12 Lane, 3 Port Gen2 PCIe® Switch Performance Report

Notes

Overview

This document presents performance measurements and benchmarking results for IDT's 89PES12T3G2 12-lane, 3-port PCI Express® Gen2 switch, a member of IDT's PRECISE™ family of PCI Express Switching solutions. The PES12T3G2 has one upstream port and two downstream ports. Ports are 4 lanes wide each. The switch is compliant with PCI Express (PCIe®) base specification revision 2.0.

The test vehicle for the PES12T3G2 is the evaluation board IDT89EBPES12T3G2 which hosts the PES12T3G2. 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 PES12T3G2 with Gigabit Ethernet endpoints attached to its downstream ports. Two dual GE NICs, one per downstream port, are used for this test.

Bidirectional performance comparisons with and without the PCIe switch in the traffic path are provided for Linux environment. SmartBits™ SMB600 is used to generate controlled Ethernet traffic which is looped back between the GE NICs.

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.

Revision History

October 30, 2007: Initial version.

SECTION I: PCIe Performance Basics

The PES12T3G2 primarily serves the purpose of high-performance I/O connectivity expansion in a typical system. Simply put, the PES12T3G2 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 PES12T3G2, 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 introduction to some basic terminology may clarify what ought to be important when selecting a switch for your system design.

Throughput

"Raw throughput" refers to the total number of bits that pass through the switch in a given period of time, regardless of function, source, or destination. The PES12T3G2 is designed to handle 5 Gigabits per second (Gbps) of raw throughput in each direction on each of its lanes. This switch is primarily designed for IO expansion (or fan-out) where the traffic flows to and from the root complex via the switch, and as such the maximum width of the upstream port indicates the maximum throughput that this switch can achieve. This results in (5 Gbps) x (2 directions) x 4 (lanes) = 40 Gbps of raw switching capacity. In reality, the switch is not required to "switch" this amount of data, as seen below.

PCI express data bytes undergo 8b/10b encoding. 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 stripped off before the data enters the switch core. Therefore, this 20% overhead must be deducted from the raw throughput that the switch must support in terms of actual switching capacity. For the PES12T3G2, the ideal "switch throughput" now becomes 80% of 40 Gbps, i.e. 32 Gbps, assuming simultaneous bidirectional traffic on all ports.

However, there is more overhead at play. Every payload packet (actual user data) is preceded and followed by a variable number of bytes as required by the PCIe protocol. These bytes include the frame K-code, sequence number, TLP header, optional ECRC, and LCRC. This is the "framing" overhead (see Figure 1).

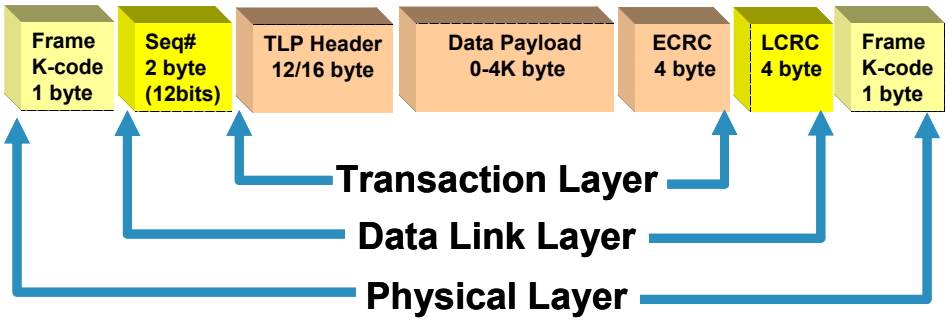

Figure 1 Framing Overhead in a Typical Transaction Packet

Figure 2 shows the effect of framing overhead on useful bandwidth for payloads at different PCIe link widths. A 20 byte overhead is assumed per payload packet for the purpose of this chart. This includes 1 byte of start of packet code, 2 bytes of sequence number, 12 bytes of TLP header, 4 bytes of LCRC, and 1 byte of end of packet code.

