THIRD I/O
INCORPORATED

An Analysis of 8 Gigabit Fibre Channel & 10 Gigabit iSCSI

in Terms of Performance, CPU Utilization & Power Consumption

Key Findings

Third I/O found the efficiency of 8 Gb Fibre Channel to be higher than 10 Gb iSCSI in the following areas:

3.9x greater **bandwidth power efficiency**

10x greater **bandwidth CPU efficiency**

3.2x greater **database power efficiency**

2.8x greater **database CPU efficiency**

Introduction

For the last decade, Fibre Channel has been the primary choice for the majority of enterprise storage area network (SAN) infrastructures. Specifically Fibre Channel has been the default SAN technology for many medium to large sized businesses and data centers; especially those that have high bandwidth, high transaction, or low latency needs. Over the last six years, iSCSI has begun to carve out a market in the small to medium sized business (SMB) computing environment while running on commodity 1 Gigabit Ethernet networks. Although significantly slower than Fibre Channel, iSCSI has filled the needs of many SMB environments as it has measurably lowered the cost of entry of an initial SAN purchase.

However, in recent months, the emergence of 10 Gigabit Ethernet has some enterprise users wondering if iSCSI can now be considered for evaluation or deployment in enterprise environments. As Fibre Channel and iSCSI are dissimilar technologies, there are several considerations that need to be discussed before this question can be factually answered.

The Basics of iSCSI & Fibre Channel technologies:

8 Gigabit Fibre Channel

Fibre Channel began to make major inroads into enterprise storage with 1 Gigabit products that began shipping in 1997. There have been three major performance enhancements to Fibre Channel, the most recent being 8 Gigabit technologies.

The main concept of Fibre Channel was that it was designed to be a low latency and high performance network storage technology. It followed the same basic operating system (OS) dynamics that had been well established by the parallel SCSI technology, plus it had the added benefits of increased distance and massive scalability through its networking capabilities.

The majority of Fibre Channel protocol and I/O operations are handled by a highly efficient application specific integrated circuit (ASIC). This ASIC communicates directly with the operating system kernel via a device driver. This efficient communication allows for native direct memory accesses (DMA) from the SCSI layers to the host computer's memory subsystem. In addition, the cleanliness of Fibre Channel's architecture, plus the protocol being managed by the ASIC allow for extraordinary I/O capabilities while being extremely conservative in regards to CPU utilization. As Fibre Channel is CPU efficient, this allows for more CPU cycles to be devoted to enterprise applications. Also, the above mentioned efficiencies give an architectural advantage to Fibre Channel in terms of scalability. In other words, Fibre Channel scales in a highly efficient manner when additional ports are added to a system.

Eight Gigabit (8 Gb/s) Fibre Channel uses an 8b/10b encoding scheme and has a single port theoretical data transfer of approximately 800 MB/s in read or write operations or 1,600 MB/s in full duplex operations.

10 Gigabit iSCSI

iSCSI has been a component of deployed SAN technology since approximately 2002. iSCSI has primarily been shipping on 1 Gigabit network interface cards (NICs) as well as specialized iSCSI host bus adapters (HBAs). 10 Gigabit Ethernet is the first speed stepping to the iSCSI architecture.

iSCSI was created based on the idea that a block level storage technology could co-exist on well established 1 Gigabit Ethernet TCP/IP networks. The driving force behind this technology was the concept that iSCSI could run entirely on existing commodity networking components such as embedded NICs, CAT-5/6 cabling, and lower cost Ethernet switches and routers.

Because iSCSI was based on this concept, it was architected to reside on top of the TCP and IP layers of the network stack of multiple operating systems. The overhead of this new iSCSI layer, plus the well known inefficiencies of TCP/IP have been a known deficiency of iSCSI for many years. These inefficiencies manifest commonly as lower bandwidth, lower IOPS, higher latency, and increased CPU utilization.

However, the networking industry has been addressing this inefficiency since approximately 2000 with the release of TCP/IP Offload Engines, many times referred to as TOE technology. TOE technology is simply a hardware and software collaboration to decrease iSCSI/TCP/IP CPU overhead and to allow for a higher bandwidth and transaction environment, while decreasing CPU utilization. TOE technology is primarily ASIC based, but is also recently being supported in several newer operating system service packs and releases. TOE technology is a necessity of 10 Gigabit Ethernet adapters as a 10 Gigabit adapter that is not offloaded would seriously saturate the host system CPU resources.

Ten Gigabit (10 Gb/s) Ethernet uses a 64b/66b encoding scheme and has a single port theoretical data transfer of approximately 1,225 MB/s in read or write operations or 2,450 MB/s in full duplex operations.

