

An Optimal Approach: Emulex Corporation's Implementation of a Windows Server 2003 Storport Miniport Driver

"Emulex has worked in close partnership with Microsoft to deliver a comprehensive implementation of the Microsoft Storport driver that brings new high value to Windows Server 2003-based SANs."

—**Claude Lorenson, PhD**
Technical Product Manager of
Storage Technologies | **Microsoft**

Abstract

The high-performance, high-availability storage capabilities of hardware-based disk arrays and storage area networks (SANs) may not be fully realized if the host operating system is not designed to fully exploit such technologies. Unlike the original SCSIport driver, the Storport driver—released with the Windows Server 2003 platform—is designed to match the high performance capabilities of SANs and hardware-based disk arrays. When implemented in combination with the Emulex Storport miniport driver—designed from the ground up to deliver the full range of Storport features—SAN performance and manageability are optimized.

This paper outlines the improvements that Storport offers over the preexisting Windows driver, and details the Emulex implementation of a Storport miniport driver. The Emulex miniport implementation delivers a distinctly enhanced level of functionality and stability to storage devices, even in highly-complex SAN fabric configurations. The combination of Microsoft's Storport port driver and the Emulex Storport miniport driver provides customers with high-performance, efficient and flexible storage solutions.

Introduction

Storage arrays are aggregates of disks that together provide higher performance and better availability than standalone disks. Disk management of such RAID arrays is either software-based (under host control) or hardware-based (managed by a RAID controller). Disk arrays can be directly attached to the host or on a SAN, a dedicated high-speed storage infrastructure (most commonly Fibre Channel) that connects servers to a shared pool of storage resources.

The host bus adapter (HBA) is a critical component in connecting storage arrays to the host, especially in SAN configurations. Residing in the server, the HBA contains a powerful processor facilitating high-speed data transfers between the server system memory and external storage devices. In this respect, the HBA is both a physical and logical bridge between the host and storage. Unlike network interface cards (NICs), HBAs provide significant CPU-offload capacity, enabling I/O transactions to be managed with minimal expenditure of server CPU resources, helping to ensure that application performance is not reduced.

Although storage components such as HBAs are optimized for high performance I/O processing, the full range of performance benefits may not be realized if complementary components of the host operating system (the I/O subsystem) are not themselves designed for hardware RAID or for SAN fabric technologies. Microsoft has been redesigning the core Windows operating system as SAN technologies have evolved, and one critical aspect of SAN compatibility is the release of the Storport port driver with Windows Server 2003. The Storport device driver is specifically designed to meet the performance and management needs of a high number of networked storage devices; capabilities that the preexisting port driver (SCSIport) does not provide.

Since Storport's release in 2003, Emulex has partnered with Microsoft to deliver the first full Storport miniport solution designed specifically for OEMs and business customers who demand high performance and manageability from their Windows platform storage solutions. The Emulex Storport miniport is a particularly robust solution for Fibre Channel SANs, as Emulex developed it from the ground up to take advantage of the full range of Microsoft's Storport functionality.

SCSI Drivers: A Comparison of SCSIport and Storport

Data destined for storage is typically in block format. The SCSI (small computer system interface) protocol is used with block level storage devices on both parallel and serial SCSI interconnects. The details of preparing data for storage are under the control of device drivers.

Port and Miniport Device Drivers

One of the most common connections between servers and the various classes of storage devices (disk, tape, etc) is the parallel SCSI interconnect used with direct attached storage. Communication across the interconnect is provided by host device drivers. The transfer of data and I/O requests to and from SCSI devices is under control of class-specific device drivers in the host operating system, acting in concert with adapter-specific firmware and drivers.

Port and miniport driver roles in the I/O communication process can be summarized as follows (see Figure 1):

- 1) The application sends I/O requests through the host operating system.
- 2) The SCSI device driver in the host translates application I/O requests and data into SCSI request blocks (SRBs).
- 3) The SRBs are passed down to the vendor's Miniport driver.
- 4) The device driver passes the I/O operation to the HBA.
- 5) Once at the HBA, the adapter processes the I/O operation, sends the operation to the device, and then returns the results of the operation to the host.

