iSCSI Can Allow Your Storage Network to go Global

Note: This document discusses how 10GbE ethernets enable iSCSI systems to outperform Fibre Channel and also allow them to become truly global. Many ideas in the document came from whitepapers by Force Networks and Neterion, but were edited into their current form by Thomas Jerry Scott.

Enterprise networks are often physically located in multiple locations, connected together via IP capabilities. Mobile users often need to access the IP network from anywhere in the world. SANs are being called upon to provide access on a global basis.

To provide effective global access presents great challenges for requires a network that can minimize latency successfully while increasing available bandwidth and decreasing system cost without creating a network that is difficult to manage.

Many SAN deployments today are based on Fibre Channel (FC) technology. FC has a distance limitation of 100 miles. Transmitting FC over a WAN is complicated by its sensitivity to delay. IFCP (Internet FC Protocol) and FCoIP (FC over IP) were designed to allow FC to run over IP networks. These alternatives offer piecemeal efforts, requiring specialized hardware, and introducing additional complexity and vulnerabilities into the IP network.

The Internet Small Computer Systems Interface (iSCSI) was created to enable SAN functionality over the IP network. iSCSI uses SCSI over TCP/IP, enabling any requesting node on the IP network (initiator) to contact any remote dedicated server (target) and perform block I/O on it just as it would using a local hard drive.

Because it is a native IP protocol, iSCSI has no distance limitations, can utilize existing network infrastructure, does not require specialized operator or administrator training, and can leverage the vast economies of scale of the enormous Ethernet market.

iSCSI's Slow Start

Early iSCSI deployments were incomplete and did not offer a complete iSCSI stack with the high quality transport and fault tolerance. The complete iSCSI protocol was not approved by the IETF until February, 2003. By then, many vendors were already selling incomplete iSCSI enabled devices.

When the iSCSI protocol was approved, FC systems offered predictable and reliable performance, but commanded a premium price. Enterprise storage managers who were using FC systems could not imagine iSCSI replacing FC systems because of the lack of required features.

Highly necessary items, such as standards compliance, reliability, interoperability, reliability, security, and integration were not part of early "pre-standards" iSCSI efforts. A sensible storage manager using FC systems and vitally interested in the reliability and integrity of the company's storage would hardly change from FC to iSCSI.

As a consequence, early iSCSI product deployments worked only marginally well and failed to address true Enterprise storage needs, prompting Enterprise storage managers to adopt a wait-and-see stance. So what's changed with iSCSI?

10 Gigabit Ethernet and Error Recovery Level Two (ERL 2)

10 Gigabit Ethernet (10GbE) was approved in 2003, and has been steadily gaining acceptance. Unless a 10GbE system uses a specialized adaptor, just running 10GbE will sap lots of CPU cycles. A 2003 estimate speculated that "it will take a 30GHz Pentium 4 processor to handle 10GbE efficiently." Problems like this are "money in the bank for hardware engineers" who see such "problems" as opportunities to build a sellable product that meets a

market need. One Gigabit Ethernet "offload engines" appeared in 2003 and provided about 100% more ethernet throughput while keeping the CPU load as low as 100 Mb Ethernet.

Hardware engineers have since built 10GbE offload engines to make 10GbE usable for truly high speed networks. From its inception in 2003, 10GbE has been growing, showing more acceptance each year. IT managers of large Fortune 1000 companies are already deploying 10GbE in their backbone switch and router networks and are also starting to connect the fast-speed line to their servers and storage systems. The Dell'Oro Group estimates that over 140 thousand 10GbE switch ports were shipped in 2005, and this number grew to nearly 500K in the year 2006.

