Test Electronics for a Multi-Gbps Optical Packet Switching Network

C.E. Gray[1], O. Liboiron-Ladouceur[2], D.C. Keezer[1], K. Bergman[2]
Georgia Institute of Technology[1], Columbia University[2]

Abstract

In this paper we present the design and performance characteristics for a custom test system developed to characterize a DWDM optically-routed packet switching network (called "Data Vortex"). The existing demonstration system supports aggregate data rates of 20 to 32 Gbps using 8 optical payload wavelengths each running at 2.5-4.0 Gbps. Several other optical wavelengths are used to transmit a source-synchronous Clock, Frame, and eight Routing address bits. All of the signals are transmitted in a parallel 25.6ns optical burst (packet). Switching nodes within the Data Vortex decode the optical routing bits in real-time and direct the packet to its intended destination. Unlike traditional switching networks, the Data Vortex nodes are switched "on the fly" by the routing information contained within the optical packet itself (rather than by a central control system).

1. Introduction

With the appearance of faster off-chip communication options such as PCIexpress and HyperTransport, very large-scale shared-memory supercomputers can be potentially constructed from commercial off the shelf processors. Such parallel systems, consisting of hundreds or thousands of processors, are most often bottlenecked with long latencies and delays on interprocessor or processor to memory messages [1]. The interconnection overhead and latency from conventional networking of the individual elements limits the performance benefits and scalability of the overall system.

Experimental all optical packet switching networks are a potential solution to this problem, minimizing end-to-end propagation latency as a result of minimizing or eliminating the various optoelectronic conversions in a mixed format system or buffering in electrical systems. Further delays can be reduced by leveraging dense wavelength division multiplexing (DWDM) to transmit an otherwise long sequence of data in a much shorter burst across a number of parallel wavelengths, decreasing the time required to transmit the data while also increasing the overall transmission capacity [2].

However, there are some fundamental challenges to interfacing a high-speed off-chip interconnection bus like PCIexpress to an optical packet switched network, which is the focus of an ongoing joint project between the Georgia Institute of Technology and Columbia University. Recently developed test electronics are intended to serve as a transparent bridge for tunneling PCIexpress packets across the network and provides inline test capabilities to evaluate and characterize the system as a whole. The current test electronics are designed to convert a single lane of PCIexpress traffic into a collection of eight parallel wavelengths at 2.5-4.0 Gbps each, as well as the corresponding Clock, Frame, and Routing signals included in the packet. This results in an aggregate data rate of 20-32 Gbps. Additional payload signals could be handled on other parallel wavelengths, increasing the data rate even furtther..

2. Communications across the Data Vortex

The optical packet switching (OPS) network being utilized for this project is based on the data vortex topology which is being developed at Columbia University [3]. The data vortex topology consists of concentric cylinders, each made up of circulating optical fibers and periodic switching nodes. The nodes are designed to respond to particular wavelengths which carry the routing information bits within the packet. Depending on the logic value of the routing bit, the data can be directed to a different cylinder, or continue circulating within the present cylinder. Optical packets are injected at special nodes on the outer cylinder, and eventually exit at destination nodes on the inner-most cylinder.

In the intended application, Tx and Rx optical fibers connect between the inner and outer cylinder ports and high-performance computers (which are bridged to the network through a bidirectional PCIexpress link included in the test electronics). This allows any computer to send/receive messages to/from any other computer in a low-latency manner.

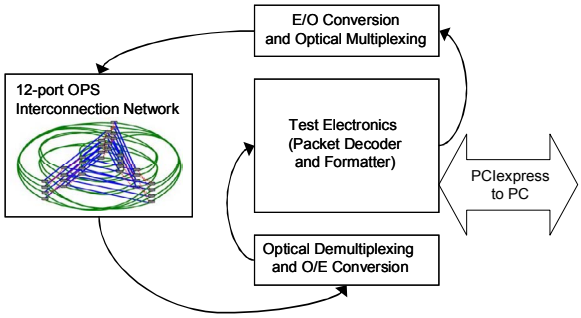

Figure 1. System topology.

