Virtualization:
Implications and Opportunities for Performance Analysis

Keynote for ISPASS 2008
Austin, Texas
April 21st, 2008

Rich Uhlig
Chief Virtualization Architect
Intel Corporation

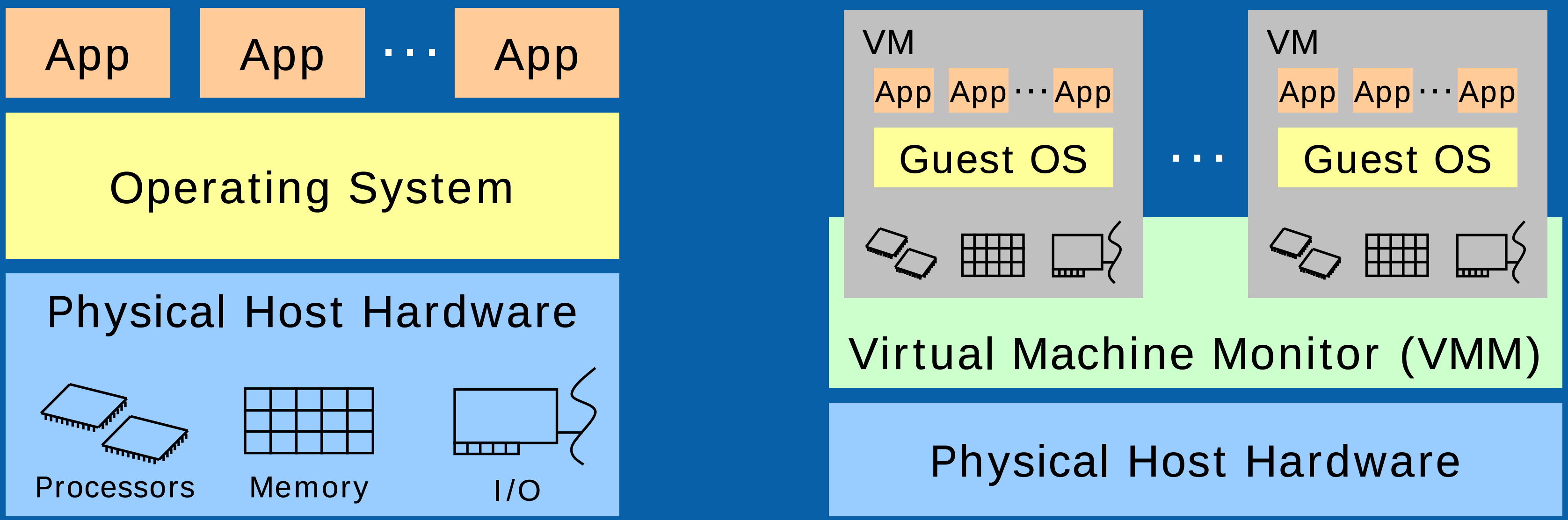
Outline

- Virtualization Overview
- Performance Implications
- New Opportunities for Performance Tools and Methods

Outline

- Virtualization Overview
- Performance Implications
- New Opportunities for Performance Tools and Methods

Virtualization Defined



Without VMs: Single OS owns all hardware resources

With VMs: Multiple OSes share hardware resources

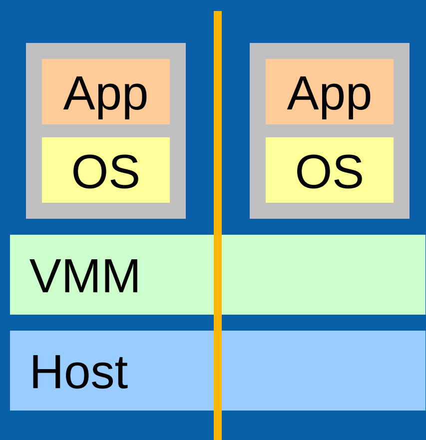
Virtualization enables multiple operating systems to run on the same physical platform

Virtualization Usage Models

- Server Consolidation
- Dynamic Datacenter
- Client Centralization
- Security
- Real-time QoS

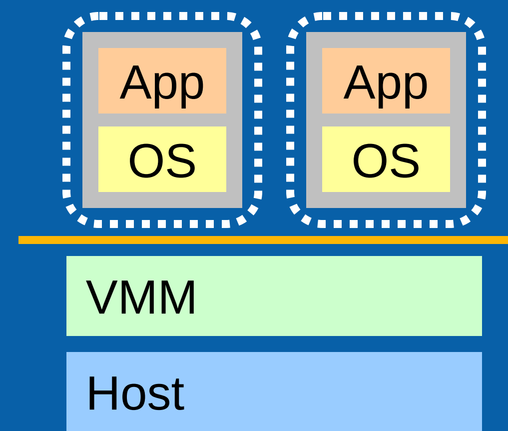
Key Properties of Virtualization

Partitioning



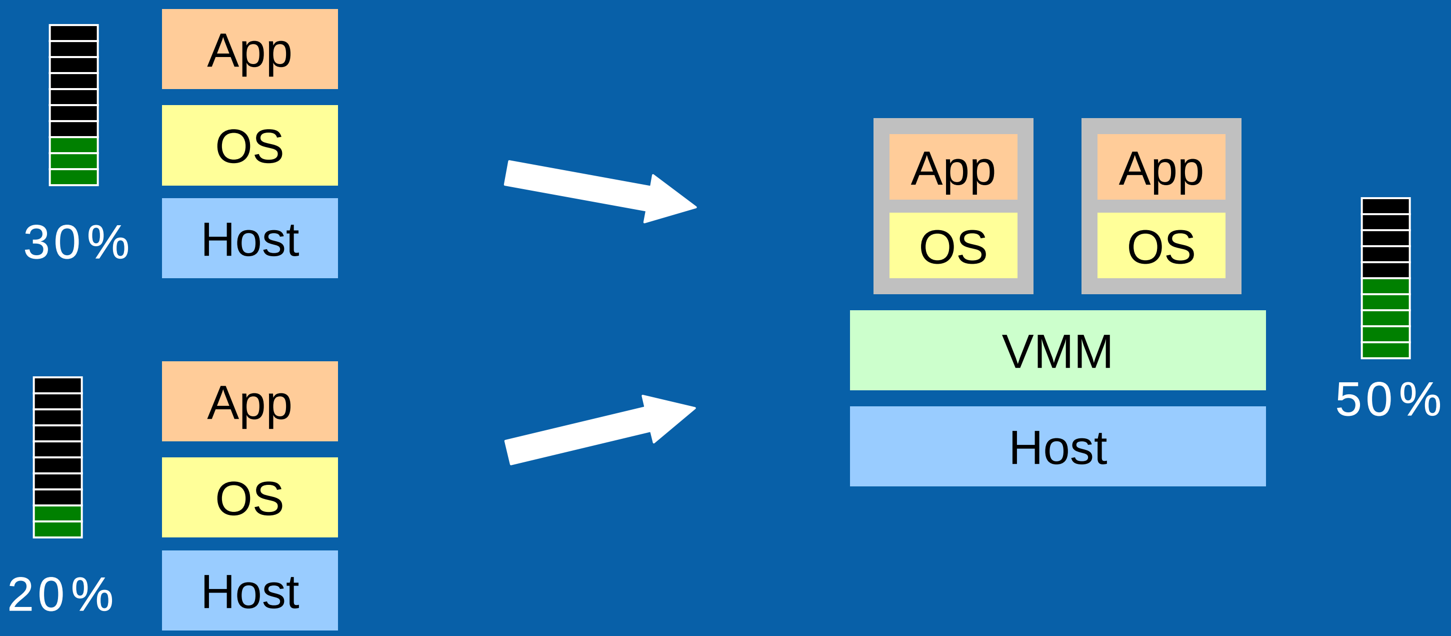
- Resource Sharing
- Isolation

Encapsulation



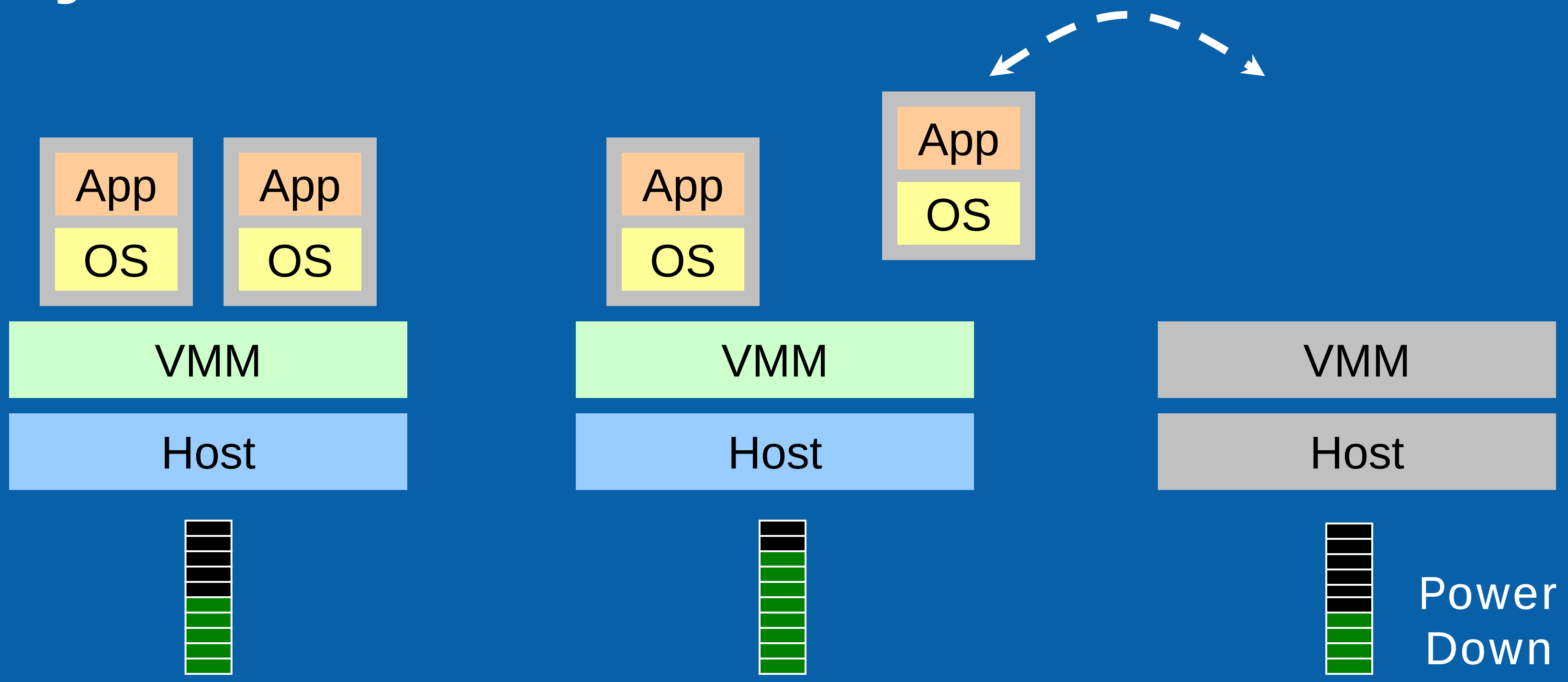
- Checkpoint / Restore
- Migration
- Execution Replay

Server Consolidation



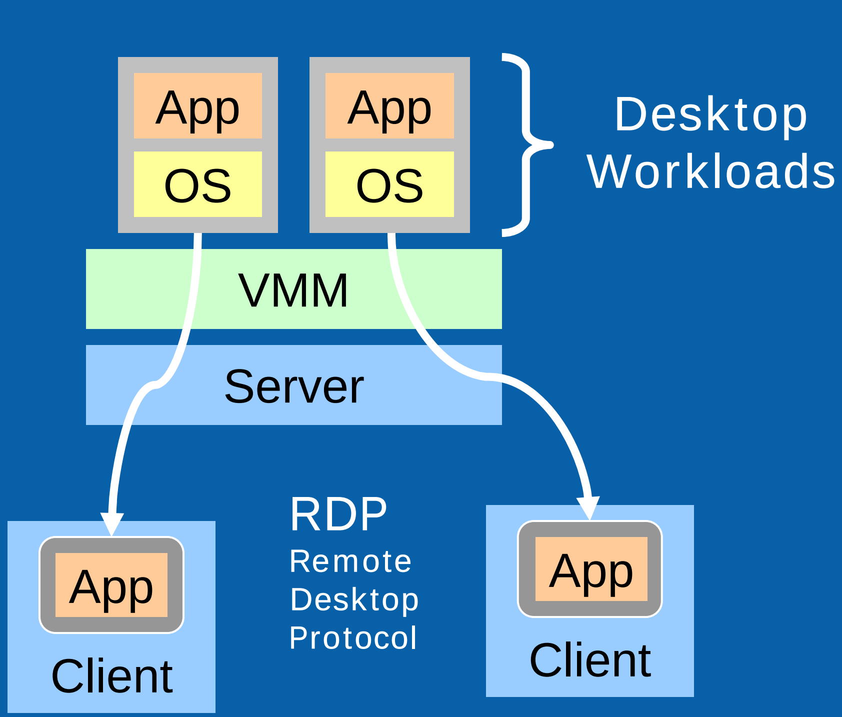
- Underutilized physical servers
- Consolidate to improve utilization / lower cost

