Performance Brief: Surpassing 1 Million I/O's Per Second from a Single Server Platform

July 16th, 2008

Introduction

The quest for higher server to data storage I/O transaction rates is the continual pursuit of the enterprise server and storage communities. Recently, there have been several extraordinary gains in both hardware and software technology that now allow for unprecedented levels of I/O transactions.

This performance brief describes Third I/O's enterprise level benchmark to realize the true I/O transaction performance potential of a single server platform connected to a single data storage device.

Overview

The term IOPS "eye ops" is well known in the system and storage performance communities. It stands for input output operations (aka transactions) per second. IOPS results are critical for database and supercomputing environments where the number of transactions per second is a foundation of a system's performance profile.

By definition, IOPS occur "on the wire"; a requirement is that they be sent to or from a data storage device. In other words, I/O accesses to OS cache or memory are not included in IOPS results. IOPS benchmarks generally focus on smaller block size test cases of meaningful I/O sizes. Some examples of these I/O sizes are:

- 512 byte—this is a standard disk sector size and is generally the smallest I/O size that can be sent to or from a storage device.
- 4 and 8 kilobyte—these two I/O sizes are generally considered the best real world I/O sizes to benchmark as they represent the most common memory page and database transaction sized measurements.

IOPS performance rates are a critical comprehensive system benchmarking metric. This is because high I/O transaction rates require several components of a system to work together efficiently in order to achieve respectable results. Small block I/O sizes can be very taxing on a system, both from a software and hardware perspective. IOPS testing has a higher overhead because every I/O is required to travel through the operating system's SCSI layers and I/O scheduler. In general, there is also a direct correlation between number of IOPS and the number of times that a storage controller will physically interrupt the CPU subsystem. With small block I/O benchmarking, the generation of more IOPS leads to significantly more accesses on the SCSI layer, OS I/O scheduler, and the number of interrupts sent to the processors. This makes IOPS analysis a great benchmark for examining the OS, the storage controller, and the host processors simultaneously.

New IOPS Friendly Technologies

Third I/O investigated several modern server and storage technologies in order to determine an optimal benchmark configuration. Prior to running our benchmark, it was our belief that the following technologies would be critical in order to achieve record breaking results.
These technologies included:

STORAGE CONTROLLERS: EMULEX 8 Gb FIBRE CHANNEL ADAPTERS WITH MSI-X

Third I/O has consistently found that Emulex's enterprise Fibre Channel host bus adapters (HBA's) have a profile of being fast, highly processor efficient, and allow for excellent multiple port scaling.

The latest generation of Emulex LPe12000 series adapters, now support MSI-X (Message Signaled Interrupts—extended) technology. MSI-X is a newly released industry standard that is an alternative to the less efficient line-based interrupt technology. Emulex supports MSI-X by incorporating the feature into its HBAs, thereby allowing enterprise users to enhance overall performance, reduce system overhead, lower interrupt latency, and improve host CPU utilization. In addition, Emulex has optimized MSI-X under Windows Server 2008 to perform exceptionally well on non-uniform memory access (NUMA) systems.

MSI-X avoids the problems that are sometimes seen when a device is forced to share an interrupt with a driver that has design, coding defects, or performance issues. Thus, using MSI-X contributes to both enhanced system performance and greater overall system reliability.

THE SERVER—AMD OPTERON BASED (NUMA)

One of the most noteworthy performance increases in x86 computing in recent years has been the inclusion of NUMA technologies in multiprocessor systems. AMD's Opteron processors with Direct Connect Architecture is unique in the fact that every processor contains a dedicated memory controller and can host system RAM on a CPU basis. This allows for an aggregate memory throughput that scales in a linear manner with every CPU added to the system. AMD NUMA systems have been the x86 performance leader in aggregate CPU to memory bandwidth.

As NUMA systems have excellent CPU to memory bandwidth, they also generally exhibit excellent PCI Express bandwidth potential. This makes them an excellent platform for high performance storage peripherals.

IOPS performance rates are eventually throttled by either high CPU utilization or the overall bandwidth capabilities of a system. With the latest Quad-Core Opteron processors, Third I/O believed that we could achieve excellent IOPS results, as Opteron processors have known characteristics of being capable of extraordinary CPU cycles while achieving superb bandwidth.

THE STORAGE SUBSYSTEM

In April of 2007, the Third I/O Iris Storage Solution was benchmarked as being the fastest Fibre Channel storage device. In these experiments, Iris exhibited 8 GB/s of full duplex bandwidth and over 778,000 IOPS. This benchmark was performed on Third I/O's Iris' previous generation 4 Gb Fibre Channel solution.

