WHITE PAPER
Intel® Virtualization Technology
Enterprise Server

Enabling Intel® Virtualization Technology Features and Benefits

Maximizing the benefits of virtualization with Intel's new CPUs and chipsets

EXECUTIVE SUMMARY

Although virtualization has been accepted in most data centers, some users have not yet taken advantage of all the virtualization features available to them. This white paper describes the features available in Intel® Virtualization Technology (Intel® VT) that work with Intel's new CPUs and chipsets, showing how they can benefit the end user and how to enable them.

Intel® Virtualization Technology Feature Brief and Usage Model

Intel VT combines with software-based virtualization solutions to provide maximum system utilization by consolidating multiple environments into a single server or PC. By abstracting the software away from the underlying hardware, a world of new usage models opens up that can reduce costs, increase management efficiency, and strengthen security—all while making your computing infrastructure more resilient in the event of a disaster.

During the last four years, Intel has introduced several new features to Intel VT. Most of these features are well known, but others may not be.

This paper describes key features of Intel VT, how they fit into Intel's platforms, and how to maximize their benefits.

Intel VT CPU-Based Features

The x86 processor architecture did not originally meet the Formal Requirements for Virtualizable Third-Generation Architectures, a specification for virtualization created in 1974 by Gerald J. Popek and Robert P. Goldberg. Thus, developers found it difficult to implement a virtual machine platform on the x86 architecture without significant overhead on the host machine.

In 2005 and 2006, Intel and AMD, working independently, each resolved this by creating new processor extensions to the x86 architecture. Although the actual implementation of processor extensions differs between AMD and Intel, both achieve the same goal of allowing a virtual machine hypervisor to run an unmodified operating system without incurring significant emulation performance penalties.

Intel VT is Intel's hardware virtualization for the x86 architecture that helps consolidate multiple environments into a single server, workstation, or PC so that you need fewer systems to complete the same tasks.

The sections that follow explain some of the key CPU-based features of Intel VT.

Intel® VT FlexPriority

Intel® VT FlexPriority is a processor extension that optimizes virtualization software efficiency by improving interrupt handling.

Marco Righini
Intel Corporation
marco.righini@intel.com

To enable Intel VT FlexPriority, you enable Intel VT extensions. Like most hardware features, Intel VT FlexPriority must be enabled by the hypervisor or virtual machine monitor (VMM), which allows multiple operating systems to run concurrently on a host computer.

Intel VT FlexPriority eliminates most VM exits due to guest task priority registers (TPR) access. This reduces the virtualization overhead and improves I/O throughput. Table 1 lists which Intel CPUs have Intel VT FlexPriority; Table 2 maps Intel VT features to CPUs. Figure 1 shows the reduction of EXITs and also looks at the I/O throughput measured (best-case scenario).

Intel® VT FlexMigration

Intel® VT FlexMigration is a feature of Intel Virtualization Technology that enables you to build one compatible virtualization pool and conduct live virtual machine (VM) migration across all Intel® Core™ microarchitecture-based servers. It gives you the power to choose the right server platform to best optimize performance, cost, power, and reliability.

Combined with support from a virtualization software provider, this feature allows IT to maximize flexibility by providing the ability to build a single live migration compatibility pool with multiple generations of Intel Xeon processor-based servers.

For details on Intel VT FlexMigration, visit http://communities.intel.com/docs/DOC-4124.

Virtual Processor IDs (VPID)

Traditionally, every time a hypervisor switched execution between different VMs, the VM and its data structure had to be flushed out of the transition look-aside buffers (TLB) associated with the CPU caches, since the hypervisor had no information on which cache line was associated with any particular VM.

With virtual processor IDs (VPID), a VM ID tag in the CPU hardware structures (e.g., TLB) associates cache lines with each VM actively running on the CPU. This permits the CPU to flush only the cache lines associated with a particular VM when it is flushed from the CPU, avoiding the need to reload cache lines for a VM that was not migrated and resulting in lower overhead.