Dynamic Datacenter

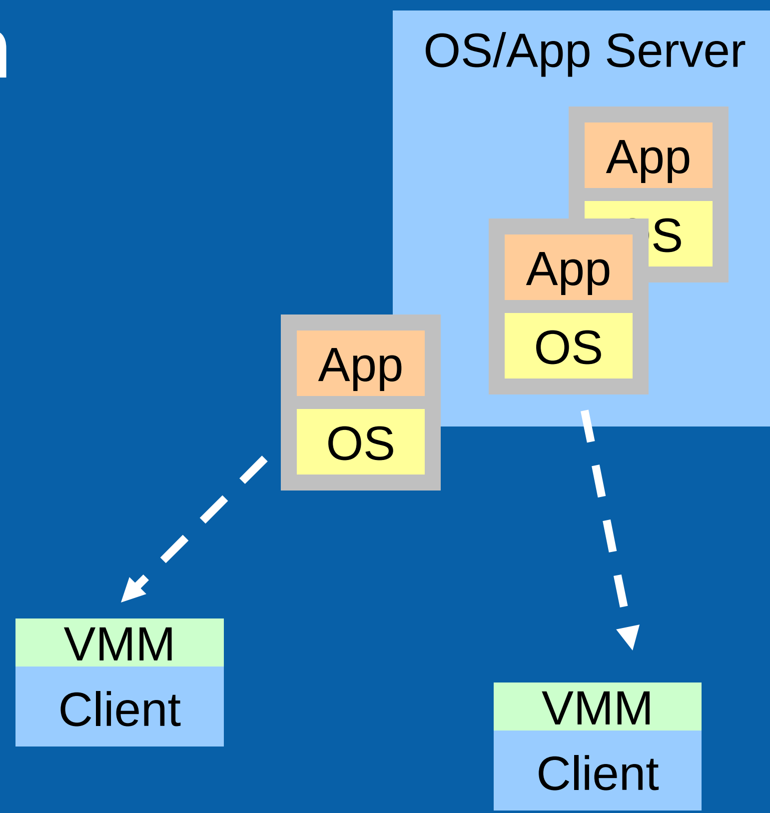


- Balance load for performance and response time
- Or, consolidate load for overall power optimization

Client Centralization



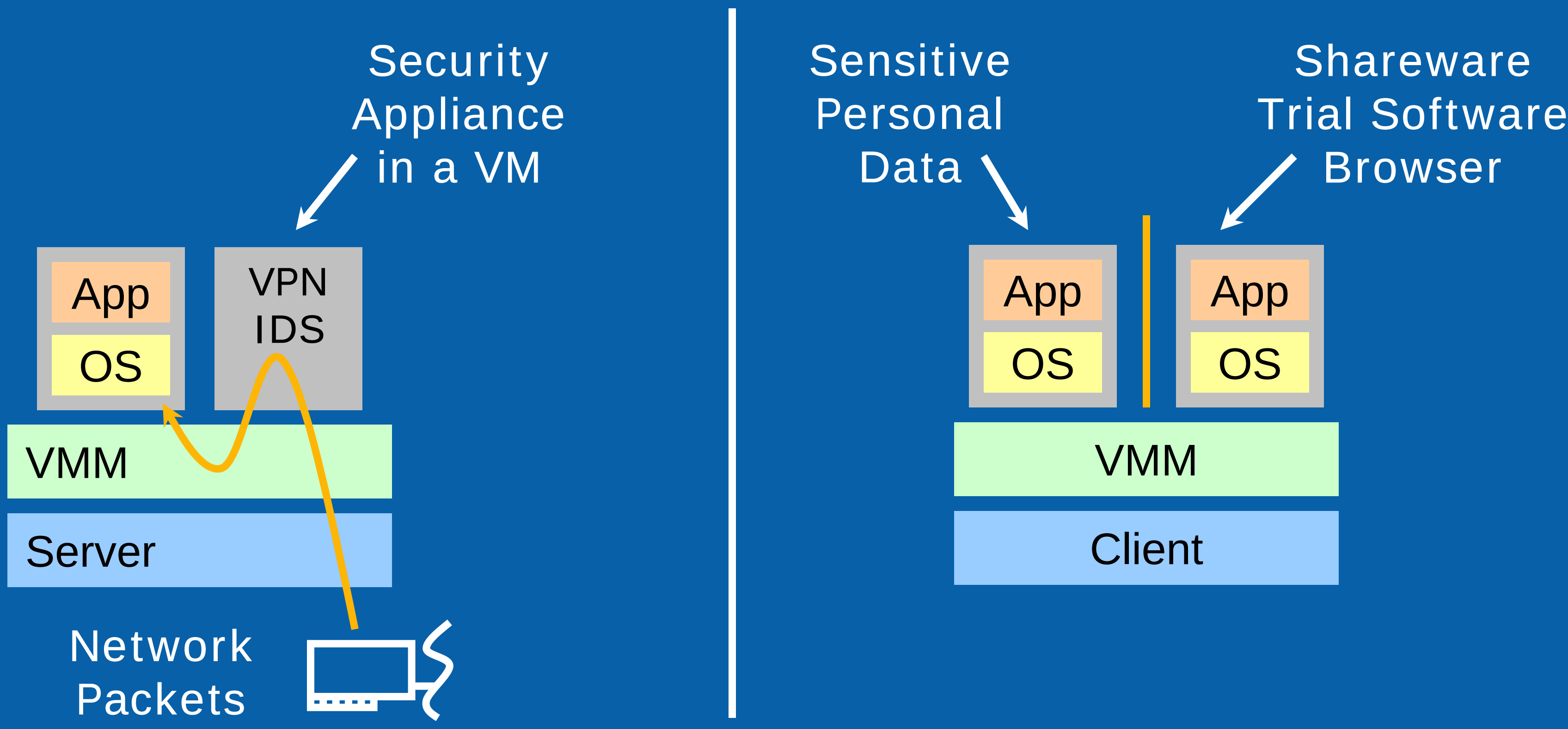
"Virtual Desktop" Consolidation



"VM Streaming" Model

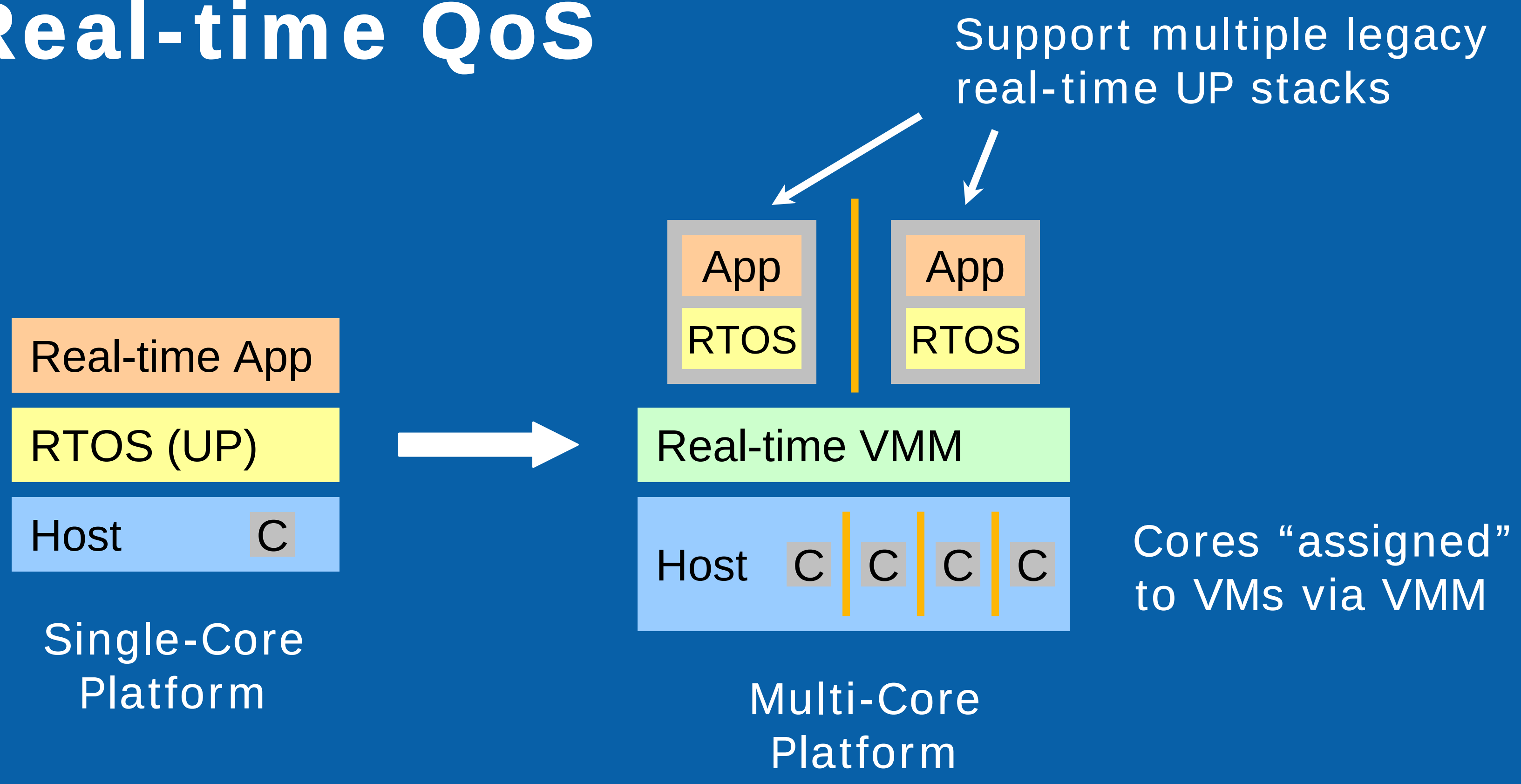
- Centralized management and backup
- Rapid provisioning of computing environments

Security



- Packet inspection through a VM security appliance
- Environment Isolation & “Disposable VMs”

Real-time QoS



- Real-time OS/App stacks are often UP-only
- Unable to leverage multi-core systems