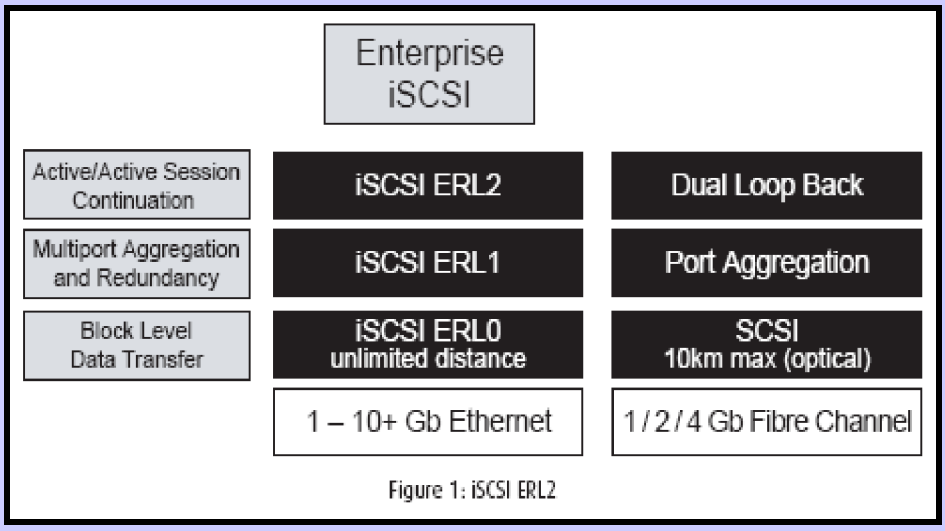


Figure 1: iSCSI ERL2

Server consolidation and virtualization, backup acceleration, large size image & video transfers over the network, etc., can benefit instantly from upgrading the end-point connection to 10 GbE. Many vertical markets like Medical Imaging, Media-Entertainment, Oil & Gas, Military and Research, are faced with exponentially increasing data volumes. And with the proliferation of Gigabit Ethernet to the desktop, the demand for bandwidth is only increasing.

With the completion of the entire iSCSI protocol as defined by IETF RFC 3720, iSCSI now has the full error recovery feature set required by enterprise storage managers. While this standard has yet to be widely embraced by the industry, the features and reliability provided by completing the upper Error Recovery Levels of the iSCSI protocol will be essential to developing Enterprise market confidence in the reliability of iSCSI.

5 key reasons to consider iSCSI with ERL2 and 10 Gigabit Ethernet	Ethernet Switch/Routers are a perfect fit for iSCSI with ERL 2 at 1 or 10 Gbps
1) Security, Reliability, Quality of Service of iSCSI ERL2 versus Fibre Channel's limited feature set	1) Unified Ethernet networks support multiple uses, saving money – both CAPEX and OPEX
2) Performance of 10 Gigabit Ethernet versus 2 Gbps Fibre Channel	2) High density switch/routers reduce the number of network devices to buy and manage
3) Cost of Ethernet vs. Fibre Channel (both from a hardware and a Total Cost of Ownership perspective)	3) Non-blocking 1 GbE and 10 GbE ports optimize storage performance
4) Proven capabilities of Ethernet, such as unlimited distance, Server disk-less booting capability, and so on.	4) High-availability switch/routers with redundant components improve uptime
5) iSCSI with ERL2 enables robust disaster recovery, back-up, and replication over remote sites via existing IP WANs	5) Protocol-rich switch/routers connect seamlessly with multiple networks

Table 2: iSCSI Reliability and Router Switch Issues

iSCSI systems before 2003 had 1 GbE iSCSI with very little Error Recovery capabilities. Even with lower costs, they just did not stack up against 2 Gb/sec Fibre Channel with full Error Recovery features. These iSCSI implementations didn't put up a great fight against their 2Gb/sec FC competitors.

However, with the maturity and completion of iSCSI standards to provide full error recovery and 10GbE's throughput, availability, cost, distance, and economy of scale advantages, iSCSI with ERL 2 presents a much more serious challenge to FC systems. Here are some reasons why the 10GbE iSCSI challenge to FC systems is real:

- Ethernet offers seamless integration with the LAN, MAN, and WAN, enabling native transport of data anywhere in the world.
- The ever-present nature of Ethernet means it provides certain economies that are not possible with FC systems
- Ethernet is an older technology than FC, and is quite robust now with proven protocols and management tools. There are also many more trained Ethernet experts than there are FC trained experts.
- 10 GbE will continue to benefit from emerging Ethernet advancements, such as hardware-based iWARP support (RDMA over TCP/IP), which will add Remote Direct Memory Access to optimize performance and reduce CPU utilization.
- All popular Operating Systems support IP stacks and essentially all modern forms of Ethernet. This inherent capability can help eliminate the need to implement OS modifications when upgrading networks.
- Modern Quality of Service (QoS) methods enable traffic to be prioritized. From our applications priorities, we can develop priority queuing or traffic shaping, if that becomes necessary for iSCSI systems.
- Single- and multi-mode fiber support gives network administrators greater throughput in short-reach environments and maximum flexibility through long-reach single-mode fiber in determining the location of server assets.
- 10 Gigabit Ethernet coexists natively with already widely deployed Gigabit Ethernet infrastructures

FC equipment has traditionally been sold with a high margin, giving FC cost flexibility and allowing FC vendors to continue to drop price to a level competitive with iSCSI. However, the majority of the cost associated with storage networks is not the initial price of adapters and disk drives; rather, the majority cost comes from the efforts to manage and upgrade the network over time.

