Technical white paper

HP SR-IOV Technology
Overview

For HP ProLiant servers and HP ProLiant server blades

Table of contents

Abstract ................................................................................................................................. 2
Overview of SR-IOV ........................................................................................................... 2
Hypervisors and I/O virtualization performance .............................................................. 4
    Emulation and paravirtualization ................................................................................ 4
    Direct assignment ......................................................................................................... 5
    Processors and I/O virtualization performance ......................................................... 6
Next generation VM solutions .......................................................................................... 6
    SR-IOV enabled network adapters ............................................................................. 6
    Server operating system recommended system configurations .............................. 8
    SR-IOV-compliant ProLiant servers ......................................................................... 8
Conclusion ....................................................................................................................... 9
Additional links ............................................................................................................. 10
Abstract

As the number of virtual machines per physical server increases, the Virtual Machine Manager (VMM) consumes an increasing portion of the available compute resources to manage network virtualization. This results in a substantial performance consequence.

As a result, there has been an industry driven effort to increase the effective hardware resource utilization using virtualization technology. This effort has produced the Single Root I/O Virtualization (SR-IOV) and Sharing Specification. This specification extends the PCI Express (PCIe) specification to enable multiple System Images (SI) to share PCI hardware resources.

SR-IOV is included in a Peripheral Component Interconnect Special Interest Group (PCI-SIG) specification for partitioning PCIe adapter bandwidth. Partitioning port bandwidth provides several benefits, including:

- Increased I/O data throughput
- Reduced CPU utilization for processing network traffic
- Reduced network path latency

The goal of the PCI-SIG SR-IOV specification is to standardize the way of bypassing the VMM’s involvement in the data path by providing independent memory space, interrupts, and direct memory access (DMA) for each virtual machine. For more information on the SR-IOV standard and industry support for the standard, go to the PCI-SIG site at pcisig.com.

This paper provides an overview of SR-IOV technology, including its benefits, how it is used, and a look ahead. It also provides an overview of the I/O virtualization technologies that HP is incorporating into HP ProLiant server designs to enable the next generation Virtual Machine (VM) solutions. For specific information on SR-IOV implementations see the papers:

- Implementing SR-IOV on HP ProLiant Servers with VMware vSphere 5.1
- Implementing Windows Server 2012 SR-IOV on HP ProLiant Servers
- Implementing SR-IOV for Red Hat Enterprise Linux on HP ProLiant Servers

Overview of SR-IOV

SR-IOV technology reduces the overhead of I/O processing and I/O latencies by allowing each VM to access a portion of the physical I/O device directly. This reduces the need for software emulation of I/O devices in memory. The SR-IOV specification defines the architecture for a new generation of I/O devices that have multiple Virtual Functions (VFs), each of which shares common properties of a PCI physical function (PF). This approach allows each VM to see its own separate virtual device while multiple VMs share the actual physical device.

In order to fully appreciate the advantages of this new class of I/O devices, let’s examine how an SR-IOV-enabled network adapter would work in a VM environment. We’ll start by looking at a block diagram of the entire VMM stack for network I/O, including the network adapter, see Figure 1.

With SR-IOV, there is still a network adapter device driver that runs in VMM space. This driver is referred to as the PF driver, since its job is to manage the global functions for the SR-IOV devices and configure shared resources. More importantly, each VM gets its own VF driver that can communicate directly with the VFs on the network adapter. Each of the VFs has its own transmit and receive buffers as well as interrupts and MAC address. This arrangement reduces the need for the VMM to create “virtualized” network adapters in system memory, which saves system resources. It also improves performance, since the VMM no longer performs the memory-to-memory copies between the virtualized network adapter and physical network adapter buffers.
Figure 1. SR-IOV PF and VF Conceptual Overview

The figure shows a hybrid environment where two of the VM network adapter functions (VM1 and VM2) are acting as virtualized network adapters through the Hypervisor. The other two VM network adapter functions (VM0 and VM3) route to their VFs on the SR-IOV network adapter directly. This direct access provides the following key benefits:

- SR-IOV network adapters and VMMs can enable more network adapter functions than are physically available.
- SR-IOV allows multiple VMs to share a single SR-IOV-capable network adapter.
- SR-IOV-capable network adapters retain the performance benefit of one PCIe device to one VM association.

