Building High-Performance iSCSI SAN Configurations

An Alacritech and McDATA Technical Note

Building High-Performance iSCSI SAN Configurations

An Alacritech and McDATA Technical Note

Internet SCSI (iSCSI) Reaching Maximum Performance

IP storage using Internet SCSI (iSCSI) provides opportunity for many organizations looking to extend existing Fibre Channel SANs to stranded servers, or looking to deploy Ethernet-based SANs. It can reduce costs by allowing IT managers to take advantage of existing, familiar Ethernet networks. Often the biggest complaint with new technology like iSCSI is its ability to provide more than simple functionality and connectivity. That argument is addressed by leveraging fast, wire-speed TCP/IP products common to networking today with fast storage systems. This combination can offer maximum performance and efficiency comparable to many Fibre Channel-based solutions.

iSCSI encompasses several components of storage networking. This includes host-based connection devices commonly referred to as initiators, and storage systems known as targets.

iSCSI Initiators can take several forms including:

- Host Bus Adapters (HBAs) with the iSCSI initiator implemented in the hardware adapter card
- Software initiators running over standard network interfaces, such as Network Interface Cards (NICs), TCP Offload Engine (TOE) NICs (TNICs), or Host Bus Adapters

iSCSI targets also come in various forms, including:

- Disk storage systems
- Tape storage systems
- IP storage switches

Since completion of the iSCSI standard in early 2003, a number of standards compliant products have entered the IP storage market. Many of these products have completed interoperability and conformance testing from organizations such as the University of New Hampshire's Interoperability Lab[1] and Microsoft's Designed for Windows Logo Program[2]. While many of these products have achieved general interoperability and functionality, most lack the levels of performance necessary to compete against other technologies in the global storage market.

Despite the lack of performance from some IP storage devices, there are a few iSCSI products that can be used to build high-performance iSCSI (Storage Area Networks) SANs to meet high transaction rates and/or high throughput requirements. This note examines the performance capabilities of the McDATA IPS Multi-Protocol Storage Switches and servers equipped with the Microsoft iSCSI Software Initiator and hardware accelerators like the Alacritech Accelerator family of TNICs or Alacritech iSCSI Accelerator family of iSCSI HBAs to build such a high-performance iSCSI SAN.

[1] http://www.iol.unh.edu/consortiums/iscsi/index.html
[2] http://www.microsoft.com/windowsserversystem/storage/technologies/iscsi/default.mspx

Performance Test Objectives

The objectives of the iSCSI performance testing aimed to show the McDATA IPS Multi-Protocol Storage Switch in conjunction with Alacritech iSCSI Accelerators and the Microsoft iSCSI Software Initiator on Windows-based servers, sustaining wire-speed iSCSI throughput at larger transaction sizes and a substantial transaction rate, measured in Input/Output Operations per Second (IOPs). In the case of saturating a single IP Storage Gigabit Ethernet link, the full-duplex wire-speed throughput is over 210 Megabytes per second. For more details please see Appendix B: Storage Networking Bandwidth.

This test looks at a balanced single test configuration tuned for performance over a broad spectrum of operation sizes. Tests and iSCSI SANs can be more specifically tuned for optimal throughput or optimal transaction rate.

Performance Configuration Details

Test Execution

Alacritech commissioned VeriTest, a division of Lionbridge Technologies, to compare the performance of a number of iSCSI initiator products. Performance reported in this note reflects the subset of testing specific to the Alacritech and McDATA configuration. The full test report is available from VeriTest at http://www.veritest.com/clients/reports/alacritech/

Equipment

The iSCSI server configuration used an SuperMicro X5DPE-G2 motherboard with dual Intel Xeon processors, and an Alacritech SES1001T iSCSI Accelerator. Alacritech's acceleration solutions are based on the company's high-performance SLIC Technology® architecture. Products based on SLIC Technology remove I/O bottlenecks for both storage and networking systems by offloading TCP/IP and iSCSI protocol processing. The server used the Microsoft iSCSI Software Initiator, version 1.02, to perform iSCSI connectivity. The iSCSI Accelerator was connected to the network through a Dell PowerConnect 5224 Gigabit Ethernet switch to the McDATA switch.