While the cost of adapters may be complementary between FC and iSCSI, the salaries of trained personnel to manage them is not. In addition, converging the storage network into the IP network can normally yield significant savings and great reductions in TCO.

Computing a storage network TCO requires evaluating more than initial hardware costs. Items such as up-front capital costs, maintenance fees, and system integration costs incurred with every network infrastructure change must be considered. Ongoing support costs, physical space, port density, power and cooling issues must also be considered. System management costs to keep servers and required components safe and available, and soft costs like slow network performance and downtime should be considered. TCO becomes an important consideration. Annual recurring costs can range from 30-50% of initial hardware investments.

Estimates vary broadly on how a yearly TCO relates to initial hardware costs, but a commonly given estimate that yearly costs represent 30-50% of initial costs. Some analyses have shown that for small storage networks, the yearly costs were 4 to 7 times the initial capital costs for drives, controllers, and the other components.

See the TCO analysis at

http://tjscott.net/storage/tco.analysis.scsi.iscsi.pdf

for more details on this subject.

Ethernet Advantages for iSCSI Storage Systems

iSCSI is ideal for storage applications that serve up data that has already been created, such as for databases and streaming AV servers, as well as clustered computing. Without any distance limitations, iSCSI enables virtualized networks with nodes located physically around the world to appear as a single logical network.

Because of its thin software layer, network administrators can turn almost any TCP-enabled device into an iSCSI initiator or target, giving near-universal access to a network while preserving existing Ethernet investments. Table 3 summarizes these advantages for a 10GbE iSCSI solution.

Such flexibility changes the way people will design their networks. The fact that iSCSI runs over Ethernet means that iSCSI can be available, almost seamlessly, over the LAN, MAN, and WAN, enabling native transport of data anywhere in the world. The fact that “Ethernet is everywhere” and carries all sorts of applications means that more and more applications can take advantage of an iSCSI SAN.

Ethernet 10GbE will benefit from emerging Ethernet advancements, which are often hardware-assists built into existing cards by well known vendors for doing important things like reducing CPU utilization. Of course, 10GbE was designed to co-exist with other forms of Ethernet, such as 1GbE or 100MbE, which are already common in our corporate networks. Ethernet capability is supplied at a low-level in modern Operating Systems; thus we do not need to implement OS modifications when upgrading networks.

Stateless offload 10 GbE adapters make a great iSCSI solution	
Best-in-class 10 Gigabit Ethernet stateless adapters include comprehensive offloads and assists. They handle in hardware a significant amount of the TCP/IP processing road blocks, including Checksum Offloads, Large Send Offloads, Large Receive Offloads, and UDP "Checksum Over Fragment". As the benchmark results show in Table 4 show, these adapters are fully capable of carrying 10 GbE iSCSI traffic today with limited CPU utilization rates. With 2.8GHZ twin AMD Opterons, running Xframe 10GbE PCI X2.0 adaptors in a back to back environment, the sending machine's average CPU utilization was less than 15%. The receiving system's CPU utilization was approximately 51%. Additionally, these 10 GbE adapters offer the following benefits:	1) Full compliance with all operating systems without an overhaul of TCP/IP stacks. 2) Fast adoption and deployment of 10 Gigabit Ethernet technology by major server and storage vendors. 3) Preserving the security and reliability of established TCP/IP stacks. 4) Taking advantage of Moore's Law and the dozens of billions of dollars of R&D spent by Intel, AMD, IBM, and Sun/Fugitsu to improve processor speeds and performance. As a consequence, processor speeds will always eventually match networking speeds, making stateless adapters faster than full-offload approaches. 5) There is no need to wait for emerging TOE and RDMA incremental benefits since 10 Gigabit iSCSI is deployable with the current stateless, OS-compliant adapters available today, like Neterion's Xframe 10 GbE adapters.

Table 3: Ethernet 10GbE Offload Capabilities

Ethernet's maturity thus brings proven diagnostic and management tools and many trained experts, all of which are benefits to an iSCSI approach to network storage. There is already a large base of available personnel trained to work with Ethernet. Their learning curve to learn the particulars of storage area networks is significantly less than that required to become an expert with Fibre Channel. Leveraging existing personnel expertise contributes to a lower total cost of ownership (TCO) for end-user customers.

The simplification that stems from a converged network on Ethernet results in great reductions in TCO (downtime, problems tracking, network management, etc.). These hidden "opportunity costs" are difficult to quantify but real for datacenter managers.

