A Primer on Virtualization

Agenda

Westmere EP on Intel Roadmap

Adv. Enc. Std: New Instructions (AES-NI) in Westmere EP

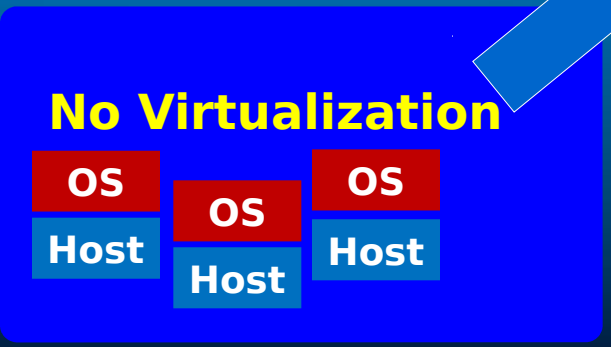
Intel® Virtualization Technology – a primer

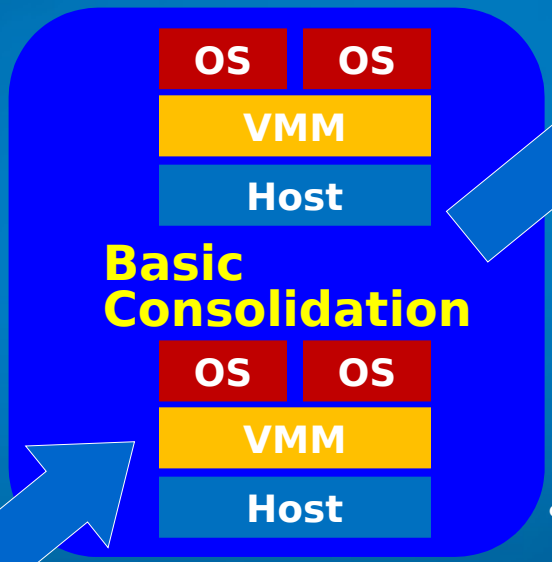
- Evolution of Virtualization
- Intel Virtualization Technologies
- A Framework for Optimizing Virtualization – improving efficiency and scaling
- Reducing Virtualization Overheads: Processor, Memory, I/O
- Improving VM scaling
- VMM readiness
- Improving VMM security through TXT

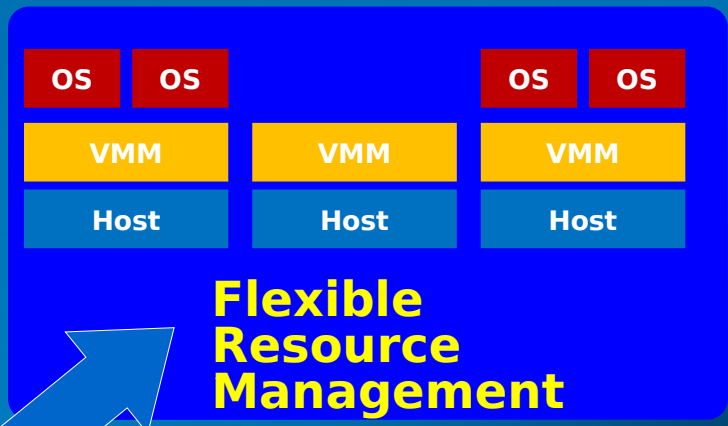
Summary

Content condensed from: "*Intel® Virtualization Technology in the Enterprise*", Rich Uhlig, Intel Fellow, IDF 2009 San Francisco USA.

Evolution of Virtualization







Drives Priorities:

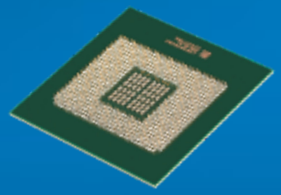
- Lower VMM Overheads
- Richer Workloads in VMs
- Seamless VM Migration
- Power Efficiency
- Increased Scaling (VMs, vCPUs)
- Improved Reliability

Intel® Virtualization Technologies

Intel® VT-x

Processor

Intel® VT-x
Hardware assists for robust virtualization
Intel® VT FlexMigration – Flexible live migration
Intel® VT FlexPriority – Interrupt acceleration
Intel® EPT – Memory Virtualization



Intel® VT-d

Chipset

Intel® VT for Directed I/O
Reliability and Security through device Isolation
I/O performance with direct assignment