The SR-IOV standard builds on the following PCI standards to extend the PCI configuration space and define access to (VFs:

- PCI-e
- Alternative Routing ID (ARI)
- Address Translation Services (ATS)
- Function Level Reset (FLR)

The Physical Function (PF) is a PCIe function of a network adapter that enables SR-IOV. The PF includes the SR-IOV Extended Capability for configuring and managing one or more VFs. Exposed as a virtual network adapter, the PF provides access to the networking I/O resources. This allows the management operating system to communicate with the external physical network. PF characteristics include:

- SR-IOV Extended Capability as part of its configuration space
- Configuration information for the associated VFs
- Base Address Registers (BAR) configuration and VF type for any associated VFs
The VF is a PCIe functional network adapter that enables SR-IOV and is associated with a PF. It represents a virtualized instance of the network adapter. Each VF has its own PCI configuration space and shares one or more physical resources on the network adapter. Resources associated with data transfer are available to the VF directly, and cannot be used by the other VFs or the PF.

By assigning a VF to each VM, multiple VMs can share a single SR-IOV PCIe network adapter that may have just one physical network port. Each partition of the PCIe network adapter appears in the PCI configuration space as a VF with its own configuration space and BAR. It shares the bandwidth of the PCIe network adapter with other PFs or VFs. Characteristics of a VF include:

- Enables native I/O virtualization
- A PCIe configuration space that defines a unique set of Message Signaled Interrupts (MSI) and MSI-eXtended (MSI-X) vectors

SR-IOV allocates multiple VFs to a PF. The VFs appear as PCI devices that are part of the physical PCI device by their resources (queues and register sets). Because SR-IOV enumerates VFs in a hardware-based PCI configuration space, it is well suited for virtualization of fixed-function devices.

Instead of using the VMM for VM processing, SR-IOV works with I/O memory management unit (IOMMU) technology to address vSwitch performance by avoiding the vSwitch; using dedicated hardware queues and a round-robin scheduler to reduce processor overhead. To accomplish this, the PCI-SIG SR-IOV specification defines independent memory space, interrupts, and DMA for each VM. Processor (such as AMD-v and Intel® VT-x) and chipset virtualization (such as AMD-Vi and VT-d) permit a VM and Guest OS to share a PCIe network adapter allowing the VMM to assign one or more VFs to a VM.

### Hypervisors and I/O virtualization performance

With the growth of VM technology, the servers that manage virtual machines, known as Hypervisors, also continue to evolve. The hardware industry continues to look for better ways to enable VMM software to improve efficiency, reliability, and security.

One of the core technical challenges that all VMM software must address is making the virtualization of I/O devices more efficient. The goal is to have each individual VM on the host server access the physical I/O devices of the underlying server hardware directly. Software services can use the following methods to service virtual I/O:

- Emulation and paravirtualization
- Direct assignment

### Emulation and paravirtualization

Emulation and paravirtualization use a software layer, or device model, that emulates each I/O device within a VM. Guest OSs in each VM communicate with this device model and not directly with the physical I/O device in the server, see Figure 2. The VMM communicates with the physical I/O device. For each virtual machine request, the VMM translates the request into the actual I/O operations on the physical device, and then manages and schedules the requests.
Both emulation and paravirtualization do a good job of insulating the various VMs from the server hardware as well as from each other. This introduces extra layers of software emulation (the device models), translation, and arbitration into the system. The consequences of this are significantly reduced performance and an increase in I/O latency and CPU utilization.