So, for example, on a x4 link, at 5 Gbps per lane per direction, raw bidirectional bandwidth is 40 Gbps. Upon removing the 8b/10b overhead, the useful theoretical maximum bandwidth available is 32 Gbps. Similarly, the theoretical maximum useful bandwidth for a x2 link is 16 Gbps and for a x1 link it is 8 Gbps.

To understand the calculations behind the chart shown in the figure, let us pick an example of a 64 byte payload packet on a x4 link to see how we come up with the corresponding data point on the chart. Total packet size with overhead becomes 84 bytes on account of the 20 byte overhead explained above. 32 Gbps (giga **bits** per second) of useful bandwidth is the same as 4 GBps (giga **bytes** per second). This is the same as 4000 MBps (mega bytes per second). In terms of packets, this means 4000/84 (i.e. 47.61) million packets. Payload bandwidth for 47.61 million packets is 47.61 multiplied by 64 bytes per packet, or a payload bandwidth of 24.38 Gbps. This is the 64 byte payload data point plotted on the x4 link chart in Figure 2). As seen in the chart, it is possible to achieve close to 32 Gbps (the theoretical maximum for a x4 link) under ideal conditions for payloads larger than 512 bytes.

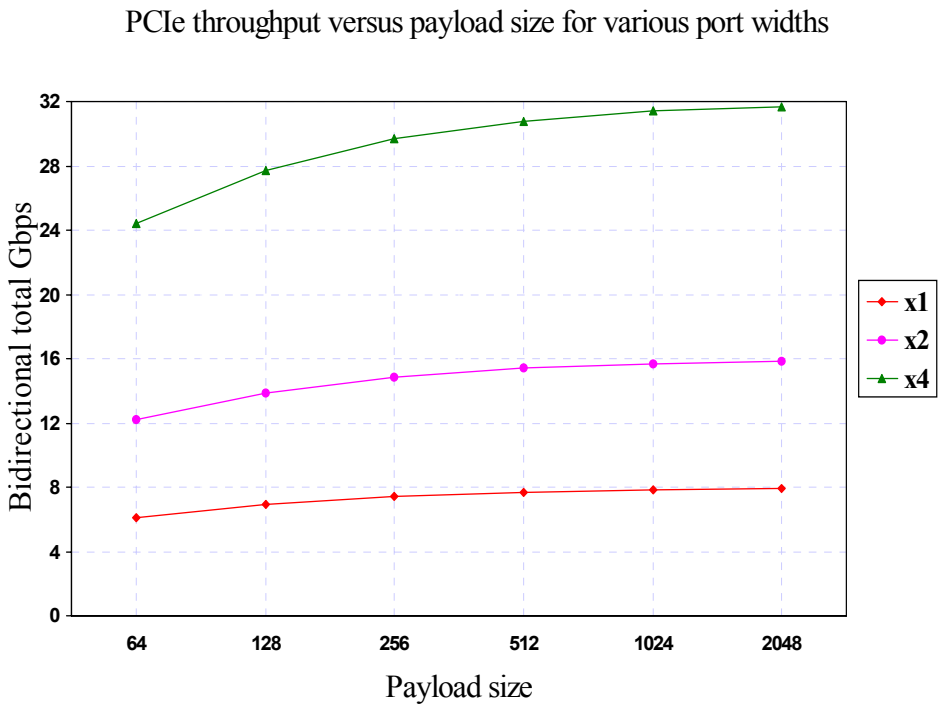

Figure 2 Effect of framing overhead on link efficiency

As calculated previously, at 64 byte payloads, the maximum throughput achievable on a x4 link is a bit over 24 Gbps. This means that out of the 32 Gbps useful bandwidth available, approximately 8 Gbps is spent on PCI Express framing overhead and approximately 24 Gbps on payload of 64 bytes payload per packet. This implies close to 75% efficiency on the “wire” (link). Since this framing overhead is constant irrespective of the link width, the wire efficiency is independent of link width in ideal conditions. For those who like to think in terms of wire efficiency as opposed to actual bytes or bits per second bandwidth, Figure 3 can help.