However, Iris has recently been upgraded to support 8 Gb/s Fibre Channel. Early test results have shown that Iris at 8 Gb has twice the bandwidth potential and more than 50% greater IOPS advantage over the previous 4 Gb version.

Prior to investigating this benchmark, it was Third I/O's belief that the Iris Storage Platform was capable of

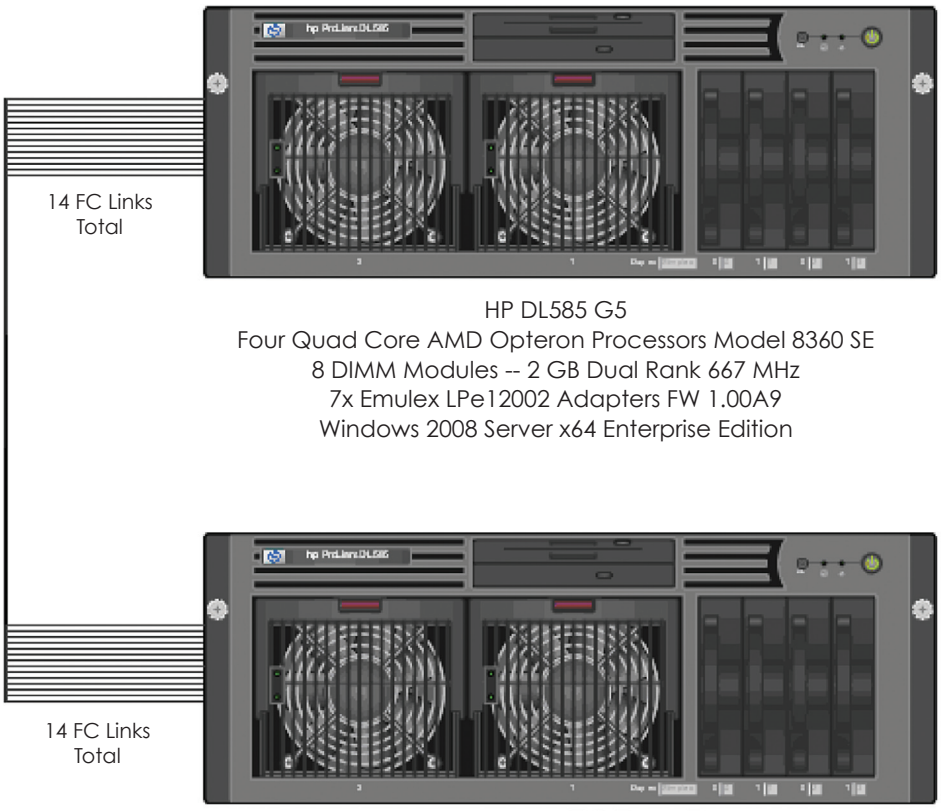
achieving over 1 million IOPS. This accomplishment would be nearly twice the performance of any advertised Fibre Channel storage product.

The Benchmark and Results

Before setting up this benchmark, Third I/O approached AMD and Emulex. We realized that the benchmark would require an extraordinary level of system hardware and lab resources. Fortunately, AMD and Emulex agreed to investigate this solution. Our combined resources allowed for optimal configuration and expertise in creating a solid benchmarking environment.

The systems chosen for the benchmark were two HP DL585 G5 enterprise servers; one to act as the Windows server platform and the second to operate as the Third I/O Iris Storage Platform. These systems allowed for seven PCI Express slots, as well as extraordinary PCI Express bandwidth.

Our benchmark configuration was laid out in the following manner:



HP DL585 G5
Four Quad Core AMD Opteron Processors Model 8360 SE
8 DIMM Modules -- 2 GB Dual Rank 667 MHz
7x Emulex LPe12002 Adapters FW 1.00A9
Windows 2008 Server x64 Enterprise Edition

HP DL585 G5
Four Quad Core AMD Opteron Processors Model 8360 SE
8 DIMM Modules -- 2 GB Dual Rank 667 MHz
7x Emulex LPe12002 Adapters FW 1.00A9
Third I/O Iris Version 3.3 B1

The top system in the diagram above was used as the Fibre Channel initiator. In storage, the initiator system is the system that sends commands to attached storage devices. The initiator system was installed with Windows 2008 x64 Enterprise Edition in order to achieve the full benefits of 64 bit technology, MSI-X, and NUMA. The benchmark application ran on this system was Iometer version 2006.07.27. Iometer, originally developed by Intel, is capable of generating a highly configurable, extremely intense IO workload from the server system.