VPID is available on all new Intel Xeon processors starting with the 5500, 5600, and 7500 series.

VGuest Preemption Timer

The Guest Preemption Timer is a mechanism that enables a VMM to preempt the execution of a guest OS.

Programmable by VMM, the timer causes the VM to exit when the timer expires. It has no impact on interrupt architecture.

Table 1. Intel Processors with Intel® VT FlexPriority

Intel® Microarchitecture	Enhanced vMotion* Compatibility (EVC) Setting	Example
45nm Intel® Core™ processor family	Intel® Xeon® processor (45nm) Intel Core 2 processor	Intel Xeon processor 5400 or 7400 series
Next-generation Intel® microarchitecture	Intel Xeon processor Intel Core i7 processor (45nm)	Intel Xeon processor 5500 or 7500 series
Intel Xeon processor 5600 series	Intel Xeon processor Intel Core i7 processor (32nm)	Intel Xeon processor 5600 series

Table 2. Intel Virtualization Technology Feature and CPU Mapping

	Intel® Xeon® Processor					
	7400 Series	7500/ 6500 Series	5500 Series	5600 Series	3300/ 3100 Series	3400 Series
VT-x Base	✓	✓	✓	✓	✓	✓
Intel® VT FlexPriority	✓	✓	✓	✓	✓	✓
Intel® VT FlexMigration	✓	✓	✓	✓	✓	✓
Extended Page Tables (EPT)		✓	✓	✓		✓
Virtual Processor ID (VPID)		✓	✓	✓		✓
Guest Preemption Timer		✓	✓	✓		✓
Descriptor-Table Exiting		✓	✓	✓		✓
Pause-Loop Exiting		✓				
TXT				✓		
Real Mode Support				✓		

This feature helps VMM vendors fulfill flexibility and quality of service guarantees. It can help when you need to switch tasks or allocate a certain amount of CPU power to a task. For telecom and networking applications, it makes virtualization a useful tool—and possibly a must-have feature.

Descriptor Table Exiting

This feature allows a VMM to protect a guest OS from internal attack by preventing relocation of key system data structures.

Pause-Loop Exiting

This feature is a hardware assist to enable detection of spin locks in guest software and avoid lock-holder preemption. It helps to reduce overhead and improve performance.

Real Mode Support

This feature allows guests to operate in real mode, removing the performance overhead and complexity of an emulator.

Uses include:

- Early VMM load
- Guest boot and resume

Extended Page Table (EPT)

Typical Intel® architecture 32-page tables (referenced by control register CR3) translate from linear addresses to guest-physical addresses. With the Extended Page Table (EPT) feature, a separate set of page tables (EPTs) translate from guest-physical addresses to host-physical addresses that are used to access memory. As a result, the guest OS can be allowed to modify its own page tables and directly handle page faults.

This allows a VMM to avoid the VM exits associated with page-table virtualization, which is a major source of virtualization overhead without EPT.

Figure 2 shows how the EPT works.

Intel® Trusted Execution Technology

Intel® Trusted Execution Technology (Intel® TXT) provides a hardware-based security foundation on which to build and maintain a chain of trust to protect the platform from software-based attacks.

The goal of Intel TXT is to provide an accurate measurement, at launch, of the measured launch environment (MLE) through the hardware features built into the CPU and chipset. This hardware-based security provides a foundation on which trusted platform solutions can be built to protect the platform from software-based attacks.

Figure 3 shows how Intel TXT works.

Features of Intel TXT include:

- **Verified Launch.** An Intel TXT hardware-based chain of trust enables launch of MLE into a known, expected state. Changes to MLE can be detected via hash-based measurements.
- **Protected Configuration.** Intel TXT hardware protects the launched configurations from malicious software, maintaining the integrity of the measured launched environment's identity.
- **Secret Protection.** Intel TXT hardware removes residual data at improper MLE shut-down, protecting data from memory snooping software.