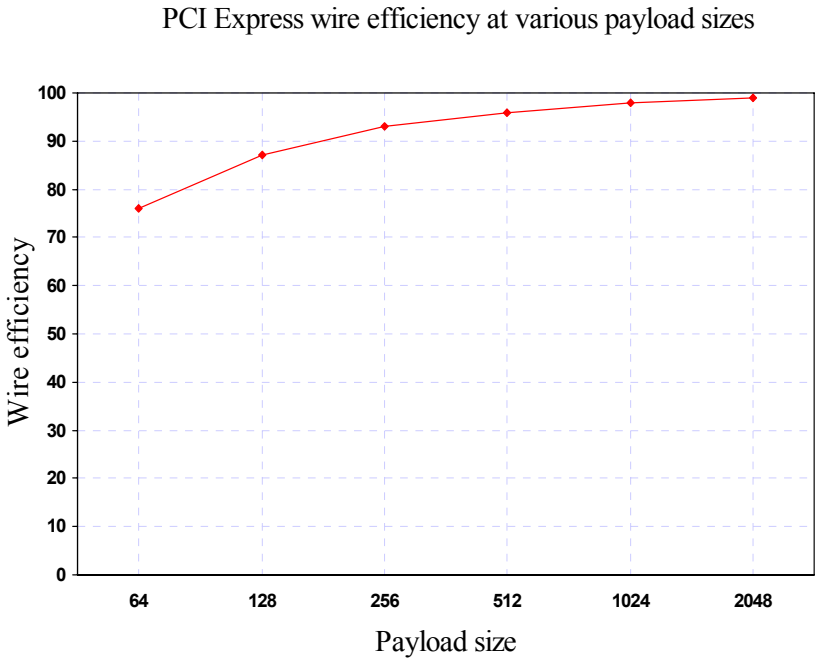

Figure 3 Data path efficiency of a PCI Express link

There is more overhead to be considered in addition to the framing overhead. “Switch utilization” is the “switch throughput” described up until this point, 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 or vendor defined messages) 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. Examples of such settings are, the ratio of ACK/NAKs to total packets, frequency of flow control updates, etc. In general, one can expect this overhead to be up to as much as 15% of switch throughput in several real life systems. So, for example, in a x4 link across the switch, for 64 byte payload size, starting from raw bits entering the switch as the base count, 20% is lost in 8b/10 encoding, 25% is lost in framing overhead and approximately 15% may be lost in other protocol overhead as described above.

A pictorial representation of the impact of this additional overhead is shown in Figure 4. This is similar to Figure 3 but also adds another line to the chart showing the effect of the additional DLLPs. The assumption here is that there are two DLLPs of 8 bytes each sent for every 4 TLPs. This equates to 16 bytes worth of DLLPs per 4 TLPs, or on average 4 bytes of DLLP overhead per TLP. This adds to the 20 bytes of framing overhead used previously as an example.

Clearly, the impact of fixed overheads such as these is minimized when the payloads are larger than 256 bytes.

