24 Lane, 3 Port PCI Express® Switch Performance Report

89PES24N3A

Notes

Overview

This document presents performance measurements and benchmarking results for IDT's 89PES24N3A 24-lane, 3-port peripheral chip, a member of IDT's PRECISE™ family of PCI Express Switching solutions. The PES24N3A has one upstream port and two downstream ports. Ports are up to 8 lanes wide. The switch is compliant with PCI Express (PCIe®) base specification revision 1.1.

The test vehicle for the PES24N3A is the evaluation board IDT89EBPES24N3A which hosts the PES24N3A. Accompanying the throughput performance metrics are descriptions and methodologies outlining the test setup and procedures.

The nature of tests and the equipment used for these tests varies significantly across the spectrum of tests performed. In the interest of readability and searchability the document is divided into various sections. Each section represents a single test suite that employs a single test setup. A single test suite is capable of high-lighting several features of the switch device under test.

Section I provides some insight into issues that can affect the performance of a PCIe device. This includes overhead derived from the protocol, as well as the architectural decisions made while implementing the PCIe device.

Section II describes the performance of the PES24N3A with Gigabit Ethernet endpoints attached to its downstream ports. Bidirectional performance comparisons with and without the PCIe switch in the traffic path are provided for both Windows and Linux environments. SmartBits™ SMB600 is used to generate controlled Ethernet traffic which is looped back between the GE NICs.

Section III provides a throughput report using PES24N3A with Fibre Channel (FC) traffic. Reads and writes to an array of disk drives are done with the IOmeter software tool. IOmeter is also used to to gather and analyze the performance data.

Section IV provides a throughput report using PES24N3A with Serial-Attached-SCSI (SAS) traffic. Reads and writes to an array of disk drives are done with the IOmeter software tool. IOmeter is also used to to gather and analyze the performance data.

Section V comprises a performance report for graphics traffic. The software tool 3DMark05 is used to generate and analyze graphics data through the PES24N3A.

Section VI demonstrates the switch behavior in a mixed endpoint scenario with a SAS HBA on one down-stream port and a dual GE NIC on the other.

Appendix A gives a brief introduction to the SmartBits traffic generator and analyzer and the SmartFlow™ test software package used in conjunction with this test equipment.

Appendix B is an introduction to the software tool called IOmeter that is used in generating some of the storage and networking test results presented in this report.

Appendix C offers a brief description of the software tool 3DMark05, which is used in benchmarking graphics applications.

Revision History

November 14, 2006: Initial version.

December 13, 2006: Content added and typos corrected.

January 12, 2007: Miscellaneous corrections.

August 27, 2007: Added multi-function testing.

SECTION I: PCIe Performance Basics

The PES24N3A primarily serves the purpose of high-performance I/O connectivity expansion in a typical system. Simply put, the PES24N3A uses one existing PCIe port in a system and offers two ports in its place. Given that nothing ever comes for free, it is presumed that the addition of a port has some "cost" associated with it in the form of real estate on the system board, power/heat, design complexity, support circuitry/devices (clocks, hot plug controllers, EEPROMs, power regulators, jumpers, etc.), signal integrity, or adverse effects on throughput/latency. All but the last item in this list are unavoidable to some extent. It is the impact on throughput and latency (system performance in general) that is the least intuitive to predict without a reasonable understanding of the system and switching device architecture, the usage model of the switching device, and some basic understanding of the PCIe protocol itself. In this section, some of these elements are introduced to the users of the PES24N3A, specifically those users who are new to PCIe and switching. Advanced users of PCIe and switches may skip the reminder of this section.

What Does Performance Mean

PCIe switch performance can mean different things to different users. The following is an introduction to some basic terminology.

"Raw bits" refers to the total number of bits that go through the switch in any given period of time, regardless of function, source, or destination. The PES24N3A is designed to handle 2.5 Gigabits per Second of raw throughput in each direction on each of its lanes. This results in (2.5 Gbps) x (2 directions) x 24 (lanes) = 120 Gbps of raw switching capacity.

"Switch throughput" is calculated as the useful bits passing through the switch per second after subtracting the 8b/10b encoding/decoding overhead from the total raw bits. Discussion of the 8b/10b mechanism is beyond the scope of this document. It is sufficient to note that two out of every ten bits passing across a PCIe link do not contribute to any meaningful user data and are, therefore, subtracted from the throughput measurement. It must also be noted that this overhead is a feature of the PCIe protocol itself and is not uniquely associated with a switch device per se. For the PES24N3A, the "switch throughput" becomes (120 / 10) x 8 = 96 Gigabits per second.

"Switch utilization" is the "switch throughput" less the overhead associated with the PCIe protocol infrastructure. Examples of this type of overhead traffic are TLPs containing no user data (messages related to interrupts, errors, hot plug, power management, vendor defined messages, etc.) and eight types of DLLPs (Ack/NAK, flow control, etc.). This overhead is variable in nature and can sometimes be fine-tuned to meet system requirements by modifying the switch settings, such as the ratio of ACK/NAKs to total packets, etc. In general, however, expect this overhead to be about 15% of switch throughput for the majority of real life systems. "Switch utilization" brings us one step closer to estimating how much user data goes through the switch in a given period of time, but there is one more overhead to consider.

Every data packet is preceded and followed by a variable number of bytes. These bytes include the frame K-code, sequence number, TLP header, ECRC, and LCRC. Once this "framing" overhead (see Figure 1) is deducted from the "switch utilization" number, the resulting performance metric is called the "switch efficiency".

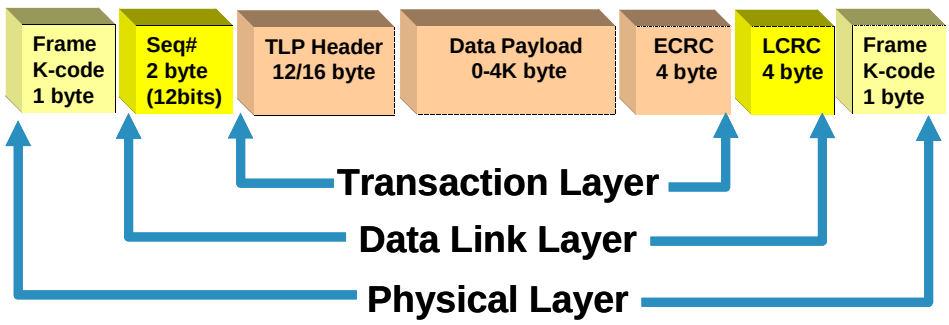

Figure 1 Framing Overhead in a Typical Transaction Packet

A different indicator of the performance of a switch is the switch “latency”, which is defined as the time spent by a bit within the switch from the moment it enters the switch to the moment it exits. The latency number, typically low hundreds of nanoseconds, can be affected by several parameters including, but not limited to, switch architecture, traffic pattern, state of the switch in terms of loading, width of the ingress port, and width of the egress port.