Because the SCSIport port driver has been so successful in parallel SCSI environments, and because Fibre Channel developers have been able to use it in storage networks, it is not immediately apparent why SCSIport performance is less than adequate in Fibre Channel environments. The next section details some of the limitations of the legacy SCSIport driver in a Fibre Channel environment.

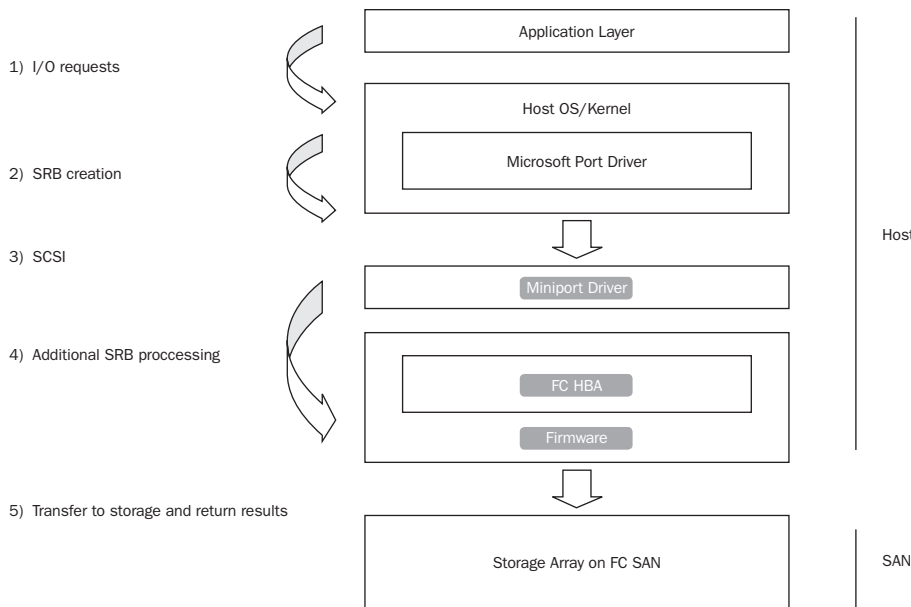


Figure 1: The path of the application I/O from host to storage device. Gray boxes indicate vendor-supplied components

SCSIport

Until Windows Server 2003, the SCSIport driver was the standard device driver used to transport SCSI commands between server and storage devices. While the SCSIport driver performs well in direct attached storage environments that support relatively few devices, it has several critical limitations when used in complex SAN environments designed to support hundreds or even thousands of devices.

- **Error Handling & Bus Resets.** In the SCSIport driver model, if an I/O operation encounters a command timeout, then the miniport driver is directed to perform a bus reset operation. In complex SAN fabrics, a bus reset—which can impact hundreds or thousands of attached devices—can be extremely time consuming and highly disruptive to operations, essentially

compromising the goal of highly available data. Tape backup operations can fail due to a bus reset being issued for an I/O timeout on a disk device.

- **Limited Queue Control.** SCSIport queues I/O requests to the HBA to which the devices are connected. It does not, however, provide a mechanism for HBA miniport drivers to control how I/O is queued to their devices. Such control was not necessary with direct attached storage. In SAN environments where devices are added and removed from the network fairly regularly, I/O queues must be paused and resumed without accumulation of errors.
- **Performance.** SCSIport was designed to meet the I/O transfer needs of a relatively limited number of devices attached to the parallel SCSI bus. Increasing the device load on SCSIport in a SAN environment will not deliver the expected performance, for a number of driver architectural reasons discussed below.
- **Adapter I/O Limit.** SCSIport supports a maximum of 254 outstanding I/O requests per HBA, regardless of the number of targeted storage devices. In Fibre Channel SAN configurations, which commonly support hundreds of targets and potentially thousands of LUNs, this is a serious bottleneck in I/O processing.
- **Sequential I/O Processing.** SCSIport cannot support the simultaneous issuing and completion of I/O requests. The impact of this limitation is seen in systems with multiple processors (quite common in high performance SANs), since their ability to simultaneously start and complete commands cannot be exploited.