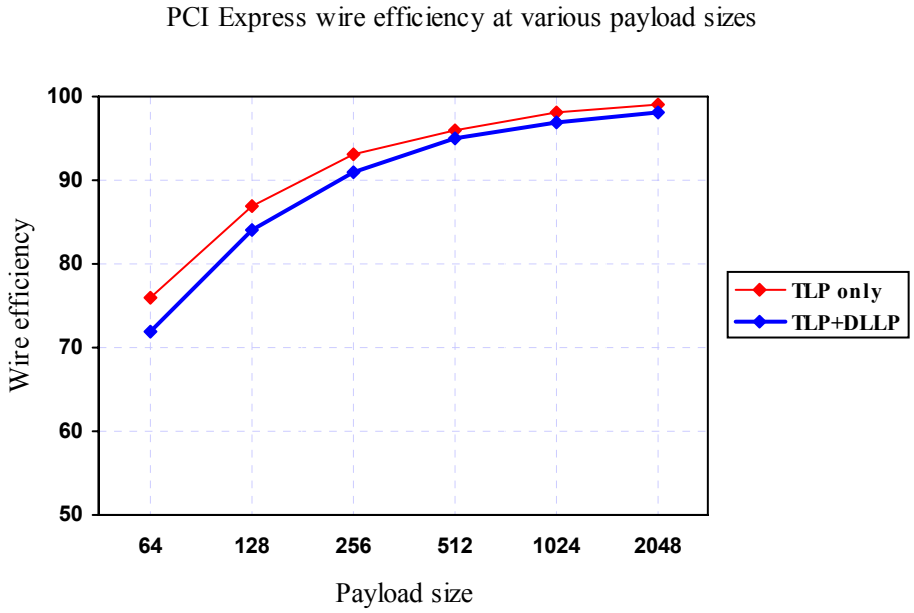

Figure 4 Effect of DLLP overhead on Data path efficiency of a PCI Express link

Latency

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.

It is crucially important to understand what matters and what does not matter when it comes to selecting a PCIe switch on the basis of latency. In general there is little correlation between the latency of a switch and the total throughput it can sustain across all its ports at the same time, which is the metric that truly matters for any system performance. An uninformed chase for a switch with the lowest latency number supplied by a switch vendor can inevitably lead to a wrong decision if no attention is paid to other performance metrics of a switch. Here is why...

Focus on the port width that matters to the application:

Some switch vendors tend to mislead customers by providing latency numbers which can only be realized when the switch is configured for the largest port width a switch can offer. In general, wider the port width, lower the latency. For a 16 lane switch, a vendor may offer a low latency number for data passing through the switch from an 8 lane ingress port to an 8 lane egress port. This information is worthless if your application requires data to move form a 4 lane or 1 lane port to a 4 lane or 1 lane port. The key is to focus on latency for the port widths actually required by your application.

Focus on simultaneous multi-port activity:

If your application requires data to flow simultaneously between several ports of the switch, what matters is the total latency experienced by the last packet within the set of packets attempting to pass through the switch at the same time. A 3 port switch may have up to 3 different packets trying to get through the switch at the same instance, one from each port. If a vendor provides the latency for one data transfer across an empty switch, that information is worthless in a scenario such as this. Insist on the total latency for the last bit of the data attempting to go across the switch in a fully loaded condition.

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 PES12T3G2 is implemented in a shared bus 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 PES12T3G2 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 PES12T3G2 are found in the 89HPES12T3G2 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 PES12T3G2 with Gigabit Ethernet endpoint devices. Test results are obtained both with and without the PES12T3G2 device in the data path, so as to measure the impact of the switch on data throughput. Dual GE NICs are attached to each of the three downstream ports.

Bidirectional performance comparisons with and without the PCIe switch in the traffic path are provided for the Linux environment. SmartBits™ SMB600 is used to generate controlled Ethernet traffic which is looped between the three GE NICs.

Hardware Setup

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

- Intel X38 Express chipset based motherboard
  - 2 available x16 PCIe Gen2 slots
  - 1 GB DDR3 Memory
  - Pentium 4 Processor, 3 GHz, 800 MHz FSB
- Fedora Core 7 - Linux Kernel
  - Broadcom GE-NIC Linux driver
- IDT PES12T3G2 - x4 upstream, two x4 downstream ports used on account of GE-NIUC availability
  - Max Payload Setting 128 bytes
- Broadcom Dual GE NICs - PCIe Gen2 interface to host
- MPS (Max Payload Size) set to 128 bytes

Figure 5 is a logical representation of the hardware setup used for GE throughput measurements with the PES12T3G2.