Hands-On Performance Analysis

In theory, 10 Gigabit iSCSI with TOE technology should be significantly faster than 8 Gigabit Fibre Channel. However, this theory is based heavily upon the concept that the TOE hardware and software layers have solved the complex issue of TCP/IP/iSCSI overhead and inefficiencies.

In order to see the reality of the performance of these two technologies, hands-on analysis is required to determine true expectations on the wire.

Performance analysis has become much more complex in recent years. It is no longer just about data transfer, it also incorporates CPU and power utilization as well as their respective efficiency ratings. Our analysis will examine maximum bandwidth, maximum database transactions (IOPS), CPU efficiency and power efficiency.

This comprehensive analysis is critical to today's SAN customers as the datacenter of today has many selection criteria to consider beyond raw performance. Below is a listing of our selected performance metrics and the reasons why they are critical to SAN evaluation.

1.) Bandwidth. Bandwidth is the amount of data that can be transferred on a specific technology in a specific time period. In storage testing, the transfer rates are usually described in Megabytes or Gigabytes per second; MB/s and GB/s respectively. Bandwidth is critical to many applications, but some primary examples are: backup and restore, continuous data protection, RAID, video streaming, file copy, and data duplication. Fibre Channel and iSCSI are both serial technologies, this means that they can transfer (write or TX), receive (read or RX), or perform full duplex (combination of reads and writes occurring simultaneously on the TX and RX channels). In theory, an iSCSI or Fibre Channel link can transfer data at two times the rated speed as the TX and RX channels are independent of one another. This is a major advantage over previous generation parallel SCSI, which had no Full Duplex capabilities.

2.) I/O Operations Per Second (also known as IOPS). IOPS are many times referred to as small block I/O's. They generally range in size from 512 byte to 8k and are a staple of database, email, and supercomputing applications. IOPS have a known performance profile of raising CPU utilization from a combination of CPU interrupt and wait times.

3.) Bandwidth and IOPS CPU efficiency. This is one of the hot topics of today. This metric examines the ratio of either bandwidth or IOPS divided by average CPU utilization. This ratio illustrates the efficiency of a given technology in terms of CPU utilization. Higher CPU efficiency shows that the given technology is friendlier to the host system's processors. Higher bandwidth or IOPS with lower CPU utilization is the desired result. This is important as users are trying to maximize their investments and CPU utilization is much better spent on CPU hungry applications.

4.) Bandwidth and IOPS power efficiency. This is an even hotter topic and it's a major component of today's green data center initiative. This metric examines the ratio of either bandwidth or IOPS divided by WATTS consumed. This ratio illustrates the efficiency of a given technology in terms of power consumption, as well as indirectly the thermal or heat generation ratio as well. Higher values of power efficiency show that the given technology is friendlier to the host system's overall power consumption and thermal generation profile. Higher bandwidth or IOPS with lower power consumption and thermal generation is the desired result. This metric is very important in terms of Total Cost of Ownership (TCO) as higher power consumption means higher utility bills. Higher power consumption also leads to hotter systems, which requires additional cooling in a data center or server room. In many regions cooling costs are significantly greater than server power consumption.

Test Setup

INITIATOR SYSTEM

- Dell Poweredge 2970
- 2x Quad Core AMD Opteron Processors at 2.2 GHz
- Windows 2003 Server R2 SP2 x64 with full updates as of 5/15/2008

NOTE: After running our initial experiments, it was determined that we should re-run our tests on an Intel system that utilized Intel I/O AT technology. Please see the Testing Addendum for an explanation. All testing and analysis mentioned below was performed on the Intel Star Lake Server Platform.

INITIATOR SYSTEM #2

- Intel Star Lake System Model S5000PSL
- 2x Quad Core Intel Xeon Processors at 2.33GHz
- Microsoft® Windows Server® 2008 x64
- Additional OS Version 6.0.6001
- Full Updates as of 6/3/2008