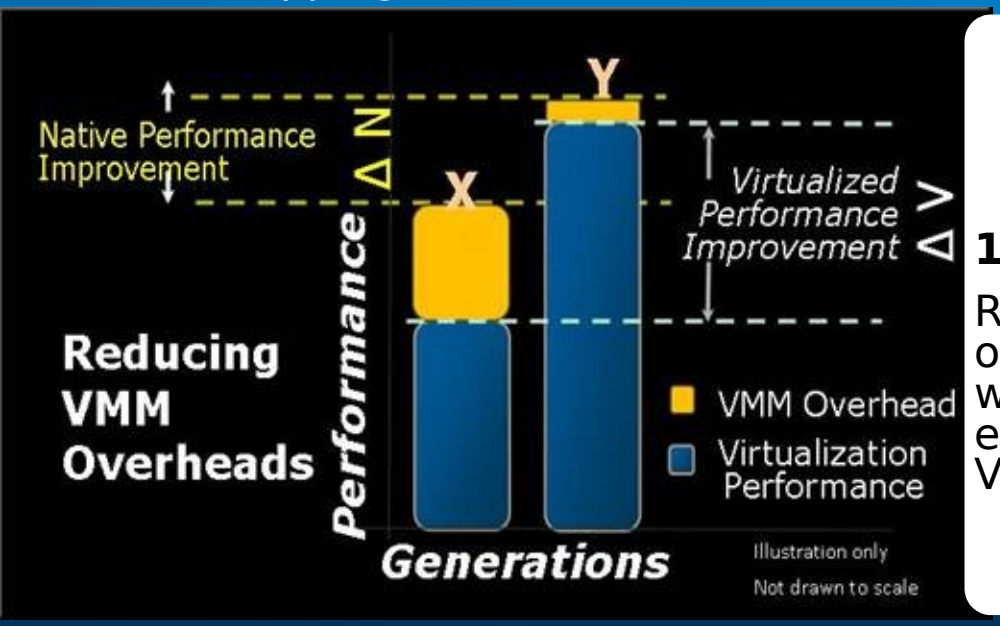
Intel® VT-c

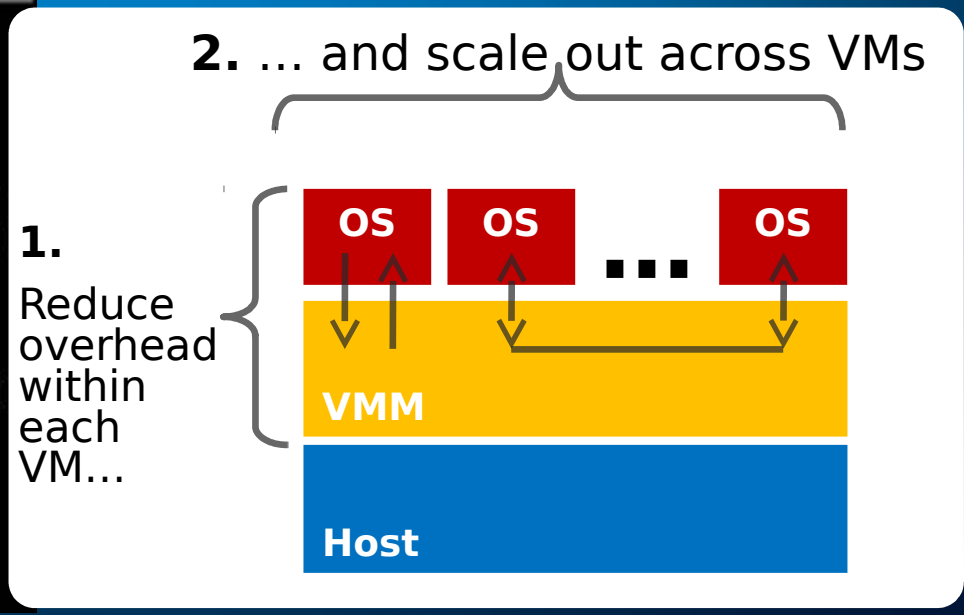
Network

Intel® VT for Connectivity
NIC Enhancement with VMDq
Single Root IOV support
Network Performance and reduced CPU utilization
Intel® I/OAT for virtualization
Lower CPU Overhead and Data Acceleration

A Framework for Optimizing Virtualization

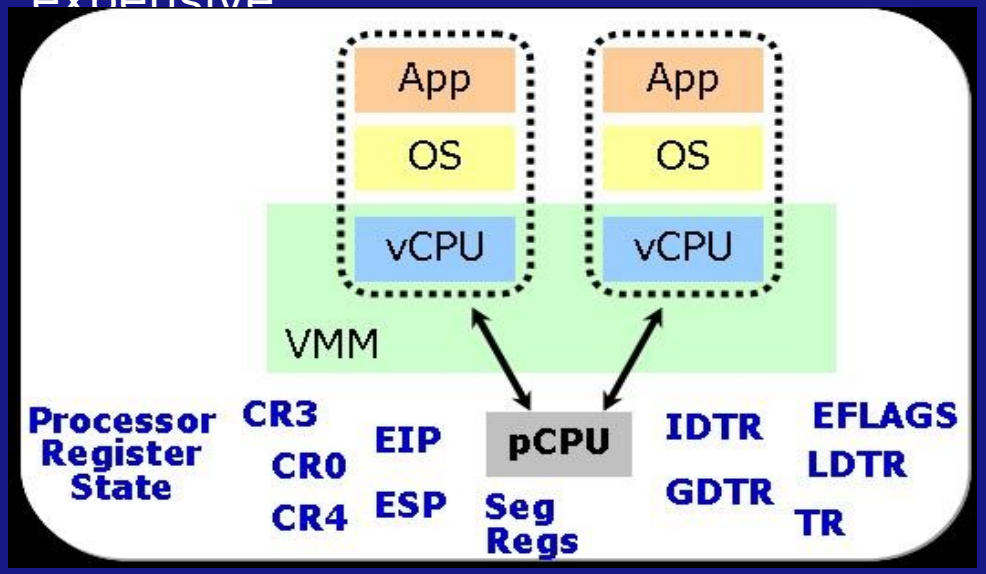
- Reduce overheads from virtualization
  - Intel® VT-x Latency Reductions
  - Extended Page Tables
  - Virtual Processor IDs
  - APIC Virtualization (Flex Priority)
  - I/O Assignment via DMA Remapping
- Introduce capabilities that increase scaling out across VMs
  - Intel® Hyper-Threading Technology
  - PAUSE-loop Exiting
  - Network Virtualization





