

MSI and MSI-X Deliver Performance Gains Through Emulex and Microsoft Partnership

Al Gamarra
Emulex Corporation

July 7, 2008

Abstract

Emulex Corporation is a Microsoft Gold Certified Partner. This prestigious level of partnership ensures that next generation Storage Area Networks (SAN) in a Windows environment with Emulex Host Bus Adapters (HBA) can take advantage of new software and hardware product features that improve overall performance. One such feature is Message-Signaled Interrupts (MSI) and its extended version, MSI-X. This white paper provides details on how Microsoft and Emulex support MSI and MSI-X to help you enhance overall performance, reduce system overhead, lower interrupt latency, improve host CPU utilization, and increase system reliability.

Products



EMULEX[®] Emulex LightPulse[®] Fibre Channel
Host Bus Adapters



Operating Systems
Microsoft Windows Vista[®]
Microsoft Windows Server[®] 2008

Contents

Overview	3
PCI-SIG.....	3
Figure 1 - Evolution of MSI and MSI-X.....	4
History of Interrupts.....	4
MSI Overview.....	4
Table 1 – Interrupt Handling Differences	5
System Enhancements with MSI	5
Dynamic Interrupt Redirection.....	6
NUMA and Directed I/O Support.....	6
Autoclear.....	6
TPC-E Benchmark.....	7
MSI/MSI-X System Requirements	7
Conclusion	7

Author's Disclaimer and Copyright: Reasonable efforts have been made to ensure the validity and accuracy of the contents of this whitepaper. Emulex Corporation and any contributors to this whitepaper, are not liable for any error in this published white paper or the results thereof. Emulex and its contributors specifically disclaim any warranty, expressed or implied, relating to the white-paper and their accuracy, analysis, completeness or quality.

This document refers to various companies and products by their trade names. In most, if not all cases, their respective companies claim these designations as trademarks or registered trademarks. This information is provided for reference only. This report is the property of Emulex Corporation and may not be duplicated without permission from the Company.

Overview

Emulex Corporation and Microsoft have a long history of working together to develop and deliver SAN solutions targeted at many users. As a Microsoft Gold certified partner and member of the Microsoft Interop Vendor Alliance, Emulex delivers robust SAN solutions that are integrated and interoperable with Microsoft technologies. Microsoft and Emulex solutions address the full range of the networked storage market (from enterprise data center environments to the emerging small-to medium business SAN market) with a common vision of simplifying the deployment and management of SANs, while increasing their performance and availability.

Both Microsoft and Emulex are longtime members of PCI-SIG, the organization that sets and maintains the Peripheral Component Interconnect (PCI) standard. Message-Signaled Interrupts (MSI) and its extended version, MSI-X, are interrupt-handling mechanisms included in the PCI standard that replace and enhance traditional line-based interrupts. Instead of using a dedicated pin to trigger interrupts, devices that use MSI trigger an interrupt by writing a value to a particular memory address.

As interrupts have evolved from pin-based to MSI, and then to MSI-X, both Microsoft and Emulex have been at the forefront of delivering solutions that embrace these open standards and improve overall system performance. External published benchmark testing has demonstrated a 14 percent increase in overall system performance in MSI-X supported systems configured with Emulex LightPulse Fibre Channel HBAs and Windows Server 2008. Online transaction processing and database environments—such as electronic banking, order processing and e-commerce—benefit from enhanced I/O scalability and improved performance for multiprocessor and multi-core systems including those with Non-Uniform Memory Access (NUMA) architectures by using Emulex HBAs and Microsoft operating systems with MSI-X support.

This white paper provides details on MSI-X, its predecessor MSI, and how these interrupt-handling mechanisms, when supported by Emulex HBAs and Microsoft operating systems, enhance overall performance, reduce system overhead, lower interrupt latency, improve host CPU utilization, and increase system reliability.

PCI-SIG

The PCI bus was created for I/O devices that have high bandwidth requirements. The PCI architecture is the most common method to extend systems by using plug-in adapters, such as Emulex's Fibre Channel HBAs. Formed in 1992, PCI-SIG is the industry organization chartered to develop and manage the PCI standard. PCI-SIG has over 900 members who effectively maintain and enhance the PCI specifications which deliver I/O functionality for computers ranging from servers to workstations, PCs, laptop PCs and mobile devices. The PCI bus has evolved into three basic standards that the PCI-SIG identifies as PCI Conventional, PCI-X, and PCI Express (PCIe). Go to pcisig.com for more information on the PCI standard.

Of the many important features coming out of the PCI standard, one is MSI. Figure 1 shows the evolution of MSI from its inception to the extended version, MSI-X, available today.