iSCSI SAN configuration for high-performance

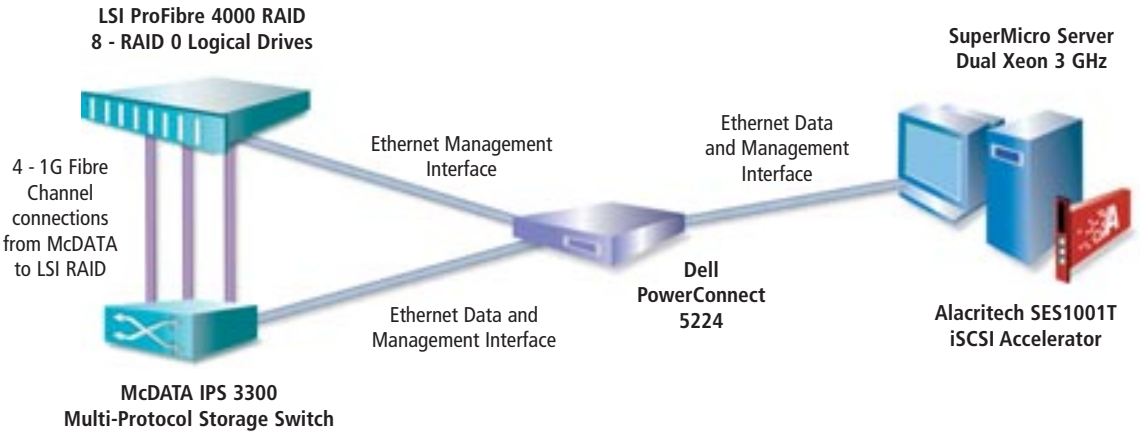

Figure 1: iSCSI Performance Configuration

A McDATA IPS3300 Multi-Protocol Storage Switch acted as the iSCSI target device. The McDATA switch connected the incoming iSCSI traffic from the iSCSI server to a LSI ProFibre 4000 Fibre Channel RAID subsystem. The LSI RAID storage included high-performance 15K RPM disk drives from Seagate.

Traffic Generation

Iometer, a server performance analysis tool developed by Intel, drove traffic for the tests. The server was running an iSCSI/TCP/IP session across the Gigabit Ethernet link. Version 2003.12.16 was used in the configuration. For more information on how Iometer was used during these performance tests, see Appendix A: Iometer Information.

Building Maximum Performance with a Software Initiator Solution

Initiators can be hardware-based or software-based. Major research firms such as International Data Corp. and Gartner Dataquest are expecting deployments of both types of solutions.

Network copies on receive with NICs penalizes performance on servers

Software iSCSI initiators can run over standard NICs. Using a NIC is a sufficient connectivity solution due to the ability to use standard Ethernet failover and link aggregation techniques, to perform in-band management of other iSCSI SAN devices and operating system vendor support of the iSCSI protocol implementation, leading to better protocol conformance and interoperability.

NIC performance is generally acceptable for write operations, but underperforms significantly for read operations. This is due to NICs not being able to perform direct memory access (DMA) operations direct to destination memory due to protocol processing required on the host prior to knowing the destination memory address. Servers tend to favor storage read operations rather than write operations, so this limitation is significant.

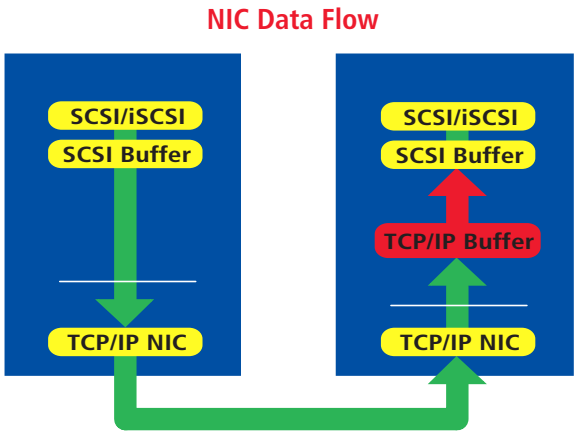

Figure 2: NIC Data Flow

A hardware solution uses an HBA that implements the iSCSI protocol on the card and appears to the system as a SCSI device. HBAs do not have the read DMA limitation of NICs due to the on-board protocol processing of the HBA.

Most HBAs lack expected Ethernet functionality

The HBA approach has a number of drawbacks such as:

- Impossible to use standard Ethernet failover and port aggregation techniques;
- Because this solution is dedicated to block transport only, it does not transport network traffic, including in-band management traffic with other iSCSI SAN and network switch devices;
- Dependence on the HBA vendor and not the operating system vendor for iSCSI protocol conformance and interoperability.