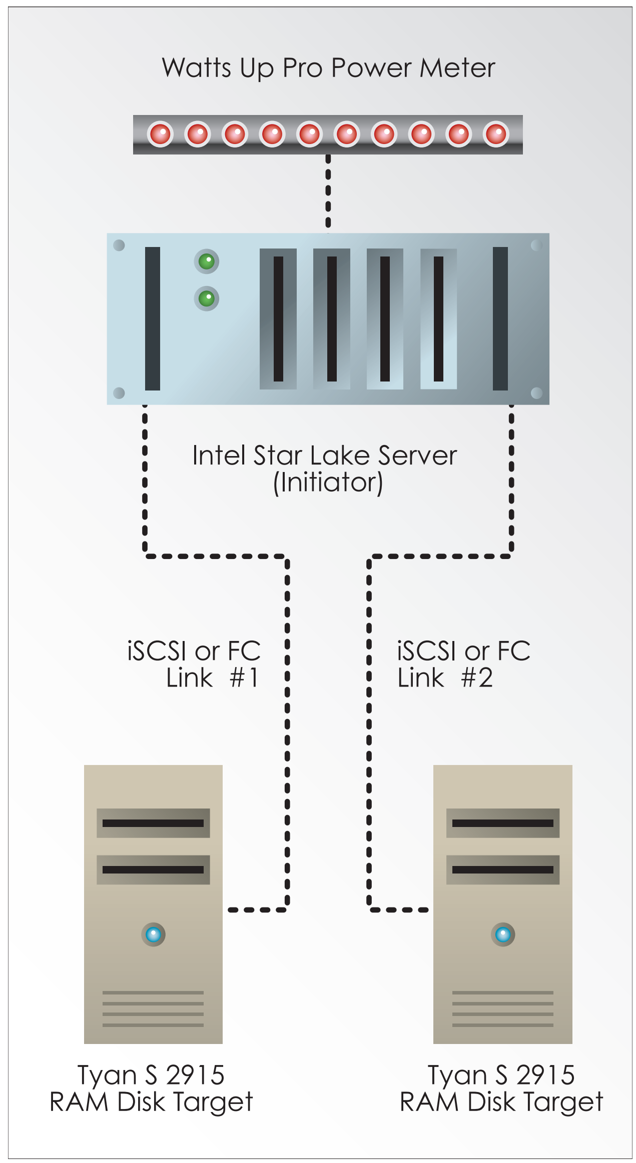
SAN ADAPTERS:

Emulex LPe12002 8 Gb Fibre Channel Adapter
This is the latest dual ported enterprise Fibre Channel adapter offered by Emulex.
We used the latest 1.00a9 firmware and the Storport Miniport Driver Version 2.01a4.

Intel® 10 Gigabit XF SR Server Adapter
This is the latest dual ported enterprise Gigabit Ethernet driver offered by Intel.
Network adapter driver and Intel I/O AT driver version 13.0 dated 4/10/21008

During testing, the Intel or Emulex adapter was placed in PCI Express slot #6 of the Star Lake Server. We used the following configuration for our testing and analysis:

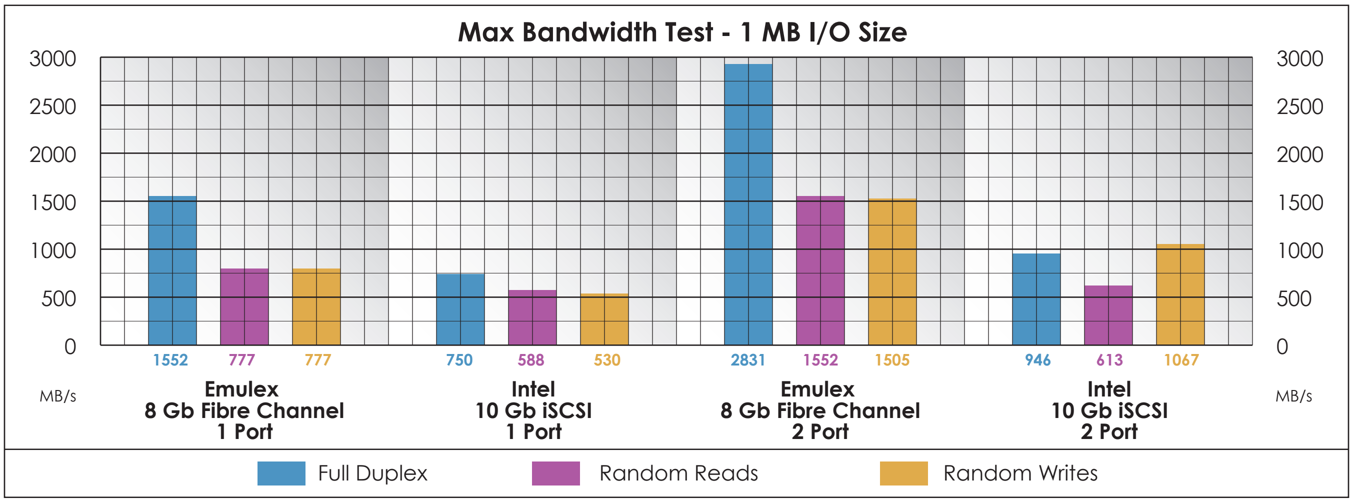
NOTE: We ran our experiments with direct connections from the initiator to the target systems. No switches were placed in the data path. In an ideal setup, switches and multiple target systems would be used. At the time of our testing, enterprise 8 and 10 Gb switches were not readily available and cost effective for these experiments.



