THIRD I/O
INCORPORATED

Maximum IOPS and Port Scalability on Fibre Channel Storage

January 6, 2010

Key Findings on a Single Third I/O Iris Platform:

Maximum Results Exceeded 2.7 Million IOPS
Aggregate Bandwidth of over 10 GB/s (80 Gb/s) Observed
Database Benchmark Performance Exceeded 2.2 Million IOPS
Amazing Port Scalability — 98% Achieved with 24 Ports
Low CPU Utilization — 33% Average at Maximum IOPS

Introduction

In July 2008, Third I/O became the first company to demonstrate 1 million I/O operations per second (IOPS) between a single server platform connected to a single data storage device using Fibre Channel connectivity. And later that same year, Third I/O publicly showcased even better results at Supercomputing Conference SC08 with a 14-port Fibre Channel solution that achieved over 1.4 million sustained IOPS.

Over the course of 2009, a number of hardware and software innovations have allowed Third I/O's technology to scale significantly in both port count and IOPS. This performance brief describes our first benchmark of 2010, where we attempt to achieve even greater IOPS and Fibre Channel port count scalability.

Overview

The term IOPS ("eye ops") is well-known in the system and storage performance communities. It stands for input/output operations (a.k.a. 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 operating system (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 (K): 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.

Scalability has become a critical metric of performance computing as well. Port scalability refers to the amount of performance that is retained as additional ports are added into a system. The significance of port scalability has increased in recent years as virtualization has become commonplace in computing. Higher port scalability allows for greater computing efficiency and resource utilization, especially in intensive virtualized environments.

Port scalability can be analyzed using either bandwidth or IOPS as the baseline. However, it's important to note that IOPS can only scale to the allowable bandwidth of a system. For example, 1 million 8 K IOPS requires a bandwidth potential of nearly 8 GB/s (gigabytes per second) from the PCI Express busses. In other words, for an ambitious scalability test, the throttling of IOPS due to bandwidth limitations is a variable of critical importance.

New Technologies for Port Scalability and Increased IOPS

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 LPe12004 Adapters with MSI-X

Third I/O has consistently found that Emulex LightPulse® Fibre Channel Host Bus Adapters (HBAs) have a profile of being fast, highly processor-efficient and allowing for excellent multiple-port scaling.

One of the latest innovations from Emulex is the quad-port LPe12004 8 Gb/s Fibre Channel adapter. This advanced peripheral is a quad-port adapter that is designed for second-generation x8 PCI Express. It allows for maximum port density and high-performance operations, while only consuming the computing real estate of a single PCI Express slot.

As with all Emulex LPe12000 series adapters, the LPe12004 supports Message Signaled Interrupts eXtended (MSI-X) technology. MSI-X is a modern 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 — Intel® Nehalem-based (NUMA Architecture)

One of the most noteworthy performance increases in x86 computing in recent years has been the inclusion of NUMA technologies in multiprocessor systems. Intel's Xeon® Nehalem processors with QPI Architecture allows for extraordinary performance in dual-socket server systems. With Nehalem, every processor contains a dedicated triple-channel 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. Intel Nehalem NUMA systems currently are the x86 performance leaders in aggregate CPU to memory bandwidth in 1- and 2-CPU socket systems.

Intel Nehalem systems have excellent CPU to memory bandwidth; they also generally exhibit superb PCI Express bandwidth potential. This makes them an excellent platform for high-performance storage peripherals such as the LightPulse LPe12004 quad-port adapter.

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

The Storage Subsystem

Since 2007, the Third I/O Iris Storage Solution has been the Fibre Channel performance leader in both IOPS and bandwidth results. Our goal has always been to be the fastest and most scalable RAM-based storage platform in the industry. In addition, we have always achieved these results using standard shipping computer systems and peripherals. We never use experimental or prototype systems or software in our benchmarks.

For the storage side of our solution, Third I/O has been able to achieve these results by working directly with Emulex Corporation. Third I/O has been able to tap into Emulex's target mode technology using Emulex's TargetConnect™ Software Developer Kit (SDK). Emulex has the only Fibre Channel/Fibre Channel over Ethernet (FCoE) storage ASIC that provides a publicly available and well-documented SDK and support subsystem for storage development.

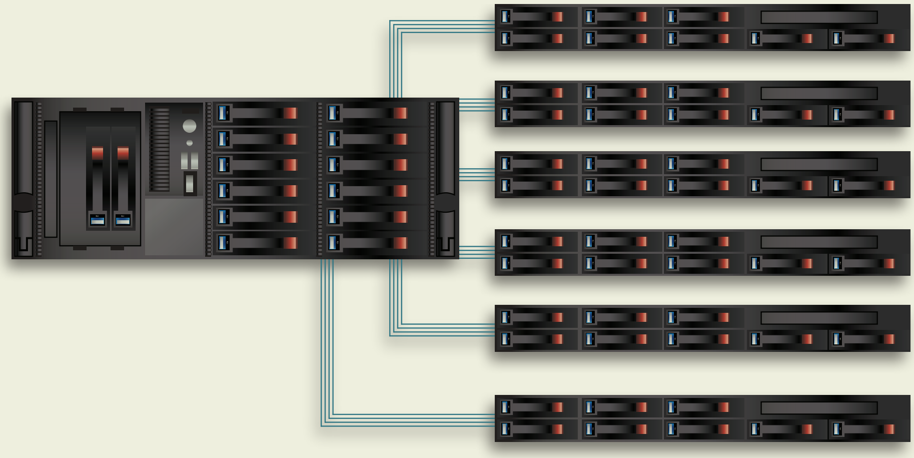
The TargetConnect SDK's support mechanisms have allowed Third I/O and Emulex engineers to work together on three primary features: performance, scalability and reliability. Working with the TargetConnect SDK and Emulex's engineers has proven to be a very successful formula for Third I/O's Iris Storage Solution.