**Direct assignment**

Direct assignment reduces the need for a device emulation layer in software. A direct assignment designates a physical I/O device to a given VM. The designated VM (and only that VM) performs I/O operations directly to the device, see Figure 3. The result is significantly improved I/O performance in VM environments. However, with Direct Assignment, you cannot migrate VMs because the VM is tied directly to the hardware. If the HW were to be removed, the VM would stop functioning.
With direct assignment, the guest VM sees a virtualized subset of system memory while the devices themselves communicate directly with the host physical memory addresses when transferring data. This can affect I/O devices that use DMA. VMM addresses this issue by employing software mechanisms to perform the translation between the guest VM and the real physical addresses.

In order to address these types of issues regarding I/O virtualization, Intel and AMD have announced enabling technologies for I/O virtualization in VMM environments.

**Processors and I/O virtualization performance**

VMMs and OSs can utilize DMA virtualization to improve performance and I/O virtualization security. The following processor technologies enable DMA virtualization:

- Intel Virtualization Technology for Directed I/O (VT-d)
- AMD I/O Virtualization Technology (IOMMU)

These technologies use DMA virtualization to solve two important challenges of I/O virtualization under direct assignment: DMA and interrupt virtualization.

DMA virtualization enables the VMM (or the OS) to create multiple memory protection domains for DMA accesses. Each protection domain represents a separate environment. The VMM assigns a subset of the host physical memory to a protection domain. It can then restrict access to a given protection domain to only the assigned I/O device. This isolates the DMA. Because the DMA is isolated, the VMs operate in a more secure environment. DMA virtualization also performs address translation, which consists of the following functions:

- Translating (or remapping) the virtualized memory addresses in DMA requests from the I/O devices into the actual host physical address
- Caching the most frequently used remapping entries

By allowing VMM software to offload the execution and management of address translations to the processor, this functionality improves performance significantly.

**Next generation VM solutions**

HP is enabling the next generation of VM solutions.

**SR-IOV enabled network adapters**

SR-IOV technology utilizes a vSwitch in the network adapter hardware. The vSwitch is a virtual Ethernet switch implemented in a virtualized server environment. It is anything that mimics a traditional layer 2 (L2) switch or bridge for connecting VMs. vSwitches can communicate between VMs on a single physical server, or they can connect VMs to the external network. Moving vSwitch functionality from the VMM (hypervisor) to the hardware reduces the performance issues associated with vSwitches. Common implementations are software-based vSwitches built into hypervisors. However, vendors can use the PCIe SR-IOV standard to build hardware-based vSwitches in network adapters.

The SR-IOV standard provides hardware accelerated I/O virtualization for shared PCIe devices on a single physical server. Today, the typical SR-IOV-enabled network adapter provisions 1 to 16 physical functions (full-featured PCI functions) and 32 to 256 virtual functions (lightweight PCI functions focused primarily on data movement). However, the SR-IOV standard allows for thousands of virtual functions in a device, providing headroom for future capabilities.

SR-IOV-enabled network adapters let the virtual network throughput bypass the hypervisor vSwitch by exposing the virtual network adapter functions to the guest OS directly. Thus, the network adapter reduces latency between the VM to the external port significantly. The hypervisor continues to allocate resources and handle exception conditions, but it doesn’t need to perform routine data processing for traffic between the VMs and the network adapter. Figure 4 illustrates the traffic flow of an SR-IOV-enabled network adapter.
Figure 4. In a vSwitch implemented as an SR-IOV network adapter, traffic can switch locally inside the vSwitch (gray line) or go directly to the external network (blue line).

The benefits to deploying vSwitches as hardware-based SR-IOV-enabled network adapters include:

- Reduction of CPU and memory usage compared to software-based vSwitches. With direct I/O, vSwitches are no longer part of the data path.
- Removes potential bottlenecks in the hypervisor, freeing it up to perform other functions such as, VM migration.
- Facilitating up to 256 functions in a typical low-cost network adapter. It significantly increases the number of virtual networking functions for a single physical server.
For example, our preliminary testing shows that when compared to traditional network adapters, SR-IOV-enabled network adapters reduce processor usage by about 10% to 15%. Other vendors report similar performance gains by moving to SR-IOV-enabled network adapters. For a list of supported network adapters, see Table 1.