Because the system is still under development, testing requires both a source of packet data (for input to the outer cylinder nodes) as well as a way to analyze the signals output from the inner cylinder. With this in-mind, we have developed a custom test system shown in Fig. 1. The payload, Frame, and routing signals are produced electronically, using either signals coming from a source computer or synthesized on-board, and converted to a range of wavelengths. The Frame and routing signals are used within the OPS network for routing decisions while payload wavelengths are transparently passed through the system without being processed. This architecture eliminates the need for complex optical buffering, and the parallel presence of the routing information eliminates the need to further process the packet for routing decisions. These features combine to reduce the latency and time of propagation of the packets considerably.

1-4244-0665-X/06/$20.00 ©2006 IEEE

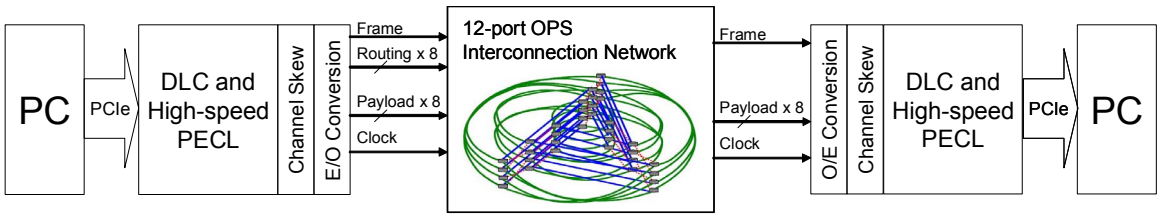

Figure 2. System data flow

While Figure 1 shows the system looping back upon itself, all communications across the OPS network are one way as shown in Figure 2. A transaction originating at a personal computer is received by the test electronics, formatted into a time-aligned packet, converted to optical wavelengths and injected into a network input port. The packet circulates to the appropriate destination port where the Frame, Payload, and Clock signals are converted back to electrical signals, deskewed (to partially compensate for any chromatic dispersion that may have occured within the Data Vortex or part variation between channels), processed by the test electronics and transmitted to the destination personal computer (or similar system, such as a high-speed memory node).

If a response is required, such as in a memory-read request, a similar but independent transaction would occur with the destination and source nodes reversed. For testing purposes, the same board is often used as the source and destination but such a configuration would be unlikely in the target application. The test system has many built-in features (controlled through a standard USB port) that allow us to adjust performance parameters (such as channel-to-channel skew) for transmission and to analyze the patterns and relative timing of the returned signals from the Data Vortex. Timing accuracy, in particular, is a major concern since it must be controlled on a picosecond scale.