Storport Port Driver

Microsoft developed Storport specifically to meet the specialized needs of hardware RAID and SAN environments. The release of Storport provided Emulex an opportunity to partner with Microsoft to develop a complementary Storport miniport driver designed to enable the HBA to confer greater manageability, control, and higher performance capabilities to the attached storage devices.

Table 1 presents a comparison of the capabilities of the SCSIport driver and the new Storport driver. Because many of these functions require support of the miniport driver to deliver a complete solution, a number of the Storport capabilities listed in the table may not in fact be implemented if the miniport driver has not been designed to

support such functionality. As is detailed in the next section, the Emulex Storport miniport driver, because of its unique design, supports full Storport functionality. (Note that a discussion of the details and significance of the Storport changes is deferred until the section, *Emulex Storport Miniport Solution*.)

Emulex Miniport Design Approach

Microsoft designed their Storport port driver to enable miniport developers to do minimal modification to existing SCSIport miniport drivers or to develop a completely new Storport miniport driver.

Ground-up Design

Emulex did not take the approach of modifying the existing SCSIport miniport code to fit the new Storport port driver. While such an approach has the potential advantage of requiring fewer development cycles, Emulex recognized that it would be difficult to capture the full range of Storport present and future functionality by simply back-porting the existing SCSIport miniport driver to Storport.

Two of the biggest challenges to miniport driver compatibility were accommodating Storport's new approach to handling I/O request processing and ensuring that error conditions during I/O operations on the SAN are resolved efficiently and successfully. This required a far more sophisticated approach to driver design than simply adding new code to the existing driver. By writing a new miniport driver designed to take full advantage of Storport's features and capabilities, Emulex has produced a Fibre Channel HBA driver that is robust and flexible.

Variable	SCSIport	Storport	Significance of Storport Change
Number of outstanding I/Os per HBA	Maximum of 254, irrespective of the number of attached devices	254 per LUN x number of LUNs	I/O throughput scalability is no longer limited by the driver; throughput limited by HBA capabilities. Management of I/O queues is now critical.
I/O processing	Sequential (half-duplex); parallel processing cannot be exploited	Simultaneous (full duplex), if implemented in miniport driver	I/O start and completion routines are uncoupled, increasing I/O throughput; power of multiple processors can be exploited.
Command timeout error handling	Error response at level of bus; whole bus resets	Hierarchical error handling: resets at level of LUN, target and bus, if implemented in miniport driver	Error response time dramatically lowered since typically only the affected LUN requires resetting; SAN fabric disruptions are minimized. A major improvement in cluster failover behavior results from this capability.
Queue management	No miniport queue control	Queue management under miniport control, if implemented in miniport driver	Under high load conditions, miniport driver can limit I/O requests to a backlogged device, keeping I/O flow in progress and preventing other devices from stopping.

Table 1: A comparison of SCSIport and Storport capabilities

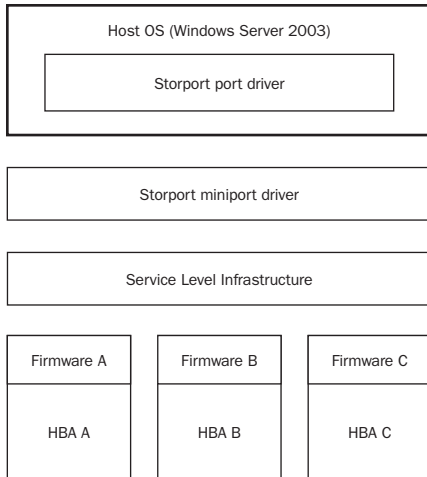


Figure 2: Windows host with Storport driver and the Emulex Storport miniport driver. Note the SLI layer between the Storport miniport driver and the HBA firmware.