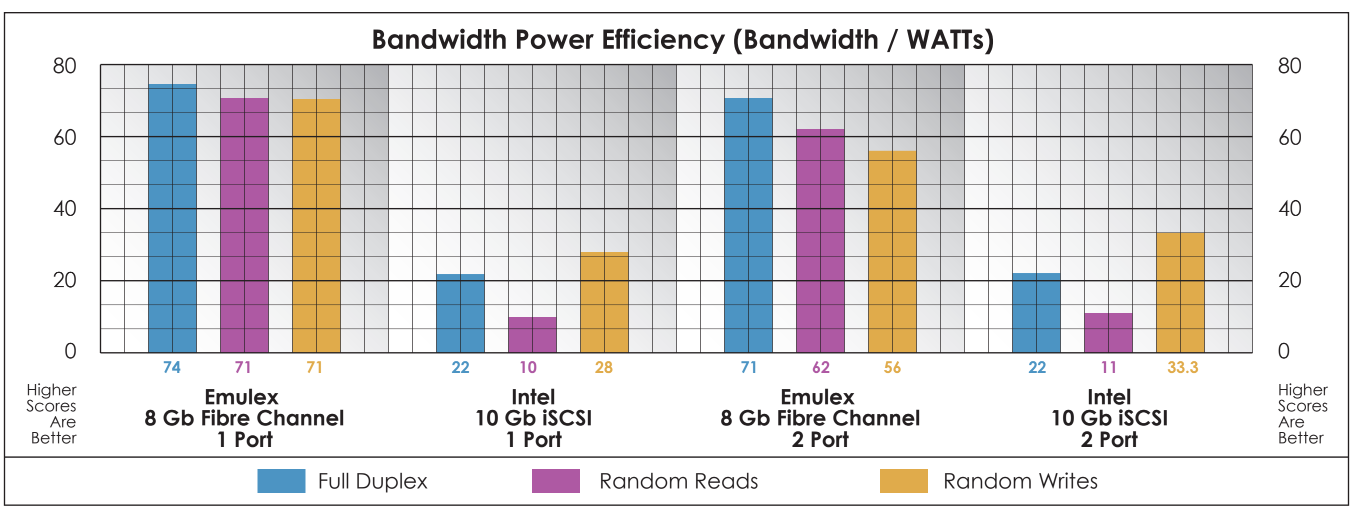
Bandwidth Testing Results

For bandwidth testing and performance analysis, we used Finisar's Medusa Labs Test Tools to verify the maximum Read, Write, and Full Duplex capabilities of both iSCSI and Fibre Channel. The Medusa tools were chosen as they run in native 64 bit mode and are very precise in terms of I/O size, distribution and Queue Depth. These tools are a staple in enterprise storage environments as well and are used by a number of top tier storage, server, and network equipment providers.

Our tests for bandwidth concentrated on a 1 MB I/O size with a constant Queue Depth (thread count) of 12 running to 4 target LUNs per port. The results are as follow:



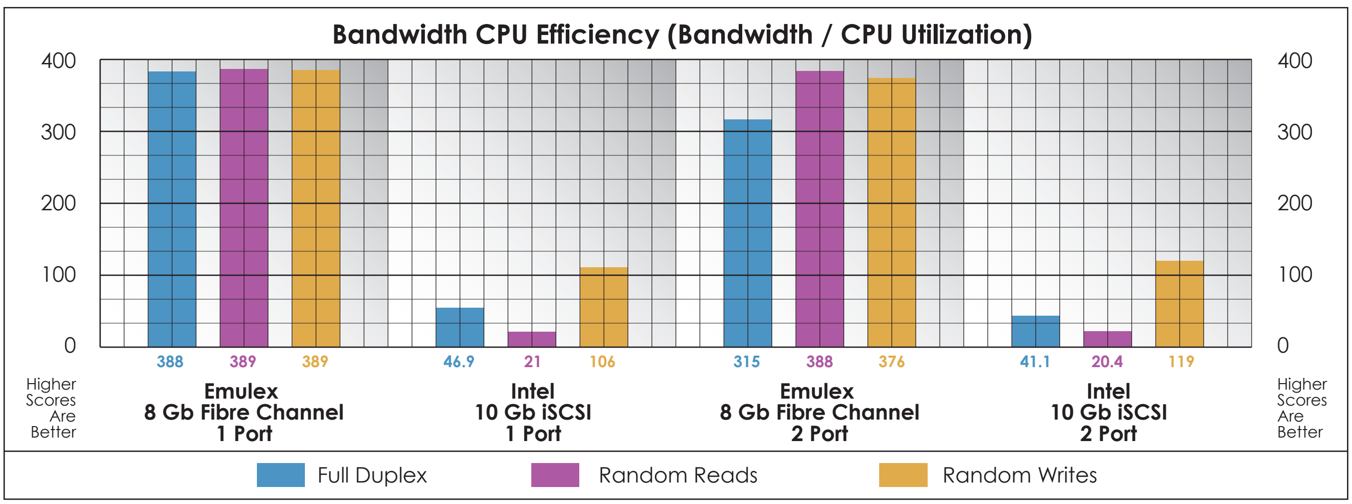
The maximum bandwidth tests show a commanding lead for Fibre Channel. In fact, all three test metrics across either one or two Emulex ports are all very close to the maximum bandwidth potential of this adapter. Intel iSCSI exhibits unexpectedly low single and dual port performance here as well. Across all bandwidth tests, Emulex Fibre Channel is on average of 2 times faster than Intel iSCSI with a maximum 3x bandwidth gain when looking at dual port full duplex performance.