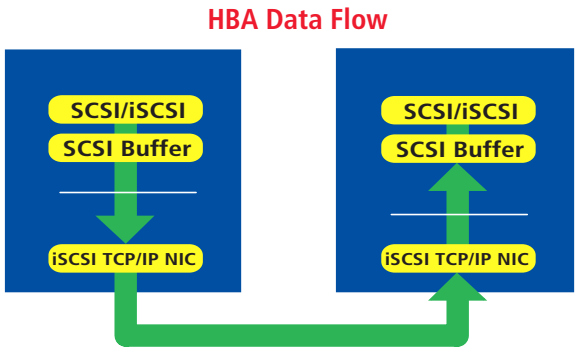

Figure 3: HBA Data Flow

Alacritech SES1001 preserves NIC benefits and HBA performance

An alternative HBA approach involves the use of TOE/iSCSI hardware with a software iSCSI initiator. This preserves the benefits of the NIC, while also sharing the same data flow as other HBAs. Alacritech's SLIC Technology architecture includes patented methods for receive processing that allow DMA operations directly to the destination SCSI buffer on the host mirroring the HBA data path. The end result is the same as a conventional iSCSI HBA of direct placement of the data in the desired host buffer without a host data copy. Alacritech also utilizes patented methods for port aggregation and failover while performing TCP offload, which HBA solutions cannot do.

Throughput Performance Configuration Results

Bi-directional throughput of over 210 MB/s

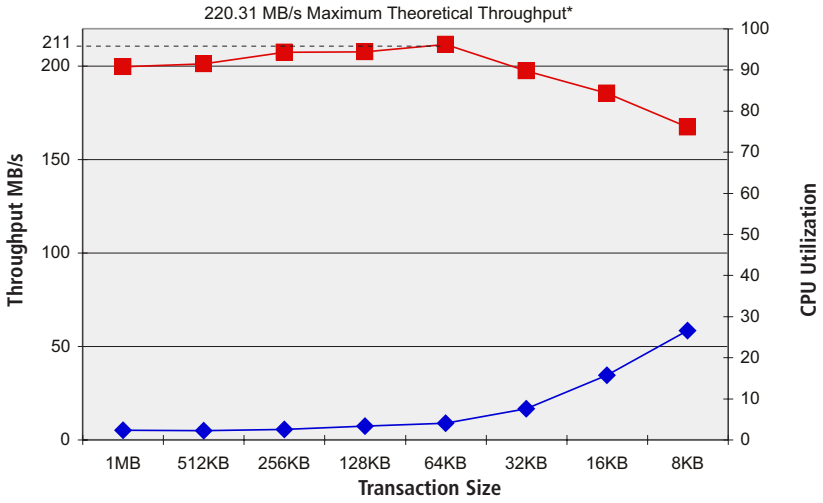

Figure 4: Bi-Directional Throughput Performance

Results of the performance testing show that average bi-directional throughput of over 210 Megabytes per second was reached across a single Gigabit Ethernet link for frame sizes of 64KB and larger.

* see Appendix B

Average unidirectional read throughput of 113 MB/s

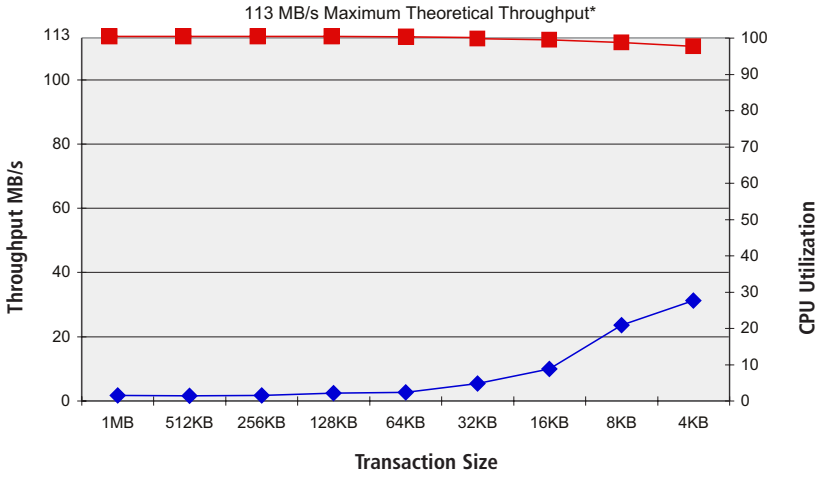

Figure 5: Read Throughput Performance

Results of the performance testing show that average unidirectional read throughput of 113 Megabytes per second was reached across a single Gigabit Ethernet link for frame sizes of 8KB and larger.