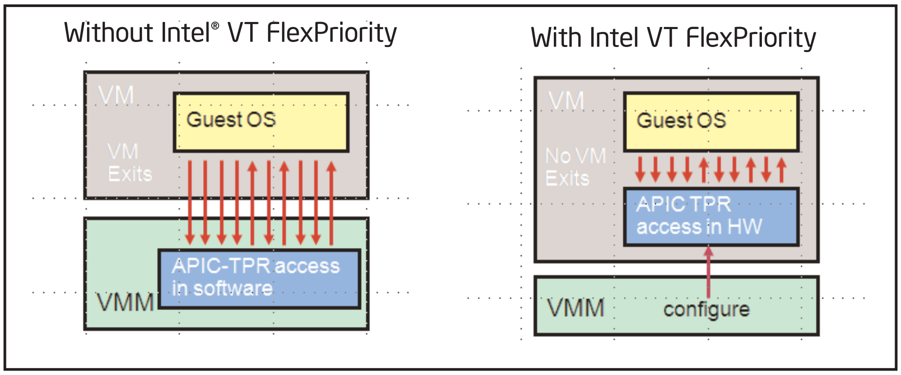
Without Intel® VT FlexPriority
VM
Guest OS
VM Exits
APIC-TPR access in software
VMM
With Intel VT FlexPriority
VM
Guest OS
No VM Exits
APIC TPR access in HW
VMM
configure

Figure 1. Reduction of EXITs with Intel® VT FlexPriority

Intel® VT for Directed I/O (Intel® VT-d)

In computing, an input/output memory management unit (IOMMU) is a memory management unit (MMU) that connects a digital media adapter (DMA)-capable I/O bus to the main memory. Like a traditional MMU, which translates CPU-visible virtual addresses to physical addresses, the IOMMU takes care of mapping device-visible virtual addresses (also called device addresses or I/O addresses in this context) to physical addresses. Some units also provide memory protection from misbehaving devices.

Intel® VT-d is a feature integrated into the chipset and therefore not related to the CPU. Before Intel VT-d and hypervisors supporting it, any VM running on top of a VMM was seeing emulated, or para-virtualized, devices. Figure 4 shows how Intel VT-d works.

No matter what type of hardware was physically present in the server, the VM itself sees a virtualized device. So, for example, on VMware vSphere*, you would typically see a VMXnet* network card instead of the real network interface card (NIC) installed on the server.

This has both pros and cons:

- **Pros:** This hides any type of change between the hardware vendors and makes it possible for VMs to migrate easily.
- **Cons:** Performance takes a hit. This is true even if the emulated device is based on a para-virtualized or synthetic driver, either in terms of CPU utilization, bandwidth, or latency.

