A Fiber Optic Local Area Network Demonstrator

Albert Au [1], Boonchuay Supmonchai [1], and Ted H. Szymanski [2]
[1] Department of Electrical and Computer Engineering, McGill University, Canada.
[2] Communications Research Laboratory, McMaster University, Canada.

Abstract

The design of a fiber optic local area network (LAN) demonstrator is described. The demonstrator will be used as a testbed for research in high speed networking technologies, lean protocols, and bandwidth-intensive network-oriented applications, performed at McMaster University in Canada. A complete LAN system would consist of an array of 16 Pentium-based Personal Computers (PC). Each PC has a "Network Interface Card" (NIC), with a parallel fiber optic datalink to a centralized electrical switch core. The centralized core switches the data generated by 16 NICs, up to 128 Gbit/s of bandwidth, roughly 2 product generations ahead of current 1 Gigabit Ethernet LAN technology. The demonstrator is designed to scale to Terabits of bandwidth using an emerging optoelectronic technology, i.e. integrated CMOS substrates with VCSEL optical I/O. A subset of the complete system has been constructed. We have developed a prototype NIC card, using the Motorola Optobus VCSEL transceivers for the optical datalinks, along with a prototype high speed bipolar switch core, using statically configurable electrical ECL 16x16 crossbar switches, CMOS FPGAs and Motorola Optobus transceivers. We have successfully demonstrated the transmission of high speed packetized data from one NIC card, through 10 meters of parallel fiber ribbon and the centralized switch core, and back to the NIC. This paper will summarize the design and testing of our first demonstration system, and our development towards a Terabit switch core.

Keywords: optical networks, local area networks, Networks of Workstations, parallel fiber ribbon, Motorola Optobus, field programmable gate arrays.

1. Introduction

Over the past two decades, the performance of computing systems has been increasing by roughly 55% per year compounded, or roughly a factor of 10 improvement every 4 years. Driven by Networks of Workstations [1], the performance of LAN systems has been increasing at the rate of roughly 60% per year compounded, or roughly a factor of 10 increase every 4 years. For example, Ethernet technology has progressed from 1 Mbit/s in the early 1980s, to 1 Gbit/s in the late 1990s. All projections indicate that the trend will continue for the next 2 decades. Hence, within a decade 100 Gbit/s LAN technology is expected. According to [2], 100 Terabit interconnect technology is expected in high performance electronic systems by the year 2003, such as the 100 TeraFlop computing systems planned by the US Accelerated Strategic Computing Initiative (ASCI) project.

This paper describes the design of a fiber optic LAN demonstrator. The demonstrator will be used as a testbed for research in high speed networking technologies, lean protocols, and in bandwidth-intensive network-oriented applications, being performed at McMaster University in Canada. With the rapid growth in network capacity, it is difficult to predict what new bandwidth-intensive applications may evolve over the next decade. However, traditional applications such as scientific computing will always demand very high bandwidth networking [2], and new applications such as teleconferencing and virtual reality will likely accelerate the need for high bandwidth communications networks. High speed switching and networking is being explored at various locations [3][4]. The basic system architecture of our LAN demonstrator has been described in some recent papers [5][6]. Ref. 5 proposed a scalable optical network architecture for multiprocessing systems and described how the system scales to Terabits of bandwidth. This paper describes our first operational prototypes.

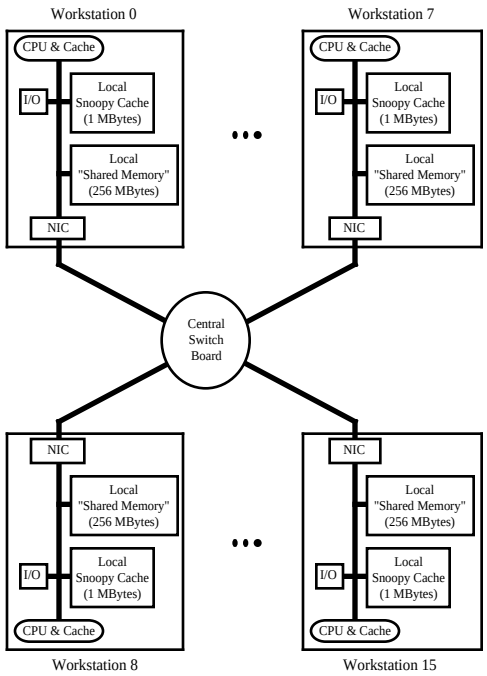


Fig. 1. Fiber Optic LAN used in Network of Workstations (NIC is Network Interface Card).

