

White Paper

I/O Virtualization

Intel® Virtualization
Technology for
Directed I/O

Achieving Fast, Scalable I/O for Virtualized Servers

With Intel® Virtualization Technology and the PCI-SIG*
Single Root I/O Virtualization and Sharing Specification

Introduction

IT organizations are taking advantage of virtualization to consolidate server infrastructure, reduce power, cooling and management costs, and provide simpler and more affordable solutions for high availability, load balancing and disaster recovery. Servers based on multi-core Intel® Xeon® processors with Intel® Virtualization Technology† (Intel® VT) help to magnify the benefits of virtualization, by enabling higher consolidation ratios and better application performance.

Recent enhancements to Intel VT are directed at solving the next virtualization challenge: delivering fast and scalable I/O bandwidth for virtualized servers. These new technologies can help IT organizations further increase consolidation ratios, virtualize a wider range of applications, and manage workloads more effectively. They also provide a necessary prerequisite for next-generation cloud computing models, which will ultimately deliver another major leap in data center efficiency through enhanced automation and more dynamic control of hardware and software assets.

I/O Challenges in Virtualized Servers

The cost benefits of virtualization are roughly proportional to consolidation ratios. However, as the number of virtual machines (VMs) per server increases, so does the volume and complexity of I/O traffic. This can introduce a number of challenges for IT organizations, by:

- Creating data access and networking latencies that negatively impact application performance.
- Introducing I/O bottlenecks that limit the number of VMs that can be hosted per physical server.
- Slowing-down or preventing live migration (the transfer of a running VM from one physical server to another).

To address these I/O challenges, Intel has extended Intel VT to provide hardware-based assistance for I/O virtualization processes and to complement the Single Root I/O Virtualization and Sharing (SR-IOV) specification created by the Peripheral Component Interconnect Special Interest Group* (PCI-SIG*). SR-IOV enables efficient sharing of a single I/O device among multiple VMs. Coupled with Intel® VT for Directed I/O (Intel® VT-d), it provides a foundation for efficiently utilizing I/O resources while achieving near-native I/O performance (i.e., nearly the same I/O performance as in a non-virtualized server environment). It can help IT organizations increase consolidation ratios, improve application performance, and create a more dynamic data center through fast, reliable live migration – all while reducing the cost and complexity of I/O solutions.

The Limits of Software-only I/O Virtualization

In most virtualized servers today, I/O traffic is processed and I/O resources are managed by the Virtual Machine Manager (VMM) software (Figure 1A). The VMM emulates a complete I/O hardware device for each VM. This is a very flexible approach, enabling multiple VMs to share a single I/O port with a high level of isolation. However, it limits server performance and scalability in two key ways.

Increased I/O latency. With software emulation, the VMM must process and route every data packet and interrupt. This additional processing time can negatively impact application response times.

Scalability limitations. Since software-based I/O processing consumes CPU cycles, it reduces the processing capacity available for business applications.

Near-Native Performance through Hardware-assisted Virtualization

The first step in fully optimizing I/O performance in virtualized servers is provided by Intel VT-d, which enables Direct Memory Access (DMA) between VMs and physical I/O devices (Figure 1B). Once the VMM assigns an I/O port to a VM, DMA allows the I/O stream to bypass the VMM. Intel VT-d provides memory address translation in silicon to accelerate I/O processing. It also helps to ensure that each VM accesses only its assigned memory space.