### Table 1. SR-IOV compliant network adapters

<table>
<thead>
<tr>
<th>Network adapter</th>
<th>Supported on SR-IOV compliant server</th>
</tr>
</thead>
<tbody>
<tr>
<td>HP Ethernet 10Gb 2-port 560FLB Adapter</td>
<td>ProLiant BLGen8 servers</td>
</tr>
<tr>
<td>HP Ethernet 10Gb 2-port 560SFP+ Adapter</td>
<td>ProLiant DL, ML, or SL (G7 or Gen8 servers)</td>
</tr>
<tr>
<td>HP Ethernet 10Gb 2-port 560M Adapter</td>
<td>ProLiant BLGen8 servers</td>
</tr>
<tr>
<td>HP Ethernet 10Gb 2-port 560FLR-SFP+ Adapter</td>
<td>Select ProLiant DL or SL Gen8 servers</td>
</tr>
<tr>
<td>HP Flex-10 10Gb 2-port 530FLB Adapter</td>
<td>ProLiant BLGen8 servers</td>
</tr>
<tr>
<td>HP Flex-10 10Gb 2-port 530M Adapter</td>
<td>ProLiant BLGen8 servers</td>
</tr>
<tr>
<td>HP Ethernet 10Gb 2-port 530SFP+ Adapter</td>
<td>ProLiant DL, ML, or SL (G7 or Gen8 servers)</td>
</tr>
<tr>
<td>HP Ethernet 10Gb 2-port 530FLR-SFP+ Adapter</td>
<td>Select ProLiant DL or SL Gen8 servers</td>
</tr>
<tr>
<td>HP Ethernet 10Gb 2-port 530T Adapter</td>
<td>ProLiant DL, ML, or SL (G7 or Gen8 servers)</td>
</tr>
</tbody>
</table>

### Server operating system recommended system configurations

The PCI subsystem and the server kernel contain the core SR-IOV implementation, so both must be SR-IOV capable. In addition, the PCIe network adapter driver must support VFs and PFs.

Processor virtualization (such as AMD-v and Intel VT-x) and chipset virtualization (such as AMD-Vi and Intel VT-d) allow a VM and Guest OS to share a SR-IOV-capable PCIe network adapter. This allows the VMM to assign one or more VFs to each VM.

The following lists supported operating systems.

- Microsoft® Windows® Server Hyper-V 2012
- Red Hat Enterprise Linux 6.2 KVM (x64)
- Red Hat Enterprise Linux 6.3 KVM (x64)
- VMware ESXi 5.1

See the “SR-IOV-compliant ProLiant servers” section for details on our servers that include these processors.

Not all of the supported OSs enable SR-IOV for all of the ProLiant servers. For specific information on SR-IOV implementations see the papers:

- Implementing SR-IOV on HP ProLiant Servers with VMware vSphere 5.1
- Implementing Windows Server 2012 SR-IOV on HP ProLiant Servers
- Implementing SR-IOV for Red Hat Enterprise Linux on HP ProLiant Servers

### SR-IOV-compliant ProLiant servers

Table 2 lists the HP ProLiant servers and minimum System ROM version that are required for SR-IOV support.

#### Important

SR-IOV requires both hardware and System ROM compliance to work. Your server must have the minimum System ROM version listed in the table to enable SR-IOV. If necessary, update the System ROM before enabling SR-IOV.