A complete system would consist of an array of 16 Pentium-based PCs, running the Linux operating system, as shown in Fig. 1. Each PC would have a "Network Interface Card" (NIC), with a parallel fiber optic datalink to a centralized switch core. The centralized switch core switches the data generated by 16 NICs, for an aggregate bandwidth of up to 128 Gbit/s. With LAN capacity increasing by a factor of 10 times every product generation (3-4 years), this

current demonstrator is roughly 2 product generations ahead of current 1 Gigabit Ethernet LAN technology. However, the demonstrator is designed to scale to Terabits of bandwidth capacity by using emerging optoelectronic technologies [7], such as integrated CMOS substrates with modulator (SEED) based optical I/O, or integrated CMOS substrates with VCSEL optical I/O.

A subset of the complete demonstrator has been constructed. We have developed a prototype NIC card shown in Fig. 2 and 3, using the Motorola Optobus VCSEL transceivers for the optical datalinks, along with high speed bipolar electrical serial-to-parallel converters, and CMOS FPGAs for the Message-Processors. We have developed a prototype electronic switch core shown in Fig. 4, using statically configurable electrical bipolar ECL 16x16 crossbar switches, CMOS FPGAs and Motorola Optobus transceivers. We have successfully demonstrated the transmission of high speed data over the LAN, from one NIC card, through 10 meters of parallel fiber ribbon and the centralized electronic switch core, and back to the NIC. This paper will summarize the design and testing of our first prototype demonstration system, and will outline some future directions of the project.

Our current demonstrator emphasizes high bandwidth electrical switching, rather than optical switching. Much of our current research involves exploring very high speed switching structures in CMOS VLSI [5][6][8]. Our network architecture could scale to higher bandwidths, i.e. several hundred Gigabits per second, using a custom designed CMOS VLSI switch core. As the next step in this project, we plan to replace the switch core board with a single-chip optoelectronic CMOS IC, integrated with SEED-based or VCSEL-based optical I/O. The optical signals from the parallel fiber-ribbons would be fed directly on to the optoelectronic IC through an imaging system, as described in [5]. A prototype optoelectronic switch core has been designed and fabricated through the 1997 Lucent/ARPA/COOP workshop [7], and the chip is fully functional thus far. We hope to include this switch core into our demonstrator within 2 years.

This paper is organized as follows. Section 2 provides an architectural overview of the proposed optical LAN system. Sections 3 and 4 describe the design and implementation of the workstation Network Interface Card and the Centreal Switch Core Board, respectively. Section 5 presents experimental results. Finally, section 6 provides concluding remarks.

2. System Overview

To develop the prototype, we have used samples of the Motorola Optobus transceivers, which are described in detail in [10]. These devices support parallel fiber-ribbons with 10 fibers with optical clock rates of up to 800 Mhz. The Optobus is not a commercial product, and this transceiver was used in our prototypes to demonstrate proof-of-concept. In the future, we plan to use an alternative VCSEL/parallel fiber ribbon technology.

The VCSEL transceivers can support clock rates of up to 800 Mhz. The electronic memory bus within the workstations or Personal Computers (PCs) will run much slower, typically at 66 or 100 Mhz. The *Network Interface Card* (NIC) shown in Fig. 2 interfaces between these two domains with different clock rates. It is intended to sit directly on the high bandwidth CPU-memory bus, rather than the low bandwidth I/O bus. The NIC shown in Fig. 2 consists of 4 basic modules, the *Message-Processor,* the *Transmitter Module,* the *Receiver Module* and the *Optobus Transceiver* module. The Message-Processor (MP) implements the communication protocols required for the LAN in FPGA hardware, and provides a 64 bit-wide CMOS datapath for data to be transmitted. The Transmitter module takes the slow wide CMOS datapath, and generates a fast narrow bipolar datapath clocked at 800 Mhz to be transmitted. The Optobus transceiver module includes the Motorola transceiver IC, as well as peripheral bipolar ICs which perform electrical signal translation. The Receiver module takes a fast narrow bipolar datapath from the Optobus module, and generates a slow wide CMOS datapath, which is fed to the MP for processing.

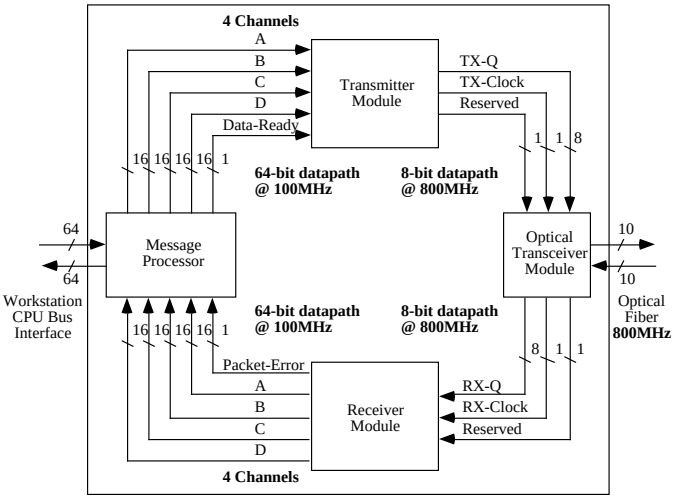


Fig. 2. Design of the Network Interface Card.