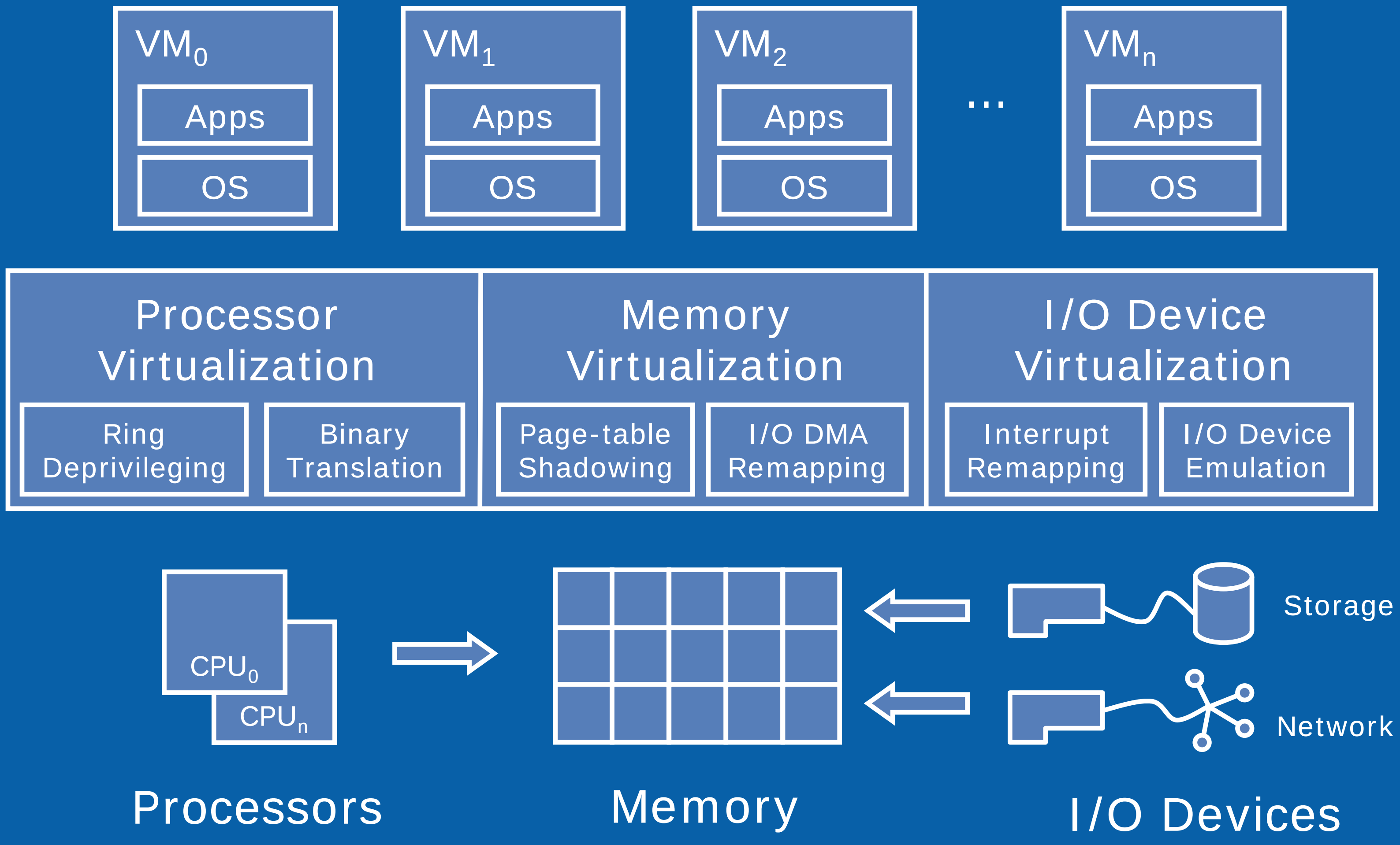
Recap: Virtualization Usage Models

Server Client Embedded

- Server Consolidation
- Dynamic Datacenter
- Client Centralization
- Security
- Real-time QoS

Not just a “mainframe” technology anymore…

Inside the VMM…



Intel® VT Roadmap: Overview

Vector 3:
I/O Focus

VT-c

I/O Endpoint Support
- IOV Standards Definition
- Sharable I/O (Networking)

Vector 2:
Platform Focus

VT-d

Core Platform Infrastructure
- DMA protection and remapping
- Interrupt filtering / remapping

Vector 1:
Processor Focus

VT-x

VT-i

Core Processor Virtualization Support
- VT-x: Intel® 64 ISA extensions for CPU virtualization
- VT-i: Intel® Itanium® ISA extensions for virtualization

VMM
Software
Evolution

SW-only VMMs
- Binary Translation
- Paravirtualization
- Device Emulation

- **Simpler** VMMs through foundation of virtualizable ISAs
- Enhanced **functionality** and legacy software compatibility
- Improved **performance** through hardware assists

Past
No Hardware
Support

Today →

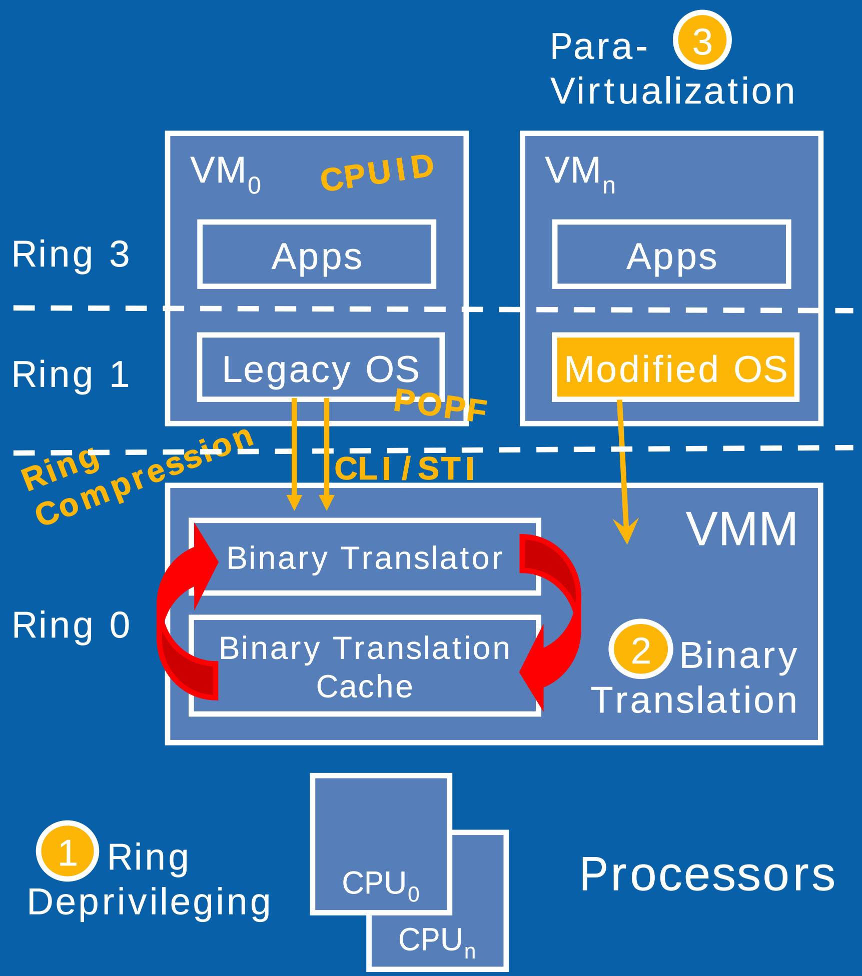
VMM software evolution over
time with hardware support

CPU Virtualization Challenges

Ring Deprivileging
- Run guest OS above ring 0
- Control privileged state access

Virtualization Holes
- Ring Compression
- Non-trapping operations
- Excessive trapping

Software Methods
- Binary Translation
- Paravirtualization



CPU Virtualization with VT

New CPU Operating Mode

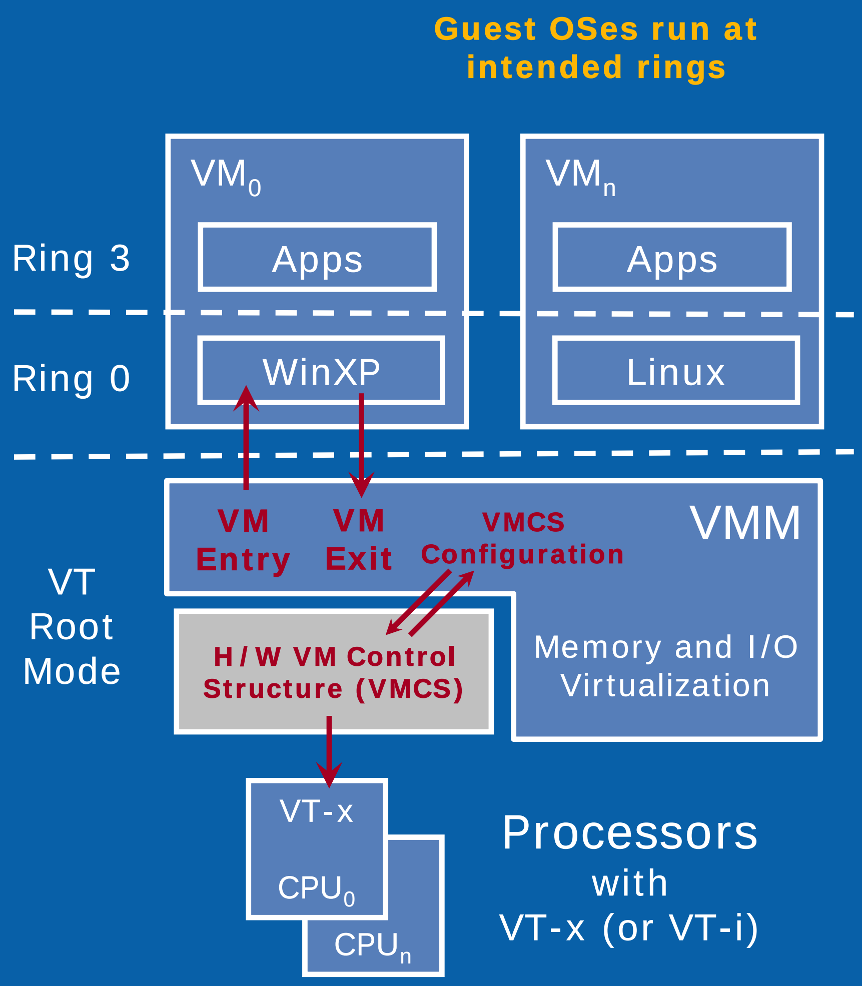
- VT Root Operation (for VMM)
- Non-Root Operation (for Guest)
- Eliminates ring compression

New Transitions

- VM entry and exit
- Swaps registers and address space in one atomic operation

VM Control Structure (VMCS)

- Configured by VMM software
- Specifies guest OS state
- Controls when VM exits occur



Memory Virtualization Challenges

Address Translation

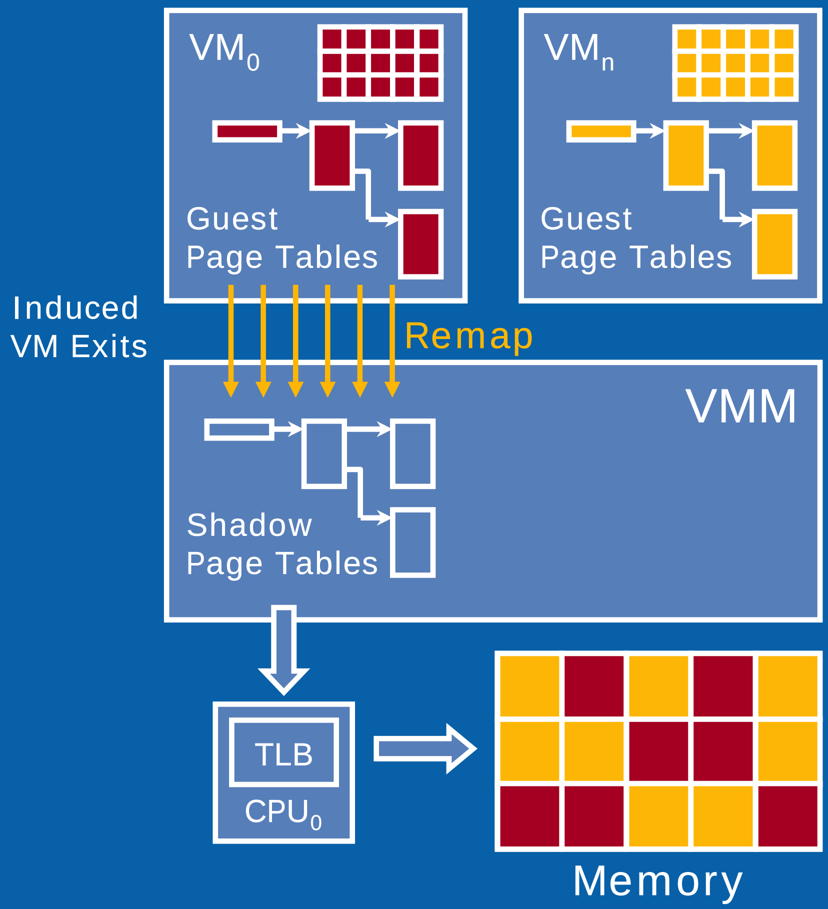
- Guest OS expects contiguous, zero-based physical memory
- VMM must preserve this illusion

Page-table Shadowing

- VMM intercepts paging operations
- Constructs copy of page tables

Overheads

- VM exits add to execution time
- Shadow page tables consume significant host memory