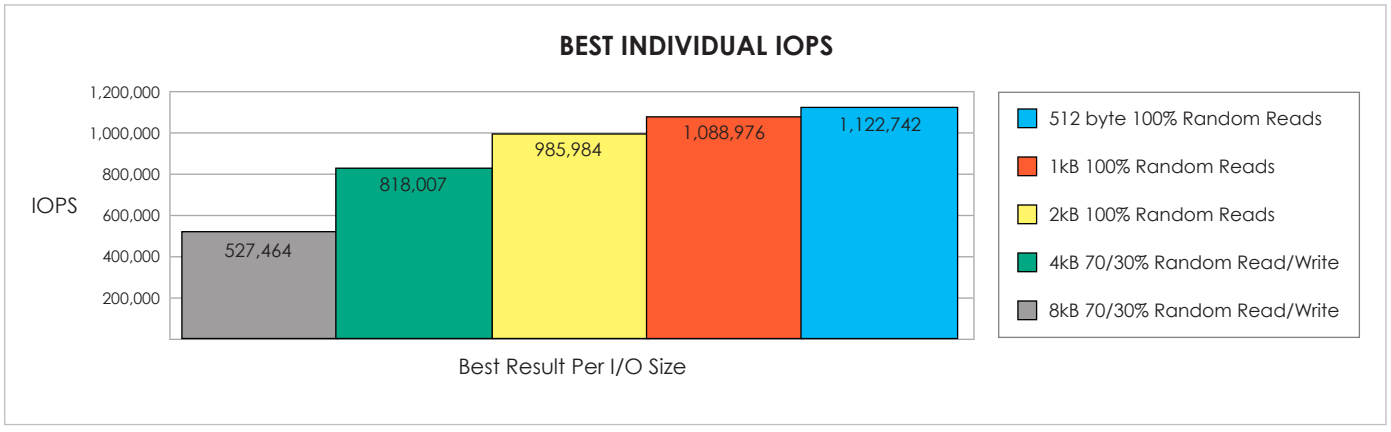
The second system was used as the Fibre Channel storage device. Using Third I/O's Iris software, we configured this platform as an 8 Gb Fibre Channel solid state disk. This system was configured as having one logical unit (LUN) per Fibre Channel port.

The benchmark configuration was basic. All Fibre Channel connections were directly connected in a point-to-point technology; no Fibre Channel switch was used in the experiment.

A default operating system installation was used with one notable exception. We used the Microsoft Interrupt Affinity Policy Configuration Tool to verify that every storage controller interrupt resided on a unique CPU core. Also, Windows Task Manager was used to set the processor affinity of Iometer to reside on the first 14 CPU cores. This number corresponded to the number of workers used by Iometer.

The Iometer setup consisted of 14 workers or one per every Fibre Channel port. An outstanding I/O or queue of 38 was set for each worker. Third I/O then tested I/O sizes of 512 byte, 1 KB, 2 KB, 4 KB, and 8 KB. 100% random reads and writes were investigated, as were 30/70% distributions of write/read I/O mixes to examine full duplex performance.

Benchmark Results:



As the goal of this benchmark was to determine maximum IOPS potential, Third I/O's analysis focused on the best-case benchmark results per I/O size.

We were pleasantly surprised to see that both 512 byte and 1 KB I/O size results comfortably surpassed the 1 million IOPS threshold. We recorded 1,122,742 and 1,088,976 IOPS on 512 byte and 1 KB I/O sizes respectively. To the best of our knowledge, this is the first time that 1 million IOPS have ever been observed on a Fibre Channel platform. The fact that we were able to observe these results on both the server and storage components makes these results even more impressive as this showcases excellent end-to-end performance.

In addition, the 4 and 8 KB I/O size results were also significantly above our expectations. The reason for this is because these results show a combination of extraordinary IOPS and bandwidth capabilities. For example, the maximum 8 KB IOPS result was 527,464 in a 70/30% Read Write mix. This is extraordinary in the sense that this level of I/O required bandwidth of 4,120 MB/s, while using an I/O size that is not bandwidth efficient.

Conclusion

The technologies of Third I/O, Emulex, and AMD formed the 'perfect storm' to realize the true potential of each, surpassing the 1 Million IOPS mark for the first time in the industry. In addition, these results were obtained on a 100% shipping hardware and software configuration. However, we fully expect that future enhancements to Third I/O's software, Emulex's hardware and software, and AMD's processors and chipsets will exceed these benchmark results and achieve significantly higher performance and scalability.

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.

 First Publication October 2008, Benchmark date July 16th, 2008.