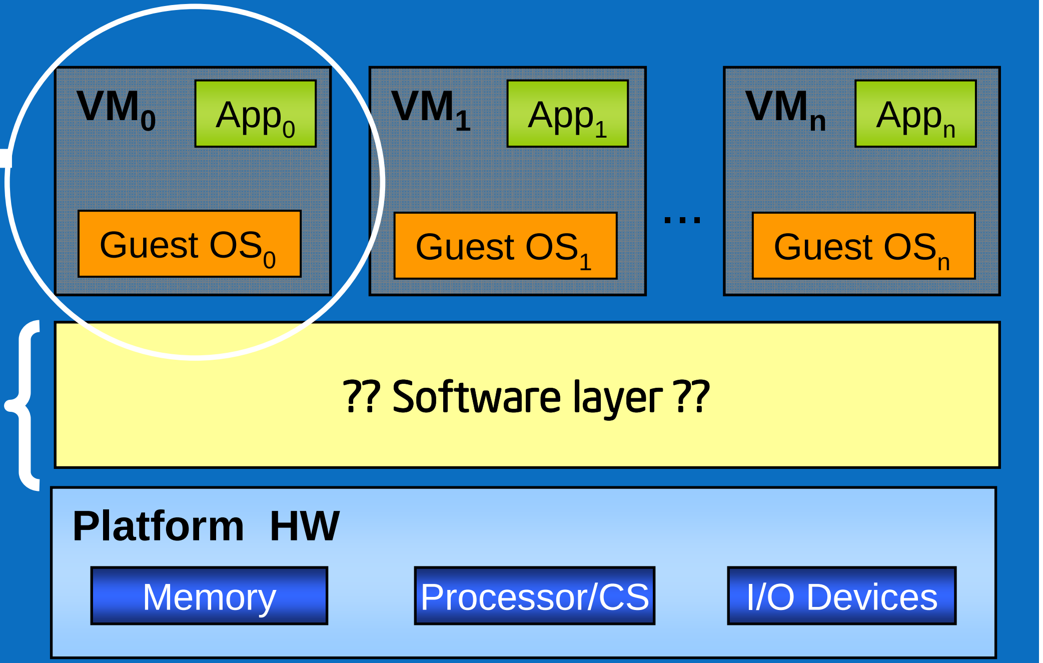


Virtualized Environment

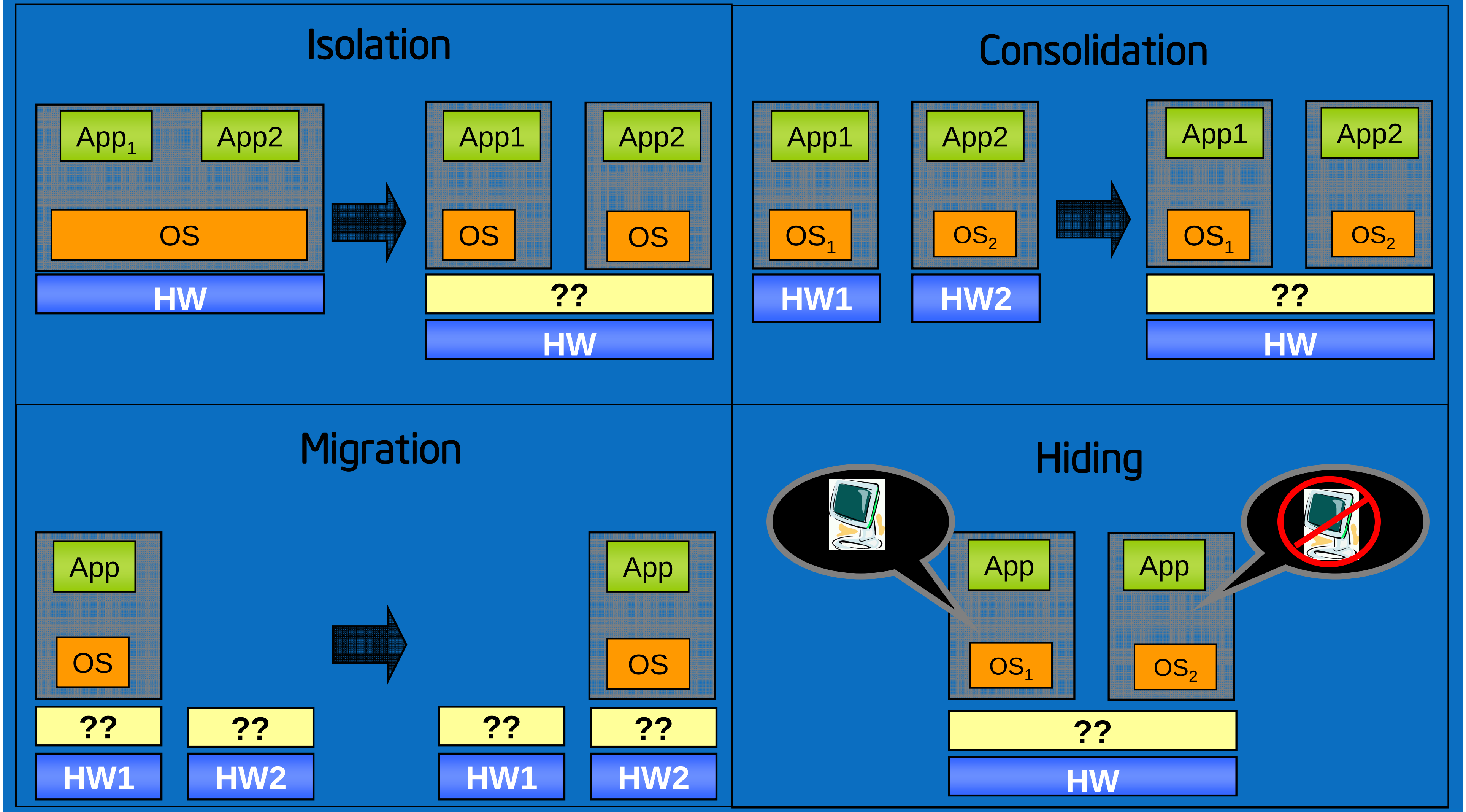
- Run multiple operating systems on a single physical platform