Memory Virtualization with VT

Extended Page Tables (EPT)

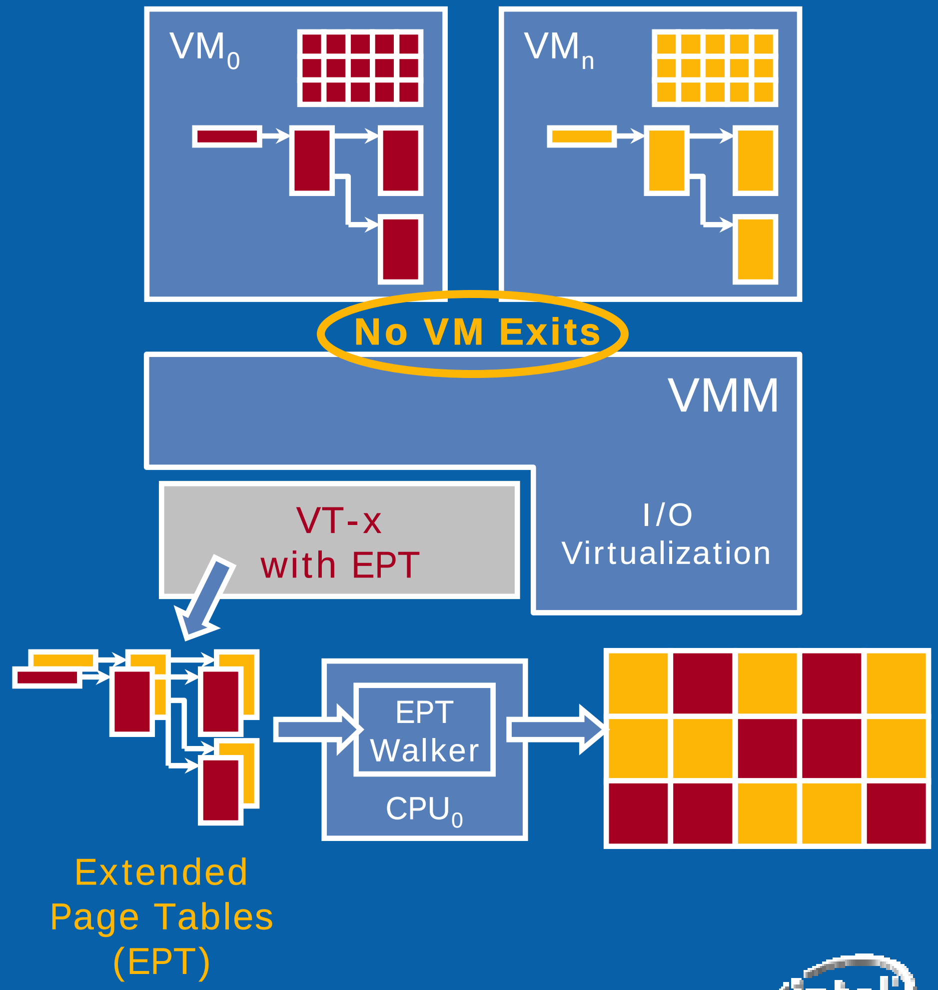
- Map guest physical to host address
- New hardware page-table walker

Performance Benefit

- Guest OS can modify its own page tables freely
- Eliminates VM exits

Memory Savings

- Shadow page tables required for each guest user process (w/o EPT)
- A single EPT supports entire VM

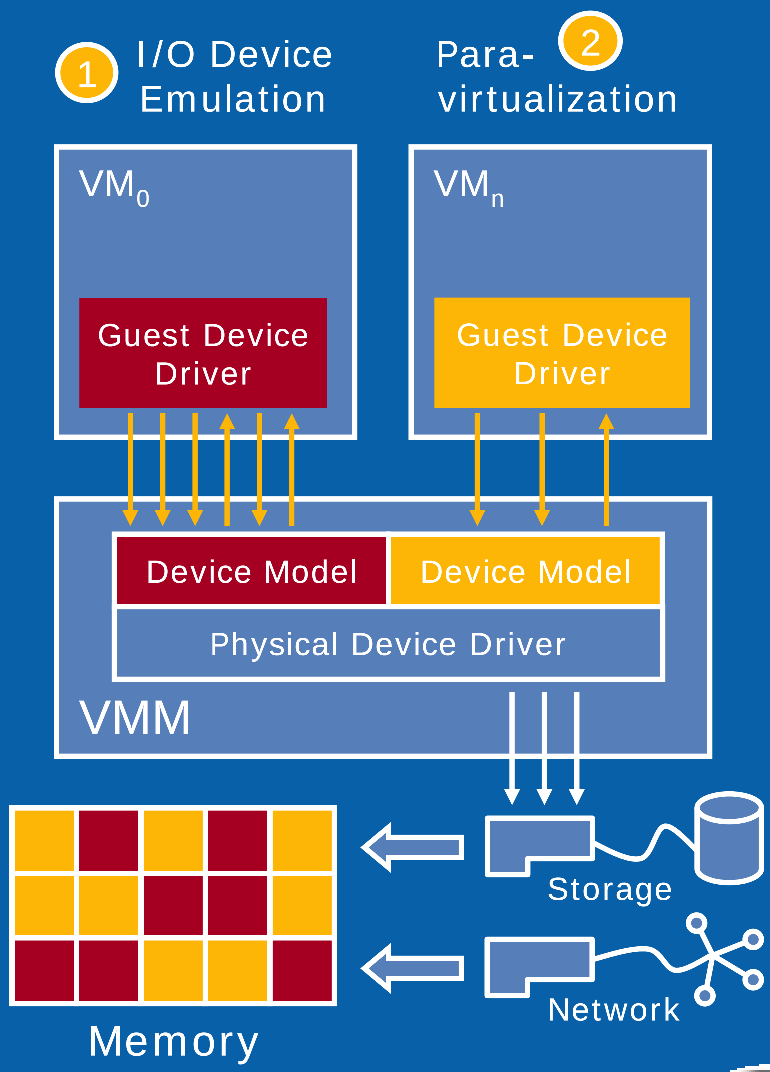


I/O Virtualization Challenges

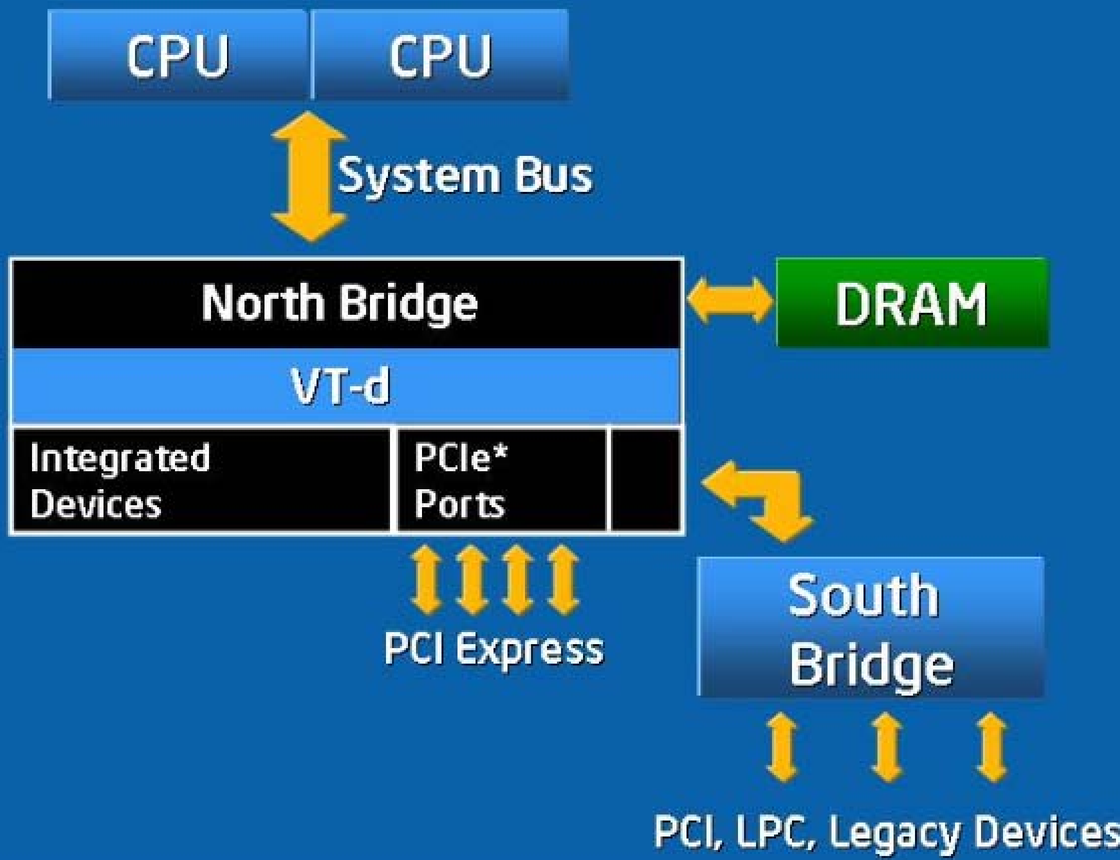
Virtual Device Interface
- Traps device commands
- Translates DMA operations
- Injects virtual interrupts

Software Methods
- I/O Device Emulation
- Paravirtualize Device Interface

Challenges
- Controlling DMA and interrupts
- Overheads of copying I/O buffers

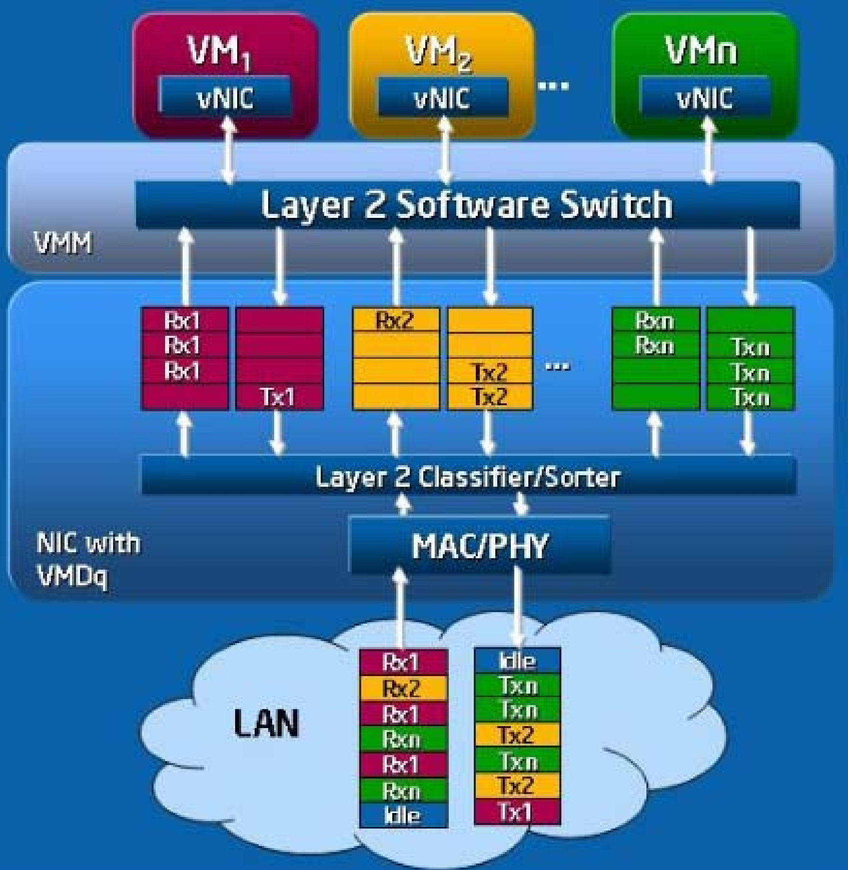


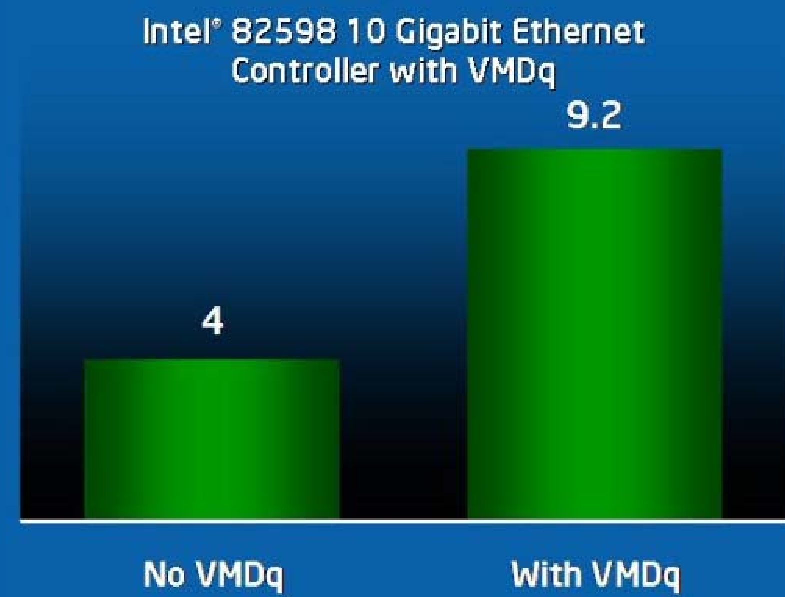
VT Core Platform Support for I/O



- VT-d provides DMA and interrupt-remapping support
- Supports I/O device assignment to VMs, security, etc.