Impact of Architecture on Switch Performance

Two high-level architectural decisions which will have the biggest impact on switch performance are “how” the data is forwarded from one port to the other within a switch and “when” the data is forwarded. System designers must make these decisions at the very beginning of the design process. The architectural choices available for the “how to forward” question are: Shared bus, Crossbar, and Shared memory, or a hybrid of some combination of the above. The PES24N3A is implemented in a Crossbar style architecture. Explanation of these different types of switching architectures is beyond the scope of this document.

The architectural choices available for the “when to forward” question are: Cut-through (start forwarding a packet while it is being received) or Store and Forward (start forwarding only after an entire packet is received). The PES24N3A uses the Cut-through forwarding method.

There are several other micro-architectural features or implementation details of a switch that can also have noticeable impact on the performance of a switch. Discussion of the relationship between a feature choice and its impact on performance are beyond the scope of this document. It is relevant to note that several implementation details, such as the transmit retry buffer sizes, ingress buffer sizes, flow control mechanism, allowable maximum payload size (MPS), and controllable frequency of DLLPs including flow control updates and ACK/NACK, have an impact on the performance of the switch. Specifications related to these implementation details for the PES24N3A are found in the 89HPES24N3A User Manual, available by contacting IDT.

SECTION II: GE Throughput Measurements

The goal of this set of tests is to demonstrate the behavior of the PES24N3A with Gigabit Ethernet endpoint devices. Test results are obtained both with and without the PES24N3A device in the data path, so as to measure the impact of the switch on data throughput.

Hardware Setup

Following is a list of system components used for this test:

- Asus A8N-SLI Deluxe Motherboard
  - AMD Athlon 64 3200+ (64-bit)
  - 1GB of DDR-RAM)
  - PCIe slots - Two x16 and Two x1
  - Max Payload Setting 128B
  - Linux O/S: Fedora Core 3 - Linux Kernel 2.6.9 - 1.667 SMP
  - Windows O/S: XP
- IDT PES24N3A - PCIe upstream (x8), PCIe downstream (two x4 used)
  - Max Payload Setting 128B
- Ethernet Cards: Intel Pro/1000 PT Dual Port Server Adapter [x4 PCIe]

Figure 2 is a logical representation of the hardware setup used for GE throughput measurements with the PES24N3A. In this case, only one PCIe slot on the motherboard is used.

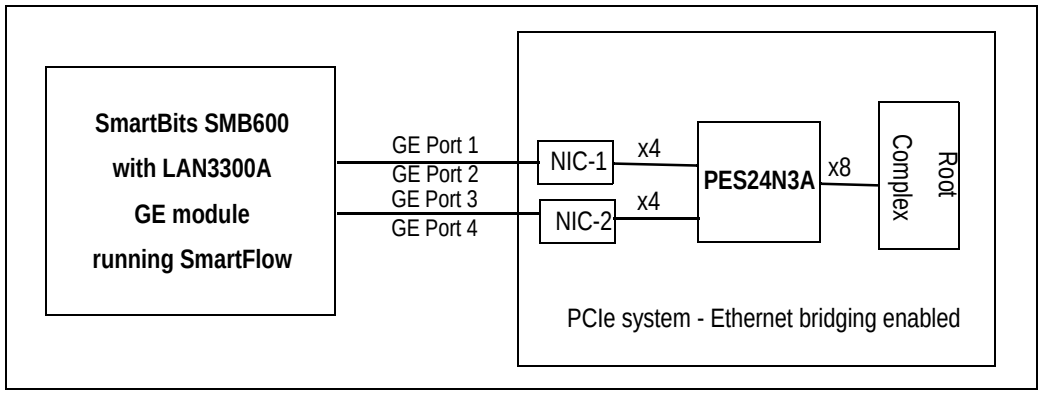

Figure 2 GE Throughput Measurement Setup with the PES24N3A

Figure 3 is a logical representation of the hardware setup used for GE throughput measurements without the PES24N3A in the data path. In this setup, two PCIe slots on the motherboard are used by the endpoints since the fan-out provided by the PCIe switch is no longer available. The GE NIC cards are plugged directly into the PCIe slots on the motherboard.

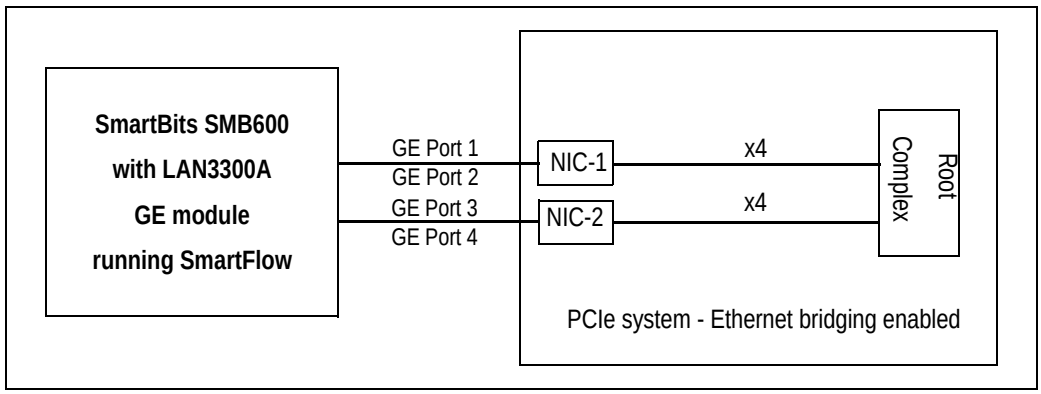

Figure 3 GE Throughput Measurement Setup without the PES24N3A

Software Setup

The SmartBits 600 Gigabit Ethernet traffic generator is controlled by the SmartFlow software package to generate and sink Ethernet traffic in a loopback mode. Details related to SmartBits setup can be found in Appendix A. The PCI Express-enabled server system is controlled by the operating system (Linux or Windows) and implements bridging of Ethernet traffic from one Ethernet port to another.

Test Procedure and Methodology

Each port of the SMB600 transmits Ethernet packets of predefined sizes targeted at the other port. Each packet transmitted by Port 1 travels through the corresponding NIC in the PCIe system, through the PCIe switch, if present, through the memory in the PCIe system, gets bridged over to the other NIC via the PCIe switch, if present, and returns to Port 2 of the SMB600. Packets starting at Port 2 of the SMB600 traverse the exact opposite path described above. GE ports 2 and 3 have the same relationship. Combined throughput measurements of these two flows for each packet size, with and without the PCIe switch in the path, are recorded in Tables 1 and 2 below. No data loss is permitted along the entire data path in either direction.