* see Appendix B

Throughput Performance Configuration Results Continued

Average unidirectional write throughput of 113 MB/s

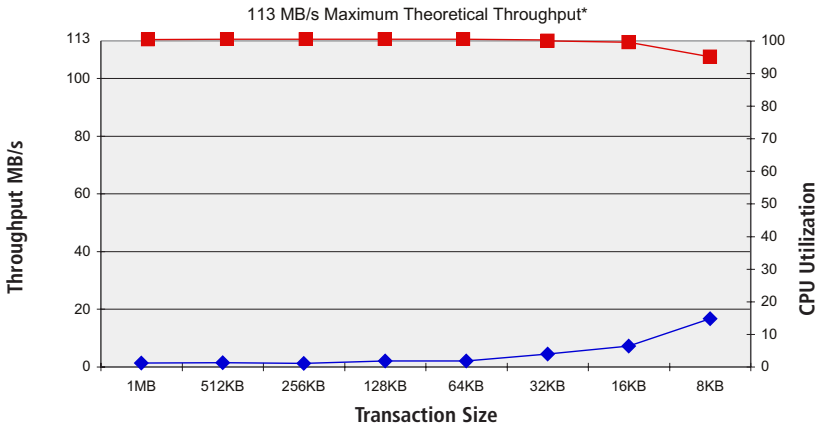

Figure 6:
Write Throughput Performance

Results of the performance testing show that average unidirectional write throughput of 113 Megabytes per second was reached across a single Gigabit Ethernet link for frame sizes of 8KB and larger.

* see Appendix B

Small Transaction Rate Performance Configuration Results

Bi-directional average transaction rate of over 47,600 operations/s

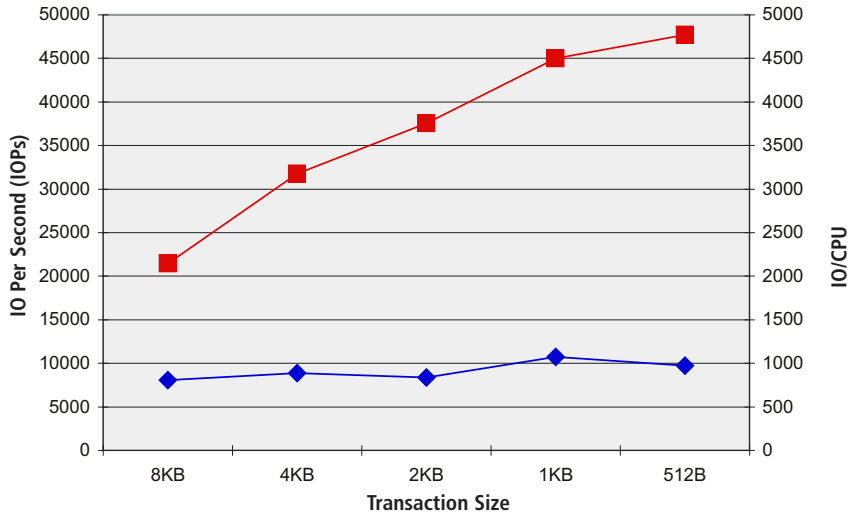

Figure 7:
Bi-Directional Transaction Performance

Results of the transaction performance testing show that a bi-directional average transaction rate of over 47,600 operations per second was reached across a single Gigabit Ethernet link for a frame size of 512B.

Unidirectional read average transaction rate of over 46,000 operations/s

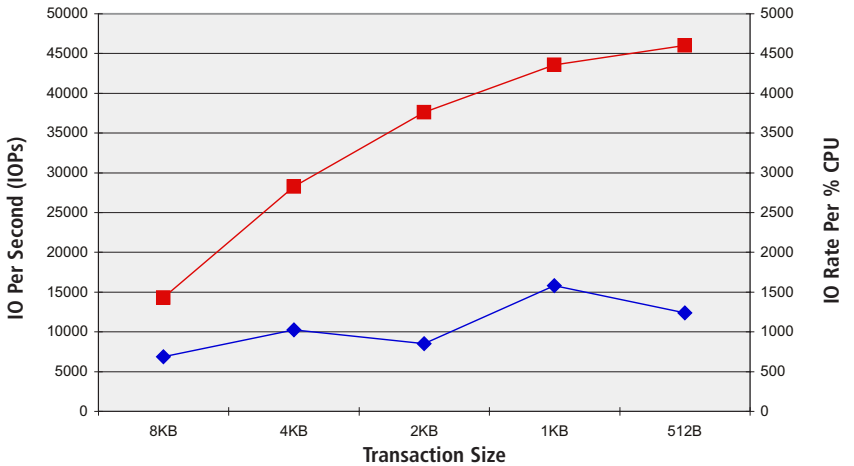

Figure 8:
Read Transaction Performance

Results of the performance testing show that unidirectional read average transaction rate of over 46,000 operations per second was reached across a single Gigabit Ethernet link for a frame size of 512B.