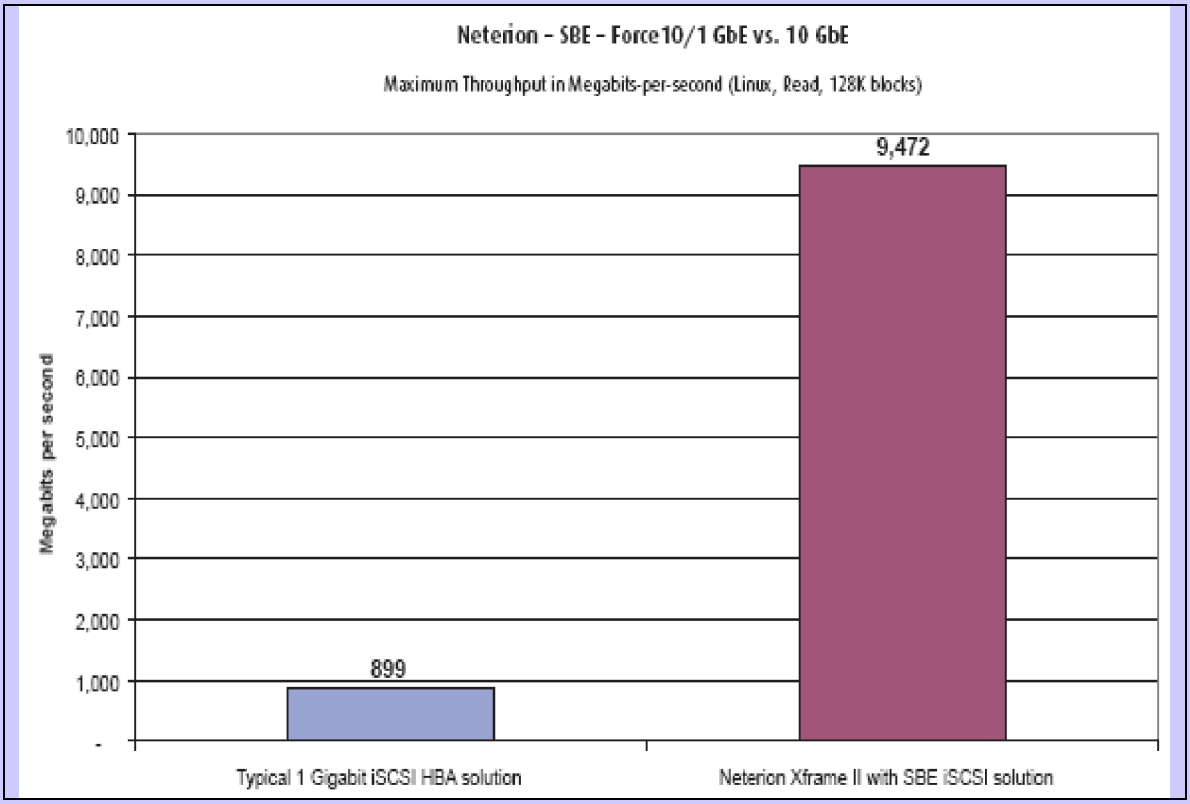


Table 4: Neterion 1GbE and 10GbE Ethernet Throughput Analysis

Why Consider iSCSI with ERL2 and 10GbE?

The following are five key reasons to consider 10GbE and iSCSI:

1) Security, Reliability, Quality of Service of iSCSI ERL2 versus Fibre Channel's limited feature set
2) Performance of 10GbE Ethernet versus 2 Gb or 4GB Fibre Channel systems
3) Cost of Ethernet vs. Fibre Channel (both from a hardware and a Total Cost of Ownership perspective)

4) Proven capabilities of Ethernet, such as unlimited distance, Server disk-less booting capability, and so on.
5) iSCSI with ERL2 enables robust disaster recovery, back-up, and replication over remote sites via existing IP WANs

iSCSI with ERL 2 brings important optimizations to storage networks. Interoperability drops deployment costs to commodity levels and enables network administrators to select best-in-class components rather than having to purchase an entire system solution from one vendor. Today, all of the pieces are available for making deployment of ERL 2 iSCSI IP SANs a reality.