Results

	Throughput in Megabits/Second						
Packet size (bytes)	**64**	**128**	**256**	**512**	**1024**	**1280**	**1518**
Mbits/S Without PES24N3A	62	130	242	512	836	977	1103
Mbits/S With PES24N3A	62	107	197	377	709	850	962

Table 1 Throughput versus Ethernet Packet Size — Windows-XP

	Throughput in Megabits/Second						
Packet Size (bytes)	**64**	**128**	**256**	**512**	**1024**	**1280**	**1518**
Mbits/S Without PES24N3A	743	1245	2170	4000	4000	4000	4000
Mbits/S With PES24N3A	711	1231	2100	3513	4000	4000	4000

Table 2 Throughput versus Ethernet Packet Size — Linux

Analysis

The goal of this test is to show the effect of the PES24N3A PCIe switch on Ethernet traffic throughput. A quick review of the results reveals that the bridging performance of the operating system determines how stressful the test will be for the PCIe switch under test. It is clear that Linux offers better Ethernet bridging performance and, therefore, stresses the switch more than Windows-XP. It is clear that introduction of the PCIe switch in the datapath does not have any meaningful impact on the system throughput while offering an additional PCIe port to the system.

SECTION III: Fibre Channel Throughput Measurements

The goal of this set of tests is to demonstrate the behavior of the PES24N3A with FC storage controllers as endpoint devices connected to the PES24N3A downstream ports. Reads and writes to an array of disk drives are done with the IOmeter software tool. IOmeter is also used to to gather and analyze the performance data.

Hardware Setup

Following is a list of system components used for this test:

- SuperMicro X6DH8-G2 motherboard
  - Dual Intel Xeon 2.8 GHz
  - Intel E7520 (Lindenhurst) North Bridge
  - 1GB RAM
  - PCIe slots:Two x8 and one x4
- Windows 2003 Server
- IDT PES24N3A - PCIe upstream (x8), PCIe downstream (two x4 used)
- QLogic QLE2462 Dual Port FC HBA (x4)
- JMR Marlin FC-to-SAS JBOD Chassis (8 SAS drives each)

The FC controller cards were plugged into the downstream port slots of the PES24N3A evaluation board hosting the PES24N3A switch. A JBOD cluster was connected to each FC controller card. The upstream port of the PES24N3A is at the edge connector of the PES24N3A evaluation board and is plugged into a x8 port slot of the motherboard. In this way, the PES24N3A switch consumes one PCIe slot on the motherboard and creates a fan-out of two slots where two FC controller cards can be used.

Figure 4 and Figure 5 show the system configurations with and without the switch respectively.

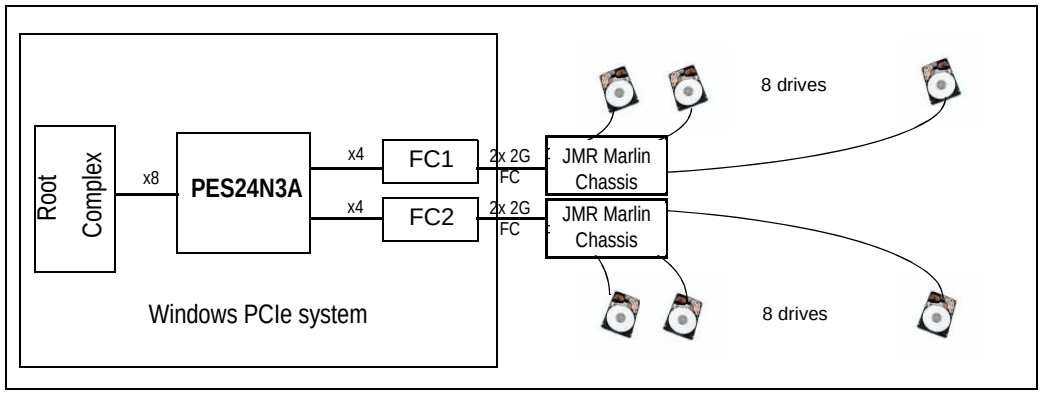

Figure 4 FC Throughput Measurement Setup with the PES24N3A

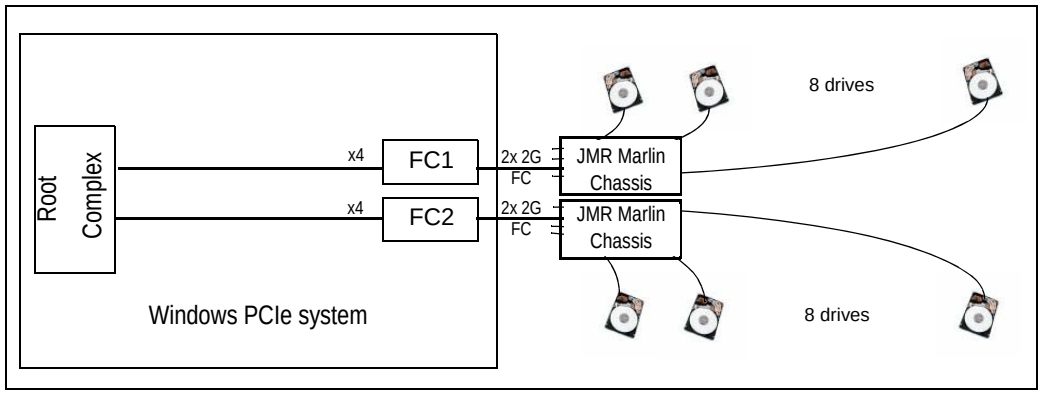

Figure 5 FC Throughput Measurement without the PES24N3A

Software Setup

Traffic was generated and measurements were taken using the IOmeter software package running on the PCIe host system under Windows-XP. Details related to IOmeter software package can be found in Appendix B. IOmeter version 2004.07.30 was used.

Test procedure and Methodology

IOmeter settings were as follows:

- **100% sequential or random "reads"**

  Sequential 310KB transfers set as 100% Read & 0% Write; 1 Manager (Dynamo) with 4 Workers and 16 Logical Drives (1168GB). Run time was 15 seconds.

- **100% sequential or random "writes"**

  Sequential 2MB transfers set as 0% Read & 100% Write; 1 Manager (Dynamo) with 4 Workers and 16 Logical Drives (1168GB). Run time was 15 seconds.