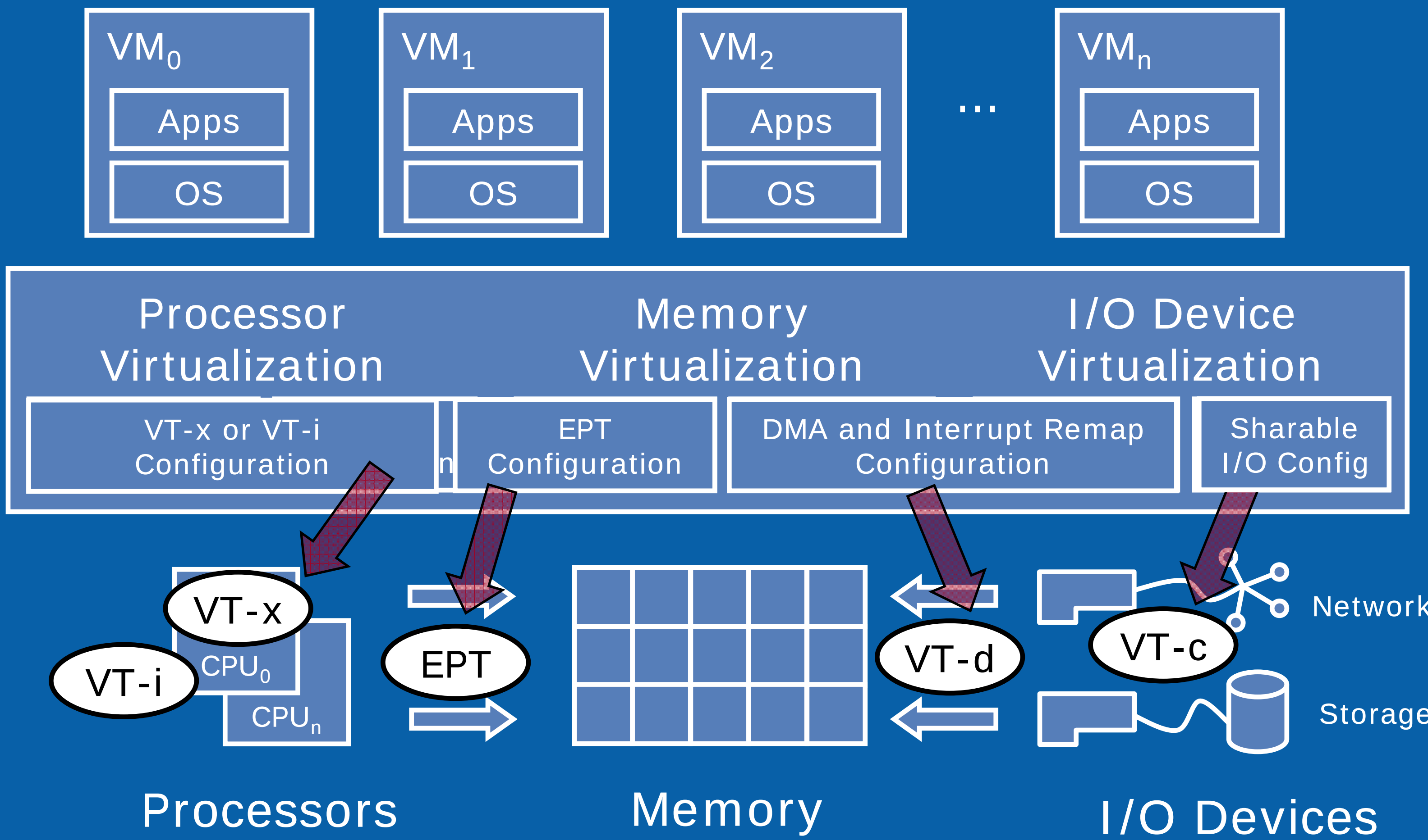
Network Virtualization with VT





- Multiple send/receive queues pre-sort packets for SW
- Reduces CPU utilization, increases throughput

Putting it all together…

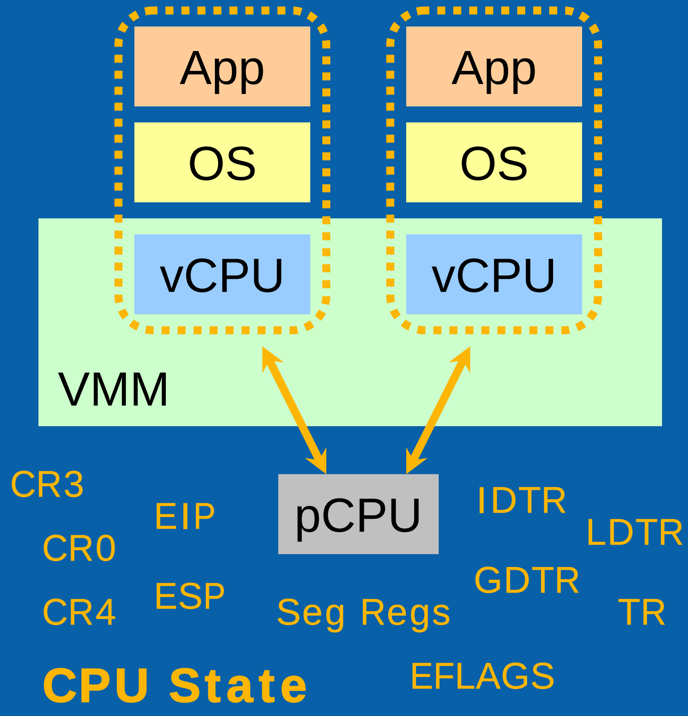


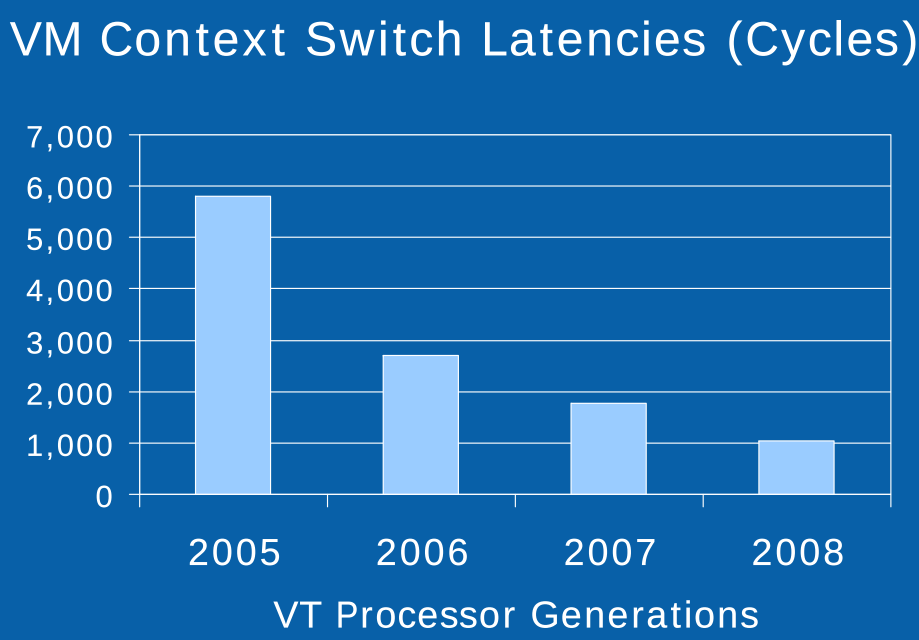
Replumbing the entire platform for virtualization…

Outline

- Virtualization Overview
- Performance Implications
- New Opportunities for Performance Tools and Methods

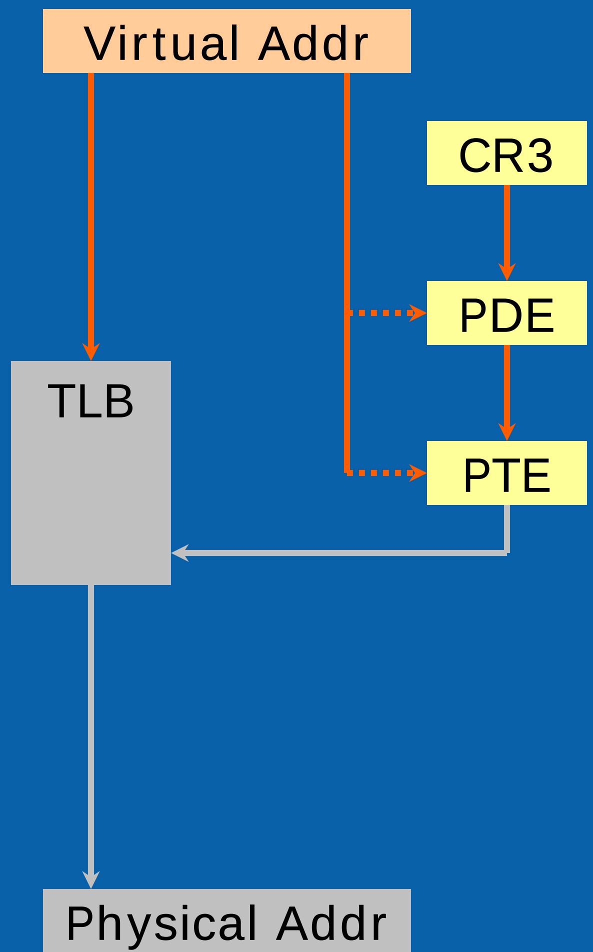
Core CPU: Context Switching





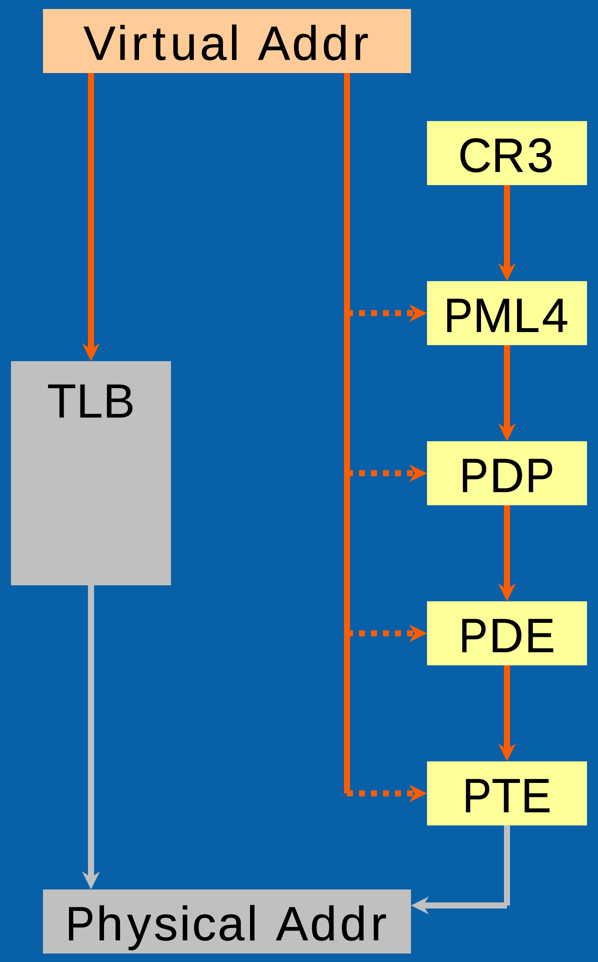
- With Virtualization: Entirely new state to switch...
- Privileged CPU state that normally doesn't change