During all testing, a WATTS Up Pro power meter was connected inline to the Windows initiator system. This allowed us to see how much power was being utilized by the system during each of our test cases. The Intel Star Lake server averaged 211 WATTS idle when the Emulex adapter was running in the system. However, the system ran 217 WATTS idle when the Intel Ethernet adapter was placed in the system. Based on these results, we used 211 WATTS as our baseline and recorded the average power increase during our test runs for our calculations.

Bandwidth power efficiency was determined by simply dividing the bandwidth results by the average WATTs consumed by the specific adapter during each test run. Emulex Fibre Channel showed an average of 3.9 times the power efficiency when performing bandwidth tests against Intel iSCSI with a maximum 7.1x power efficiency gain when looking at single port random read performance.

Fibre Channel consumed significantly less power during these experiments. In fact, iSCSI required an average 38 additional WATTS while running the same tests as Fibre Channel. The most extreme example was in single port random read performance where iSCSI required 47 additional WATTS of power while performing 189 MB/s slower than Fibre Channel.



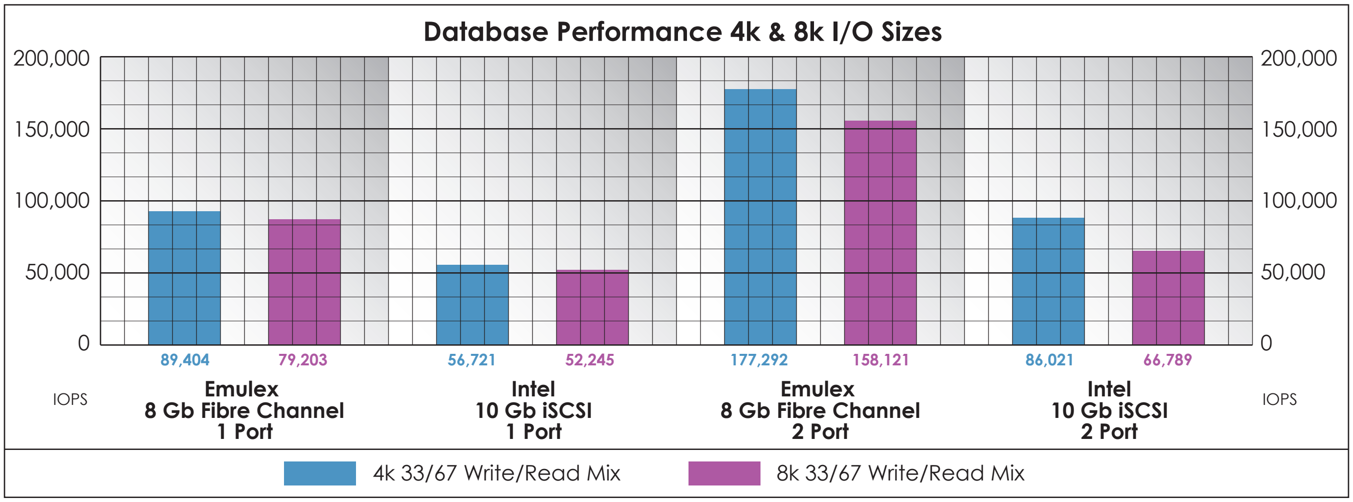
Bandwidth CPU efficiency was determined by simply dividing the bandwidth results by the average CPU utilization of the system taken from Perfmon during each test run. Emulex Fibre Channel showed an average of 10 times the CPU efficiency when performing bandwidth tests against Intel iSCSI with a maximum 19x CPU efficiency gain when looking at dual port random read performance.

These results showcase Fibre Channel's architectural design of high bandwidth and high efficiency. Furthermore, they are a major indicator that TOE technology has a long way to go towards solving the problem of efficient TCP/IP/iSCI offload.