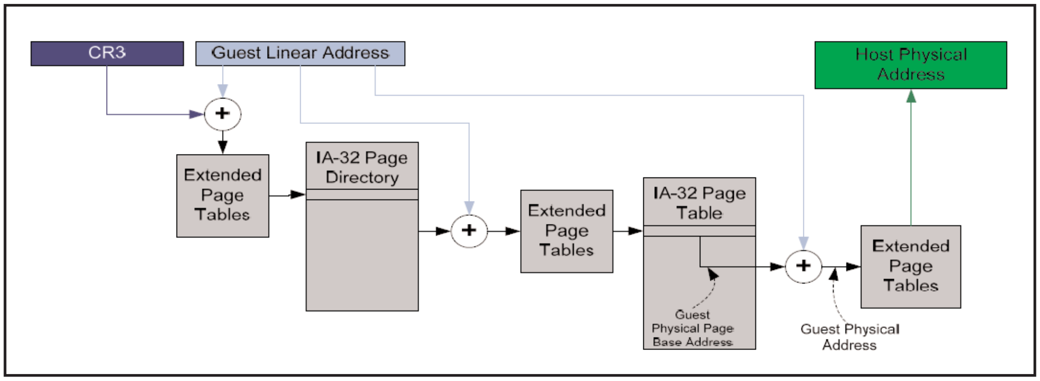

Figure 2 Extended Page Table

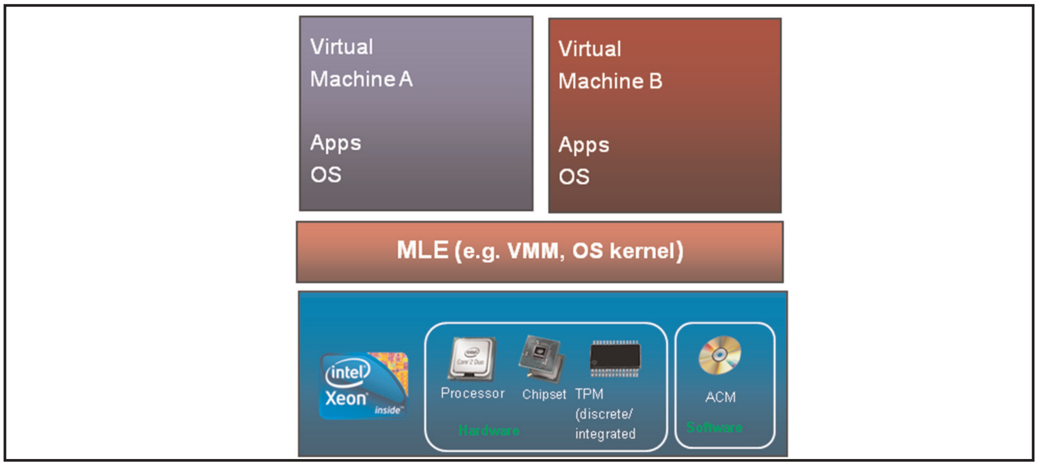

Figure 3. Intel® Trusted Execution Technology

When Intel VT-d is enabled, the guest OS can choose to use either the traditional approach or, as needed, pass-through devices.

In pass-through mode, the PCI* device is not allocated by the hypervisor and, therefore, the device can be allocated directly by a VM which now sees the physical PCI device. (Of course, a portion of the memory of that device is also remapped to the VM through the DMA remap engine.) Intel VT-d needs to be enabled in the BIOS and is a separate flag.

Interrupt-Remapping Support

The Interrupt-Remapping feature enables the VMM to isolate interrupts to CPUs assigned to a given VM and to remap/reroute physical I/O device interrupts. When enabled, this feature helps ensure an efficient migration of the interrupts across CPUs.

Queued-Invalidation Support

Queued-Invalidation enables the VMM to batch digital media translation invalidations. This gives the end user better performance.

Address Translation Services (ATS) support

Address Translation Services (ATS) is a PCI-SIG specification that allows PCI-e devices to cache the IOTLB entries (used for DMA remapping) of that device directly in the device itself. This helps the performance of high-end devices, since the translations can be cached at the device level and the device need not depend on the chipset IOTLB cache. This BIOS feature needs to be enabled to permit proper Intel VT-d implementation.

Large Intel VT-d Pages

The Large Intel VT-d Pages feature enables 2MB and 1GB pages in Intel VT-d page tables. It enables the sharing of Intel VT-d and EPT page tables.