Small Transaction Rate Performance Configuration Results Continued

Unidirectional write average transaction rate of over 23,500 operations/s

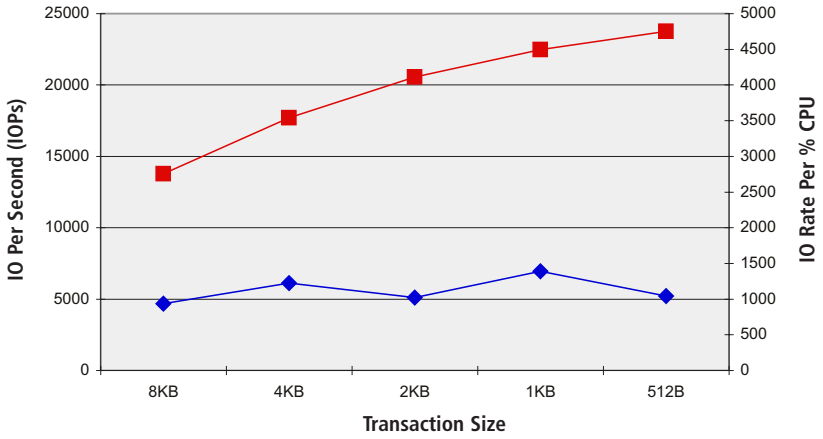

Figure 9: Write Transaction Performance

Results of the performance testing show that unidirectional write average transaction rate of over 23,500 operations per second was reached across a single Gigabit Ethernet link for a frame size of 512B.

Conclusions – High-Performance iSCSI

The results of the tests show that full duplex throughput with iSCSI can reach line rates of over 210 megabytes per second (MB/s) in both IP Storage switches from McDATA and HBAs such as Alacritech's iSCSI Accelerator. In throughput tuned configurations, the configuration has been validated at closer to the 220 MB/s theoretical maximum iSCSI payload bandwidth in a number of tests subsequent to the initial January 2002 Alacritech, Hitachi, and Nishan Systems wire-speed demonstration[3]. McDATA subsequently acquired Nishan Systems in 2003, with the Nishan Systems products and technologies forming the foundation of the McDATA Multi-Protocol Storage Switch product family.

The capability of the McDATA Multi-Protocol Storage Switch to handle high throughput wire-speed conversion and high transaction rate conversion between iSCSI and Fibre Channel is clearly demonstrated through these results. This performance confirmation with Alacritech iSCSI Accelerators helps facilitate the adoption of iSCSI in the enterprise and the emergence of large-scale, high-performance IP Storage networks.

[3] http://www.alacritech.com/iscsi/pr/index.html

Appendices:
Building High-Performance iSCSI SAN Configurations

Appendix A: Iometer Information

What were the Iometer settings?
Iometer was set for operations at block sizes from 512 bytes to 1MB.

Is the block size representative of applications?
Applications run block sized anywhere from 512 bytes up to 16MB. Smaller block sizes mean that at any given moment, additional time is spent awaiting receipt of the block than sending the block itself. Use of larger block sizes reduced the relative overhead and allows more data to be in the network pipe.

What was the setting for outstanding I/O operations?
Iometer was set to six (6) outstanding I/O operations. Typical applications range from one to sixteen outstanding I/O operations. One outstanding I/O operation is used when the application requires confirmation of each I/O before executing the next I/O operation. The one outstanding I/O setting typically is used to guarantee the highest integrity, with a tradeoff in performance. With only one outstanding I/O at a time, the recovery time is minimized. By allowing more outstanding I/O operations at any given time, more data can be placed in the network pipe and overall link utilization will increase. Recovery of multiple I/Os also is possible, but will take slightly longer. These effects are specifically detailed in the McDATA white paper *Data Storage Anywhere, Any Time.*

What version of Iometer was used?
Version 2003.12.16

Appendix B: Storage Networking Bandwidth

To verify that full wire speed is achieved, one must first determine the theoretical limits of the technology under test. The signaling rate of 1000BASE-SX Gigabit Ethernet is 1.25 Gigabits per second in each direction. After accounting for 8B/10B coding overhead and the 1460 byte payload size, the usable bandwidth is 113.16 megabytes per second. Bi-directional bandwidth is approximately 220.31 megabytes per second. Note that the bi-directional bandwidth accounts for acknowledgements (see below). SCSI and iSCSI protocol overhead is not included in these calculations.