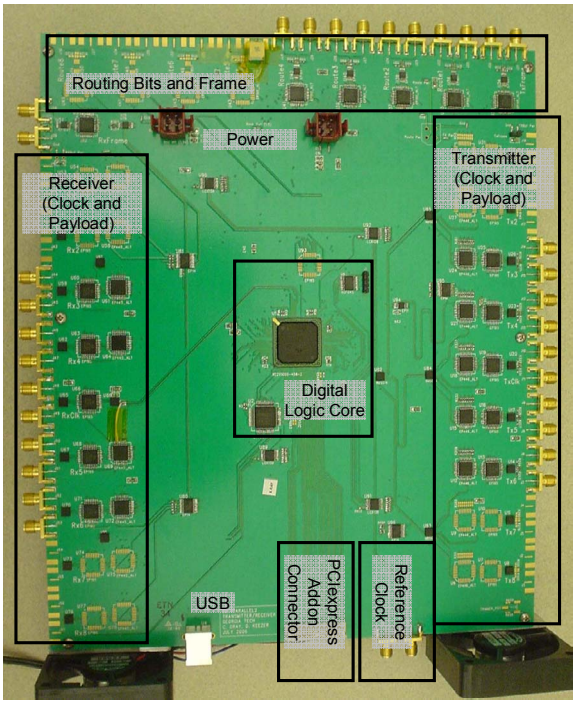

Figure 3. Photograph of the test electronics

3. Test Electronics

Figure 3 shows a photograph of the custom test electronics. The board is generally similar to an earlier design previously reported in [6] [7], but has been upgraded to nearly double the previous high-speed channel capacity. The central element is a CMOS based FPGA, flash configuration PROM, and USB connector collectively referred to as a digital logic core (DLC) [4] [5]. This DLC, in conjunction with an addon card not pictured, allows for the interconnection of the board to a PCIexpress expansion port on a personal computer. Short data packets recieved from the computer are buffered briefly whithin the FPGA before being formatted into an eight payload wide parallel packet that iw then injected into the Data Vortex. Each channel can be independently time-skewed to assure proper alignment or to intentionally stress the system for timing-margin characterization. In the absence of a personal computer PICexpress interface, or to support additional tests the FPGA can independantly synthesize a variety of test patterns.

The PECL circuits for the transmitting and receiving functions can be seen surrounding the DLC components. The high-speed output data channels, including the source clock, are to the right, while the slower speed frame and routing signals are at the top. Input logic for the frame and high-speed data/clock is located to the right and the RF clock enters from the bottom. Flexible coaxial cables and SMA connectors are used to interface the differential PECL signals to removable electro-optic modules (not shown).

Every outgoing and incoming signal (including all payload data and control) has the capability for the addition of up to 10ns of additional programmable delay in 10ps steps. This delay can be used to align outgoing channels, deskew incoming signals, or simulate various sources of system timing inaccuracy.

4. Packet Composition

The current test electronics and the earlier version have been used to evaluate various packet structures that are compatible with the OPS network and the packet structure shown in Figure 4. Eight data signals (each 32 bits in length)

are decoded from the PCIexpress packet or synthesized locally. These are precisely aligned in time with a source synchronous reference clock which is used to sample and recover the data at the destination. The length of this clock signal is slightly longer due to the setup and pipeline flush requirements of the deserializer used in the receiver. The Frame bit signals when the data is valid and the eight header channels carry the routing address data, which is used within the data vortex along with the frame to transparently route the message to the desired port.

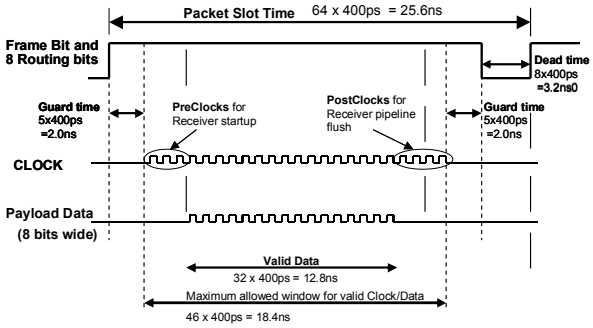

Figure 4. Basic packet structure

The individual fiber-optic interconnections of the data OPS network are sized to match a packet injection once every 25.6ns or 64 bit periods. The extra 32 bits of length relative to the embedded data accommodate a setup time before the data, a pipeline flush after the data, and guard times around either edge. The guard times are required due to the routing nature of the data vortex. The network does not buffer packets, instead deflecting them away from occupied segments. A virtual buffer is created by allowing a packet to circulate around a cylinder until the exit port or routing path is available. A side effect, however, is a slight trimming at the extreme edges of the packet as each routing decision is made. The guard time is provided for this purpose.

However, by comparison the length of a PCIexpress packet is quite large, with even the relatively short 128ns packet used for this project being five times larger than the slot injection time. Due to the limited payload space within a network packet, this information must be distributed across eight parallel data channels, effectively time compressing the signal to fit within the available data window (Figure 5). The packet is further augmented with a ninth channel that is utilized as a sampling clock at the destination. Routing information is extracted from the PCIexpress packet and translated into the appropriate routing bits used within the OPS network itself. Due to the decoding necessary to process the data, extract the routing information, and map the data onto the eight channels, the signal must be buffered temporarily, increasing slightly the amount of latency. Once the packet is passed through the E/O conversion modules, no additional processing or buffering occurs to the signals, which *would* be required in an electrically based network.