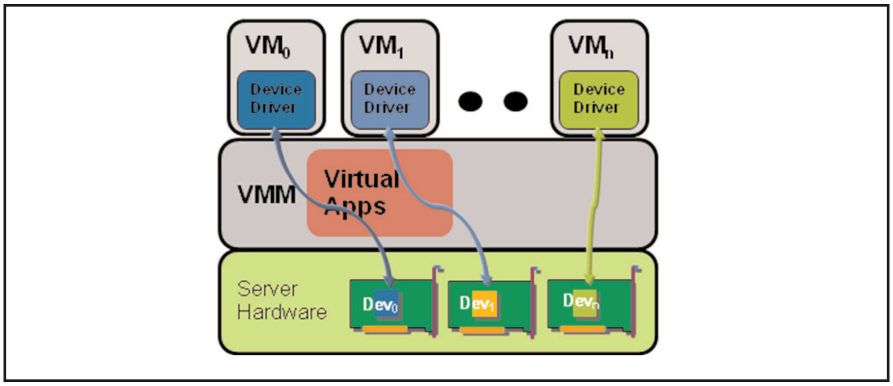

Figure 4. How Intel VT-d works

Table 3. Intel VT-d Feature and Chipset Mapping

Intel VT-d Feature	Intel® Itanium® Proc. 9000 Series	Intel® Xeon® Proc. 7300 Series	Intel Xeon Proc. 7500 Series	Intel Xeon Proc. 5500/ 5200 Series	Intel Xeon Proc. 3200/ 3100 Series	Intel Xeon Proc. 3400 Series
Intel VT-d Base	✓		✓	✓	✓	✓
Interrupt Remapping Support	✓		✓			✓
Queued Invalidation Support	✓		✓	✓		✓
Address Translation Services Support			✓	✓		
Large Intel VT-d Addresses Support for PCI-SIG I/O Virtualization Standards			✓	✓		
Support for PCI-SIG I/O Virtualization Standards	✓		✓	✓		✓

I/O Hardware Assist Features of Intel® Virtualization Technology for Connectivity (Intel® VT-C)

Virtual Machine Device Queue

The Virtual Machine Device Queue (VMDQ) feature is a hardware assist in the Intel networking silicon that improves data processing performance by improving throughput and lowering CPU utilization. This is a more effective way of sorting and grouping data packets at the NIC instead of the VMM.

Intel VMDQ on Intel® Ethernet controllers can lower CPU utilization and improve LAN throughput by supporting:

- Reduced decisions/data copies by VMM switch
- VM transmit fairness with round-robin servicing
- Operation that is independent from Intel VT-d

On the new Intel® 82576 and 82599 10-Gigabit Ethernet controllers, Intel VMDQ provides:

- Flexible bandwidth allocation per VM (only on Intel 82599 Ethernet Controller)
- Hardware support for VM-to-VM loop-back
- Broadcast/multicast replication in hardware

Table 4. OEMs with Intel SR-IOV Cards

	Cisco	Dell	Fujitsu	HP	IBM	Sun
10 GbE NIC	✓	✓	✓			✓
10 GbE Mezz	✓	✓	✓		✓	✓
Quad-Port GbE	✓	✓	✓	✓	✓	✓

Note: HP decided to implement its proprietary technology, called Flex-10, to make this possible.*

Intel VMDQ works with VMMs to remove some virtual network overhead when moving traffic from the network adapter to the VM. Nothing needs to be enabled on the card itself.

Single-Root I/O Virtualization (SR-IOV)

The Single-Root I/O Virtualization (SR-IOV) feature is a PCI Special Interest Group (PCI-SIG) specification. Intel, along with other industry leaders, is actively participating in the PCI-SIG working group to define new standards for enhancing virtualization capabilities of I/O devices. SR-IOV provides a standard mechanism for devices to advertise their ability to be simultaneously shared among multiple virtual machines. It also allows for the partitioning of a PCI function into many virtual interfaces for the purpose of sharing the resources of a PCI Express* (PCIe*) device in a virtual environment. Intel plans to support the SR-IOV specification in its networking devices.

Each virtual function can support a unique and separate data path for I/O-related functions within the PCIe hierarchy. Use of SR-IOV with a networking device, for example, allows the bandwidth of a single port (function) to be partitioned into smaller slices that may be allocated to specific virtual machines, or guests, via a standard interface. A common methodology for configuration and management is also established to further enhance the interoperability of various devices in a PCIe hierarchy. This resource sharing can increase the total utilization of any given resource presented on an SR-IOV-capable PCIe device, potentially reducing the cost of a virtual system.

Intel-enabled NICs are:

- Intel® 82576 Gigabit Ethernet Controller
- Intel® 82599 10-Gigabit Ethernet Controller

PCI-SIG SR-IOV ecosystem requirements include:

- A SR-IOV-capable NIC
- Intel VT-d

- BIOS support
- VM ability to support this feature

Table 4 lists OEMs with Intel SR-IOV cards.