Address Translation and TLBs



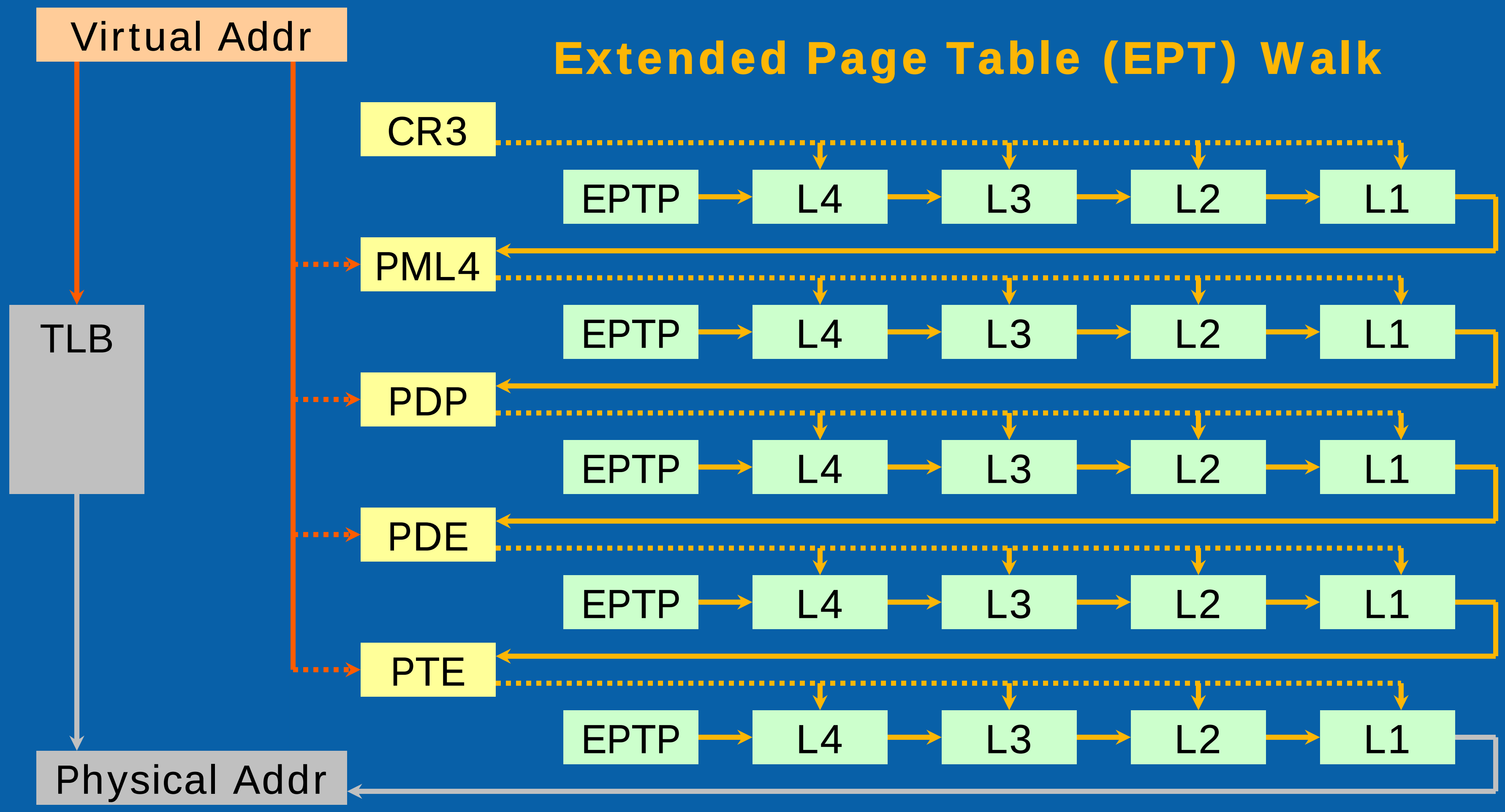
With 32-bit Addressing: 2-level Walk

Address Translation and TLBs (2)



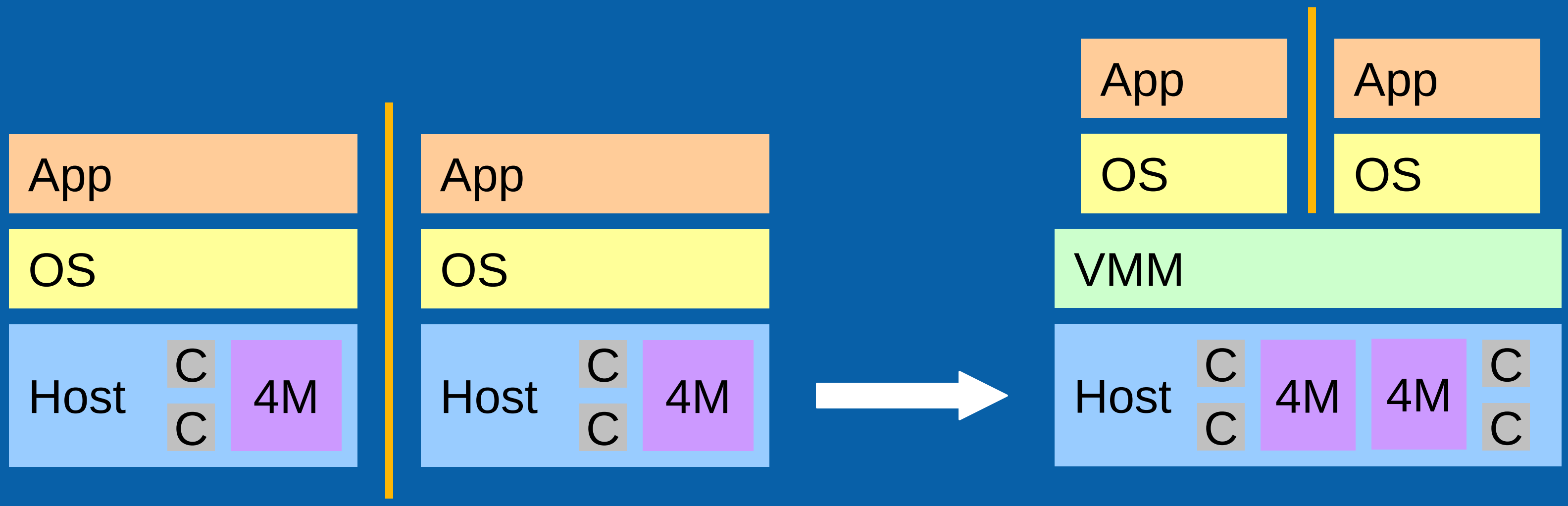
With 64-bit Addressing: 4-level Walk

Address Translation and TLBs (3)



With Virtualization: 24 Steps in Walk!

Cache Interference

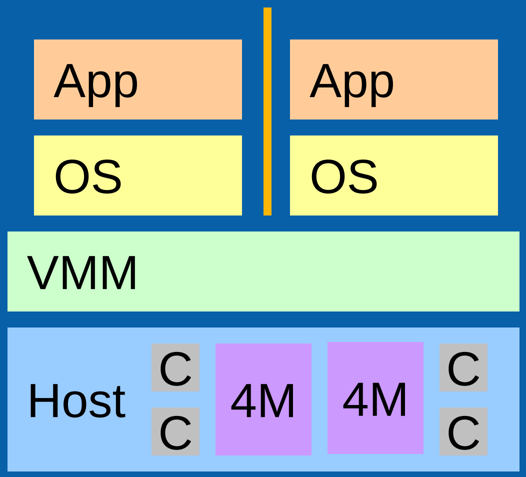


Single Socket
Dual Core
4MB LLC

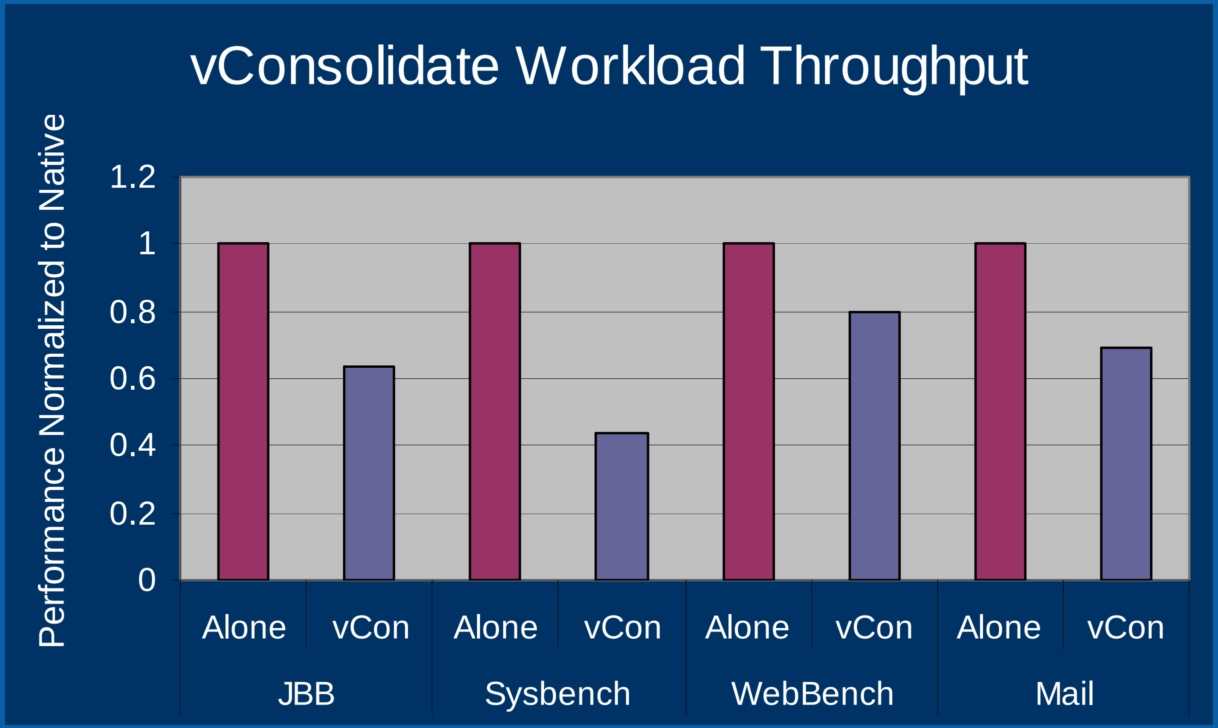
Dual Socket
Dual Core
4MB LLC

Server consolidation mixes working sets in cache...