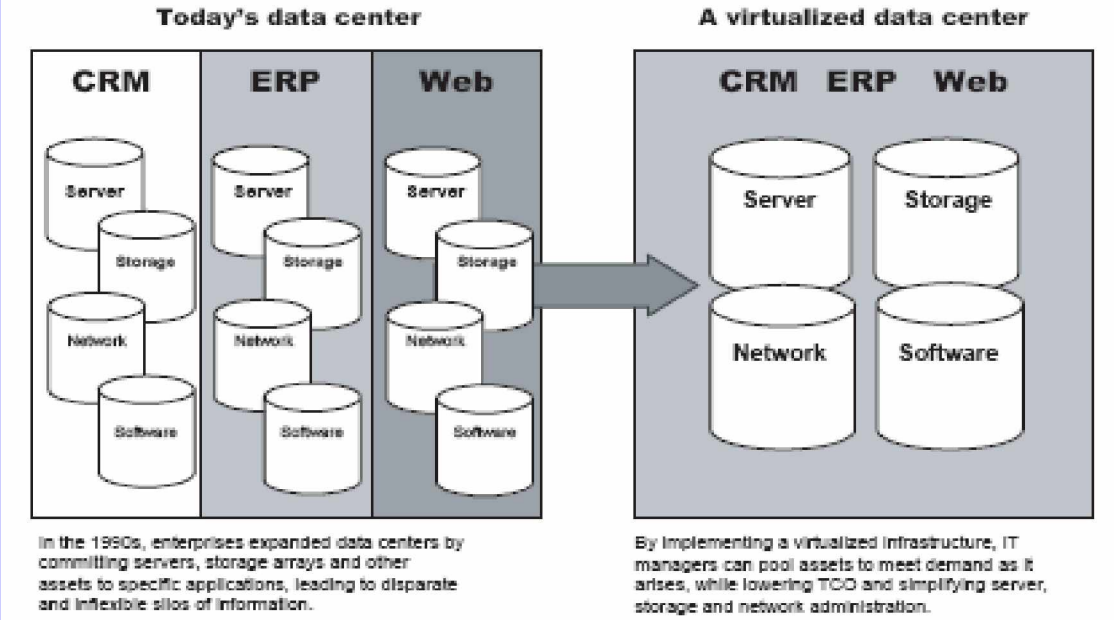


Figure 2: Maximizing Bandwidth in the Virtualized Data Center

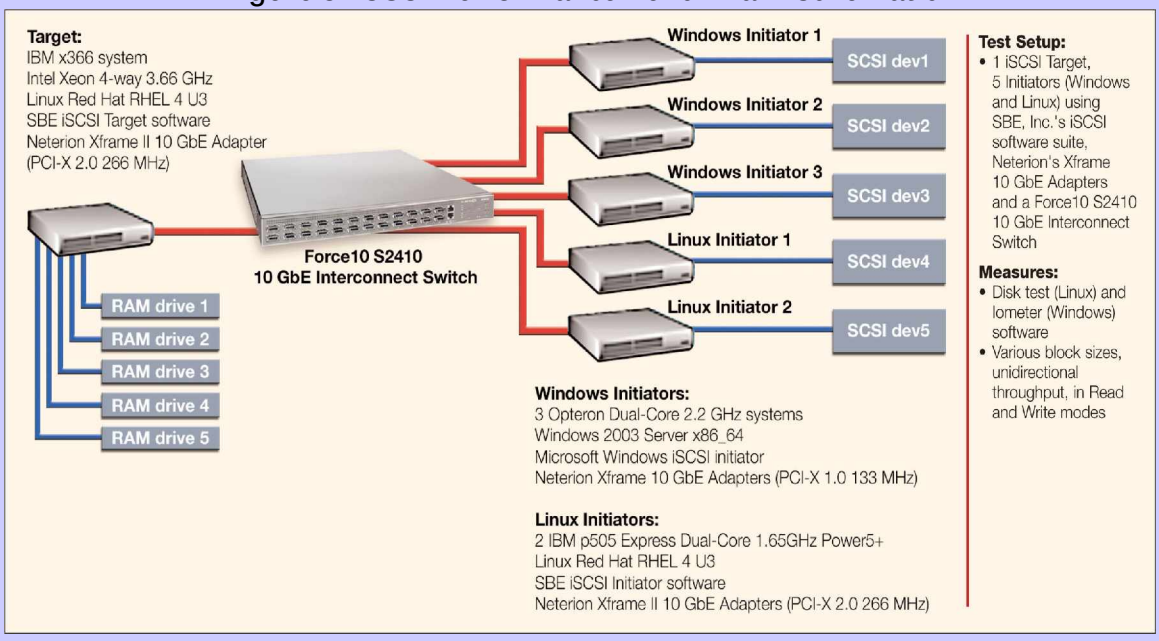
Evaluating an Available 10GbE iSCSI System

For global storage area networks, the cost savings of iSCSI using ERL 2 versus FC can be tremendous. With modern iSCSI, administrators will still be able to leverage their existing FC investments. However, since iSCSI with full error recovery at 10GbE offers much more functionality and value, as well as bridging to FC networks, new deployments will certainly favor iSCSI ERL 2. FC's 4Gbps spec is an available, but does not perform as well as 10 GbE iSCSI systems.