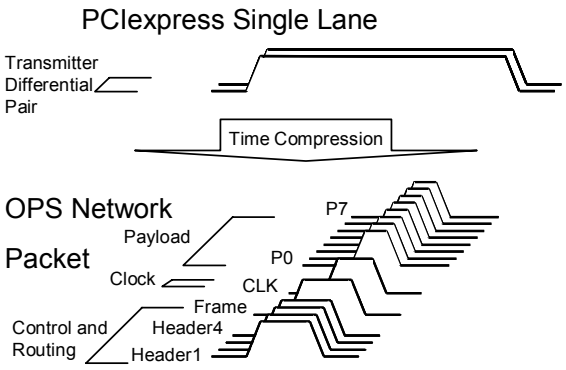

Figure 5. Time compression and relative mapping of a PCIexpress packet to a OPS network packet

The system is presently configured for only one single lane of 2.5 Gbps PCIexpress traffic, but due to the parallel nature of PCIexpress lanes, additional data can be added by using more wavelengths within the available spectrum. Only a single set of Frame and Routing signals are needed for a network packet, while many collections of eight payloads, corresponding to a PCIexpress lane, can be added to the packet, allowing for future expansion to two, four, or even more lanes of traffic.

Figure 6 shows the Frame, Clock, and a single Payload channel from a fully formatted packet. Since the payload (including the clock) is passed through the OPS network unchanged, the exact positioning within the packet is variable so long as the clock-to-data relative timing remains consistent. Experimentally, we've discovered that moving these signals earlier into the packet, eroding the guardtime at the left of the image, works better than centering the data in the middle of the packet slot time.

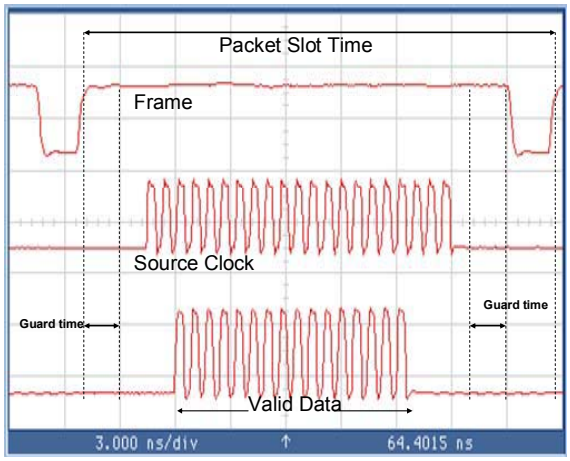

Figure 6. Test stimuli signals used for the Optical Test Bed application

5. System Performance

Figure 7 shows an eye diagram of a payload channel operating at the current project target rate of 2.5 Gbps. For this test, the output waveform is a pseudo-random bit pattern produced by a linear-feedback shift register (LFSR) encoded into the DLC. In addition to fast rise and fall times, the silicon germanium (SiGe) output buffers used on these channels introduce very little jitter, which was measured at the crossover point. For a 2.5 Gbps signal, jitter was measured to be 46.7ps peak-to-peak (including DDJ and 6-sigma RJ), resulting in a usable eye opening of 0.88 unit intervals (UI). While the current system has been designed for operations at 2.5Gbps, the test electronics have been used to generate sample signals at 4.0Gbps with similar results. The measured jitter at the crossover point was 47.2ps p-p with a usable eye opening of 0.81 UI and no significant signal attenuation. We independently measured the 20% to 80% rise and fall times on a single edge and found them to be in the range of 70 to 75ps. The jitter on individual transitions was only 24ps peak-to-peak (about 3.2ps rms).