“Virtual Machines” running “Guest Operating Systems”

A software layer that manages the underlying physical resources

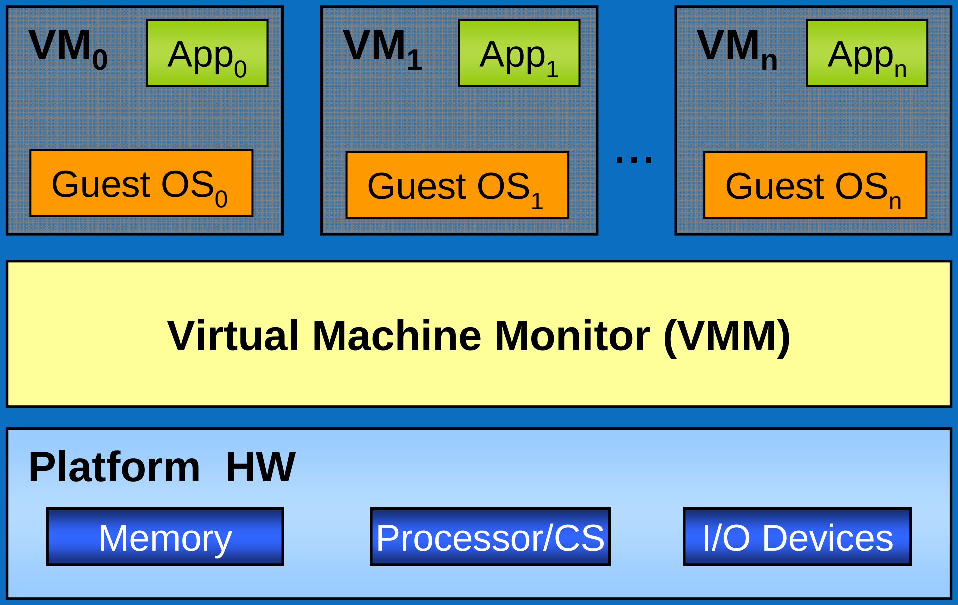


Virtualization Usage Models



The virtual machine monitor (VMM)

- The Virtual Machine Monitor is the software layer that controls the access to all hardware resources.



Challenges of running a VMM

- OS and Apps in a VM don't know that the VMM exists or that they share CPU resources with other VMs.

- VMM should isolate Guest SW stacks from one another.