Intel VT ISV Support

Tables 5 through 7 show ISVs that support the features of Intel VT.

Summary

All the features in Intel Virtualization Technology help expand its usefulness by enabling either new virtualized environment usage models or better performance.

When you enable Intel VT and Intel VT-d, you enable all of its major features. (The sub-features of Intel VT-d need to be enabled separately if the target VMM supports it. You can enable pass-through, with or without SR-IOV needs, after Intel VT-d enablement within the software stack. VMDq is enabled by default if the NIC supports it.)

By learning to enable and use all the features of Intel Virtualization Technology, you can reduce costs, increase management efficiency, and strengthen security—all while making your computing infrastructure more resilient in the event of a disaster.

To learn more about Intel Virtualization Technology, visit **www.intel.com/technology/virtualization**.

Table 5. ISV Support for Intel VT

Intel VT-d Feature	VMware	Microsoft	Xen	KVM	Citrix	Red Hat	SuSE	Oracle	Parallels
Intel VT-d Base	✓	✓	✓	✓	✓	✓	✓	✓	✓
Intel VT FlexPriority	✓	✓	✓	✓	✓	✓	✓	✓	✓
Intel VT FlexMigration	✓	✓	✓	✓	TBD	TBD	TBD	TBD	TBD
Extended Page Tables (EPT)	✓	✓	✓	✓	✓	✓	✓	✓	✓
Virtual Process ID(VPID)	✓	✓	✓	✓	✓	TBD	✓	TBD	✓
Guest Preemption Timer	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD
Descriptor Table Exiting	TBD	TBD	✓	TBD	TBD	TBD	TBD	TBD	TBD
Pause-Loop Exiting	TBD	✓	✓	WIP	TBD	TBD	TBD	TBD	TBD
Real Mode Support	TBD	TBD	✓	✓	TBD	TBD	TBD	TBD	TBD

Table 6. ISV Intel VT-d Support Matrix

Intel VT-d Feature	VMware	Micro-soft	Xen	KVM	Citrix	Red Hat	SuSE	Oracle	Parallels
VT-d Base	✓	TBD	✓	✓	TBD	✓	✓	TBD	✓
Interrupt-Remapping Support	✓	TBD	✓	✓	TBD	TBD	TBD	TBD	TBD
Queued-Invalidation Supporet	✓	TBD	✓	✓	TBD	TBD	TBD	TBD	TBD
Address Translation Services	TBD	TBD	✓	✓	TBD	TBD	TBD	TBD	TBD
Large VT-d Pages	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD	TBD

Table 7. ISV VMDq and Intel NIC Support Matrix

I/O Silicon	Oplin		Niantic		Zoar		Kawela	
	Base Driver	VMDq	Base Driver	VMDq	Base Driver	VMDq	Base Driver	VMDq
ESX 3.5	Now	Now	Now	Now	Now	TBD	Now	TBD
ESXi 3.5	Now	Now	Now	Now	Now	N/A	N/A	N/A
ESX 4.0	Now	Now	Now	Now	Now	TBD	Now	TBD
Xen* Kernel	Now	TBD	Now	TBD	Now	TBD	Now	TBD
Microsoft* Hyper-V	Now	N/A	Now	N/A	Now	N/A	Now	N/A
Microsoft* Hyper-V R2	Now	Now	Now	Now	Now	Now	Now	Now

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. UNLESS OTHERWISE AGREED IN WRITING BY INTEL, THE INTEL PRODUCTS ARE NOT DESIGNED NOR INTENDED FOR ANY APPLICATION IN WHICH THE FAILURE OF THE INTEL PRODUCT COULD CREATE A SITUATION WHERE PERSONAL INJURY OR DEATH MAY OCCUR.

Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing. For more information on performance tests and on the performance of Intel products, visit Intel Performance Benchmark Limitations.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined." Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information. The products described in this document 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. Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order. Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or by visiting Intel's Web site at www.intel.com.