NICs 1 & 2 are plugged into x4 downstream port slots of the PES12T3G2 evaluation board (89EBPES12T3G2) hosting the PES12T3G2 switch. The upstream port of the PES12T3G2 is at the x4 edge connector of the PES12T3G2 evaluation board and is plugged into a x4 port slot of the motherboard. The PES12T3G2 switch uses one PCIe slot on the motherboard and creates a fan-out of three slots where the GE NIC endpoints can be used in this system.

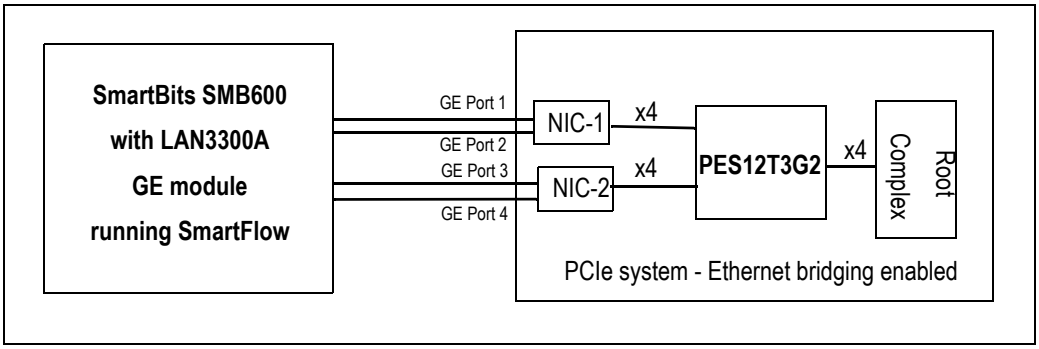

Figure 5 GE Throughput Measurement Setup with the PES12T3G2

Figure 6 is a logical representation of the hardware setup used for GE throughput measurements without the PES12T3G2 in the data path. In this setup, three 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.

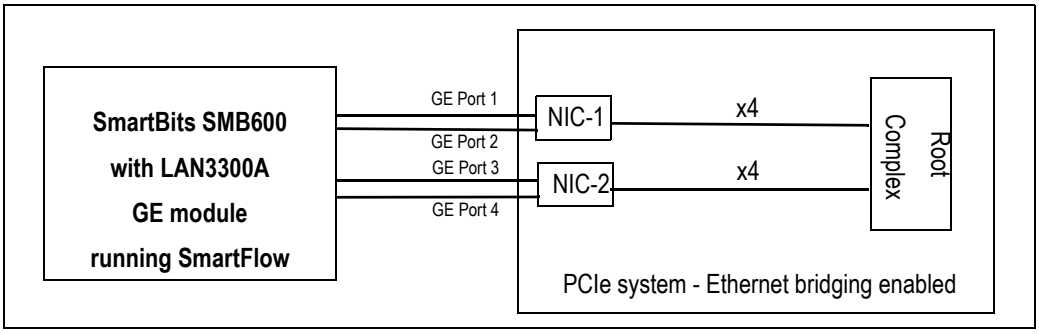

Figure 6 GE Throughput Measurement Setup without the PES12T3G2

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 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. The same occurs on another pair of the GE ports. Combined throughput measurements of these four flows for each packet size, with and without the PCIe switch in the path, are recorded in Table 1 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 PES12T3G2	197	355	681	1328	2509	3156	3691
Mbits/S With PES12T3G2	197	355	653	1328	2509	3156	3662

Table 1 Throughput versus Ethernet Packet Size — Linux

Analysis

The goal of this test was to show the effect of the PES12T3G2 PCIe switch on Ethernet traffic throughput. It should be noted that given that the upstream port is a x4 wide port capable of close to 16 Gbps throughput, the limited bandwidth offered by the endpoints does not come close to stressing the switch. With these limitations on the ability to stress the switch, it can however be proven that the IDT switch does not present a bottle-neck to the wire-speed ethernet traffic. The switch is able to sustain the maximum throughput allowed by the endpoint and the operating system.

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.