- **67% "reads" and 33% "writes"**

  Sequential 310KB transfers set as 100% Read & 0% Write with 67% of Access and Sequential 64KB transfers set as 0% Read & 100% Write with 33% of Access; 1 Manager (Dynamo) with 4 Workers and 16 Logical Drives (1168GB). Run time was 15 seconds.

- **100% "reads" one HBA-1, 100% "writes" on HBA-2**

  1 Manager (Dynamo) with Sequential 310KB transfers set as 100% Read & 0% Write with 2 Workers and first 8 Logical Drives (584GB), and Sequential 2MB transfers set as 0% Read & 100% Write with 2 Workers and last 8 Logical Drives (584GB). Run time was 15 seconds.

Results

Type of workload	Throughput without Switch (MB/S)	Throughput with Switch (MB/S)
100% Sequential or Random "read"	775	778
100% Sequential or Random "Write"	1384	1363
67% "read", 33% "write"	491	499
100% "reads" on HBA1 100% "writes" on HBA2	1006	985

Table 3 Fiber Channel Endpoint Performance data

Additional Results

The same set of experiments was repeated on a different motherboard with different type of storage. The motherboard was a Tyan S4885 with 4 AMD Opteron 854 CPUs and nVidia Krush K8-04 root complex chipset. The storage was high speed RAM disks emulating FC disk arrays (PMC Sierra 67xx-DE acting as storage). The results of this test are shown in Table 4 below.

Type of workload	Throughput without Switch (MB/S)	Throughput with Switch (MB/S)
100% Sequential or Random "read"	1400	1400
100% Sequential or Random "Write"	1340	1335
50% "read", 50% "write"	2200	2210

Table 4 Fiber Channel Endpoint Performance data on fastest server and storage

Analysis

System performance with the switch is almost identical to the system performance without the switch. Introduction of the PES24N3A in the path between the root complex and FC cards does not impact the system performance. Without sufficient amount of storage it is not possible to test the limits of switch performance. This was achieved by using RAM disks for highest performance. Switch withstood the stress.

SECTION IV: SAS Throughput Measurements

The goal of this set of tests is to demonstrate the behavior of the PES24N3A with SAS storage controllers as endpoint devices connected to the PES24N3A downstream ports. Reads and writes to an array of disk drives are done with the IOmeter software tool. IOmeter is also used to to gather and analyze the performance data.

Hardware Setup

Following is a list of system components used for this test:

- SuperMicro X6DH8-G2 motherboard
  - Dual Intel Xeon 2.8 GHz
  - Intel E7520 (Lindenhurst) North Bridge
  - 1GB RAM
  - PCIe slots:Two x8 and one x4
- Windows 2003 Server
- IDT PES24N3A - PCIe upstream (x8), PCIe downstream (two x8 used)
- LSI SAS3801E Dual SAS Controllers (x8)
- SAS Hard Drives (8 drives per card)

The SAS controller cards were plugged into the downstream port slots of the PES24N3A evaluation board hosting the PES24N3A switch. Eight drives were connected to each SAS controller card. The upstream port of the PES24N3A is at the edge connector of the PES24N3A evaluation board and is plugged into a x8 port slot of the motherboard. In this way, the PES24N3A switch consumes one PCIe slot on the motherboard and creates a fan-out of two slots where two SAS controller cards can be used. This is shown in Figure 6.

Tests were also performed without the PES24N3A; first with the SAS controllers using x4 port width and then using x8 port width. Changes to port width were accomplished using a slot reducer which physically connects only the lower 4 lanes to the link partner, leaving the remaining 4 lanes unconnected. Figure 7 shows the set up without the switch in the data path.

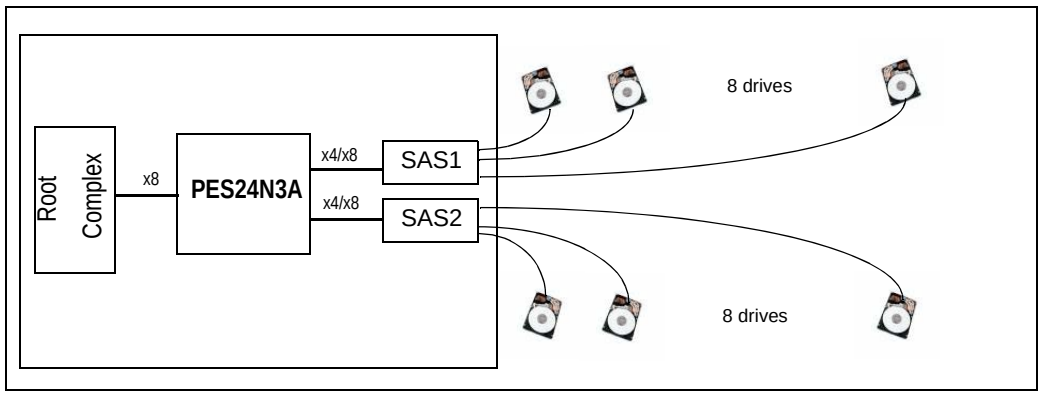

Figure 6 SAS Throughput Measurement Setup with the PES24N3A

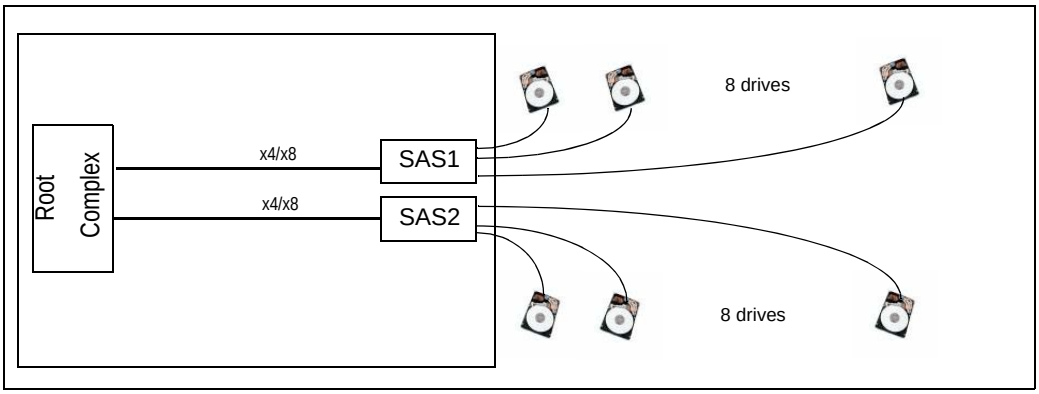

Figure 7 SAS Throughput Measurement without the PES24N3A using x4 SAS Connections

Software Setup

Traffic was generated and measurements were taken using the IOmeter software package running on the PCIe host system under Windows-XP. Details related to IOmeter software package can be found in Appendix B. IOmeter version 2004.07.30 was used.