Integration with SLI Architecture

The Emulex design for the Storport miniport driver follows the same highly-successful architecture used for all Emulex devices and HBAs produced since 1995: that of an interface layer lying between an operating system-specific Emulex device driver and the device's firmware and hardware. This Service Level Interface (SLI) acts to abstract the software driver components from firmware/hardware, making it possible to upgrade both the software and the firmware/hardware independently of each other.

By allowing the SLI layer to deal with the specifics of the Storport driver interface to Emulex Fibre Channel HBA specific hardware and firmware, the Storport port and miniport driver can evolve to incorporate new functionality and features, and still remain compatible across both firmware and hardware. Not only can any generation of Emulex hardware and firmware operate with any generation of Storport driver, but servers on the SAN can also contain multiple generations of HBAs running with the same version or multiple versions of drivers.

Emulex Storport Miniport Solution

In order for a Storport miniport driver to work with the Storport port driver, a number of core routines controlling basic functionality (such as `HwStartIo` and `HwInterrupt`) must be implemented. In the Emulex Storport miniport, these routines are executed in the upper layer module that interfaces directly with Storport.

In addition to these core routines, a number of supplemental Storport routines conferring additional functionality can be implemented in the miniport driver, *provided the miniport driver has been designed to accommodate*

the routines. The design of the Emulex miniport driver not only allows the full range of Storport's current functionality to be implemented, but also positions the miniport driver to readily accommodate future Storport developments.

Parallel Processing of Multiple I/O Requests

The new Storport port driver enables the I/O start routine to be decoupled from the I/O completion routine, such that the driver supports sending I/O requests and receiving I/O completions simultaneously.

The ground-up design of the Emulex Storport miniport complements the Storport redesign by parsing I/O operations into functionally separate phases. The separation of these two phases, command initiation and command completion, enable the benefits of parallel processing to be realized.

- **Command Initiation.** This phase of I/O processing prepares the command for the HBA firmware and queues the command. The command phase has two sub-phases: *Preparation* (which Storport sees as `HwBuildIo`), and *Submission*, during which a start I/O lock is held, so that no other commands in the submission phase can be simultaneously executed.

Once command initiation is complete, the I/O request is built, the target storage device is identified, and the I/O is sent to the target so that further processing can occur.

- **Command completion.** This phase begins when the HBA interrupts the host to indicate that the device has finished executing the command and I/O is complete. Any errors during this phase will trigger Storport to handle the error appropriately.

Because the miniport completion phase can be executed at the same time as command initiation, a greater number of I/O requests (relative to SCSIport) can be processed in the same amount of time on a multiprocessor system.

Device I/O & Device Discovery Limitations

Storport eliminates the SCSIport limitation of 254 outstanding I/Os per HBA, supporting instead 254 I/Os per LUN times the number of LUNs per HBA. This advance in port driver technology dramatically increases the underlying scalability, essentially shifting the I/O throughput limitation away from the underlying drivers to the HBA itself.

The current generation of Emulex HBAs supports a theoretical limit of 2048 outstanding I/Os and subsequent generations of HBAs will allow for an even greater number of I/Os to be processed. For Emulex HBAs, device login is currently limited to a maximum of 510 targets; this limit will increase with future firmware releases. Adoption of Storport is an effective means by which to increase I/O throughput capabilities, especially in large SAN environments. Nevertheless, it is important to ensure that the HBA supports queue management functionality (see below), since application performance will suffer without a way to deal with I/O saturation at the HBA level.

Hierarchical Bus Resets

One of the common goals of installing a SAN infrastructure is to ensure high data availability. In a large SAN, common events can cause serious disruption when using the SCSIport driver. Unlike SCSIport, Storport does not default to resetting the entire bus when a command timeout is encountered.

Instead, Storport will retry the command. When a reset is required, Storport instructs the miniport driver to reset at the least intrusive level possible.