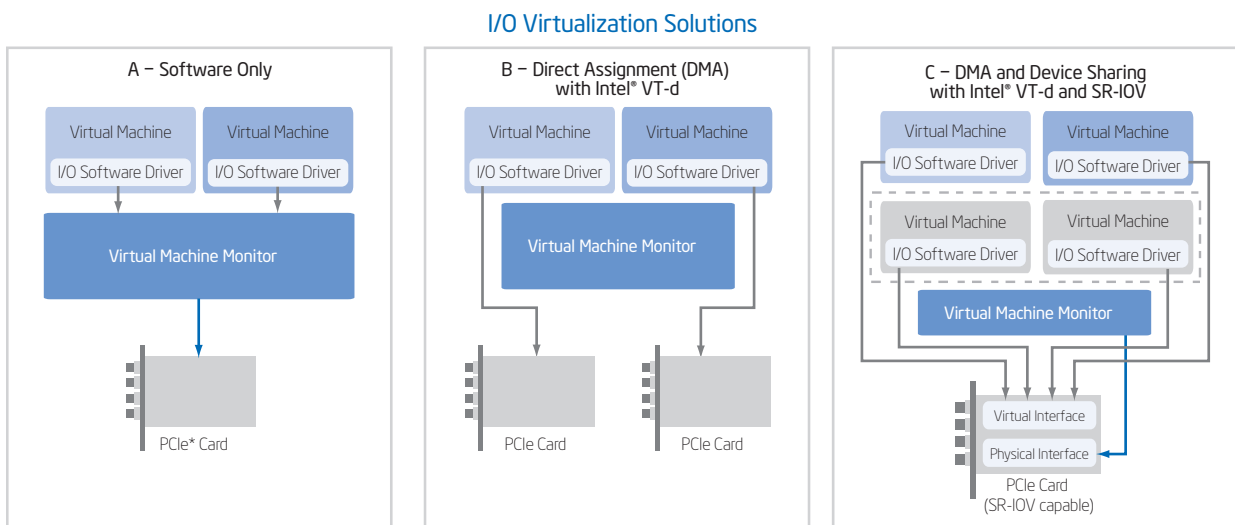


Figure 1. A: With software-only I/O virtualization, all I/O traffic must be processed and routed by the VMM, which can significantly limit I/O performance and scalability. B: Intel® Virtualization Technology for Directed I/O enables Direct Memory Access (DMA) for near-native I/O performance. C: SR-IOV, together with Intel® VT-d, shown with additional virtual machines, enables DMA and efficient device sharing to optimize I/O performance and scalability, while reducing I/O-related cost and complexity.

VMM vendors are already taking advantage of Intel VT-d to improve I/O performance.

According to performance tests by Intel engineers, these solutions can provide raw I/O performance in a virtualized server that is up to 99 percent of native I/O performance! This corresponds to an increase in raw I/O performance of roughly 6x compared with software-based I/O virtualization,¹ and can help boost total application performance by as much as 4.9x² (Figure 2). As a result, IT organizations can now successfully virtualize I/O-intensive applications, such as backup and recovery, data warehousing, Web applications, and time-sensitive online transaction processing. This can significantly extend the benefits of virtualization.

Scalability challenges remain, however, since these I/O virtualization solutions require a dedicated physical I/O port for each VM. Dual- and quad-port I/O devices can help improve utilization and reduce hardware costs, but the total number of VMs per server is still limited by the number of available I/O slots.

Solving the Scalability Challenge with SR-IOV

The next step in fully optimizing I/O in virtualized environments is to provide truly scalable high-performance, by enabling a single I/O device to provide DMA for multiple VMs. To address this need, the PCI-SIG developed the SR-IOV specification, an extension to the PCI Express* specification suite. Intel actively participated in development and is now working with leading hardware and software vendors to deliver comprehensive support.

SR-IOV provides a standards-based foundation for efficiently sharing a PCIe* card among multiple VMs. Physical I/O resources are virtualized within the PCIe card, so each card presents multiple virtual I/O interfaces (Figure 1C). A compliant PCIe card provides two function types:

- **Virtual Functions (VF)** provide all the resources necessary for data movement, along with a minimized set of configuration resources. A VM can interface directly with a VF to perform data transfer operations without VMM intervention. Each VF provides dedicated resources to its assigned VM, including an isolated memory space, a work queue, interrupts and command processing. VF functionality is compatible with Intel VT-d, which enables DMA for high-speed I/O transfers.
- **Physical Functions (PF)** provide full PCIe functionality, including the SR-IOV Extended Capability. When a PCIe card advertises its SR-IOV functionality, the VMM interfaces with a PF to configure and manage I/O resource sharing among the multiple VMs. This allows the VMM to resolve issues that impact more than one VM, such as a Guest OS request for a reset of the network interface or the addition of a new disk to a networked storage device.

By supporting Intel VT-d and the SR-IOV specification, independent hardware vendors (IHVs) can design PCIe cards that deliver near-native I/O performance for multiple VMs, while also providing memory and traffic isolation for security and high availability. IT organizations can use these cards to provide fast and scalable I/O for their virtualized servers, to improve consolidation ratios, to accelerate live migrations, and to reduce the cost and complexity of their I/O solutions.

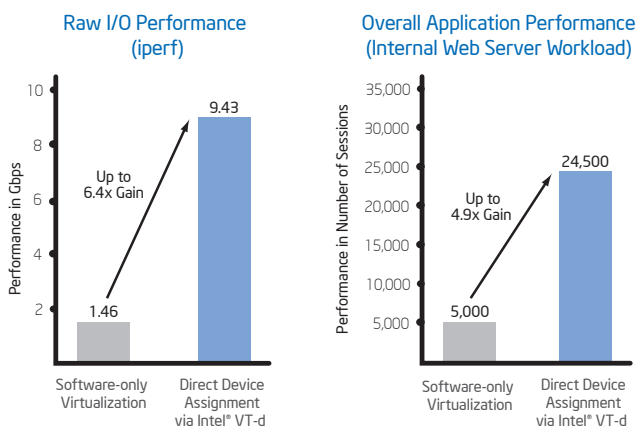


Figure 2. Intel engineers performed a number of tests to validate the performance benefits enabled by Intel® VT-d vs. software-only I/O virtualization. Results showed up to 6.4x increase in raw I/O throughput (using the iperf benchmark), and up to 4.9x improvement in overall application performance.