Reducing VT Latencies

What makes VM Context Switching expensive

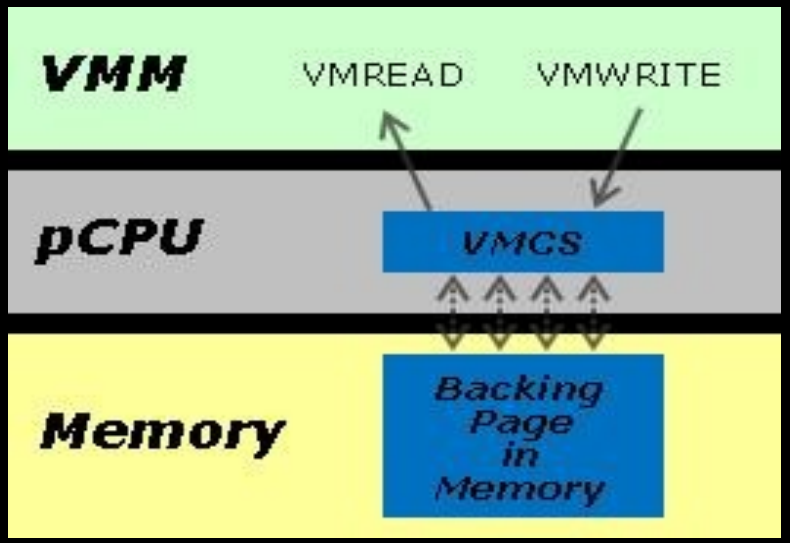


Saving-Loading privileged state

- Compounded by consistency checking

Addressing Context changes

- Translation Lookaside Buffer (TLB) flushes



Virtual Machine Ctl Structure (VMCS)

- Maintains Guest & Host reg. state
  - "Backed" by host physical memory
- Accessed via architectural VMREAD/VMWRITE
- Enables caching of VMCS state on-die

Virtual Processor IDs (VPIDs)

- Tag µarch structures (TLBs)
- Removes need to flush TLBs

Intel® Virtualization Technology (VT-x)

- VT-x® provides architected assists to allow guest OSes to run directly on hardware
- On Nehalem and Westmere VT-x is extended with:

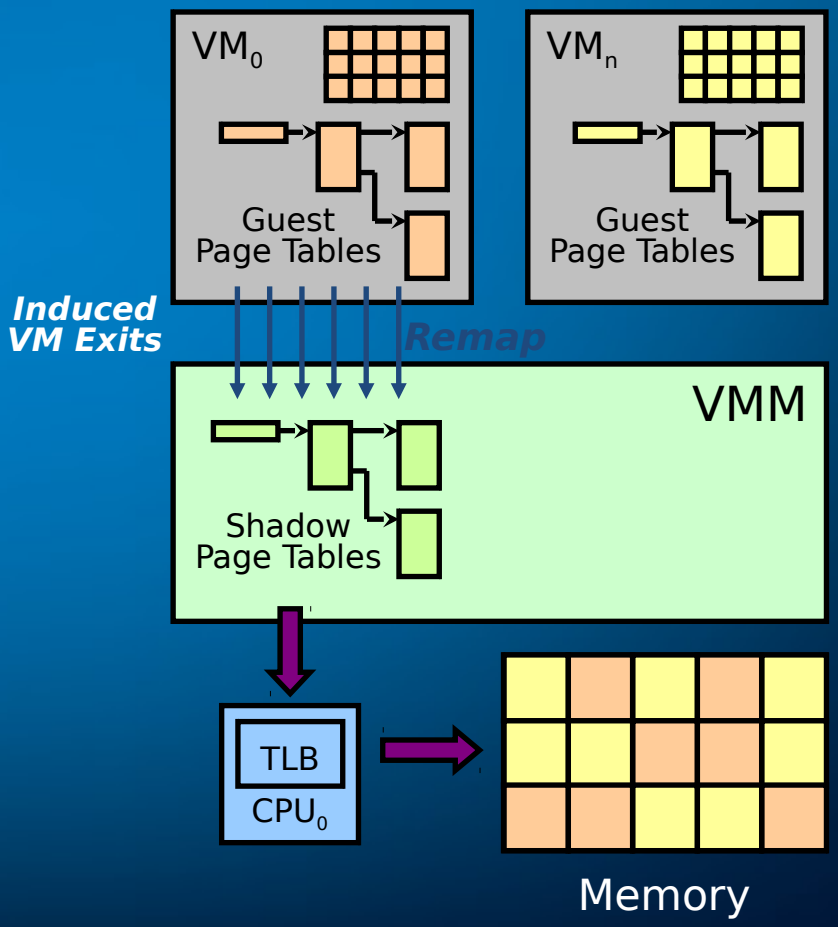
Extended Page Tables (EPT)	Eliminates VM exits to the VMM for shadow page-table maintenance
Virtual Processor IDs (VPID)	Avoid flushes on VM transitions to give a lower-cost VM transition time
Guest Preemption Timer -- lets a VMM preempt a guest OS	Aids VMM vendors in flexibility and Quality of Service (QoS)
Descriptor Table Exiting –Traps on modifications of guest DTs	Allows VMM to protect a guest from internal attack
Transition Latency reductions	Continuing improvements in microarchitectural handling of VMM round trips