Test procedure and Methodology

IOmeter settings were as follows:

- **100% sequential or random "reads"**

  Sequential 310KB transfers set as 100% Read & 0% Write; 1 Manager (Dynamo) with 4 Workers and 16 Logical Drives (1168GB). Run time was 15 seconds.

- **100% sequential or random "writes"**

  Sequential 2MB transfers set as 0% Read & 100% Write; 1 Manager (Dynamo) with 4 Workers and 16 Logical Drives (1168GB). Run time was 15 seconds.

- **67% "reads" and 33% "writes"**

  Sequential 310KB transfers set as 100% Read & 0% Write with 67% of Access and Sequential 64KB transfers set as 0% Read & 100% Write with 33% of Access; 1 Manager (Dynamo) with 4 Workers and 16 Logical Drives (1168GB). Run time was 15 seconds.

- **100% "reads" one HBA-1, 100% "writes" on HBA-2**

  1 Manager (Dynamo) with Sequential 310KB transfers set as 100% Read & 0% Write with 2 Workers and first 8 Logical Drives (584GB), and Sequential 2MB transfers set as 0% Read & 100% Write with 2 Workers and last 8 Logical Drives (584GB). Run time was 15 seconds.

Results

Type of Workload	Throughput without Switch SAS x4 (MB/S)	Throughput with Switch SAS x4 or x8 (MB/S)	Throughput without switch SAS x8 (MB/S)
100% Sequential or Random "read"	1634	1596	2387
100% Sequential or Random "Write"	1149	1169	1182
67% "read", 33% "write"	1472	1527	1642
100% "reads" on HBA-1 100% "writes" on HBA-2	1379	1926	2003

Table 5 SAS Performance Data with and without PES24N3A

Analysis

Introduction of the switch in the data path does not seem to impact the system performance adversely, especially when there is a balance in the upstream lanes and downstream lanes. In the x8 Read (endpoints to root direction) test, direct connection to the root complex shows significant advantage over introduction of a switch as a result of the direct connectivity over 16 lanes between the root and the endpoints.However, in real life scenarios where there is a mix of read and write transactions, the impace of the switch is negligible.

SECTION V: Graphics Performance

The goal of this set of tests is to demonstrate the behavior of the PES24N3A with graphics cards connected to the PES24N3A downstream ports.

Hardware Setup

Following is a summary of the main system components for this test:

- Asus A8N-SLI Deluxe Motherboard
  - AMD Athlon 64 3200+ (64-bit)
  - 1GB of DDR-RAM)
  - PCIe slots - Two x16 and Two x1
  - Max Payload Setting 128B
  - Linux O/S: Fedora Core 3 - Linux Kernel 2.6.9 - 1.667 SMP
  - Windows O/S: XP
- PES24N3A PCI Express Switch (3 x8 ports)
- WinFast PX6600GT TDH 128MB graphics cards (two used in SLI mode.)
  - Chipset - NVIDIA GeForce 6600 GT

Figures 8 and 9 show the test environment.

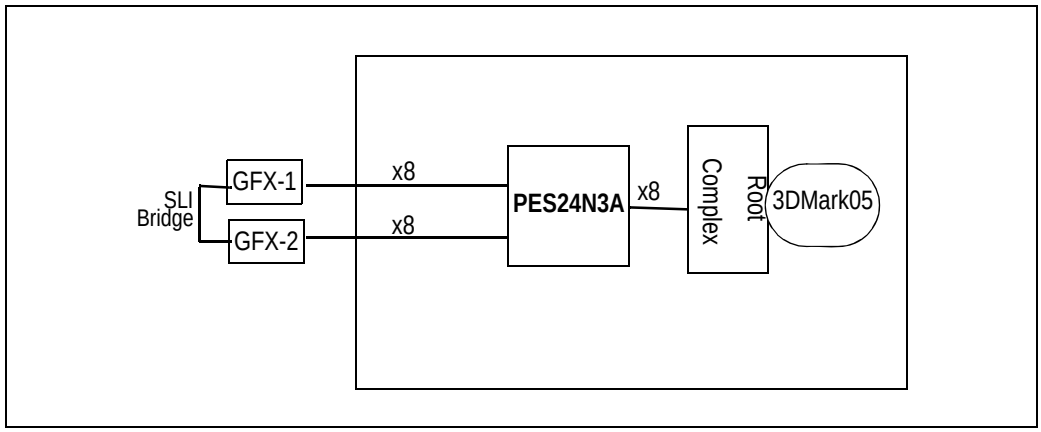

Figure 8 Graphics Setup Using 3DMark, with PES24N3A

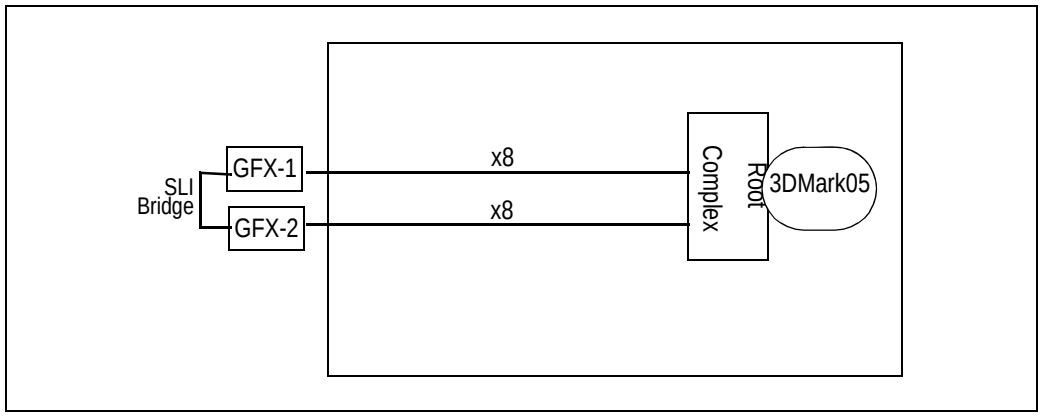

Figure 9 Graphics Setup Using 3DMark, without PES24N3A

Software Setup

The 3DMark05 software tool is used to generate the graphics and generate CPU scores with and without the PES24N3A. Please see Appendix C Introduction to 3DMark05 for more information on 3DMark05.

Figure 10 shows the 3DMark05 setup screen.