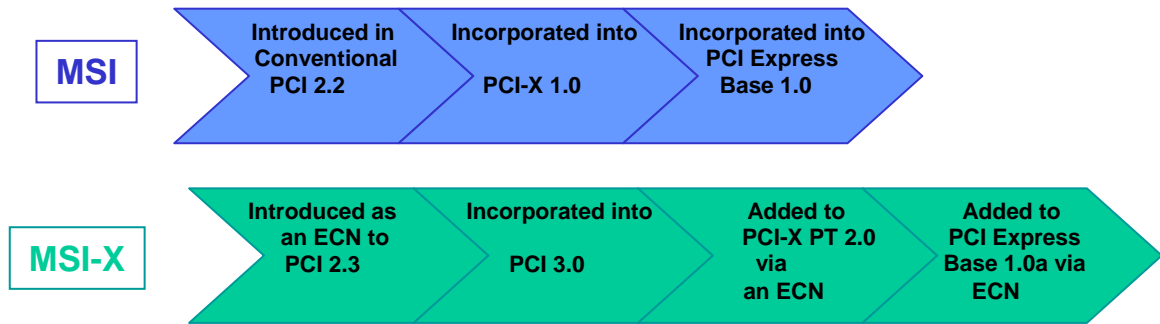


Figure 1 - Evolution of MSI and MSI-X

History of Interrupts

An interrupt is a hardware signal from a device to a CPU, informing the CPU that the device needs attention and signaling that the CPU should stop current processing and respond to the device. If the CPU is performing a task that has lower priority than the priority of the interrupt, the CPU suspends its current thread. The CPU then invokes the interrupt handler for the device that sent the interrupt signal. The interrupt handler services the device, and when the interrupt handler returns, the CPU resumes the processing it was doing before the interrupt occurred.

Interrupts in old bus technologies are referred to as “legacy” or “line-based” interrupts. With these technologies, interrupts are signaled by using one or more external pins that are wired separately from the main lines of the bus, a process known as “out of band.” Newer bus technologies, such as PCIe, maintain software compatibility by emulating legacy interrupts through in-band mechanisms. These emulated interrupts are treated as legacy interrupts by the host operating system. Line-based interrupts, as defined by the PCI standard, were limited to only four lines and, due to multiple devices, were often shared causing increased latencies.

MSI Overview

Following the PCI specifications, Emulex and Microsoft support MSI with their Fibre Channel HBAs and operating systems, respectively. An MSI is generated when the Emulex HBA delivers a “message” over the PCIe fabric to a specific system memory address. Windows provides the message content and the address to which the message is delivered to the HBA driver to identify the interrupt source. MSI supports up to 32 interrupt sources per device. Although this provides more flexibility than line-based interrupts, it can become a limitation on servers with a large number of processor cores. To overcome this limitation, MSI-X supports 2048 interrupt sources per device, enabling a more flexible interrupt model. Table 1 shows the differences between the three types of interrupt handling mechanisms.

Line-based Interrupt	<ul style="list-style-type: none"> • The number of discrete interrupts that are available on many systems is limited, which typically results in PCI devices sharing an interrupt with other devices. • Each PCI device can have up to four interrupts. • There is little control to either driver writers or system managers over which specific processor is chosen as an interrupt target.
MSI	<ul style="list-style-type: none"> • A PCI function can request up to 32 MSI messages. • They are in-band messages, instead of using side-band pins, and can target addresses via the host bridge. • They can send a small amount of data along with the interrupt message • They are not shared, thus an MSI that is assigned to a device is guaranteed to be unique within the system. • There is control over which processor is chosen as the interrupt target.
MSI-X	<ul style="list-style-type: none"> • A PCI function can request up to 2048 MSI messages rather than 32 messages. • There is support for an independent message address and message data for each message. • There is support for per-vector masking. • There is control over which processor is chosen as the interrupt target.

Table 1 – Interrupt Handling Differences

System Enhancements with MSI

With MSI, Emulex HBAs do not have to share interrupts with other devices because the number of potentially available messages is limited only to the number of interrupt dispatch table entries that are available on a system. This can result in lower interrupt latency and higher system reliability. Interrupt processing overhead, as measured on a system-wide basis is also lower. In addition, using MSI or MSI-X avoids problems that can occur when a device is forced to share an interrupt with a driver that has design or coding defects in its interrupt service routine.

Windows Server 2008 and Vista introduced IoConnectInterruptEx, a replacement of the MSI device-driver interface, to support MSI-X. When calling IoConnectInterruptEx, fewer parameters are required, thus improving system reliability. Specifically, interrupt resource information, such as vector, interrupt request level, affinity, and whether the interrupt is edge or level triggered, is not required, removing the potential for errors.

Dynamic Interrupt Redirection

With Dynamic Interrupt Redirection (DIR), Emulex and Microsoft continue to display their leadership in deploying performance enhancing features. MSI-X enables software to more flexibly target an I/O interrupt to a specific processor *dynamically* for each I/O request. Support of DIR requires additional information to be exchanged between Windows operating system and kernel-mode drivers. This information informs the driver which processor core should receive the I/O interrupt when the I/O is completed. The driver provides this information to the Emulex PCIe HBA, allowing dynamic interrupt redirection on each I/O request that is generated, thereby improving system performance while lowering interrupt latency and system overhead.