You can download System ROM updates from the SPP, which is available at [hp.com/go/spp/download](http://hp.com/go/spp/download) or the HP Support Center at [hp.com/go/support](http://hp.com/go/support).
Table 2. Supported ProLiant servers and ROM versions

<table>
<thead>
<tr>
<th>Server platform</th>
<th>System ROM family</th>
<th>System ROM Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>ProLiant BL servers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BL420c Gen8</td>
<td>I30</td>
<td>12/13/2012 or later</td>
</tr>
<tr>
<td>BL460c Gen8</td>
<td>I31</td>
<td>12/11/2012 or later</td>
</tr>
<tr>
<td>BL660c Gen8</td>
<td>I32</td>
<td>12/11/2012 or later</td>
</tr>
<tr>
<td>ProLiant DL servers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DL360e Gen8</td>
<td>P73</td>
<td>12/11/2012 or later</td>
</tr>
<tr>
<td>DL360p Gen8</td>
<td>P71</td>
<td>12/4/2012 or later</td>
</tr>
<tr>
<td>DL380e Gen8</td>
<td>P73</td>
<td>12/11/2012 or later</td>
</tr>
<tr>
<td>DL380p Gen8</td>
<td>P70</td>
<td>12/12/2012 or later</td>
</tr>
<tr>
<td>DL385p Gen8</td>
<td>A28</td>
<td>8/14/2012 or later</td>
</tr>
<tr>
<td>DL560 Gen8</td>
<td>P77</td>
<td>12/5/2012 or later</td>
</tr>
<tr>
<td>DL580 G7</td>
<td>P65</td>
<td>11/8/2012 or later</td>
</tr>
<tr>
<td>DL585 G7</td>
<td>A16</td>
<td>8/14/2012 or later</td>
</tr>
<tr>
<td>ProLiant ML servers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ML350e Gen8</td>
<td>J02</td>
<td>12/13/2012 or later</td>
</tr>
<tr>
<td>ML350p Gen8</td>
<td>P72</td>
<td>12/7/2012 or later</td>
</tr>
<tr>
<td>ProLiant SL servers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SL230s Gen8</td>
<td>P75</td>
<td>12/11/2012 or later</td>
</tr>
<tr>
<td>SL250s Gen8</td>
<td>P75</td>
<td>12/11/2012 or later</td>
</tr>
<tr>
<td>SL270s Gen8</td>
<td>P75</td>
<td>12/11/2012 or later</td>
</tr>
</tbody>
</table>

Note
This list was comprehensive as of publication. For the most up-to-date version of the paper, use the link on the first page of this document.

Conclusion

SR-IOV data handling operates in a native and direct manner, reducing processing overhead, and enabling highly scalable PCI functionality. SR-IOV’s ability to scale is a major advantage.

Another SR-IOV advantage is the prospect of performance gains achieved by removing the hypervisor from the main data path. Hypervisors are essentially another operating system providing CPU, memory, and I/O virtualization capabilities to accomplish resource management and data processing functions.

Since SR-IOV is a hardware I/O implementation, it also uses hardware-based security and quality of service (QoS) features incorporated into the physical host server.
Additional links

ProLiant technology papers
hp.com/servers/technology

HP VMware Home Page
hp.com/go/vmware

HP Custom ESXi Images
hp.com/go/esxidownload

HP ProLiant Server and Option Support Recipe for VMware
vibsdepot.hp.com

HP Windows Server 2012 Home Page
hp.com/go/ws2012

Microsoft Windows Server 2012 Home Page
microsoft.com/en-us/server-cloud/windows-server/default.aspx

HP Service Pack for ProLiant
hp.com/go/spp

HP Ethernet 10Gb 2-port 560FLB Adapter
hp.com/products/servers/networking/560FLB/index.html

HP Ethernet 10Gb 2-port 560SFP+ Adapter
hp.com/products/servers/networking/560SFP/index.html

HP Ethernet 10Gb 2-port 560M Adapter
hp.com/products/servers/networking/560M/index.html

HP Ethernet 10Gb 2-port 560FLR-SFP+ Adapter
hp.com/products/servers/networking/560FLR/index.html

© Copyright 2013 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice. The only warranties for HP products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. HP shall not be liable for technical or editorial errors or omissions contained herein.

AMD is a trademark of Advanced Micro Devices, Inc. Intel is a trademark of Intel Corporation in the U.S. and other countries. Microsoft and Windows are U.S. registered trademarks of Microsoft Corporation.

4AA4-7189ENW, June 2013