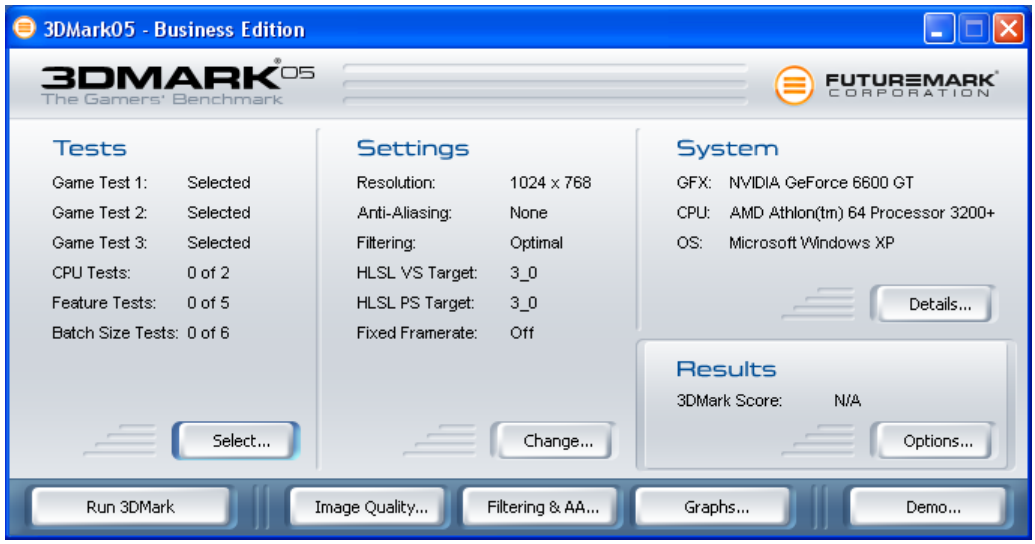

Figure 10 3DMark Software Setup Screen

Test procedure and Methodology

The 3DMark05 software tool runs three 3D games from the CPU to the graphics cards where they are rendered. If the PES24N3A is present, then the graphics traffic runs through the switch. We compare the results with and without the switch.

Results

	3DMark05 Score
without 24N3A	6466
with 24N3A	6382

Table 6 Graphics Performance with and without PES24N3A

Analysis

Introducing the PES24N3A in the path between the CPU and graphics cards results in a minimal decrease in graphics performance, just 1.3%. The PES24N3A is very efficient for graphics applications, therefore the additional PCIe slot comes at a very low performance cost.

SECTION VI: Mixed Endpoints - SAS and Dual GE

This section provides a throughput report using PES24N3A with mixed SAS and GE traffic. Reads and writes to an array of disk drives are done with the IOmeter software tool. IOmeter is also used to gather and analyze the performance data. Two different types of dual GE NICs were tested.

Hardware Setup

Following is a list of system components used for this test:

- IBM x3755
  - Dual Core AMD Opteron (Processor 8218, 2.6 GHz)
  - ServerWorks HT2100 chipset
  - 4 GB RAM
  - PCIe slots: One x16, 2 x8 (used for this test) and one x4
- Windows 2003 Server Standard Edition (SP1)
- IDT PES24N3A - x8 upstream, using two downstream ports configured for x4
- Broadcom BCM5715 Dual Port Gigabit Ethernet Server Adapter [x4 PCIe]

  (Also tested with Intel Pro/1000 PT Dual Port Gigabit Ethernet Server Adapter [x4 PCIe])
- LSI SAS3801E Dual SAS Controllers (x8, used in x4 mode through slot reducer)
- SAS storage (8 disks)

The SAS controller card was plugged into the downstream port (using x4 slot reducer) of the IDT evaluation board hosting the PES24N3A switch. The GE controller was plugged into the other downstream port slot. The upstream port of the PES24N3A is at the edge connector of the evaluation board and is plugged into a x8 port slot of the motherboard. In this way, the PES24N3A switch consumes one PCIe slot on the motherboard and creates a fan-out of two slots where two endpoint cards can be used. Figures 11 and 12 illustrate the system setups used for testing.

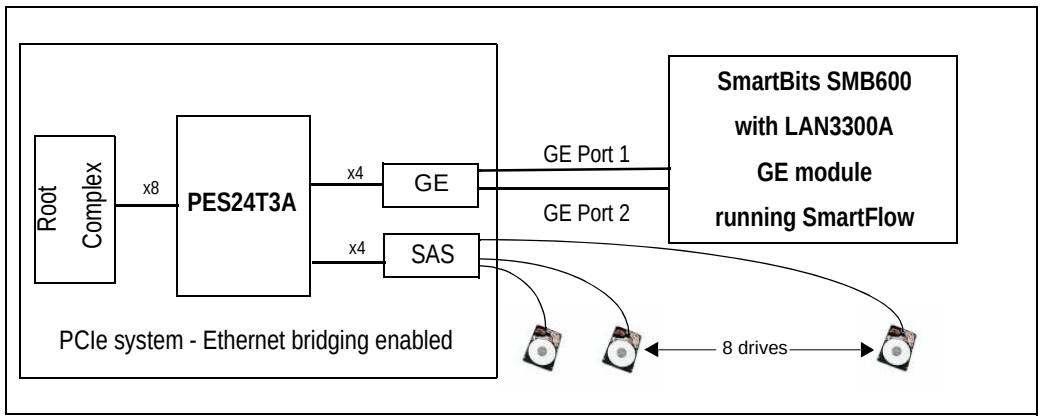

Figure 11 SAS Throughput Measurement Setup with the PES24N3A

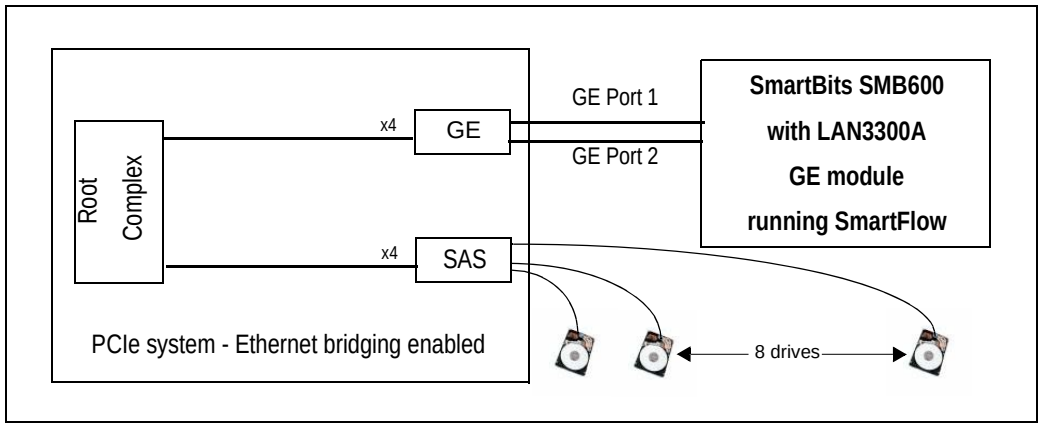

Figure 12 SAS Throughput Measurement without the PES24N3A

Software Setup

For the GE controller cards, the SmartBits 600 Gigabit Ethernet traffic generator is controlled by the SmartFlow software package to generate and sink Ethernet traffic in a loopback mode. Details related to SmartBits setup can be found in Appendix A. The PCI Express-enabled server system is controlled by the operating system and implements bridging of Ethernet traffic from one Ethernet port to another.