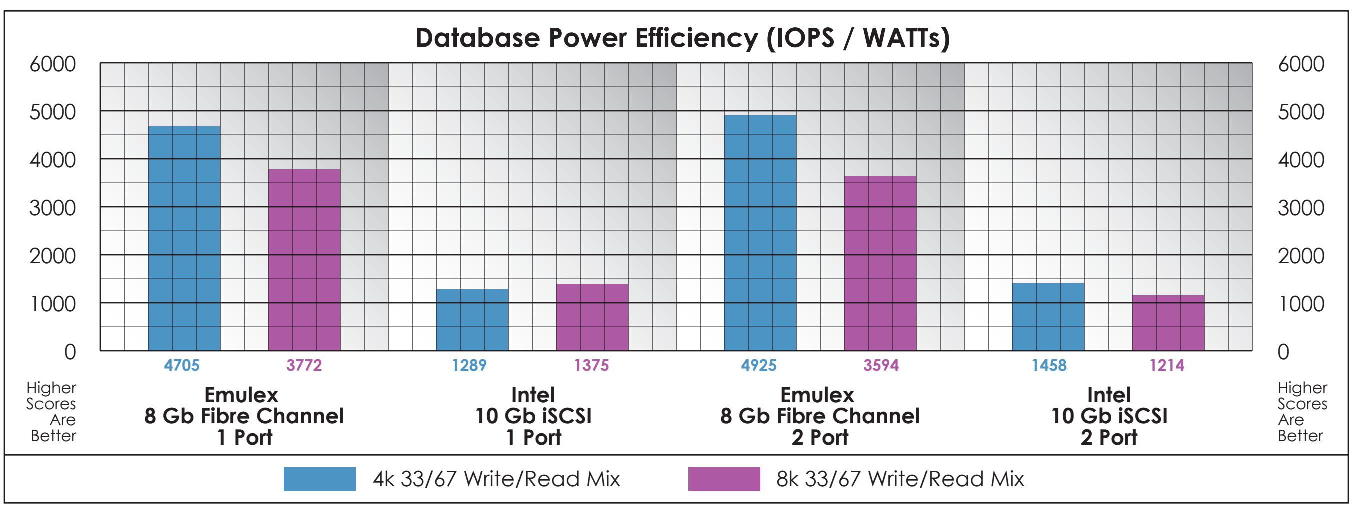
iSCSI used an average of 13% additional CPU utilization when compared to Fibre Channel. The maximum delta was in single port random read performance where iSCSI used an additional 26% of CPU resources while performing 189 MB/s slower than Fibre Channel. Fibre Channel's extraordinary CPU efficiency proves that this technology can run on lower power/cost CPU's, which will decrease initial system purchase price and will also lead to additional power and utilities savings over the life of the system.

Database IOPS Testing Results

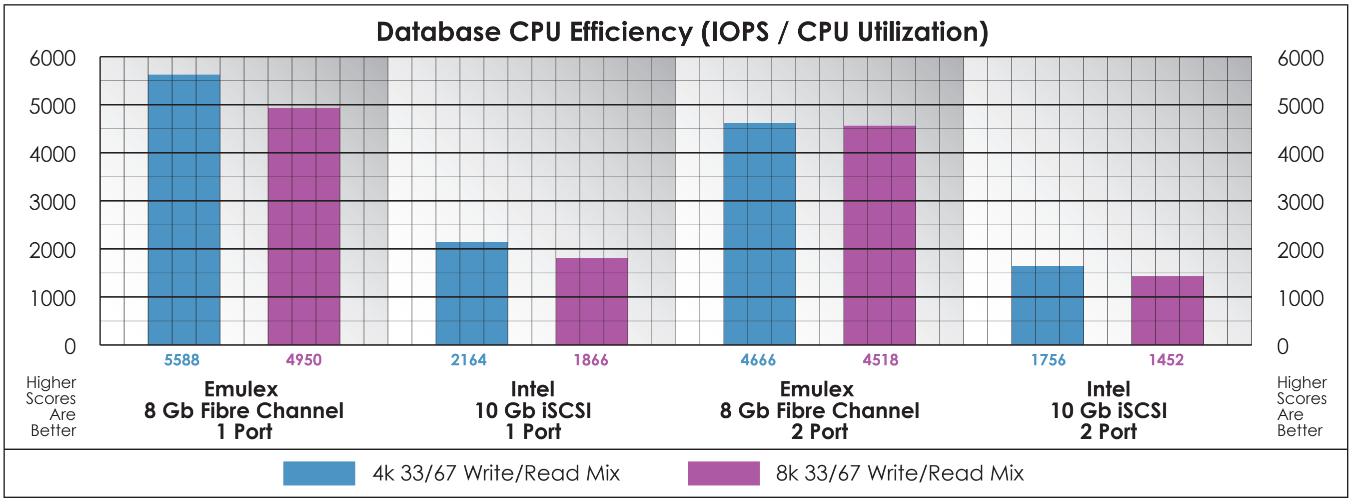
Database IOPS performance was analyzed by using the popular Iometer benchmark. As a general rule of thumb, databases follow the 1/3 2/3 rule of write/read distribution. Our benchmarks focused on the 4k and 8k I/O block sizes as they are the most commonly used in database and email application environments. We used a Queue Depth (thread count) of 12 running to 4 target LUNs per port. The results are as follow:



Fibre Channel is the clear winner in 4k and 8k I/O intensive environments. Fibre Channel outperformed iSCSI by an average of 1.9x in our database simulation testing with a maximum performance gain of 2.4x in 8k I/O testing across two ports.