We are considering the use of the 8B/10B encoding scheme for the parallel fiber ribbons. This scheme will allow for the encoding of several data and control symbols. However, our current design reserves 2 fibers in each ribbon for broadcasting of timing, synchronization and control from the core, as described in [5]. One is used for the "bit" clock, i.e., the 800 Mhz clock used by the Receiver module. The second bit is the "frame" signal used to denote the start of a packet "frame". The remaining 8 bits support bytes of data, at the 800 Mhz clock rate. These signals allow for a "self-timed" design, where each receiver uses the bit clock of the sender to sample the data. This approach simplifies synchronization within the network, since all workstations receive timing information from the low skew fibers, which are all 10 meters long in our localized network.

The CMOS Message-Processor manages the LAN communication protocols in programmable hardware rather than software [5]. Our protocols are "Lean", and will span two traditional networking functions, the "Data Link Control" protocol (DLC) for reliable bit-stream communications, and "Automatic Repeat Request" protocols (ARQ) for error and flow control, and for message fragmentation and re-assembly [11]. The DLC protocol performs error-detection and optional error-correction over a

single Optobus link. The hardware based ARQ protocol performs end-to-end functions, such as optional error-detection, error-correction, sliding window flow control, message-fragmentation and reassembly between two communicating workstations. Messages are supplied to the Message-Processors, where they are fragmented into fixed sized packets, assigned source-ID numbers and sequence numbers for error control, queued in the NIC, and then transmitted over parallel fiber optic datalinks to the centralized switch core to their destinations. The MPs also perform the receiving protocols.

The LAN is designed to support general inter-computer communications, consistent with communications patterns found in 1 Gigabit Ethernet LAN technology. However, the LAN can in principle support distributed shared memory for multiprocessors, and we are pursuing this avenue of research. The LAN currently uses a fixed size 32 byte packet format [5]. However, it is possible to use variable length packets or ATM cells, although this may increase the complexity of the switch core.

3. Design of the Network Interface Card

The NIC shown in Fig. 2 contains an Altera FPGA [12], the Transmitter/Receiver Module chipset, and the Optobus Transceiver Module. A photograph of the prototype NIC is shown in Fig. 3. The board is powered from a single 5V supply and all ICs, except the CMOS FPGA, are operated at Positive Emitter Coupled Logic (PECL) levels. All PECL signal lines are terminated into 50Ω. Bypassing and decoupling techniques [13] are applied to the 5V rail since all PECL levels are referenced to this voltage. The FPGA provides the parallel frame clock, the parallel data and the necessary control signals for the transmitter and receiver modules. The I/O signals of the transmitter module, the receiver module and the Optobus transceiver module are compatible to the PECL voltage levels (800mV swing). Conversion between CMOS and PECL levels are performed for all I/O signals between the FPGA and the transmitter and receiver modules.

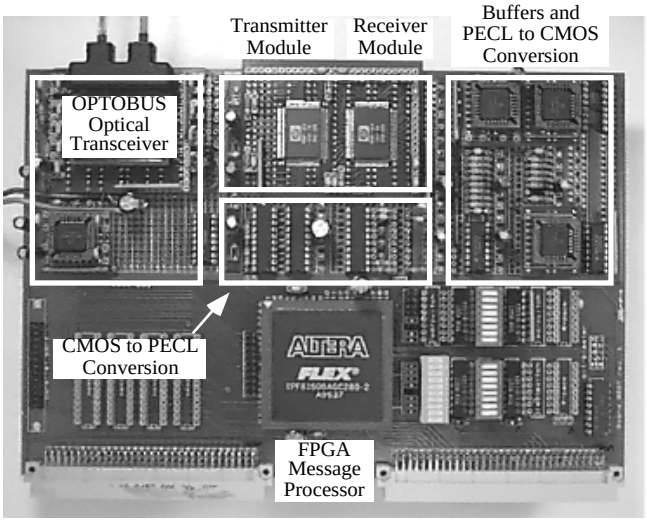


Fig. 3. Network Interface Card (NIC).