NUMA and Directed I/O Support

Non-Uniform Memory Access (NUMA) based servers were designed in the 1990s to overcome memory bottlenecks. Supercomputers of the 1980s and 90s focused on providing high-speed memory access as opposed to faster processors, allowing them to work on large data sets at very high speeds. Limiting the number of memory accesses provided the key to extracting high performance. But this is not always practical. Multiprocessor systems can make the problem worse since having memory equidistant from all processors resulted in very long latencies. NUMA addresses this problem by providing separate memory for groups of processors, i.e., a node, minimizing the performance degradation due to long latencies.

NUMA-based servers demonstrate improved system performance if the instructions and data in memory are close to the processor that will use them. Optimized I/O handling occurs with MSI-X capabilities of the system and I/O adapters such as Emulex Fibre Channel HBAs. For example, in Windows Server 2008, Emulex utilizes its MSI-X support to direct the interrupt processing back to the processor or NUMA node that initiated the I/O processing. This ability to dynamically create affinity between the CPU and interrupts lowers interrupt latency and takes advantage of cache locality thereby improving overall system performance.

Autoclear

In 2008, the Emulex LPe12000 8Gb/s Fibre Channel HBA was delivered with support for MSI-X, offering MSI-X interrupt resources and introducing the ability to target messages to a specific set of processors in multiprocessor systems. As a pioneer in this field, Emulex also developed a patent-pending enhancement that further reduces systems overhead and improves host CPU utilization in MSI-enabled systems.

Emulex developed an “autoclear” mechanism incorporated in their Fibre Channel controller ASIC that reduces CPU processing time by eliminating the need for drivers to clear attention conditions to the HBAs during the processing of interrupts while still ensuring that no interrupts are lost. The elimination of these commands greatly reduces PCIe bus traffic and increases host CPU utilization.

TPC-E Benchmark

An external benchmark completed by IBM Systems for TPC-E, featuring the technologies listed below, produced a 14 percent increase in the total number of transactions performed per second over a similar test conducted in 2007:

- Microsoft SQL Server 2008 and Windows Server 2008
- Emulex LPe12000 8Gb/s HBAs with MSI-X support
- IBM System x3850 M2 server

For more information on this benchmark, go to:

www.emulex.com/white/hba/08-882-white-TPC-E-1.pdf.

High transaction applications, such as Microsoft SQL Server and Exchange Server, benefit from enhanced overall system performance and I/O scalability by using Emulex HBAs with MSI-X support. To take full advantage of the Windows operating system device vendors should work closely with Microsoft to optimize the driver. For example, the TPC-E benchmark results relied on Emulex's Storport Miniport driver with DIR to achieve the improved transaction performance.

MSI/MSI-X System Requirements

For an Emulex HBA to take advantage of MSI and/or MSI-X in a Windows environment, there are specific system requirements. Table 2 provides a summary of these requirements.

	MSI	MSI-X
Operating Systems	Windows Vista, Windows Server 2008	Windows Vista, Windows Server 2008
Emulex Driver	Storport Miniport Driver Version 2.00a12 (or greater)	Storport Miniport Driver Version 2.01 (or greater)
Emulex HBAs	LP1000x, LP1100x, LPe1100x, LPe1200x	LPe1200x

Table 2 – MSI and MSI-X Windows Requirements

Conclusion

Emulex Fibre Channel HBAs and Microsoft Windows have always supported line interrupts. MSI is supported across Emulex's full HBA product line beginning with the LP10000 2Gb/s Fibre Channel PCI-X HBA family in Windows environments. With the LightPulse LPe12000 8Gb/s Fibre Channel PCIe HBA and Windows Vista and Windows Server 2008, Emulex and Microsoft now support MSI-X which offers extended interrupt resources and introduces the ability to target messages to a specific set of processors in multiprocessor systems.

Table 3 summarizes the many benefits delivered by Emulex and Microsoft as a result of their strong partnership in developing and delivering support for the MSI and MSI-X features of the PCI standard.

System Performance	Enhanced by DIR and Emulex autoclear innovation
System Overhead	Reduced through DIR, autoclear, no interrupt sharing across devices
Interrupt Latency	Lowered with DIR, no interrupt sharing across devices
Host CPU Utilization	Improved with autoclear and DIR
System Reliability	Increased with IoConnectInterruptEx, autoclear, no interrupt sharing across devices

Table 3 – Summary of MSI and MSI-X Benefits

Emulex and Microsoft will continue to work together to define, develop and deliver features that offer many benefits across their wide customer base, from data center users to small-medium businesses, deploying SAN solutions. As leaders and innovators in their respective technologies, Microsoft and Emulex embrace a common vision of simplifying the deployment and management of SANs while increasing their performance and availability.