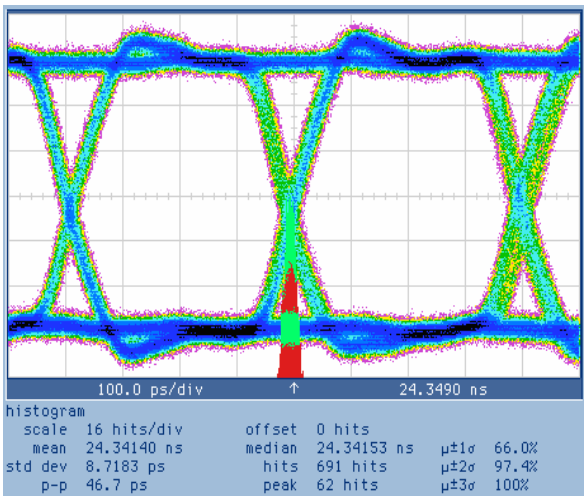

Figure 7. 2.5 Gbps eye diagram

Another timing issue, similar to jitter, that affects the optical signals is dispersion. Since the individual signals within the packet are on separate wavelengths, and the speed of propagation depends on the wavelength, the signals will drift slightly over time, proportional to the distance traveled through the network. The deskew methods available for the channels are sufficient for compensating for small variations in part delays or fiber length mismatches, but are inappropriate for and incapable of realtime adjustment for dispersion effects of incoming packets. Any such adjustment would require knowledge of the number of hops taken through the system, which is not available due to the lack of a centralized control system, nor are the parts capable of adjusting at the rates required.

As measured in the earlier system, of the 350ps available in the usable eye openings there was only a only 270ps sampling window after accounting for the setup and hold times required by the deserializer and cumulative channel-to-channel skew (see Figure 8). This window was measured by skewing the received clock, using the tunable delay, relative to the serial data and sampling the full message length of 32 bits across all 4 channels until no errors were detected in the message. This data was gathered in the ideal case, looping back the electrical data directly from the output buffers and into the recieve circuitry. Signals passed through the electro-optical components and routed through 5 hops in the optical packet switching network exhibited a 150ps timing window due to additional dispersion across the WDM bit-parallel message [9]. As with other source synchronous systems, this loss of timing precision is a central focus for future improvements as it directly impacts our ability to push to faster signaling rates. It also effects the extent to which the parallelism (data width) can be expanded and the scalability of the OPS network with respect to fiber length.

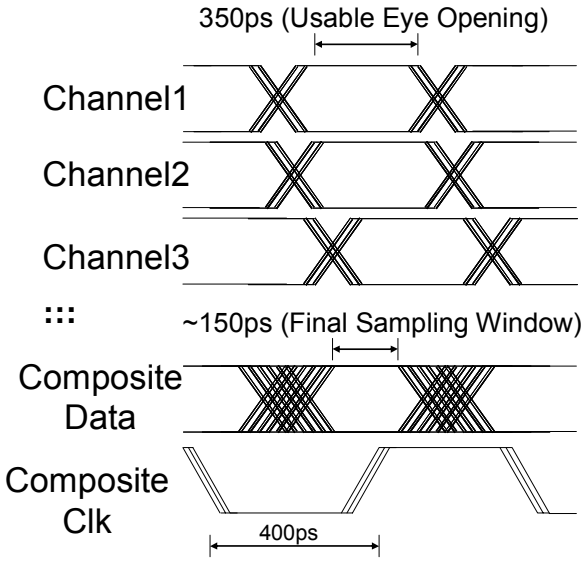

Figure 8. Cumulative effect of jitter and clock distribution on the data sampling window

Since this closing off of the sampling window has a large effect on the future expansion of the system, much of the redesign efforts and evaulations have centered on resolving or minimizing this issue. In addition to the payload channel dispersion above, the distribution of the received clock to the deserializers is also critical because the same signal is used to clock each of the eight channels. Since the clock is relatively stable, compared to the data channels, the data dependent jitter is much lower on the transmission side (measured above as 24ps peak-to-peak). However, on the receiver side the signal is fanned out after the programmable delay which allows part variation and line length mismatches to further degrade the timing of the clock signals.