Issues with abstracting physical memory

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



How Extended Page Tables help with abstracting Physical Memory

Extended Page Tables (EPT)

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

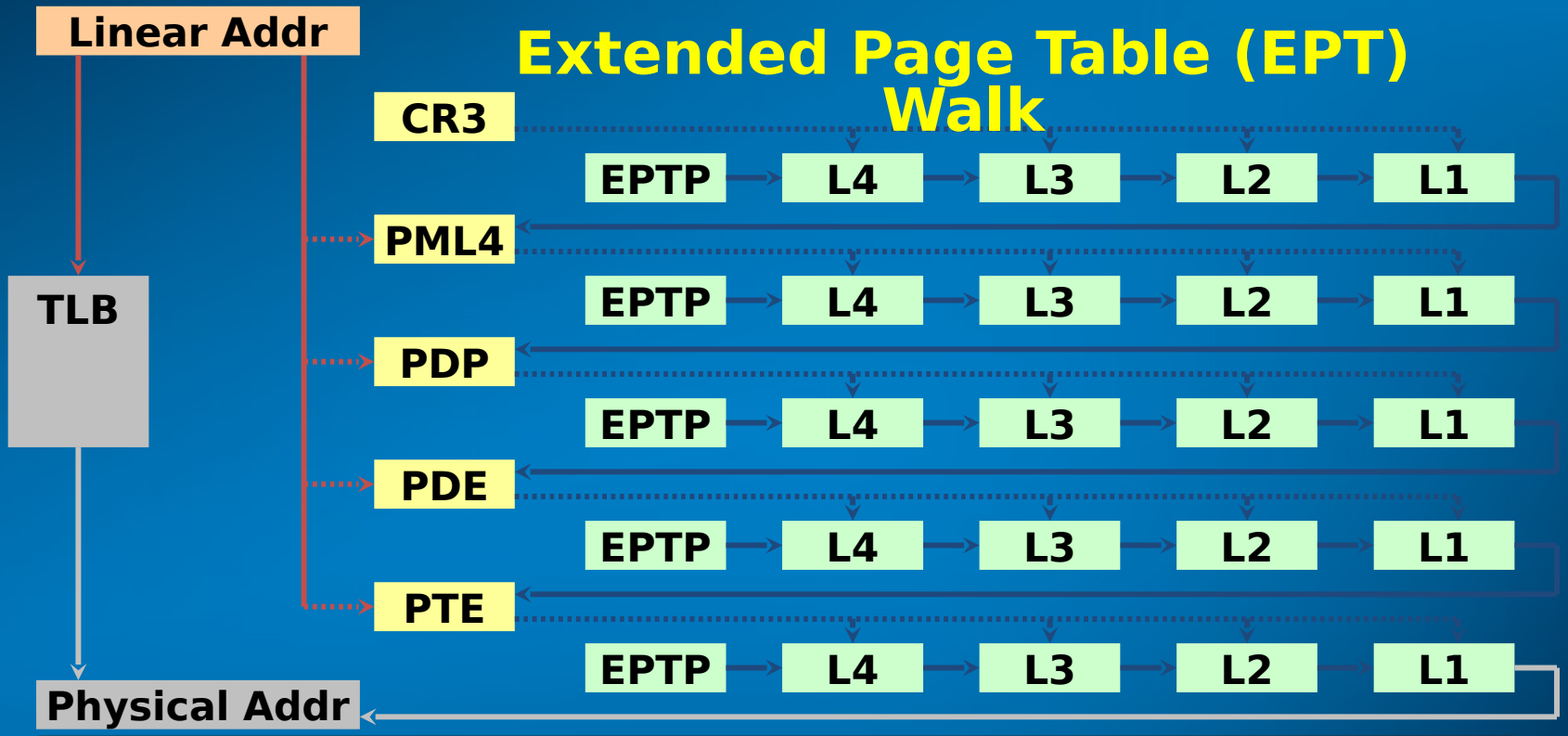
Performance Benefit

- A guest OS can modify its own page tables freely and without VM exits

Memory Savings

- A single EPT supports entire VM: instead of a shadow pagetable per guest process