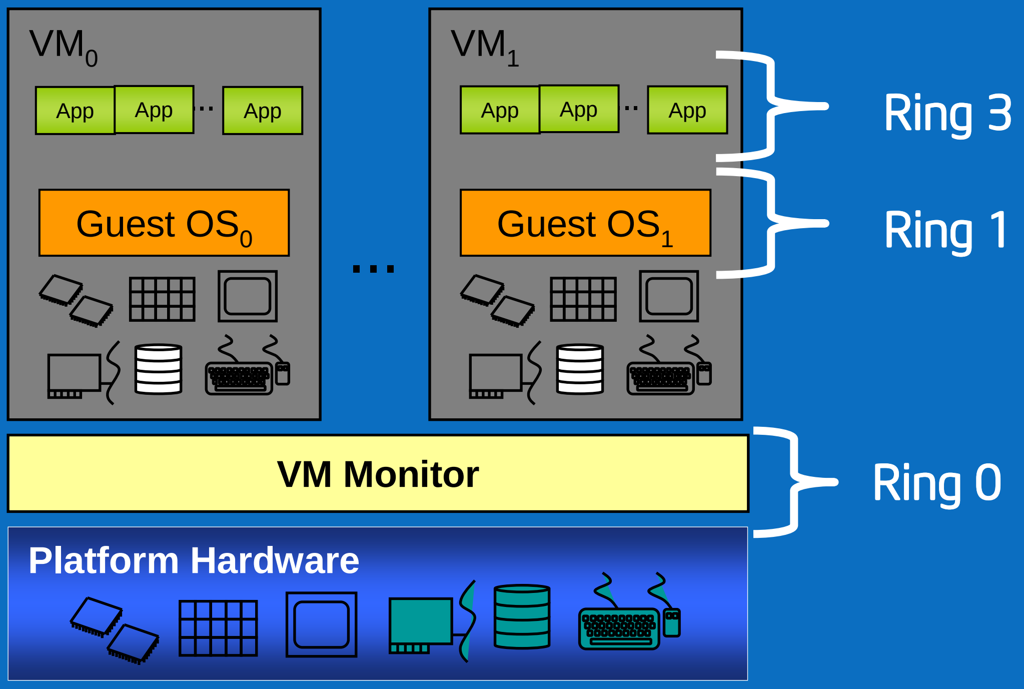
- VMM should run protected from all Guest software

.

- VMM should present a virtual platform interface to Guest SW.

SW Solution for IA-32 arch without Intel-VT

- Ring Deprivileging
  - Run Guest OS above Ring-0 and have privileged instructions generate faults
  - Run VMM in Ring-0 as a collection of fault handlers



- The VMM interprets in software privileged instructions that would be executed by an OS.
- Any non privileged instruction issued by an OS or Application Environment is executed directly by the machine.

Top Ring Deprivileging holes

- Ring Aliasing
  - Problems that arise when software is run at a privilege level other than the privilege level for which it was written.
    - Example: the SC segment register which points to the code segment. If the *PUSH* instruction is executed with the CS segment register, the contents of that register (which include the current privilege level) is pushed on the stack. A guest OS could easily determine that it is not running at privilege level 0.

- Non-trapping instructions
  - There are IA-32 instructions that access privileged state and do not fault when executed with insufficient privilege.
    - Example, the IA-32 registers GDTR, IDTR, LDTR, and TR contain pointers to data structures that control CPU operation. Software can execute the instructions that read, or store, from these registers at any privilege level.

Top Ring Deprivileging holes (cont'd)

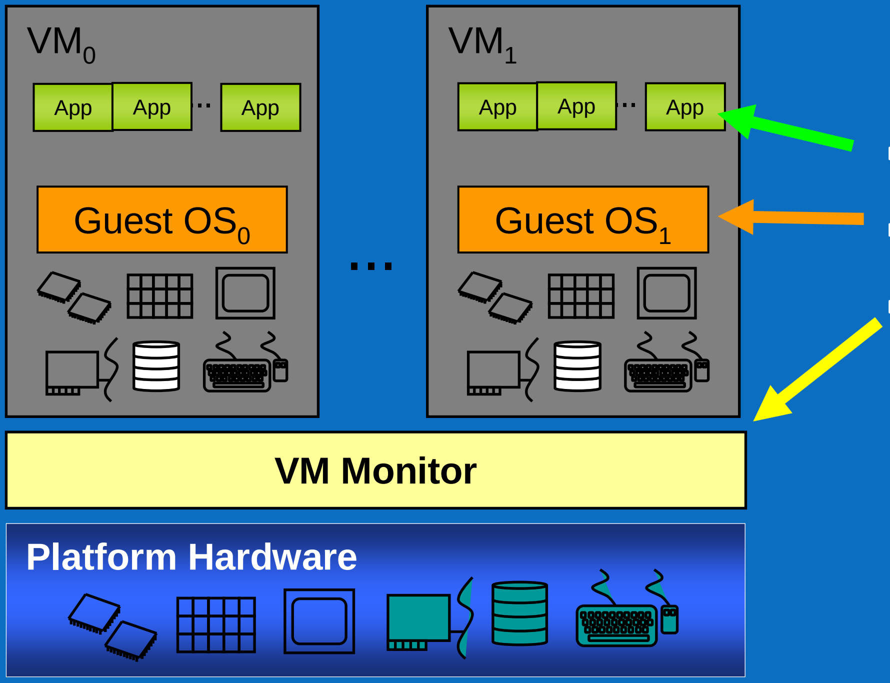
- Excessive Faulting
  - Ring deprivileging can interfere with the effectiveness of facilities in the IA-32 architecture that accelerate the delivery and handling of transitions to OS software. The IA-32 *SYSENTER* and *SYSEXIT* instructions support low-latency system calls. *SYSENTER* always effects a transition to privilege level 0, and *SYSEXIT* faults if executed outside that ring. Ring deprivileging thus has the following implications:
    - Executions of *SYSENTER* by a guest application cause transitions to the VMM and not to the guest OS. The VMM must emulate every guest execution of *SYSENTER*.
    - Executions of *SYSEXIT* by a guest OS cause faults to the VMM. The VMM must emulate every guest execution of *SYSEXIT*.
- CPU State Context Switching
  - The VMM must save the current CPU state in each context switch to be reloaded in the next VM execution. To do that, the VMM uses part of the memory assigned to the VM.
- Address Space Compression