Enabling Live VM Migration

The use of SR-IOV and Intel VT-d optimizes I/O performance and device sharing. However, it requires that each Guest OS use an I/O software driver that is compatible with the connected I/O device. This can interfere with live VM migrations if the target server has a different I/O device.

Intel is currently working with VMM vendors to solve this challenge. One approach is to have the VMM define two paths to the same I/O resource for each VM. One path uses DMA for high performance; the other uses software emulation for broad interoperability. During a live migration, the VMM can revert to software emulation if needed to maintain I/O connectivity and avoid a failed migration.

Intel Support for SR-IOV

The computing industry is still in the early stages of SR-IOV adoption. Intel is playing a central role in enabling and coordinating the vendor community to provide optimized and widely interoperable implementations. Some solutions are available today for Intel® Xeon® processor 5500^A series-based servers using Intel® Gigabit or 10 Gigabit Ethernet Controllers (implementation requires Intel VT-d support in the server chipset and Intel VT-d and SR-IOV support in the server BIOS and VMM).

With these Intel® technology-based solutions, IT organizations can support large numbers of direct assigned VMs per network card to dramatically improve scalability and slot utilization, while also reducing I/O-related power consumption and cabling requirements. Software support for SR-IOV will continue to increase throughout 2009 and 2010, providing IT organizations with more options and broader interoperability.

Conclusion

The limited performance of software-only I/O virtualization is preventing many IT organizations from taking their virtualization solutions to the next level. The combination of Intel VT-d and the SR-IOV specification offers a solution, enabling efficient and scalable I/O device sharing with near-native performance. Complete solutions are emerging today using Intel Xeon processor 5500 series-based servers, Intel Gigabit or 10 Gigabit Ethernet Controllers, and compatible software components.

Many more solutions will reach the marketplace as software support moves into the mainstream through 2009 and 2010. By optimizing I/O performance and scalability in virtualized servers, these solutions will help IT organizations virtualize a broader range of applications, improve consolidation ratios and manage workloads more effectively. This will deliver immediate value in many environments. It will also provide a foundation for next-generation cloud computing solutions that will continue to drive up the value of virtualization.

For More Information

- Technical Guidance for implementing Intel VT-d and SR-IOV: www.intel.com/network/connectivity/solutions/vmdc.htm
- Intel® Virtualization Technology: www.intel.com/technology/virtualization/server/hardware.htm
- PCI-SIG Single-Root I/O Virtualization and Sharing Specification: www.pcisig.com/specifications/iov/review_zone

¹10Gb NIC receive performance on Iperf ver 2.0.4; 9.43 on Intel VT-d vs. 9.5 native (non-virtualized). Default setting. 8K buffer size, 4 thread, TCP window size: 28.6 MB. Intel® Xeon® processor 5500 series Server System: 2 socket NHM 2.6 GHz with 8 MB LLC Cache, C0 stepping. Enable only 2 core on each socket Hardware Prefetches OFF, Turbo mode OFF, EIST OFF. RAID bus controller: LSI Logic/Symbios Logic MegaRAID SAS 1078. Intel 5500 chipset with Intel 10 Gb XF SR NIC (82598EB). Software configuration: Hypervisor: Xen 3.4 CS18711-upsteam; Native O/S Distribution: Red Hat EL5 (2.6.18-8.el5) with kernel 2.6.27.1; Guest O/S Distribution: Red (2.6.18-8.el5) with kernel 2.6.27.1; Benchmark Stack : Rock web + JSP. Intel internal measurement July 2009.

²Intel Internal Web Server Workload. Default setting. 8K buffer size, 4 thread, TCP window size: 28.6 MB. Intel® Xeon® processor 5500 series Server System: 2 socket NHM 2.6 GHz with 8 MB LLC Cache, C0 stepping. Enable only 2 core on each socket Hardware Prefetches OFF, Turbo mode OFF, EIST OFF. RAID bus controller: LSI Logic/Symbios Logic MegaRAID SAS 1078. Intel 5500 chipset with Intel 10Gb XF SR NIC (82598EB). Software configuration: Hypervisor: Xen 3.4 CS18711-upsteam; Native O/S Distribution: Red Hat EL5 (2.6.18-8.el5) with kernel 2.6.27.1; Guest O/S Distribution: Red (2.6.18-8.el5) with kernel 2.6.27.1; Benchmark Stack : Rock web + JSP. Intel internal measurement July 2009.

³Intel® Virtualization Technology requires a computer system with an enabled Intel® processor, BIOS, virtual machine monitor (VMM) and, for some uses, certain platform software enabled for it. Functionality, performance or other benefits will vary depending on hardware and software configurations and may require a BIOS update. Software applications may not be compatible with all operating systems. Please check with your application vendor.

⁴Intel processor numbers are not a measure of performance. Processor numbers differentiate features within each processor family, not across different processor families. See www.intel.com/products/processor_number for details.

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.

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.

Copyright © 2009 Intel Corporation. All rights reserved. Intel, the Intel logo, and Xeon are trademarks of Intel Corporation in the U.S. and other countries.

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