The Benchmark and Results

The benchmarked system was the HP DL370 G6 enterprise server configured as the Third I/O Iris Storage Platform. The HP DL370 G6 was chosen because it allowed for nine Gen2 PCI Express slots, as well as extraordinary memory bandwidth.

Our benchmark configuration was laid out in the following manner:

TARGET SYSTEM -- HP DL370 G6
TWO QUAD CORE INTEL NEHALEM PROCESSORS MODEL X5560
SIX DIMM MODULES -- 2 GB DUAL-RANK 1333 MHZ
4X EMULEX LIGHTPULSE LPE12004 QUAD-PORT ADAPTERS FW 1.11A5
4X EMULEX LIGHTPULSE LPE12002 DUAL-PORT ADAPTERS FW 1.11A5
THIRD I/O IRIS VERSION 4.10X2
24 FIBRE CHANNEL DIRECT CONNECTIONS



CLIENT SYSTEMS – HP PROLIANT GEN2 PCI EXPRESS SYSTEMS
INTEL XEON QUAD CORE PROCESSORS
6X EMULEX LIGHTPULSE LPE12004 QUAD-PORT ADAPTERS FW 1.11A5
WINDOWS SERVER 2008 SP2 (FULL UPDATES AS OF JANUARY 5, 2009)

The primary system (HP DL370 G6) in the diagram to the left was configured as the Third I/O Iris target or storage system. Quad-port Emulex LPe12004 adapters were placed in the four x8 and x16 PCI Express slots. In addition, four Emulex dual-port LPe12002 adapters were placed in the four remaining slots closest to the processors.

The six systems that were connected to the Third I/O platform were all HP Proliant Intel Xeon based systems. Each client was populated with a single quad-port LPe12004 adapter for maximum port density. Six client systems were used against a single storage device to prevent any bottlenecks outside of the storage platform.

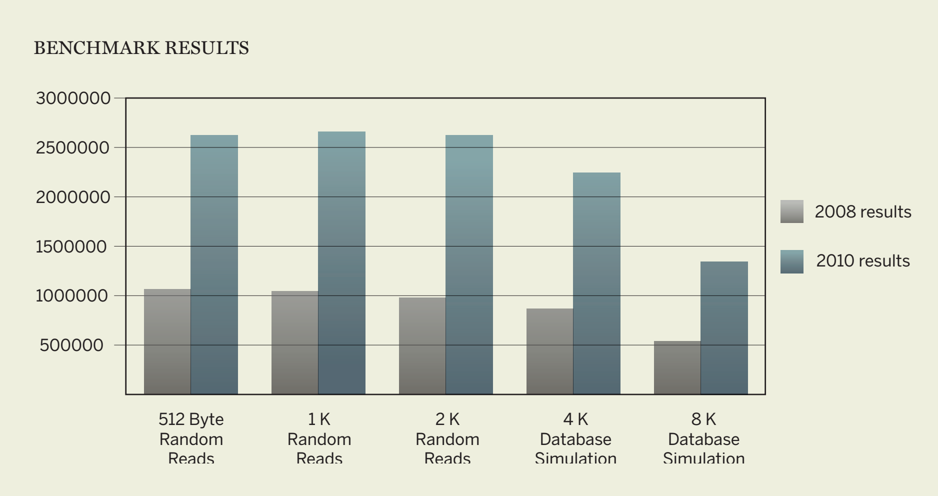
In storage, the initiator system is the system that sends commands to and from attached storage devices. The initiator systems were all running with Windows Server 2008 SP2. The benchmark application ran on these systems was Iometer version 2006.07.27. Iometer, originally developed by Intel, is capable of generating a highly configurable, extremely intense I/O workload from the server system.

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.

The Iometer setup consisted of 48 workers, or two per every Fibre Channel port. An outstanding I/O or queue of 64 was set for each worker. Third I/O then tested I/O sizes of 512 byte, 1 K, 2 K, 4 K, and 8 K. 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. As a point of reference, Third I/O is also including the details of our previous scalability and IOPS test from 2008.



We observed record setting IOPS of approximately 2.7 million in 512 byte to 2 K random reads. In addition, the Third I/O system averaged only 33% CPU utilization when running these benchmarks. Scalability in these tests showed an extraordinary 98%, which is near linear scalability whether running traffic on one or all 24 ports simultaneously.

Beyond random reads, most end users are interested in I/O traffic patterns that more closely resemble true database traffic. Toward this goal, we ran mixes of a 30%/70% random-write-read mix of both 4 K and 8 K I/O sizes. In the 4 K database access test, we observed over 2.2 million IOPS with an average target system CPU utilization of 38%. In the same test, using 8 K I/Os, we observed nearly 1.3 million IOPS while only taxing the target CPU at 28%.

NOTE: Both the 4 K and 8 K results are approaching the theoretical peripheral bandwidth limit of the HP DL370 G6. Respectively, these two tests ran at 90% and 97% of the best observed bandwidth from our entire test run. Even when running 512 K I/O sizes, we were just able to achieve 10 GB/s in read, write or combination read-write traffic.

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 Intel'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.

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.

This paper references various companies and products by their trade names. These companies claim these designations as trademarks or registered trademarks. Third I/O reserves the right to make changes or corrections without notice.

Reasonable efforts and care have been made to ensure the validity and accuracy of these performance tests. Testing Date January 6th, 2010. First Publication April 2010

Conclusion

The technologies of Third I/O, Emulex and Intel's Nehalem have showcased IOPS and scalability that are significantly higher than any previous single storage solution benchmark.