The Force Networks S2410 24-port 10 GbE is the world's first Ethernet device to deliver 300 nanosecond switching latency. The S2410 interoperates seamlessly with Neterion's Xframe®, a family of 10GbE adapters for PCI-X 1.0, 2.0 and PCI Express buses. iSCSI software from SBE provides Error Recovery Level 2. ERL2 provides for no single point of failure end-to-end, with guaranteed quality of transport supported by multipath, session continuation, and dual loop back functionality. SBE iSCSI supports clustered file system applications, scales to support N-number of storage devices, and eliminates the need for upper-level switching at the hardware level. All of these technologies are available today and interoperate seamlessly with each other.

These market leaders came together to build a network system using these off-the-shelf devices and to benchmark ERL 2 iSCSI performance at 10 Gbps. Two configurations were set up for the benchmarks: first with a couple of Linux iSCSI initiators, the second with three Windows iSCSI initiators.

Figure 3: iSCSI Performance Benchmark Schematic



In both configurations, the initiator systems accessed a single RAM disk storage target, running SBE's iSCSI target software for Linux.

Figure 4 shows the detailed hardware/software configuration. The results were measured using DiskTest and IOmeter, standard network profiling tools available for Linux and Microsoft systems.

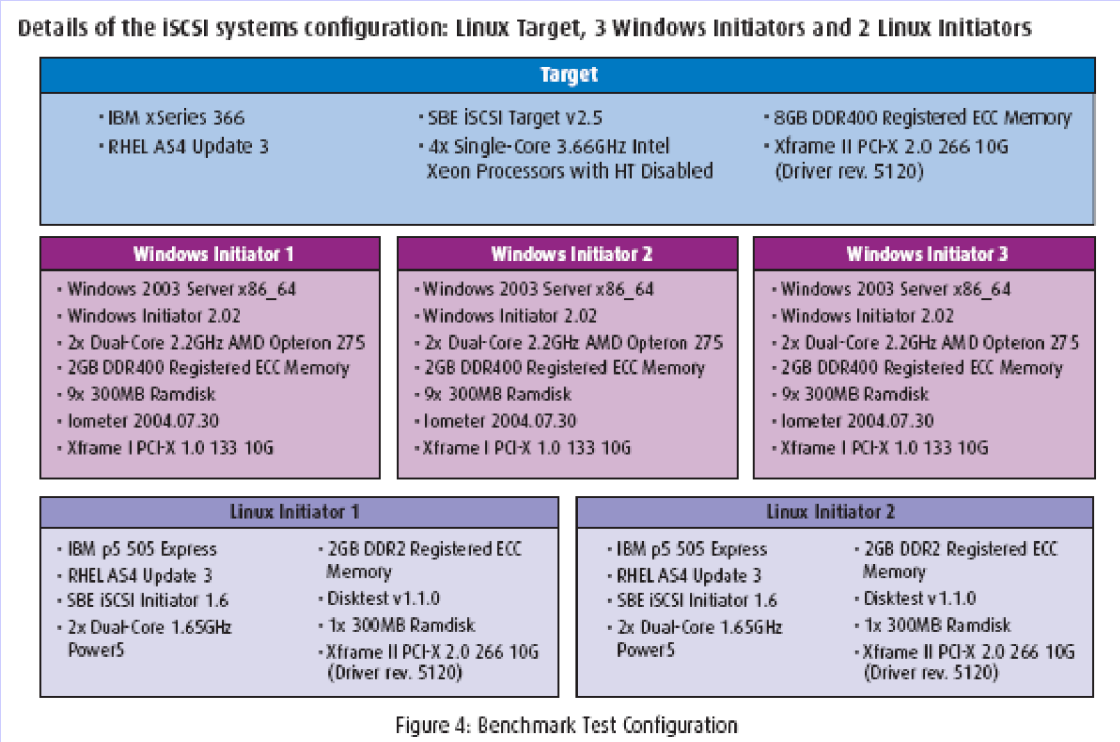


Figure 4: Benchmark Test Configuration

Detailed results of the benchmark can be found in the tables of Figures 5a and 5b. Performance measures include Throughput (in MegaByte per second), IOPS (I/O per second) and percentage of CPU utilization.

As an example, transferring sequential 128K sized blocks between the target and the Linux initiators yielded 1,184 MBps (9,472 Mbps) when reading, and 740 MBps (5,920 Mbps) when writing.