Extended Page Table (EPT) Walk



2-level TLB reduces page-table walks
- VPID tags help to retain TLB entries

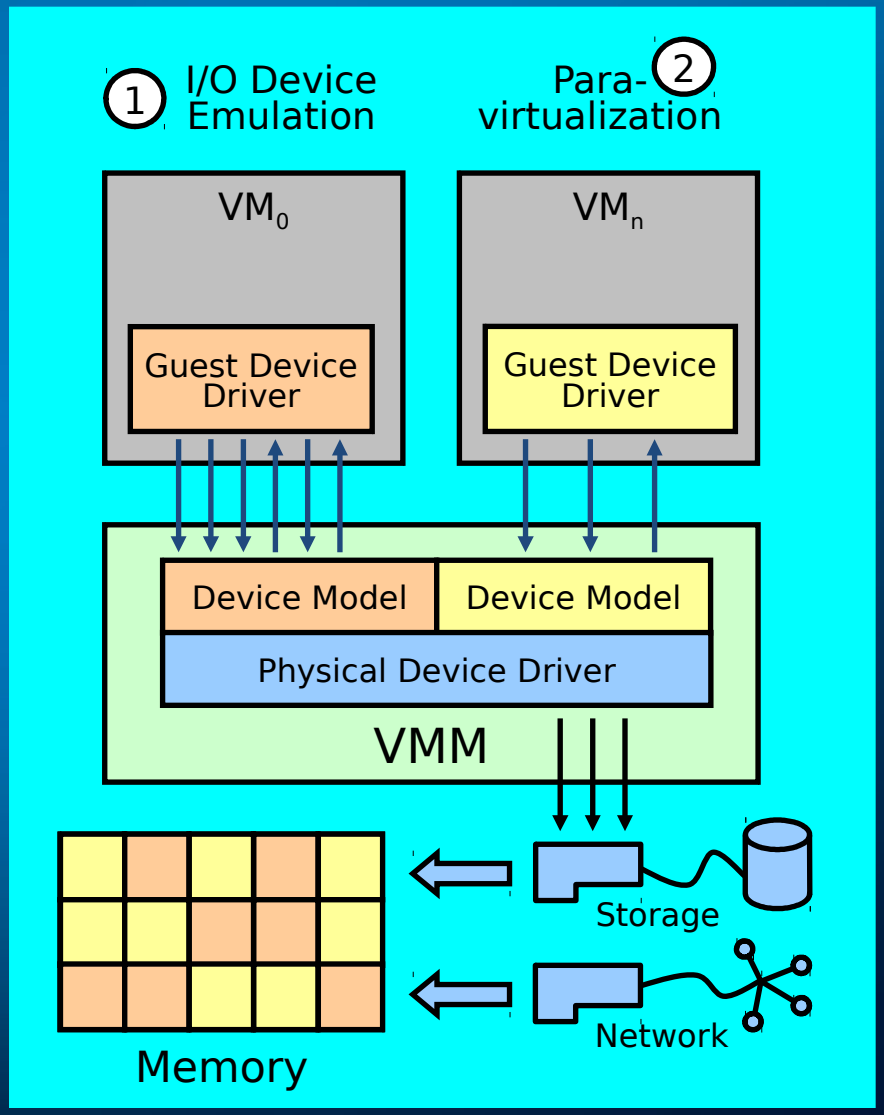
Paging-structure caches
- Cache intermediate steps in walk
- Result: Reduce length of walks
- Common case: Much better than 24 s