And database power efficiency scales with a significantly better ratio as Fibre Channel is showing a 3.2x average advantage over iSCSI with a maximum 3.7x increase in single port 4k I/O testing.



With Windows 2008, Microsoft has included support for MSI-X, which is a recent technology whose goal is to reduce CPU utilization and interrupt time for IOPS intensive environments. Both the Emulex and Intel adapters in our test bed support this feature which addresses the well known issue of IOPS CPU efficiency.

However, even with TOE, I/O AT, and MSI-X technologies, the Intel iSCSI solution cannot compete with Fibre Channel. Here Emulex shows a commanding 2.8x CPU efficiency ratio over Intel iSCSI with a maximum of 3.1x CPU efficiency gain in 8k I/O testing across two ports.

Conclusion

Based on our testing, Fibre Channel shows a definitive and clear lead over iSCSI in terms of average performance and both power and CPU efficiency. Even though iSCSI can now operate at 10 Gb/s, our testing shows that the performance of iSCSI is significantly less than the architectural expectations of 10 Gigabit Ethernet.

Our analysis and years of experience lead us to a simple conclusion as to why there is such a gulf between these two technologies. Simply put, we believe the extraordinary overhead of 10 Gigabit iSCSI, TCP, and IP is not fully addressed by today's hardware and software TOE initiatives.

We believe that this initial test shows that 8 Gb Fibre Channel is the clear winner in SAN technologies for both standalone and virtualized operating system environments. This is because virtualization is essentially the sharing of multiple operating systems and applications on a single hardware platform. The key to a successful and high performance virtualization environment is high performance and highly efficient CPU efficiency. If our testing is any indicator, it appears as though Fibre Channel is much better suited to virtual environments due to its profile of high performance and low CPU and power utilization.

1 Gb iSCSI has a place in the entry level SAN market, but it appears as though 10 Gb iSCSI requires quite a bit more maturity before it can come close to competing with 8 Gb Fibre Channel in the enterprise SAN arena.

Testing Addendum

The Ambiguous State of TCP Offload (TOE) Technology

When Third I/O first set out to run our experiments, we decided to pick a midrange high bandwidth server that had received many accolades in the industry. We selected the Dell Poweredge 2970 as it has been a well sold system and it has extraordinary benchmarked CPU to memory bandwidth of 16 GB/s, or the equivalent of four x8 PCI Express slots.

For a 10 Gigabit Ethernet NIC, we simply went with another enterprise volume leader again when we chose the Intel 10 Gigabit XF SR Server Adapter. This adapter has been fully qualified and certified by Dell to function in the Dell Poweredge 2970 as well. In addition, Intel has been one of the top sellers of Ethernet silicon and adapters for the last decade as well, so they have an extremely strong market presence and are a commonly purchased adapter.

We were under the assumption that this adapter would simply plug and play and function in the Poweredge 2970.

But, here's where the situation became confusing.

Windows 2003 Server SP2 and Windows 2008 now ship with a technology called SNP or Scalable Networking Pack. One of the foundations of Microsoft's SNP is a technology called Chimney. The foundations of Chimney technology were licensed from Alacritech Inc. and several other TOE NIC vendors now pay a licensing fee to Alacritech for Chimney development as well. Chimney is not an open or standardized technology and the licensing cost appears to be a barrier to acceptance by all Ethernet silicon vendors.

One company that went against the grain was Intel, when they chose to pursue a technology called Intel I/O Acceleration Technology or Intel I/OAT. Without going into great detail, I/O AT is essentially Intel's proprietary TOE technology which is also fully supported by Microsoft's SNP technology.

The catch here, however, is that Intel I/OAT requires specific hardware which is only available on recent Intel platforms—and it most certainly does not work on AMD server platforms. Although the Intel 10 Gigabit XF SR Server Adapter installed and ran in the Poweredge 2970, it was unclear to us how much impact the lack of I/O AT would have on our testing results.

We spent a significant amount of time researching this topic and even spent hours on the phone with tech support from both Intel and Dell. Our answers from support were always unclear, but the short story is that this adapter is fully supported on the Poweredge 2970, albeit with an unknown or undocumented level of performance deficiency.