Compared to Fibre Channel industry standard of 2 Gbps, which can typically transport up to 200 MBps, 10 GbE iSCSI with ERL2 led to an impressive sixfold increase in performance. With newer 4 Gbps FC systems, the 10 GbE system still provided a 3X performance increase.

Windows Performance Results												
	Target (x366)			Windows Initiator3			Windows Initiator2			Windows Initiator1		
	Throughput (MB/s)	IOPS	CPU	Throughput (MB/s)	IOPS	CPU	Throughput (MB/s)	IOPS	CPU	Throughput (MB/s)	IOPS	CPU
1KB Write	46.55	47668	78.9%	15.58	15953	12.6%	15.50	15877	11.3%	15.47	15839	11.2%
1KB Read	64.80	66355	65.5%	22.02	22548	17.6%	21.00	21514	16.7%	21.77	22293	16.7%
2KB Write	86.55	44313	74.6%	29.04	14869	11.4%	30.40	15565	11.2%	27.11	13880	9.7%
2KB Read	135.46	69357	65.6%	45.64	23366	18.8%	46.04	23573	18.9%	43.78	22418	17.8%
4KB Write	27.93	7149	10.6%	13.84	3542	2.5%	11.90	3046	2.5%	2.19	561	0.7%
4KB Read	147.75	37824	43.2%	48.62	12447	12.2%	52.25	13376	12.8%	46.88	12001	11.6%
16KB Write	434.89	27833	75.1%	146.18	9356	8.7%	146.24	9360	8.5%	142.47	9118	8.0%
16KB Read	429.01	27457	47.7%	140.55	8995	13.9%	150.97	9662	14.7%	137.50	8800	14.0%
32KB Write	555.70	17782	74.1%	182.98	5855	6.3%	188.33	6027	6.4%	184.39	5901	6.0%
32KB Read	869.72	27831	43.1%	291.03	9313	20.5%	291.72	9335	19.7%	286.97	9183	19.6%
64KB Write	576.86	9230	60.7%	193.34	3093	4.7%	191.30	3061	4.6%	192.22	3076	4.7%
64KB Read	1175.45	18807	34.9%	393.72	6300	23.2%	396.73	6348	24.2%	385.00	6160	22.3%
128KB Write	571.33	4571	59.1%	187.32	1499	4.0%	193.07	1545	3.8%	190.93	1528	4.1%
128KB Read	1186.30	9490	26.1%	397.10	3177	23.8%	401.37	3211	24.2%	387.83	3103	24.0%

Figure 5a: Windows Benchmark Results

Linux Performance Results									
	Target (x366)			Linux Initiator2			Linux Initiator1		
	Throughput (MB/s)	IOPS	CPU	Throughput (MB/s)	IOPS	*CPU	Throughput (MB/s)	IOPS	*CPU
1KB Write	40.85	41980	74.6%	20.4	20928	32.1%	20.45	21052	28.1%
1KB Read	66.55	68246	79.5%	33.66	34514	14.9%	32.89	33732	15.0%
2KB Write	80.43	41211	72.9%	40.22	20625	25.7%	40.21	20586	26.5%
2KB Read	128.98	65994	80.8%	65.74	33439	15.7%	63.24	32555	14.1%
4KB Write	148.03	37930	73.2%	73.36	18772	27.3%	74.67	19158	26.8%
4KB Read	232.07	59260	80.7%	116.31	29771	14.8%	115.76	29489	13.6%
16KB Write	476.65	30486	90.1%	238.23	15232	23.4%	238.42	15254	24.2%
16KB Read	666.06	42612	87.8%	333.11	21302	16.8%	332.95	21310	16.7%
32KB Write	578.39	18545	61.7%	292.91	9373	21.3%	285.48	9172	21.4%
32KB Read	868.71	27823	77.6%	434.37	13905	12.3%	434.34	13918	12.4%
64KB Write	701.87	11230	56.9%	367.55	5881	18.2%	334.32	5349	18.3%
64KB Read	1143.1	18291	69.2%	573.95	9183	14.4%	569.15	9108	10.8%
128KB Write	739.64	5912	52.2%	369.89	2959	14.1%	369.75	2953	14.3%
128KB Read	1183.9	9471	47.8%	578.29	4626	11.9%	605.61	4845	18.1%

Figure 5b: Linux Benchmark Results

10GbE is a key driver for iSCSI, delivering unparalleled performance for IP SANs, with the same quality of service and system robustness that Fibre Channel now has, but at a fraction of the total cost of acquisition and maintenance. Because all these products are available in the commercial marketplace today, the need for expensive, proprietary equipment is done away with.