Difficulties in virtualizing I/O

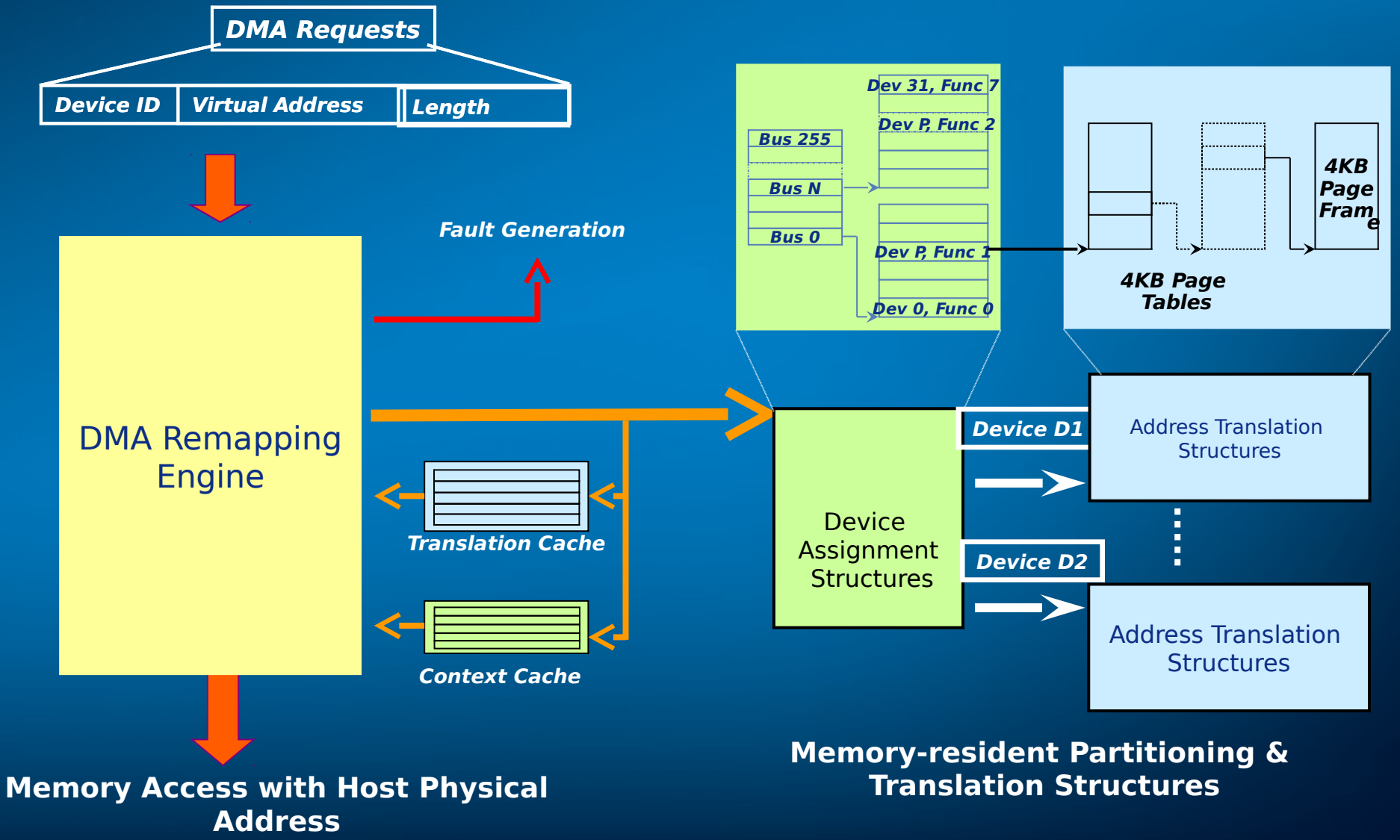
Virtual Device Interface
- Traps device commands
- Translates DMA operations
- Injects virtual interrupts

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

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



Solution: DMA remapping (Intel® VT-d)

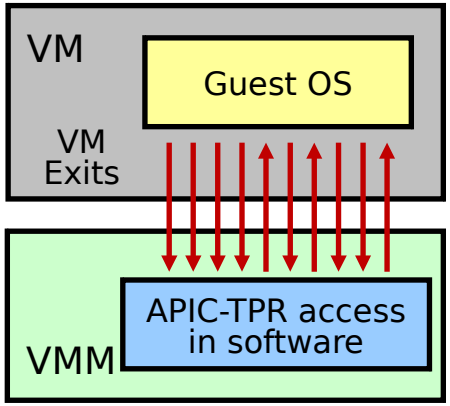


Intel® VT FlexPriority

APIC Task Priority Register (TPR)

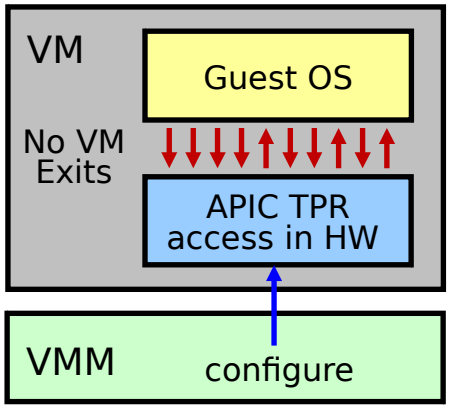
- Controls interrupt delivery through the APIC
- Accessed very frequently by some guest OSes

Without Intel VT FlexPriority



- Fetch/decode instruction
- Emulate APIC-TPR behavior
- Thousands of cycles per exit

With Intel VT FlexPriority



- Instruction executes directly
- Hardware emulates APIC-TPR access
- No VM exit in the common case

Technologies to improve VM scaling

- Hyper-threading
- Decreasing lock holder preemption impact
- Network virtualization with Virtual Machine Device Queues
- Single Root I/O Virtualization (SR-IOV)

Reducing Lock Holder preemption impact

Problem:

- In an SMP guest, a vCPU holding a lock may get preempted
- Other vCPUs that attempt to acquire that lock spin for the full quantum

Solution: Pause-Loop Exiting (PLE)