Because of this, a decision was made to re-run all of our tests on an Intel Star Lake platform, with full Intel I/O AT support. Using this platform, we can safely state that we chose a well researched platform that had all the necessary technologies for TOE, Intel I/O AT and SNP support.

Third I/O is detailing our experiences here because we believe that many end users will experience this same type of scenario and might find themselves running in a less than ideal environment. The reason for this, in our opinion, is simply the fact that TOE technology appears to not be fully mature and is most definitely not yet standardized in the industry today. Here are a few discoveries that support this opinion:

1) Microsoft's SNP technology is new and still unproven in the enterprise
    a. It is a service pack release in Windows Server 2003 and is only now native in Windows Server 2008
    b. Some of the open errata on SNP cause OS instability (I.E. Stop Messages)

2) Microsoft's SNP only covers Chimney and Intel I/OAT technologies

3) Some NIC vendors have licensed Chimney support while others are using proprietary TOE hardware and software models which are not supported by SNP.

4) Microsoft's SNP is disabled by default in both Windows Server 2003 and Server 2008
    a. Users must type in a command line parameter and reboot the system to enable SNP
    b. In some cases, the registry must be edited to enable or disable specific SNP dynamics
    c. And even then, SNP may not function (I.E. an I/O AT NIC in an AMD system or on a TOE/NIC that has not licensed Chimney)

5) Tech support at both Intel and Dell were severely lacking in general knowledge of TOE technology. For example, how it works, how to verify that it is functioning, how much performance gain is expected, and how much less CPU utilization will be observed.

Third I/O spent the majority of our project time researching TOE and iSCSI to ensure that we were comparing Fibre Channel and iSCSI on a fair and level playing field. Oddly enough, the Emulex Fibre Channel adapter installed very quickly and was able to connect to drives and run performance tests using default settings in under two minutes. However, we spent several days reconfiguring systems, researching TOE and SNP, but ultimately iSCSI still performed very poorly in our tests. We not only found iSCSI to be lower performing, we also found it significantly more difficult to fully understand and configure properly as a SAN initiator.

Testing Notes

Test tools used for performance analysis:

Finisar Medusa Test Tools Version 2.0 x64. The Medusa tools are the only publicly available native 64 bit data integrity and benchmarking tools in the marketplace today. One of the best features of these tools is that they can compare every bit of data for full data integrity while still performing at benchmark speeds.

Iometer Version 2006.07.27 Windows 32 Bit (i386). Even though Iometer is now a decade old and only runs as a 32-bit application, it is still considered to be one of the premiere benchmarks in the industry. We utilized Iometer following the guidelines that many performance teams use to baseline database performance.

For power monitoring we ran the Intel Star Lake Server into a Watts Up Pro Power Meter. The Watts Up Pro incorporates sophisticated digital electronics that enable precise and accurate measurements in an easy-to-use format. State-of-the-art digital microprocessor design utilizes high-frequency sampling of both voltage and current measurements for true power.

The Intel Star Lake Server was directly connected to the Tyan workstations.
No Fibre Channel or Ethernet switches were placed in the I/O path.

TARGET SYSTEMS

- Tyan S2915
- 8x 1 GB 667 MHz DDR
- 2x Dual Core AMD Opteron Processors at 3.2 GHz
- BIOS Version/Date Phoenix Technologies Ltd. 2.03.2915, 2/8/2008
- Windows 2003 Server R2 SP2 x64 with full updates as of 6/3/2008

ADAPTER:

Chelsio S310E-SR-C
Running Starwind iSCSI Target with Integrated RAMDISK Version 3.5.4 x64 Version
Emulex LPe12002 FW 1.00a9 running the Third I/O Iris Software Version 3.3.B1
4 RAMDISK LUNs per port were created on both the iSCSI and Fibre Channel configurations for a total of 8 LUNs.

NOTE: These targets systems were chosen as they are extremely fast in terms of PCI Express bandwidth (7,000 MB/s observed) and I/O (600,000 observed) operations. They were equipped with the fastest shipping dual core Opteron processors and with NICs that claim the best target side performance in both the iSCSI and Fibre Channel industries. As our performance analysis is based on the Windows initiator system performance, we choose to run to two target systems so that they would provide for more than sufficient bandwidth and IOPS testing.

Mark Lanteigne was the primary author and test engineer for this paper. Mark is the founder of Third I/O Incorporated and has been involved in enterprise computing test and development since 1996.

THIRD I/O, Incorporated specializes in high-speed bandwidth and supercomputing technologies. Our founder and key employees are experts in the enterprise server, storage, and networking industries. For further information, please contact us at info@thirdio.com or www.thirdio.com

The analysis in this publication is based upon the testing configuration as noted in this document. Any modification to this environment will likely yield different results.