The Emulex Storport miniport driver has a three-tiered progressive reset hierarchy, designed to ensure that the lowest level of reset is tried first. Resets are directed first to the LUN, then to the target, and finally to the SCSI bus level. Typically the LUN reset will be successful, so target and bus resets are not executed. The Emulex miniport driver also manages the pausing and resuming of I/O at these levels, as explained below.

Queue Management

The Microsoft Storport port driver now supports I/O queue management by the miniport driver, provided the miniport driver has been architected to take advantage of such capabilities.

The Emulex Storport miniport driver implements a number of queue management routines that help ensure that I/O throughput to storage devices remains high and error free. These routines can be implemented at the adapter level or the device level.

- **Adapter-level Miniport Routines.** Adapter-level routines affect the entire HBA and affect all targets and LUNs that the HBA supports. This is useful if HBA resources are limiting I/O throughput. Emulex implements the following as adapter-level routines to ensure that I/O requests to the HBA can be slowed (throttled) or halted in order to prevent the HBA from becoming overloaded, causing system and HBA inefficiencies. Without these routines in the Storport miniport, the high throughput capabilities of the Storport port driver cannot be effectively realized.

- **StorPortBusy.** This routine notifies the port driver that new I/O requests to the HBA should be temporarily halted, since the HBA is busy handling outstanding requests. The HBA can respond by allocating more resources to the process. If no additional resources are available, no new I/Os are passed to the adapter until a specified number of outstanding I/Os have been completed. This prevents I/Os from failing and the host from wasting CPU cycles.
- **StorPortReady.** Once the HBA is no longer busy, this routine sends notification to the port driver, and I/O requests are resumed.
- **Device-level Miniport Routines.** The Emulex Storport miniport driver has implemented several target- or LUN-level optional routines that provide more specific control over both the distribution of I/O load and the response to errors or other SAN fabric event conditions. Without this level of control in the HBA driver, SAN performance can be significantly disrupted.
 - **StorPortSetDeviceQueueDepth.** This routine sets the maximum number of I/O requests (queue depth) for the indicated LUN within a device.
 - **StorPortDeviceBusy.** If I/O requests to the logical unit or target have reached maximum queue depth, this routine notifies the port driver of a busy status for that specific LUN/target on the SCSI bus. Storport will not issue any new requests to that storage unit until its queue depth falls below the maximum.
 - **StorPortDeviceReady.** This routine notifies the port driver that the logical unit or target is ready to handle new I/O requests. As with the device busy command, the logical unit is identified up through storage target and SCSI bus levels. In most

cases this command is not necessary, since the device busy routine counts the outstanding I/Os remaining in the queue.

- **StorPortPauseDevice.** This routine pauses a device for a specified period of time. When the timeout expires, I/O requests to the device are resumed.
- **StorPortResumeDevice.** This routine resumes a paused device.

As will be seen below, all of these capabilities—queue management, hierarchical resets, and improvements in I/O performance processing—are critical to effective storage management in most large-scale storage scenarios.

Enhanced Support for Critical Storage Scenarios

The features and capabilities detailed above in the Emulex implementation of the Storport miniport driver support a number of critical storage scenarios that are important not just to large organizations, but also to small and medium organizations who require more cost-effective ways to protect their mission critical data.

Clustering

Clustering servers helps ensure that applications remain available to users at all times, even if a single server goes offline. In the Windows environment, the new Storport driver solution much more successfully handles clustered servers than SCSIport could. Using SCSIport, if a connection to a device (in a storage array or on a SAN) was lost, the entire bus—and all connected devices—became unavailable during the bus reset process; as a consequence, neither server can access the shared storage for a period of time.

In contrast, when using Storport in a clustering configuration, the hierarchical reset localizes the reset to the device with the problem; other devices on the bus are not impacted. The net result of using Storport in a clustering scenario is far better use of system resources (since the CPU intensive reset/reservation process will occur less frequently) and high resource availability.