To illustrate these issues, Figure 9 shows an overlay of the sampling clock recieved at four of the payload channels (two signals are overlaid on each other at either end of the shown range). These signals serve as outer bounds, with the other

four aligning between these two extremes. The separation between the two groups is about 80ps, with any one signal having approximately 40ps of peak-to-peak jitter (measured over 6 sigma). While the skew can be tuned to satisfy the clock-to-data requirements for any one channel which minimizes this comparatively large gap between channels, this further increases the difficulty of satisfying the setup and hold requirements for all channels across the system.

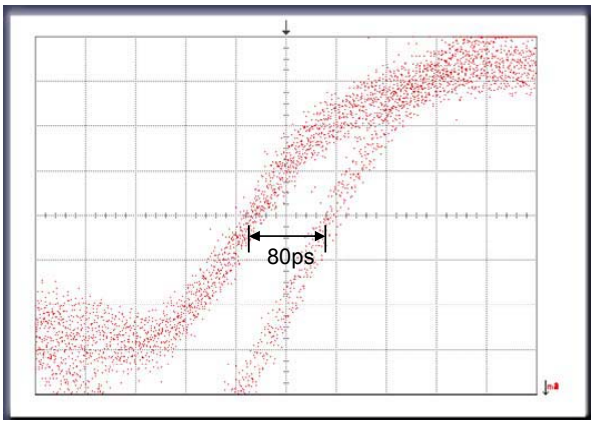

Figure 9. RX sampling clock, composite across channels

Since the amount of dispersion that a group undergoes is related to the selected wavelengths, the spectrum of signals was slightly adjusted for minimal dispersion. The physical network is tuned to a specific set of routing and frame wavelengths, so only the payload and clock signals were changed. Figure 10 shows the previous arrangement, while Figure 11 shows the revised spectrum to reduce the effect of chromatic dispersion. The nine wavelengths (8 payload + 1 clock) are distributed from 1543.73 nm to 1550.12 nm with 0.8 nm spacing between each channel, effectively reducing the timing skew. Additionally, by selecting the clock wavelength midway between the shortest and longest wavelength, the maximum clock-to-data timing skew is reduced to 58 ps/km. A very large data vortex network, having 10k x 10k input and output nodes, would incorporate approximately 200m of interconnected fiber. This results in a maximum clock-to-data timing skew of less than +/- 12 ps (for SMF28 fiber with dispersion = 18 ps-nm/km).

The O/E and E/O interfaces have also been physically optimized for compactness, flexibility, and mobility in a tower configuration. The number of components increases linearly with the number of channels used, and is therefore much more bulky than the test electronics system (wich is less than double the previous size, due mostly to the space constraints of the SMA connectors). The O/E receiver was also redesigned to accomodate true burst-mode reception, as the incoming signal is at best active for only 50% of the time.

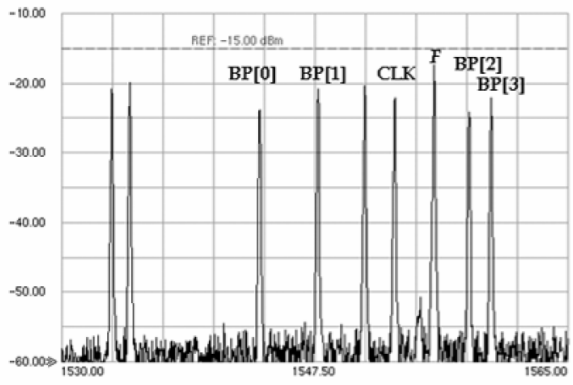

Figure 10. Optical packet spectrum from the previous design. Wavelengths labeled BP are payload signals, F is Frame, and CLK is Clock. Unlabeled wavelengths are for routing. [9].