Cache Interference



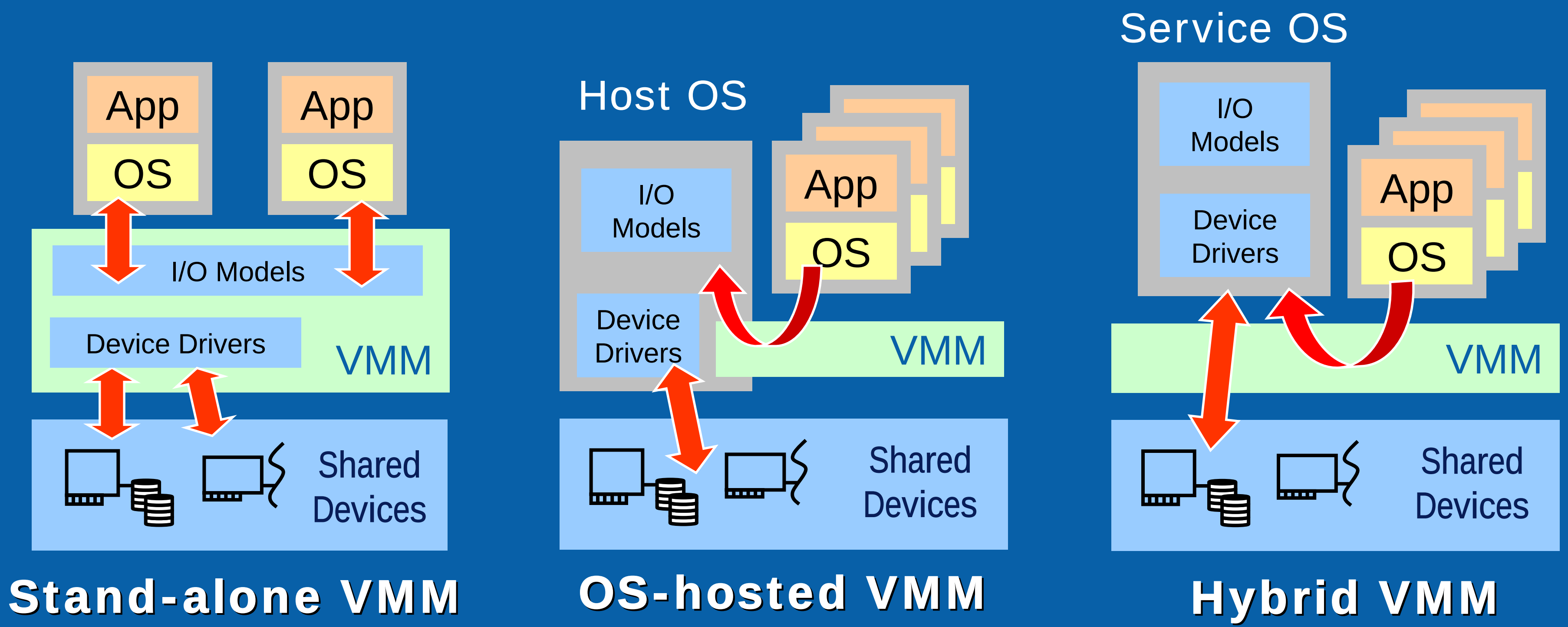
Dual Socket
Dual Core
4MB LLC



Data Courtesy: Ravi Iyer, Don Newell

… leading to new sources of performance variability

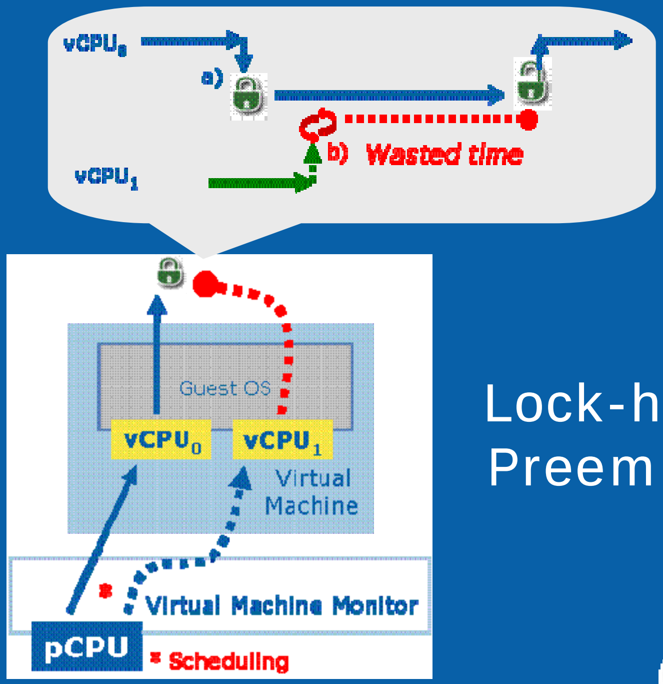
I/O Virtualization Overheads



- Common sources of I/O virtualization overhead
  - Traversal of dual I/O stacks (guest and VMM)
  - Overheads of I/O device models
  - Additional I/O buffer copies
  - Interrupt routing / processing in presence of vCPU migration

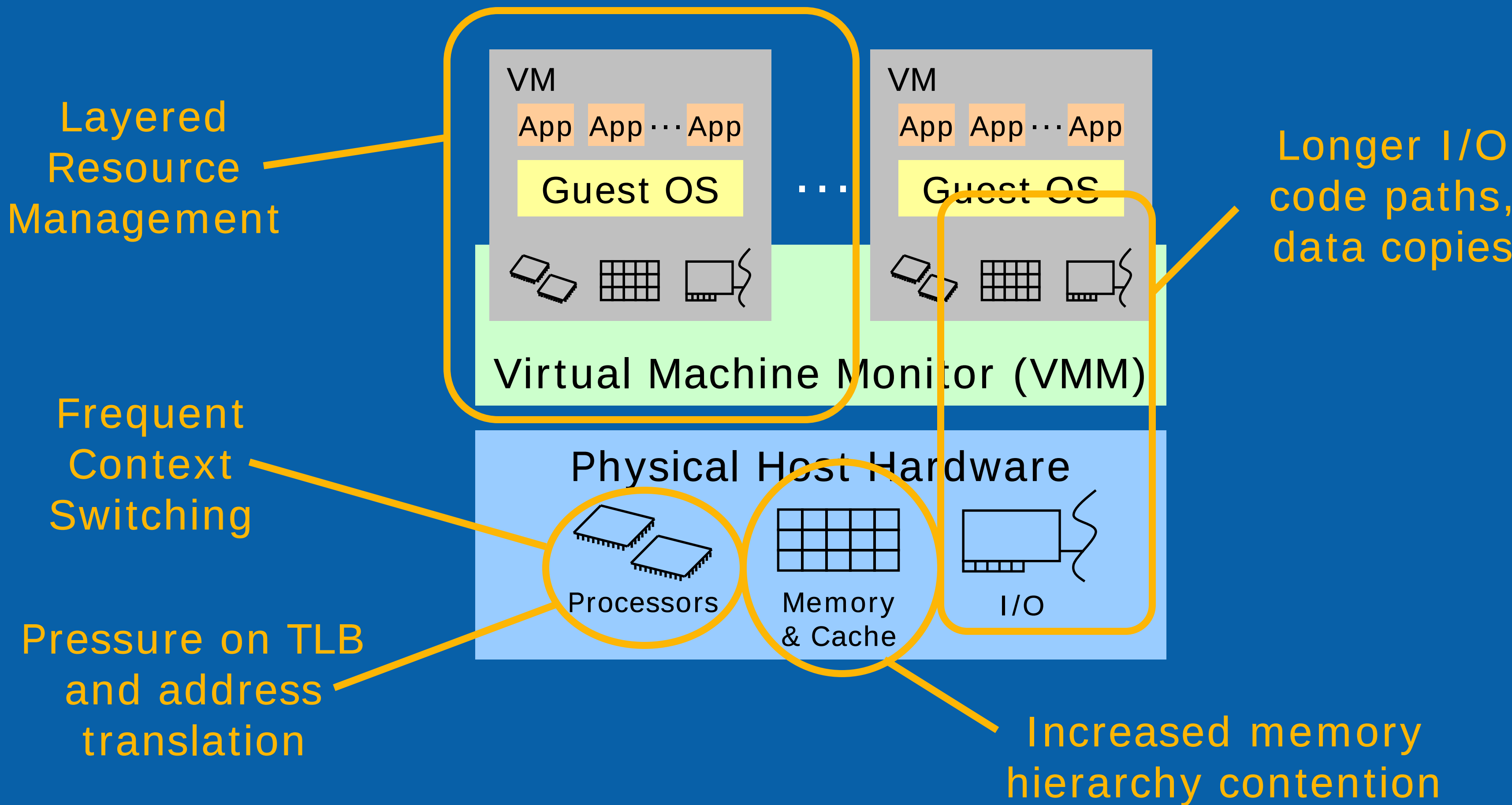
Guest OS – VMM Interactions

- Layered Resource Management
  - OS no longer manages physical resources directly
  - Can give rise to various adverse performance effects
- CPU Overcommit
  - Lock-holder preemption
- Memory Overcommit
  - Paging policy by guest?
  - Or VMM?



Lock-holder Preemption

Performance Implications



Every part of the system is affected...

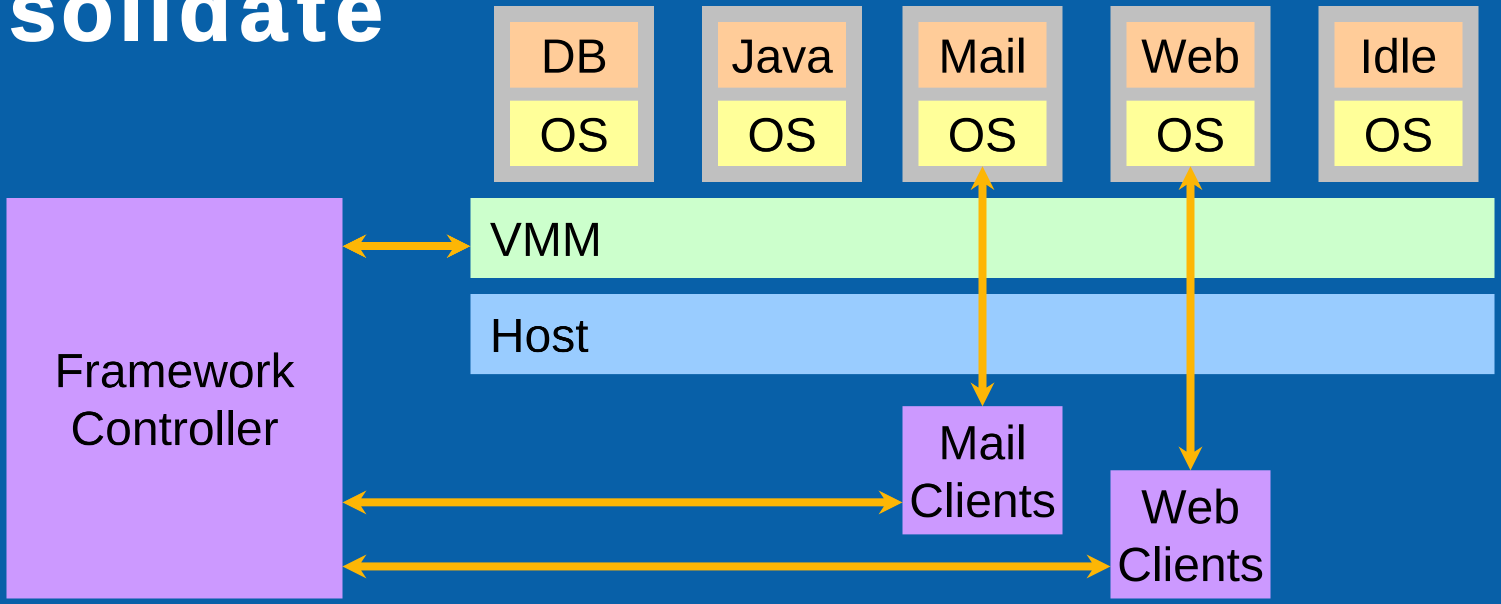
Outline

- Virtualization Overview
- Performance Implications
- New Opportunities for Performance Tools and Methods

New Tools & Workloads Needed…

- Many existing tools & methods break
  - Performance counters not visible to guest OS
  - Profiling tools can’t easily span VM address spaces
  - New tracing tools and trace content needed
- Workloads
  - Single-VM workloads not sufficient
  - Need multi-VM workloads to understand real scenarios
  - Requires new workload definition effort, run rules, etc.