The faster link speed of Gigabit Ethernet to one gigabit Fibre Channel requires the test configuration to have storage on multiple Fibre Channel ports to provide sufficient disk bandwidth to fully saturate the single Gigabit Ethernet connection between the server and switch. Additional detail is provided in Figure 10.

For the purposes of these calculations, 1 megabyte is defined as 1024^2 Bytes or 1,048,576 bytes. This correlates with the definitions in Iometer.

Condition	Gigabit Ethernet
Raw Link Bandwidth	1.250 Gbit/s
Link Bandwidth with 8B/10B Coding Overhead	1.0 Gbit/s
Unidirectional Payload Bandwidth	113.16 MB/s (1460 Byte Payloads)
Bi-directional Payload Bandwidth	**220.31 MB/s** (see detail on ACKs below)

The iSCSI / Gigabit Ethernet frame is assumed to have 78 Bytes of overhead (Ethernet 14B, IP 20B, TCP 20B, CRC 4B, Interframe Gap 20B).

These calculations ignore iSCSI Protocol Data Unit (PDU) and SCSI command overhead. In throughput configurations this overhead is insignificant, with SCSI commands and the 48 byte iSCSI PDU header occurring once per operation, i.e. one SCSI command in each direction, and one iSCSI PDU header on the SCSI response per 64KB to 1MB transaction. SCSI and iSCSI overhead become much more significant in smaller transactions.

Figure 10: Storage Networking Bandwidth

Acknowledgements (ACKs)

The TCP protocol used with iSCSI provides a reliable transport. All data sent over TCP is acknowledged by the receiving system. In the unidirectional read or write tests, the ACK traffic has no impact, since it is on the otherwise idle portion of the bi-directional link. iSCSI traffic using TCP typically requires unique ACK frames rather than piggybacking the ACK with the data. This occurs since the traffic is unidirectional in nature – a small SCSI CMD will generate lots of data in one direction, with the other direction essentially idle. The number of ACK frames can vary depending on the ACK algorithm (e.g. ACK every frame, ACK every other frame, ACK at specific intervals, etc.), however, an ACK typically will occur for every other frame. This results in a bit over six megabytes of ACK traffic on a bi-directional link, reducing the payload to 220.31 megabytes per second.

Additional Details for Figure 10

Calculation of Ethernet and iSCSI Bandwidth

Raw Link Bandwidth	1.25 Gbit/s
Net Rate (8B/10B Coding)	1.00 Gbit/s
Net Rate, in megabits	1,000 Mbit/s
Net Rate, in bits	1,000,000,000 bit/s
Net Rate, in bytes (Bps)[4]	119.21 MB/s
Total Bytes per Frame[5]	1538
iSCSI Overhead	78
Data Payload per Frame	1460
Payload percentage	94.93%
Actual Data Rate without ACKs	113.16 MB/s
Net Rate, in Bytes[4]	125,000,000
Frames per second no ACKs	81274.4
How many frames per ACK	2
Frames per second with ACKs	79113.9
Data Rate with 2 frames per ACK	110.2 MB/s
Bi-directional Rate	**220.31 MB/s**

Ethernet Frame (Bytes)

1500	
14	Header
4	CRC
20	Interframe Gap
1538	**Total Ethernet Frame**

TCP Overhead (Bytes)

14	Ethernet
20	IP
20	TCP
4	CRC
20	Interframe Gap
78	**Total TCP Overhead**

ACK Frame (Bytes)

14	Ethernet
20	IP
20	TCP
6	Ethernet Pad[6]
4	Ethernet CRC
64	Total
20	Interframe Gap
84	**Total ACK Bytes**

Notes

[4] 1 Megabyte = 1024^2 bytes = 1,048,576

[5] Includes Ethernet preamble and interframe gap (20 bytes), but no VLAN tags (4 bytes)

[6] Required for Ethernet minimum frame size of 64 Bytes

Alacritech, Inc.
234 East Gish Road
San Jose, CA 95112 USA
www.alacritech.com

toll free: 877.338.7542
tel: 408.287.9997
fax: 408.287.6142
email: info@alacritech.com