Software workarounds to support Ring Deprivileging

- Source Guest OS Modifications: *Paravirtualization*
  - Example: Xen
    - Developers of these VMMs modify the source code of a guest OS to create an interface that is easier to virtualize.
    - Paravirtualization offers high performance and does not require changes to guest applications.
    - A disadvantage of paravirtualization is that only support modified OS.
- Binary Guest OS Modifications
  - Examples: Vmware, Virtual PC
    - A VMM can support unmodified OSs by transforming guest-OS binaries on-the-fly to handle virtualization-sensitive operations.

Intel® Virtualization Technology (VT-x)

- Guest SW runs *deprivileged* in a *new operating mode*



- Apps run deprivileged in ring 3
- OSs run deprivileged in ring 0
- VMM runs in a new mode with full privilege.

- VMM preempts execution of Guest SW via new HW-based transition mechanism

New Operating Modes

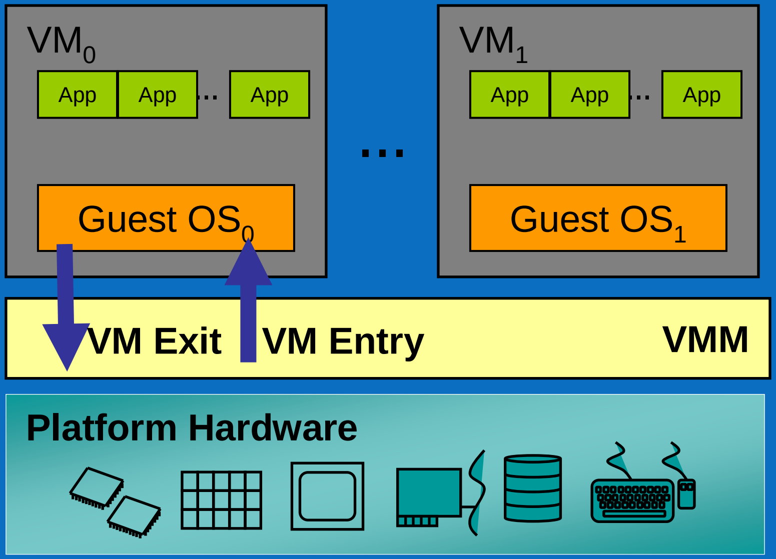
- VMX root operation:
  - Full privileged, intended for Virtual Machine Monitor

- VMX non-root operation:
  - Not fully privileged, intended for guest software

➢ Both forms of operation support all four privilege levels from 0 to 3

New transition mechanisms

- VM entry
  - from VMX root operation mode to VMX non-root operation mode
- VM exit
  - From VMX non-root operation mode to VMX root operation mode



Virtual Machine Control Structure (VMCS)

- Data structure that manages VM entries and VM exits.