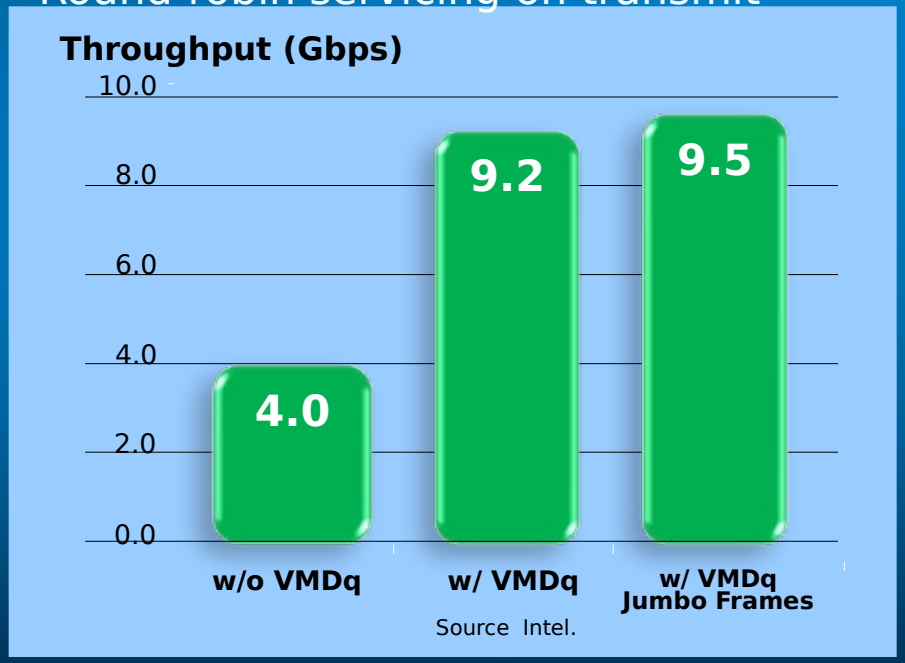
- Spin-locking code typically uses PAUSE instructions in a loop
- A longer than "normal" loop duration taken as a sign of lock-holder preemption
- When that happens, HW forces an exit into VMM
- VMM takes control and schedules some other vCPU

```
spin_lock:
   attempt lock-acquire;
   if fail {
      PAUSE;
      jmp spin_lock
   }
```

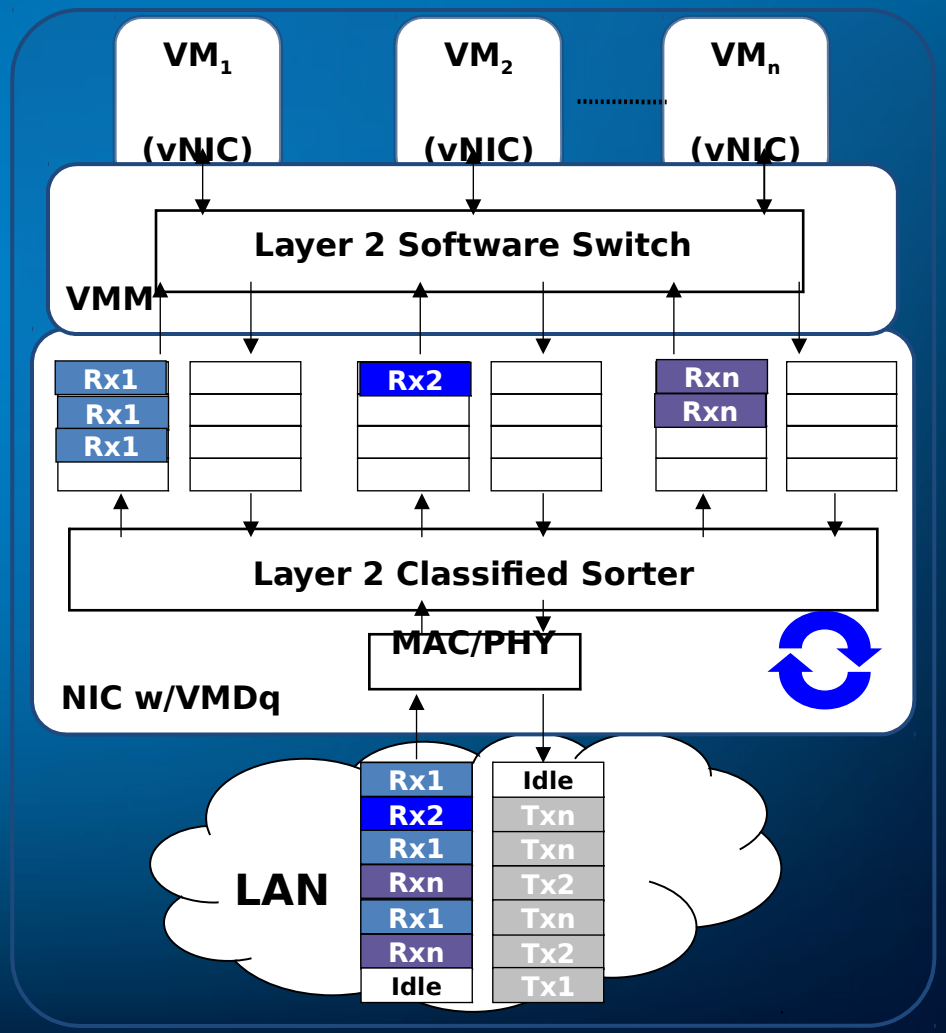
Network Virtualization: HW traffic management