The overwhelming presence of Ethernet products to drive our network faster and faster, while getting cheaper and cheaper means that expert personnel, tools, and leading edge solutions will always be available. Finally, unlike Fibre Channel, iSCSI has no distance limitations. Thus an iSCSI SAN can truly be global.

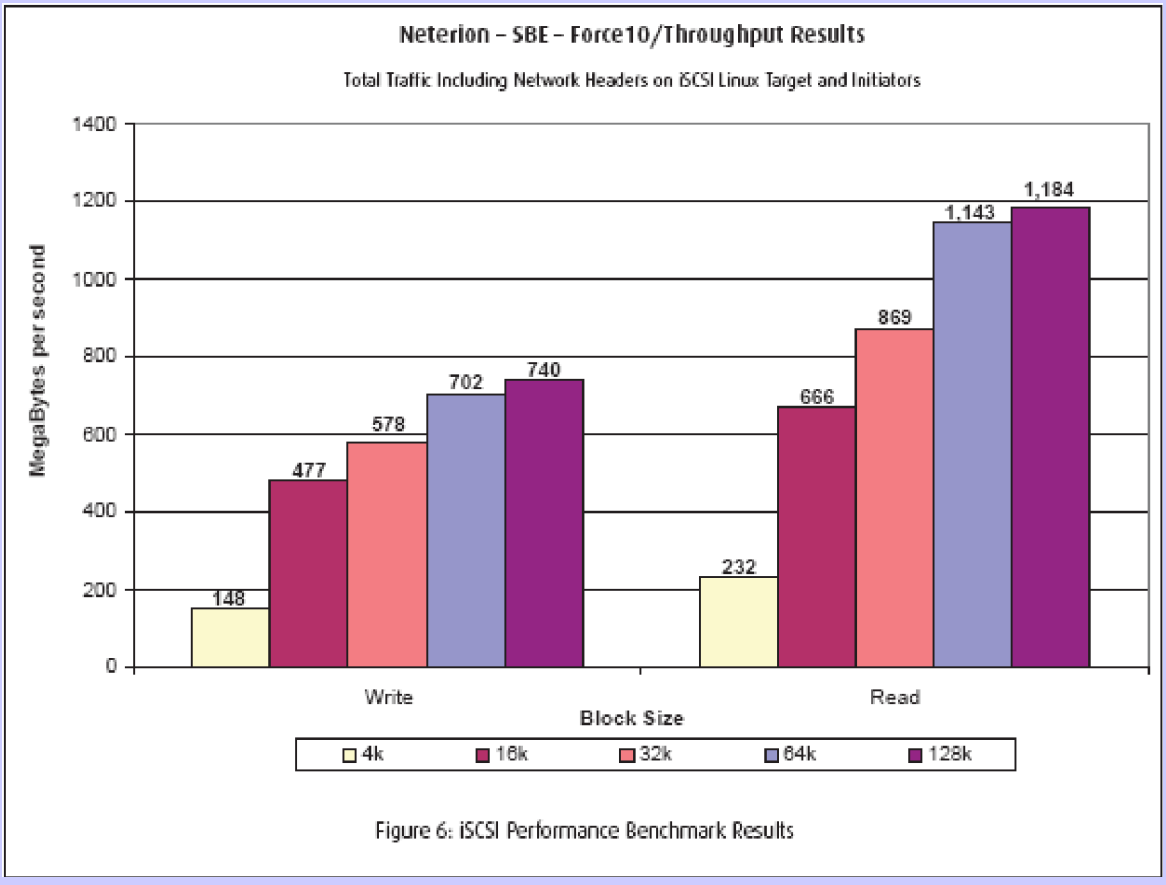


Figure 6: iSCSI Performance Benchmark Results

Summary

As the demand for bandwidth grows, so does the reach of networks. Users need to access data globally, from multiple locations around the world, quickly and transparently without breaking budgets. Data centers locked into servicing only local users unnecessarily constrain growth and expansion.

The following five factors

- Cost
- Bandwidth
- Reach
- Latency, and
- Ease-of-access

are the driving metrics behind today's storage area networks (SAN).

Fibre Channel has served the early SAN market well. However, iSCSI and Ethernet technology have reached a maturity where replacing Fibre Channel with iSCSI is a compelling proposition.

IETF RFC 3720 introduces complete error recovery mechanisms to iSCSI, providing robust and highly-available system operation across geographically diverse enterprise networks. These techniques are sometimes called ERL2, and are available in off the shelf products now.

With immediate availability of 10GbE adapters and switches, SANs can push to higher performance levels than ever before, as mass adoption drives costs down.