Disks and Tapes on the Same HBA

Many servers are limited to two built-in slots that support HBAs, which can constrain the configuration and number of attachable devices. This constraint is acute in a parallel SCSI environment with disks and tapes on the same HBA. Because tapes transfer data in larger block sizes than disks and have slower response times than sequential transfer disks, tapes obtain the bus less frequently than do disks. When tape devices do obtain the bus, however, they tend to tie it up with long data transfers.

In Fibre Channel environments the increased number of devices and the complexity of network topology can lead to error conditions that are especially detrimental to tapes. Should an error condition unrelated to a tape device cause a bus reset, the tape backup will either slow down or fail completely. In transfers involving several gigabytes of data (which can take many hours to back up) a bus reset can be highly disruptive.

The Emulex Storport miniport driver is designed to ensure continuous functioning of both disks and tapes on the network. This driver adheres to the FCP-2 specification, which defines the standards for tape error recovery on Fibre Channel, interconnects.

About Emulex

Emulex Corporation is the world leader in Fibre Channel HBAs and delivers a broad range of intelligent building blocks for next generation storage networking systems. Emulex ranked number 16 in the Deloitte 2004 Technology Fast 50.

The world's leading server and storage providers rely on Emulex HBAs, embedded storage switching and I/O controller products to build reliable, scalable and high performance storage solutions. The Emulex award-winning product families, including its LightPulse[®] HBAs and InSpeed[®] embedded storage switching products, are based on internally developed ASIC, firmware and software technologies, and offer customers high performance, scalability, flexibility and reduced total cost of ownership. The company's products have been selected by the world's leading server and storage providers, including Dell, EMC, Fujitsu Ltd., Fujitsu Siemens, Bull, HP, Hitachi Data Systems, IBM, NEC, Network Appliance, Quantum Corp., StorageTek, Sun Microsystems, Unisys and Xyratex. In addition, Emulex includes industry leaders Brocade, Computer Associates, Intel, McDATA, Microsoft and VERITAS among its strategic partners. Corporate headquarters are located in Costa Mesa, California. News releases and other information about Emulex Corporation are available at <http://www.emulex.com>.

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. Although this information is believed to be accurate and reliable at the time of publication, Emulex assumes no responsibility for errors or omissions. Emulex reserves the right to make changes or corrections without notice.

Hierarchical resets (LUN, target and bus) are implemented at the least disruptive level, allowing isolation of disk device errors from tape devices on the same HBA.

Other Scenarios

In addition to the functionality mentioned above, the Emulex Storport miniport driver can be used in combination with the Emulex Storport filter (or adjunct) driver to provide additional capabilities not included in the Storport driver.

- **Persistent binding.** Targets can be associated with specific SCSI IDs, allowing the connection between a server and storage resources to be retained following a reboot. The Emulex Storport filter driver accesses information in the Windows registry to enable persistent binding.
- **LUN masking.** The Emulex Storport filter driver supports masking of LUNs (making them unavailable) to the server. This masking or hiding of resources prevents servers from accessing storage resources that do not belong to them.
- **LUN mapping.** The Emulex HBA can be configured to enable LUN mapping. The internal operating system LUN number can be associated with any 64-bit Fibre Channel LUN address. This functionality permits communication with targets that don't use the Peripheral Addressing Method for LUN addressing.

Summary

The Microsoft Storport port driver and the new Emulex Storport miniport driver together help deliver highly manageable storage solutions in hardware RAID and storage area network environments. Built in partnership with Microsoft and architected from the ground up to take advantage of new and future functionality of the Windows port driver, the Emulex Storport miniport driver lets customers realize the advantages of full duplex I/O transfer, high volume I/O transfers, as well as sophisticated I/O queue management functions. These and other features help deliver solution providers, architects and end users higher I/O throughput and greater storage resource manageability to the Windows platform than was previously possible.

Additional Resources

All papers are available at:

<http://www.emulex.com/products/white/>

- The Critical Role of a Host Bus Adapter (HBA) in Storage Area Networks.
- Emulex SLI Architecture Simplifies Storage Management.
- Storport in Windows Server 2003: Improving Manageability and Performance in Hardware RAID and Storage Area Networks.