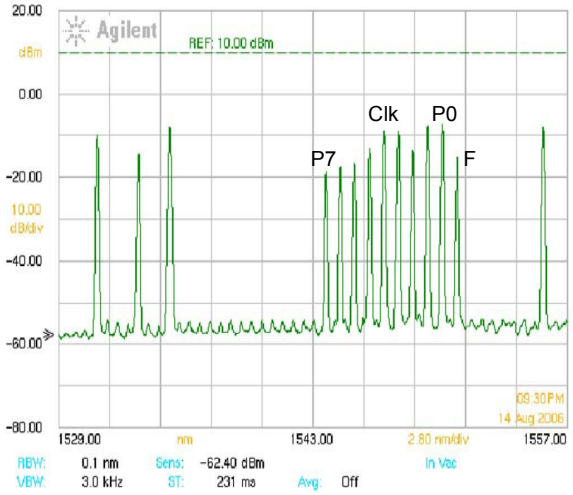

Figure 11. Rearranged packet spectrum. Unlabeled wavelengths are for routing

Preliminary measurements using the RX clock optimizations, spectrum rearrangement, and improved receivers are encouraging, with cleaner and more defined clock edges as well as generally wider sampling windows available. This is achieved despite having to sample across double the number of data channels as previous. These results are necessary to support both higher channel counts and faster data rates in the future, such as the anticipated PCIexpress 2.0 standard expected in the near future, supporting 5.0Gbps per data lane.

6. Conclusions

High-performance computing is very reliant upon high-bandwidth and low latency interconnections. Tunneled PCIexpress traffic across an optical packet switched network is a potential solution to the problem, but presents formidible technical challenges. The system presented here is designed with some approaches that are hoped will close the loop for the first time and allow complete end-to-end communication across the Data Vortex between computer systems, while continuting to allow us to characterize the performance of the system as a whole.

References

1. D. Dai, D.K. Panda, "How much does network contention affect distributed shared memory performance?," *Proc. Int. Conf. Parallel Process.*, Aug. 11-15, 1997, pp. 454-461.

2. C. Hawkins, D.S. Wills, "Impact of Number of Angles on the Performance of the Data Vortex Optical Interconnection Network," *Journal of Lightwave Technology*, Vol. 27, Issue 9, pp. 3288-3294, Sept 2006.

3. A. Shacham, B.A. Small, O. Liboiron-Ladouceur, K. Bergman, "A Fully Implemented 12x12 Data Vortex Optical Packet Switching Interconnection Network," *J. Lightwave Technol*. 23 (10) 3066-3075 (Oct 2005).

4. J.S. Davis, D.C. Keezer, "Multi-Purpose Digital Test Core Utilizing Programmable Logic," *Proc. of the Intl. Test Conf*., pp. 438-445, October 2002.

5. J.S.Davis, D.C. Keezer, K. Bergman, O. Liboiron-Labouceur, "Application and Demonstration of a Digital Test Core: Optoelectronic Test Bed and Wafer-level Prober," *Proc. of the Intl. Test Conf*., pp.166-174, Sept./Oct. 2003.

6. D.C. Keezer, C. Gray, A. Majid, N. Taher, "Low-Cost Multi-Gigahertz Test Systems Using CMOS FPGAs and PECL," pp. 152-157, *Proc. of Design, Automation, and Test in Europe*, March 2005.

7. C. Gray, D.C. Keezer, O. Liboiron-Ladouceur, K. Bergman, "Multi-Gigahertz Source Synchronous Testing of an Optical Packet Switching Network," International Mixed-Signals Test Workshop 2006, Edinburgh, Scotland (June 2006).

8. A.M. Majid, D.C. Keezer, "An Improved Low-Cost 6.4 Gbps Wafer-Leve Tester" *Proc. of the 6th IEEE Electronics Packaging Technology Conference* (EPTC), pp.814-819, December 2005.

9. O. Liboiron-Ladouceur, C. Gray, D.C. Keezer, and K. Bergman, "Bit-Parallel Message Exchange and Data Recovery in Optical Packet Switched Interconnection Networks," *IEEE Photonics Technology Letters*, Vol. 18, No. 6, pp 779-881, Mar. 2006.