Virtual Machine Device Queues (VMDq)

- Data packets grouped and sorted in HW
- Packets sent to their respective VMs
- Round-robin servicing on transmit



Tests measure Wire Speed Receive (Rx) Side Performance With VMDq on Intel® 82598 10 Gigabit Ethernet Controller



PCI-SIG SR-IOV

Description:

- PCI-SIG Single Root I/O Virtualization (SR-IOV) Standard:
- Allows for I/O devices to be simultaneously shared among VMs
- Virtual interfaces can be directly assigned to reduce routing overheads

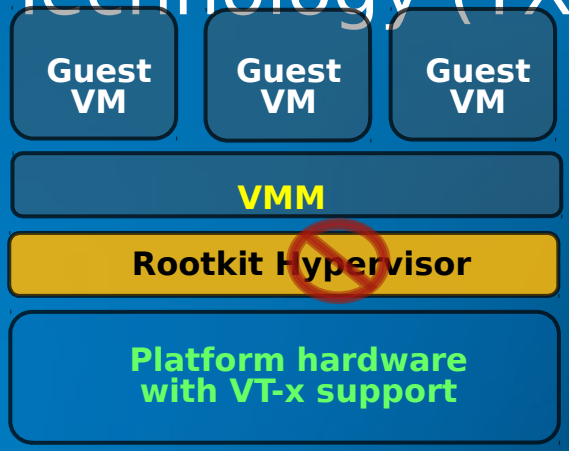
Benefits:

- Provide near native performance due to direct connectivity
- Allows direct VM control of I/O virtual functions

Requirements:

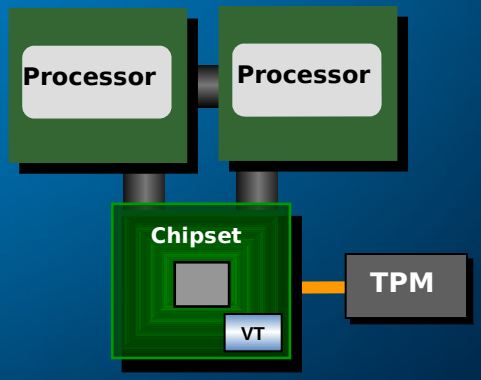
- Intel VT-d as the core platform ingredient
- BIOS support of SR-IOV

Westmere: Trusted Execution Technology (TXT) Server Extensions



Helps prevent hijacking by rootkit

- TXT uses features in processor, chipset and TPM to enable more secure platforms
- TXT works through measurement, memory locking and sealing secrets
- TXT helps prevent software attacks
  - such as, attempts to insert rogue VMM (rootkit hypervisor)
  - and, reset-attacks that are designed to compromise platform secrets in memory
  - and, BIOS and firmware update attacks



TXT technology incorporates multiple components

TXT Makes Platforms More Robust Against SW-based Attacks

Intel VT Features/VMM Support

	Feature		VMware ESX		Microsoft Hyper-V		Xen OSS		Red Hat (RHEL)		KVM		Novell (SLES)	
			Min Rev	Date	Min Rev	Date	Min Rev	Date	Min Rev	Date	Rev	Date	Min Rev	Date
Virtualization	Processor (VT-x)	Min Rev for Nehalem	3.5U4	Now	Win Svr '08	Now	Any	Now	Any	Now	Any Recent	Now	Any	Now
		FlexPriority	3.5U4	Now	Win Svr '08	Now	3.1	Now	5.2	Now	2.6.24	Now	10 SP1	Now
		FlexMigration	3.5U4	Now	Win Svr '08[2]	Now	3.3	Now	TBD	TBD			TBD	TBD
		EPT + VPID	4.0	Now	Win Svr '08 R2	1H'10	3.3	Now	5.3	Now	2.6.26	Now	10 SP2	Now
	Chipset (VT-d)	VT-d2	4.0	Now	TBD	TBD	3.3	Now	5.4	Q3'09	2.6.31	Now	11	Now
	Network (VT-c)	SR-IOV	TBD	TBD	TBD	TBD	3.4[1]	Now[3]	TBD	TBD	TBD	TBD	TBD	TBD
		VMDq	3.5U4	Now	Win Svr '08 R2	1H'10	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD
PWR		Power Mgmt (S, P, C, T States)	4.0	Now	1.0 (S-state TBD)	Now	3.3	Now	TBD	TBD			11	Now
Other		SSE 4.2	4.0	Now	TBD	TBD	Any	Now	Any	Now			Any	Now
		Turbo	4.0	Now	1.0	Now	Any	Now	Any	Now			Any	Now
		AES (GCM)	3.5U4	Now	1.0	Now	Any	Now	Any	Now			Any	Now

Notes: Table is based on latest information as of: VT-c : July 09; VT-x and VT-d: May 09 and subject to change; Please contact vendors directly for unreleased products

[1] Requires other ecosystem support
[2] Product allows capability but robustness improvements in future releases
[3] PCI-SIG SR-IOV standard is supported now


intel
Software