For the dual SAS controller card, each port was connected to 4 disk drives. Traffic was generated and measurements were taken using the IOmeter software package running on the PCIe host system. Details related to IOmeter software package can be found in Appendix B. IOmeter version 2004.07.30 was used.

Test procedure and Methodology

For the GE controller cards, each port of the SMB600 transmits Ethernet packets of predefined sizes targeted at another port. Each packet transmitted by Port 1 travels through the corresponding NIC in the PCIe system, through the PCIe switch, if present, through the memory in the PCIe system, gets bridged over to Port 2 via the PCIe switch, if present, and returns to Port 2 of the SMB600. Packets starting at Port 2 of the SMB600 traverse the exact opposite path described above. Combined throughput measurements of these two flows for each packet size, with and without the PCIe switch in the path, are recorded. No data loss is permitted along the entire data path in either direction. The tests were running by sweeping through the ethernet packet sizes while running the IOmeter program continuously to monitor any throughput changes worth recording on the storage throughput (SAS) side of the test. It was noted that the storage performance remained constant irrespective of the ethernet packet size changes.

For the SAS controller card, the IOmeter settings were as follows:

- **100% sequential or random "reads"**

  Sequential 2MB transfers set as 100% Read & 0% Write; 1 Manager (Dynamo) with 4 Workers and 8 Logical Drives (1168GB).

- **100% sequential or random "writes"**

  Sequential 2MB transfers set as 0% Read & 100% Write; 1 Manager (Dynamo) with 4 Workers and 8 Logical Drives (1168GB).

The results for the Broadcom dual GE NIC are presented in Table 7 while the results for Intel dual GE NIC are presented in Table 8.

Results

SAS throughput	Ethernet	Throughput in Megabits/Second						
100% Write (2MB)	**GE Packet size (bytes) -->**	**64**	**128**	**256**	**512**	**1024**	**1280**	**1518**
580 MB/S	**<-- Without switch -->**	54	87	144	242	425	636	566
667 MB/S	**<-- With PES24N3A -->**	76	121	200	369	622	889	1016
100% Read (2MB)	**GE Packet size (bytes) -->**	**64**	**128**	**256**	**512**	**1024**	**1280**	**1518**
573	**<-- Without switch -->**	65	99	200	397	580	777	847
572	**<-- With PES24N3A -->**	65	99	189	341	650	734	1100

Table 7 Broadcom Ethernet performance, with and without PES24N3A, for Mixed SAS and GE Traffic

SAS throughput	Ethernet	Throughput in Megabits/Second						
100% Write (2MB)	**GE Packet size (bytes) -->**	**64**	**128**	**256**	**512**	**1024**	**1280**	**1518**
667 MB/S	**<-- Without switch -->**	99	121	242	650	833	959	1044
667	**<-- With PES24N3A -->**	65	132	200	594	805	833	917
100% Read (2MB)	**GE Packet size (bytes) -->**	**64**	**128**	**256**	**512**	**1024**	**1280**	**1518**
572	**<-- Without switch -->**	110	155	166	411	622	1058	1269
572	**<-- With PES24N3A -->**	54	132	189	425	706	1016	1381

Table 8 Intel Ethernet performance, with and without PES24N3A, for Mixed SAS and GE Traffic

Analysis

Working with different types of PCIe endpoints in the same system creates some fascinating scenarios that beg a serious look into the system behavior by the system architect and the software developer. As shown by the Broadcom GE NIC test scenario above, the IDT switch actually helps the system throughput to a very large degree. This typically occurs as a result of endpoints not having sufficient flow control credits, a limitation which is masked when the PCIe switch comes into play. The switch is able to accommodate larger chunks of data from the root complex, thus leaving the root complex to service other tasks in its queue while the switch manages the limited flow with the ill equipped endpoint. The Intel GE NIC, a different design altogether does not demonstrate this limitation and thus the switch is not required to help the system performance. In the Intel scenario the switch does not adversely affect the system performance either.

Appendix A Introduction to SmartBits and SmartFlow

Note: Information contained in this section pertains to tools offered by a third party. The information is provided for the convenience of the reader and is not guaranteed to be complete or accurate.

The following document was used for reference while generating this text: Spirent Communications, Inc., 2005. “Introducing SmartFlow.” SmartFlow User Guide (5.0).

SmartFlow is a performance analysis tool to test Layers 2, 3, and 4 on Class of Service devices and networks built with Class of Service priority strategies. SmartFlow allows the setup of multiple flows of IP frames to simulate network traffic and measures latency, frame loss, and throughput. It presents results in charts and tables that include measurements for latency, frame loss, and standard deviation of flows. Results can be tracked by priority or by type of traffic to determine the effect a prioritizing Class of Service device has on the network.

Since our primary goal was to measure throughput through the PCI Express switch, we used the SmartFlow Group Wizard to simply generate flows, track them, and group them. SmartFlow is used in conjunction with a Spirent Communications SmartBits chassis and at least two SmartMetrics or TeraMetrics (or TeraMetrics-based) ports.

SmartFlow includes the following tests:

- Throughput
- Frame Loss
- Latency
- Latency Distribution
- Latency Snap Shot
- Smart Tracker

Below is a general description of the tests that were used for our measurements.

Throughput

Measures the maximum rate at which frames from flows and groups can be sent through a device without frame loss. A sequence of transmissions from one port on the SmartBits chassis to the other port on the chassis is setup. This traffic flows through the device under a test (PCI Express switch) which has Ethernet NICs connected to its downstream ports. An OS-based bridge is created between these two NIC, causing traffic entering one NIC to get forwarded to the other NIC. Bidirectional traffic is used, and each test consists of several sequential transmissions of Ethernet packets varying in size from 64 bytes to 1518 bytes with each type of packets getting transmitted in a single flow for several seconds at a time.

SmartFlow and SmartFlow Demos are available at **support.spirentcom.com**. Path: Self Service Tools –> Download Software Updates –> All Software –> SmartBits –> Applications or Demo. It is necessary to obtain a support account from Spirent to login to this site.

Appendix B: Introduction to IOmeter

Note: Information in this section pertains to tools offered by a third party. The information is provided for the convenience of the reader and is not guaranteed to be complete or accurate.