- VMCS contains fields corresponding to:
  - Processor state of the Guest area (VM state)
  - Processor state of the Host area (VMM state)
- VM entries load processor state from the guest-state area.
- VM exits save processor state to the guest-state and then load processor state from the host-state area

➢ Only one VMCS active per virtual processor at any given time

VT-x New instructions

- VMLAUNCH: Used on initial transition from VMM to Guest
  - Enters VMX non-root operation mode
- VMRESUME: Used on subsequent entries
  - Enters VMX non-root operation mode
  - Loads Guest state and Exit criteria from VMCS
- VMEXIT
  - Used on transition from Guest to VMM
  - Enters VMX root operation mode
  - Saves Guest state in VMCS
  - Loads VMM state from VMCS
- VMPTRLD
  - Establishes a pointer to a desired VMCS
- VMREAD
  - Read from a VMCS
- VMWITE
  - Write to a VMCS

Conclusions

- VT Reduces guest OS dependency
  - Eliminates need for binary patching / translation
  - Facilitates support for Legacy OS
- VT improves robustness
  - Eliminates need for complex SW techniques
  - Simpler and smaller VMMs
  - Smaller trusted-computing base
- VT improves performance
  - Fewer unwanted Guest ⇔ VMM transitions

Thank you!

Contacts to: gisela.giusti@intel.com

Reference to: duilio.j.protti@intel.com

intel®

Leap ahead™

Backup

Leap ahead™

- World leader in silicon based advanced technology innovation with more than 38 years of leadership in computer science and communications
- Foundation: 1968
- Employees: 90,000+
- Products and services: 450+
- Offices and Installations worldwide: 294

VT Client Roadmap

2005 Lyndon*

Intel® Pentium® 4 Processor
945G Chipset
HT, XD, EM64T, EIST, Intel AMT, VT

2006 Averill*

Intel Pentium 4 Processor & DC
Broadwater Chipset
2005 features plus Intel AMT2, LT

2005 Intel Centrino™ Mobile Technology

Intel Pentium M Processor
Intel 915 Chipset Family
Intel PRO Wireless Network Connection 2915ABG & 2200BG,
XD, EIST

2006 Napa*

Mobile Dual Core Processor code-named “Yonah”
Chipset code-named “Calistoga”
Wireless LAN solution code-named “Golan”
2005 features plus VT, Intel AM

VT Server Roadmap

2005 - 2006

Millington / DP Montvale

Intel® 8870, Enabled

Dual Core, MT, Foxton, Pellston, VT

2 Socket

intel inside ITANIUM 2

2005 - 2006

Montecito / Montvale

Intel® 8870 / Enabled

MT, Foxton, Pellston, VT

≥ 4 Socket

intel inside ITANIUM 2

2006 Bensley*, Glidewell*

Dempsey

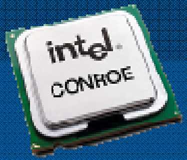
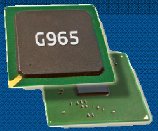
Blackford & Greencreek

2005 features plus VT, IAMT, I/OAT

2 Socket

intel inside XEON

vPro Virtual Appliance Roadmap

	Averill *Shipping*	*Weybridge* *Q3'07*	*Montevina* *Q2'08*	*McCreary* *Q3'08*
 Processor	Intel® Core™ 2 Duo processor family (Conroe)	Intel® Core™ 2 Duo (Conroe and Wolfdale) Intel® Core™ 2 Quad (Yorkfield)	Intel® Core™ 2 Duo (Penryn)	Intel® Core™ 2 Duo (Wolfdale) Intel® Core™ 2 Quad (Yorkfield)
 Chipset	Intel® Q965 Express Chipset w/ ICH8-DO	Intel® Q35 Express Chipset w/ ICH9-DO	Intel® Cantiga Chipset w/ ICH9-M	Intel® Eaglelake Express Chipset w/ ICH10-DO
 Networking	Intel® 82566DM (Nineveh)	Intel® 82566DM (Nineveh)	Intel® 82567LM (Boazman) Intel® WiFi (Shiloh)	Intel® 82567LM (Boazman)
Security		TPM 1.2 (Discrete)	TPM 1.2 (in Cantiga)	TPM 1.2 (in Eaglelake)
T *T's	Intel® Virtualization Technology (VTx) Intel® Active Mgmt Technology 2.0* EM64T, EIST *Intel® AMT2.1 FW is needed for EIT VA 2.0	Intel® Virtualization Technology (VTx+VTd) Intel® Active Mgmt Technology ver 3.0 Intel® Trusted Execution Technology EM64T, EIST	Intel® Virtualization Technology (VTx+VTd) Intel® Active Mgmt Technology ver 4.0 Intel® Trusted Execution Technology EM64T, EIST	Intel® Virtualization Technology (VTx+VTd) Intel® Active Mgmt Technology ver 5.0 Intel® Trusted Execution Technology EM64T, EIST