vConsolidate

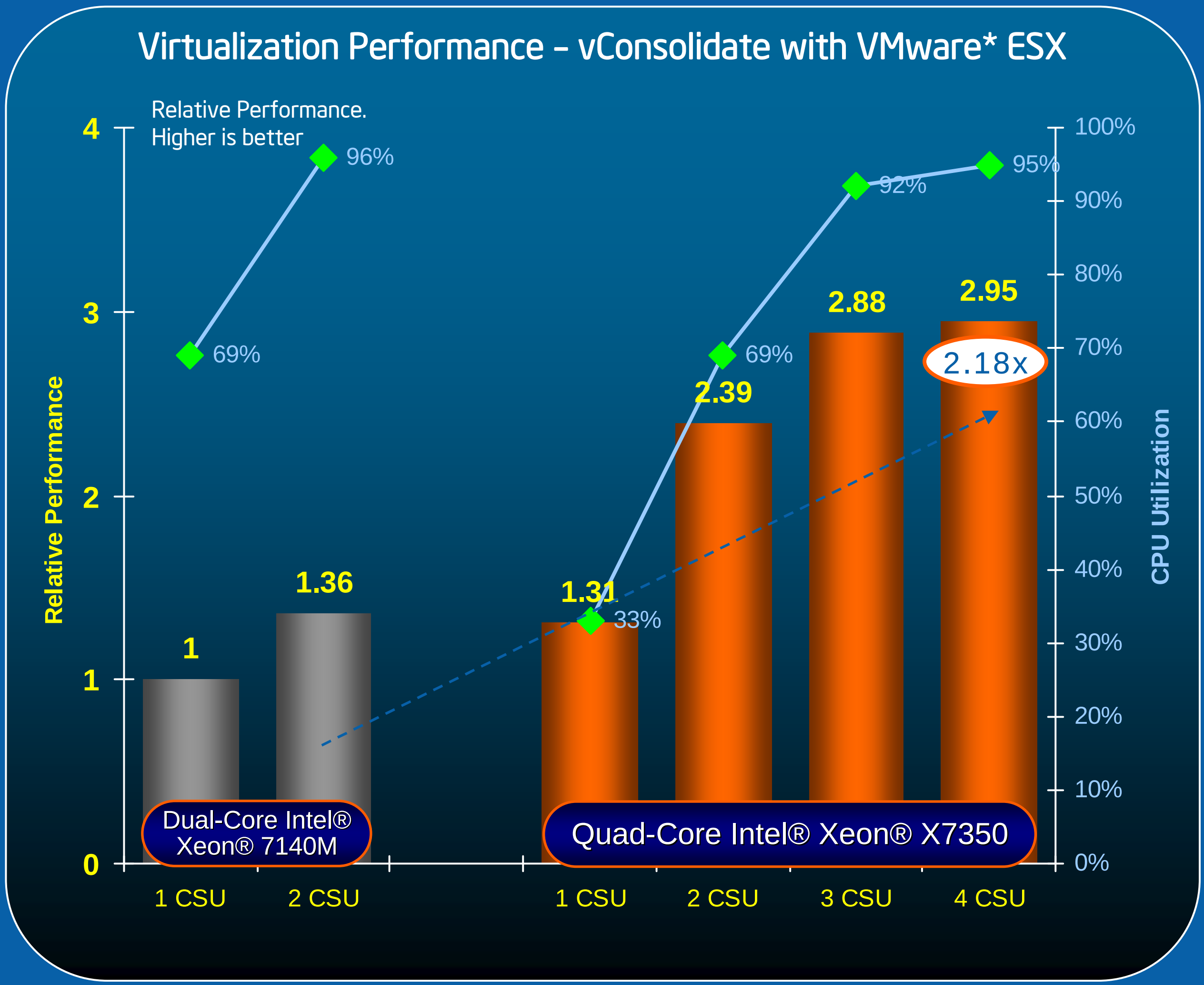


- Represents a server-consolidate usage scenario
  - Benchmark scales thru CSUs (Consolidated Stack Unit)
  - One CSU = 5 VMs (Database, Java, Mail, Web, Idle)
  - Methodology, profiles and run rules

Working with SPEC* Virtualization Committee

*Other names and brands may be claimed as the property of others

vConsolidate Example



*Other names and brands may be claimed as the property of others

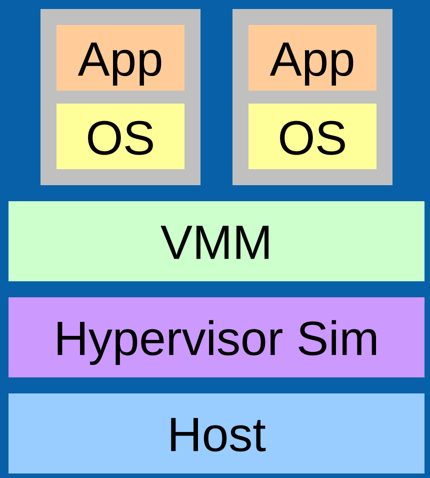
Performance Analysis: Pitfalls

- Effects of “virtual time” dilation
  - Many sources of time distortion in virtualized systems
  - Can’t always trust reported scores from guest software!
- Masking of “virtual” CPU feature set
  - CPUID reports available ISA features to guest (SSE, etc.)
  - CPUID values sometimes masked (e.g., for VM migration)
- Guest software stack configuration
  - Improperly configured guest can significantly affect results
  - OS version, “VM tools”, paravirtualized drivers, etc.

But Virtualization Can Also Help…

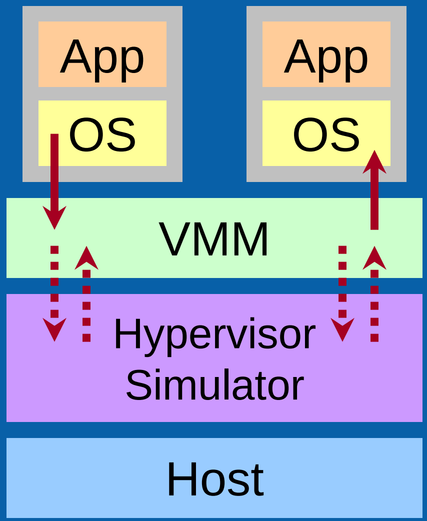
- VMM as a monitoring tool
- Workload management
- Reproducibility of measurement

Hypervisor as Simulator



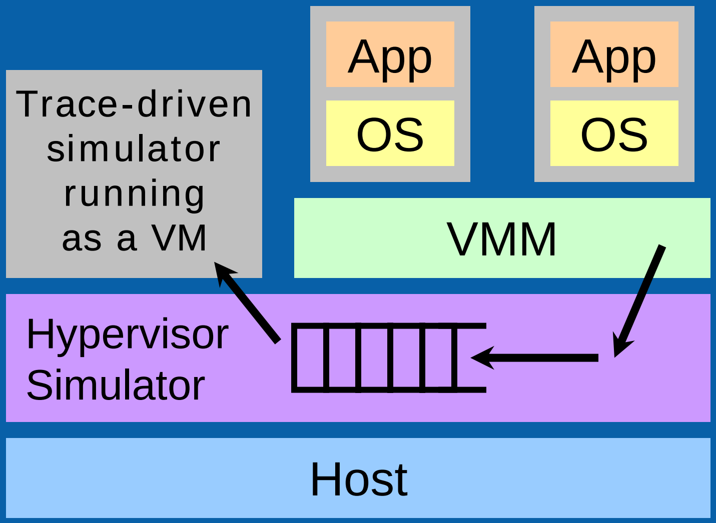
- Deprivilege VMM to run as a guest
- An example of "recursive virtualization"

Some Examples…



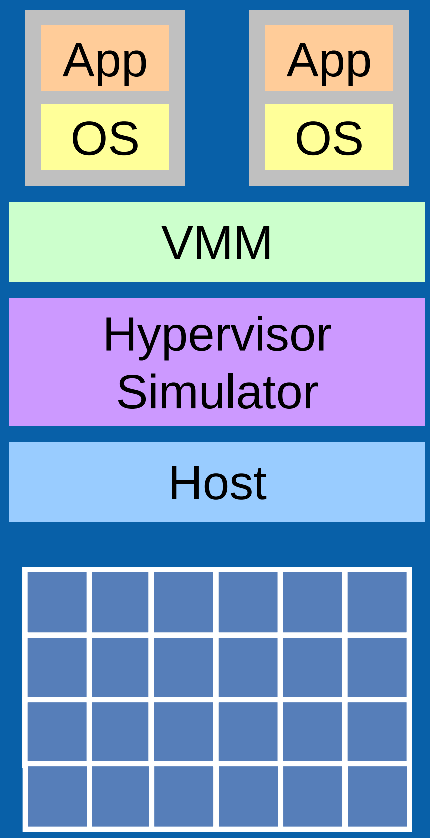
- Measuring Event Frequencies
  - Without instrumenting guest VMM by leveraging VT

Some Examples…



- Measuring Event Frequencies
  - Without instrumenting guest VMM by leveraging VT
- Trace Collection
  - VT supports guest single-stepping
  - Consume trace on-the-fly from another VM

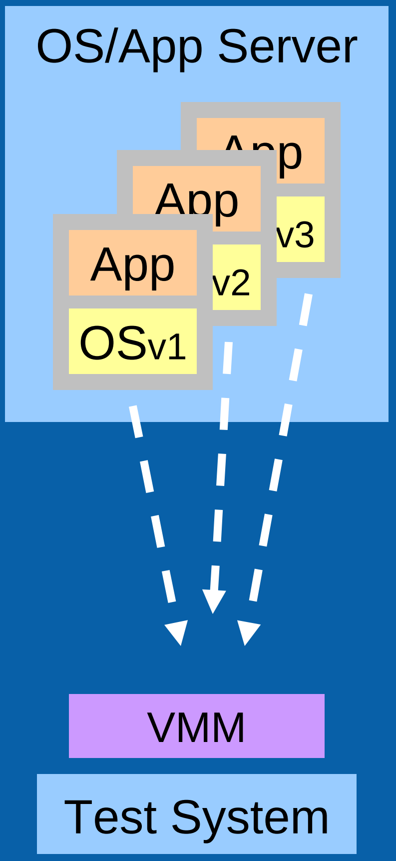
Some Examples…



- Measuring Event Frequencies
  - Without instrumenting guest VMM by leveraging VT
- Trace Collection
  - VT supports guest single-stepping
  - Consume trace on-the-fly from another VM
- Fast TLB / Mem Simulation
  - Effect of larger/smaller TLBs/mem
  - By controlling page-table contents

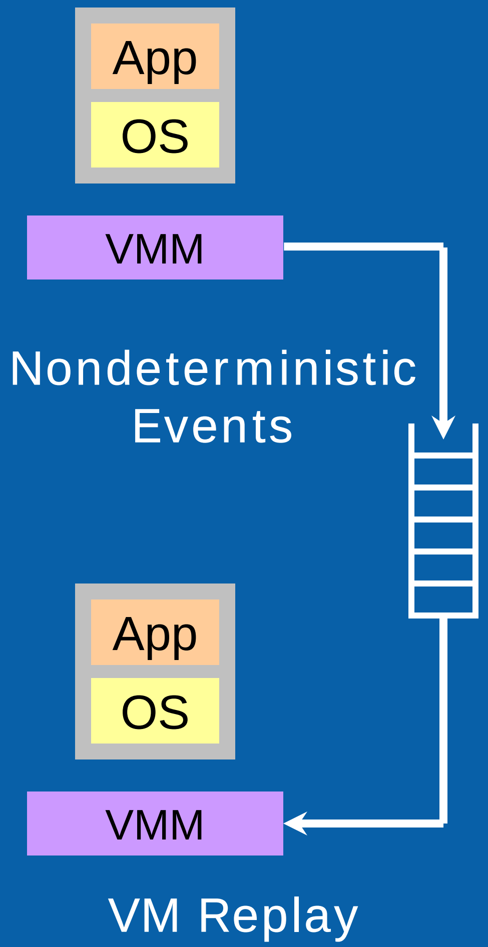
Workload Management

- Workload = OS+App+Compiler
  - Exact combo can significantly affect results
  - Tedious to manage different versions
- Encapsulate workloads in VMs
  - Simplifies workload configuration
  - Automate performance regressions
  - Archive to reproduce results in future



Reproducing Results with VM Replay

- Log nondeterministic events
  - (virtual) I/O operations
  - (virtual) hardware interrupts
- Replay events for later VM run
  - Precise, identical instruction stream
  - Remove run-to-run variability
- Some Limitations
  - Current state-of-the art: UP replay
  - Research: MP logging and replay



Summary and Conclusions

- Virtualization is not just for mainframes anymore
- Intel redesigning every aspect of the platform to support and accelerate this trend
- Many new challenges for performance analysis…
  … but virtualization itself can help with new tools

Leap ahead™

Legal Disclaimer

- INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL® PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER, AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL® PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT. INTEL PRODUCTS ARE NOT INTENDED FOR USE IN MEDICAL, LIFE SAVING, OR LIFE SUSTAINING APPLICATIONS.
- Intel may make changes to specifications and product descriptions at any time, without notice.
- All products, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice.
- Intel, processors, chipsets, and desktop boards may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.
- Intel and the Intel logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
- *Other names and brands may be claimed as the property of others.
-

Throughout this presentation:

VT-x *refers to Intel® VT for IA-32 and Intel® 64*

VT-i *refers to the Intel® VT for IA-64*

VT-d *refers to Intel® VT for Directed I/O*

VT-c *refers to Intel® VT for Connectivity*