The following document was used for reference while generating this text: Intel Corporation: Iometer User's Guide December 16, 2003, which is available at:

http://cvs.sourceforge.net/viewcvs.py/*checkout*/iometer/iometer/Docs/Iometer.pdf

The latest version of Iometer, including the documentation, can be obtained from the Iometer project Web Site at the following URL: http://www.iometer.org/

An Iometer is an I/O subsystem measurement and characterization tool for systems. It is both a *workload generator* (that is, it performs I/O operations in order to stress the system) and a *measurement tool* (that is, it examines and records the performance of its I/O operations and their impact on the system). It can be configured to emulate the disk or network I/O load of any program or benchmark, or it can be used to generate entirely synthetic I/O loads. It can generate and measure loads on single or multiple (networked) systems.

An Iometer can be used for the measurement and characterization of:

- Performance of disk and network controllers.
- Bandwidth and latency capabilities of buses.
- Network throughput to attached drives.
- Shared bus performance.
- System-level hard drive performance.
- System-level network performance.

The Iometer tool consists of two programs, *Iometer* and *Dynamo*.

Iometer is the controlling program. Using the Iometer's graphical user interface, you configure the workload, set operating parameters, and start and stop tests. Iometer tells Dynamo what to do, collects the resulting data, and summarizes the results in output files. Only one copy of Iometer should be running at a time; it is typically run on the server machine in which the devices under test are plugged.

Dynamo is the workload generator. It has no user interface. At Iometer's command, Dynamo performs I/O operations and records performance information, then returns the data to Iometer. There can be more than one copy of Dynamo running at a time; typically one copy runs on the server machine and one additional copy runs on each client machine.

Dynamo is multithreaded; each copy can simulate the workload of multiple clients programs. Each running copy of Dynamo is called a *manager*; and each thread within a copy of Dynamo is called a *worker*. A system can simulate stress conditions by deploying several managers and workers. The worst case combination can be determined by experimenting and noting the results.

Once the IOmeter program has been started, a screen similar to that in Figure 13 is displayed. The "Results Display" tab displays performance statistics while a test is running. A user can choose which statistics are displayed, which managers or workers are included in a particular run of the test, and how often the display is updated in real time. A user can change the settings of all controls in the Results Display tab while the test is running. Changes take immediate effect.

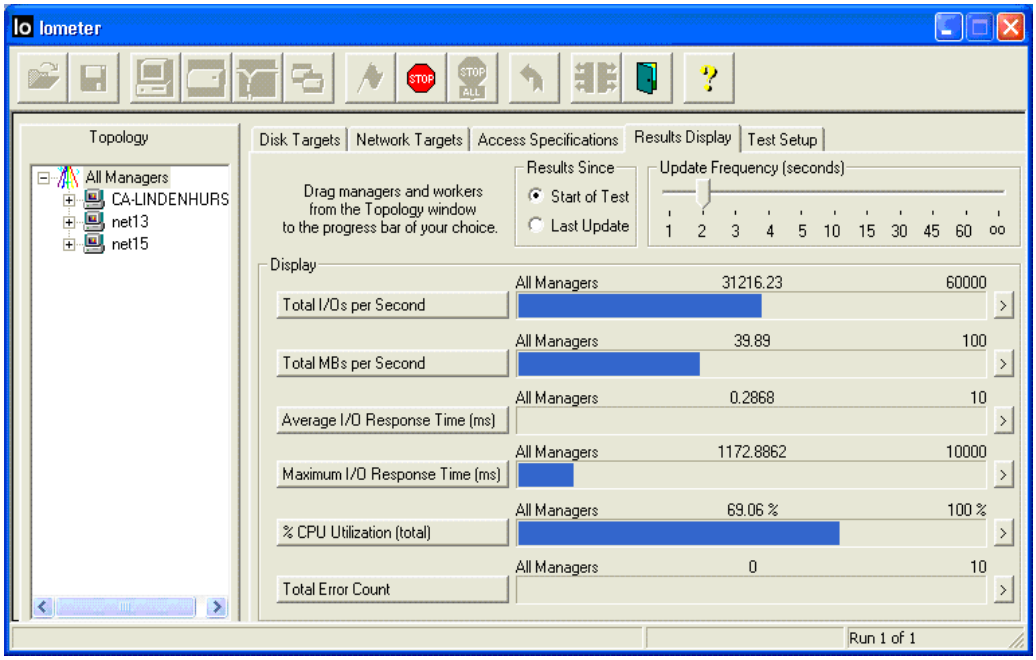

Figure 13 IOmeter main screen

As seen in Figure 13, six performance metrics can be displayed as bar charts on the program screen at one time. There are seven main categories of performance metrics. Within each main category, there are several sub-categories of more refined information. Any six out of these large number of metrics can be displayed and tracked as charts in real time. At the end of a test run, results of all sub-categories can be saved as tabulated text files in various formats.

The following is a list of the main metrics and sub-metrics within each main metric.

Operations per Second

- Total I/Os per Second
- Read I/Os per Second
- Write I/Os per Second
- Transactions per Second
- Connections per second

Megabytes per Second

- Total MBs per Second
- Read MBs per Second
- Write MBs per Second

Average Latency

- Average I/O response time (ms)
- Average Read response time (ms)
- Average Write response time (ms)
- Average Transaction time (ms)
- Average Connection time (ms)

Maximum Latency

- Maximum I/O response time (ms)
- Maximum Read response time (ms)
- Maximum Write response time (ms)
- Maximum Transaction time (ms)
- Maximum Connection time (ms)

CPU

- % CPU utilization (Total)
- % User time
- % Privileged time
- % DPC time
- % Interrupt time
- Interrupts per Second
- CPU effectiveness

Network

- Network packets per Second
- Packet Errors
- TCP segments retransmitted per Second

Errors

- Total Error Count
- Read Error Count
- Write Error Count

Appendix C Introduction to 3DMark05

Note: Information contained in this section pertains to tools offered by a third party. The information is provided for the convenience of the reader and is not guaranteed to be complete or accurate.

3DMark05 is a collection of 3D tests. These include a set of three *game tests*; these are the tests used to calculate the overall 3DMark05 score. The benchmark also includes a set of *CPU*, *feature*, *image* quality, and *batch size* tests. Each of these tests measures specific 3D-related functionality, but their result is not included in the overall score. They do not fall into the target usage, but are included to allow the user to evaluate these features.

For more detailed information on 3DMark05, paste this link into a browser:

http://www.futuremark.com/companyinfo/pressroom/companypdfs/3DMark05_Whitepaper_v1_1.pdf?m=v