The ICs are mounted on a printed circuit board (PCB), which can be seen in Fig. 3. Sockets were designed and fabricated to mount the transmitter module, the receiver module and the Optobus transceiver module ICs on to the PCB. Bypass and decoupling capacitors and terminating resistors were designed on to the sockets, so they are located as close as possible to the ICs.

3.1. Transmitter Module

The Transmitter Module [14] within the NIC shown in Fig. 3 contains circuitry for frame assembly, parallel-to-serial conversion and clock generation. The IC used in the current demonstration system accepts a 20-bit parallel data word, a frame clock and several control signals. The user-supplied frame clock is internally multiplied to create the high speed serial clock. The 20-bit data field is appended by a 4-bit control field, thus creating a 24-bit data frame for serial transmission. The control field contains frame information and provides a master transition for frequency locking at the Receiver Module. The frame mux performs parallel-to-serial conversion on the 24-bit data frame and outputs the high-speed differential signal to the Optobus Transceiver Module for transmission over fiber to the central switch core. Depending on the frame clock, the Transmitter Module can output a serial rate of up to 1 Gbit/s.

3.2. Optobus Transceiver Module

The Optobus Transceiver Module within the NIC shown in Fig. 3 converts the high-speed serial data between the electrical and optical domains. The Motorola Optobus [10] is a VCSEL-based, bidirectional point-to-point datalink, operating at a wavelength of 850nm. This device supports 10 transmit and 10 receive channels, on 2 multi-mode parallel fiber-ribbons, and has a bandwidth of 800 Mbit/s per channel.

The transmitter and receiver on the Optobus operate independently. On the transmit side, the differential PECL-level serial data signal from the Transmitter Module is buffered before entering the Optobus. The serial data is then transmitted over the parallel fiber-ribbon. On the receive side, the serial data from the central switch core is received over the parallel fiber-ribbon. The electrical outputs of the Optobus is in differential Current Mode Logic (250mV swing), which is restored to PECL levels (800mV swing) before feeding to the Receiver Module.

3.3. Receiver Module

The Receiver Module [14] within the NIC shown in Fig. 3 contains circuitry for frequency locking, serial-to-parallel conversion, data and status extraction and clock generation. The input serial data is sampled by the input sampler. The frequency/phase detect unit controls the input sampler and outputs the recovered parallel frame clock. The frame demux performs serial-to-parallel conversion on the received 24-bit data frame. The data field is extracted by the data field decoder, which outputs the 20-bit parallel data word. The control field is decoded by the control field decoder, which also outputs the data status. The error status of the high-speed point-to-point link is also reported. The data and status outputs of the Receiver Module are synchronized to the recovered parallel frame clock.

4. Central Switch Core Board

The prototype Central Switch Core Board, shown in Fig. 4, contains an Altera FPGA, the switch core IC and several Optobus Transceiver Modules. The Central Switch Core Board also communicates with a unique controller workstation via the parallel port. Static routing information is entered into the switch core using custom software running on the controller workstation. The FPGA implements the interface for the parallel port communication and the control logic for programming the switch core IC. Each input and output port pair of the switch core is connected to an Optobus Transceiver Module.

The Switch Core and the Optobus Transceiver Module ICs were mounted on the PCB in Fig. 4 using custom designed sockets. The board is powered from a single 5V supply and all ICs, except the CMOS FPGA, are operated at PECL levels. The data I/O signals of the switch core IC, and the Optobus Transceiver Modules are PECL compatible. The control signals of the switch core IC are CMOS compatible. PECL signal lines are terminated into 50Ω. As on the NIC, bypassing and decoupling techniques are applied to the supply rails.

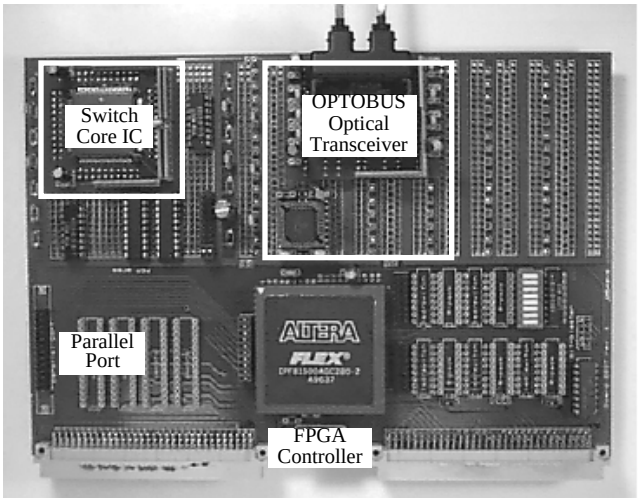


Fig. 4. Central Switch Core Board.

4.1. Switch Core IC

The Switch Core in the current demonstrator is a non-blocking 16x16 crossbar switch [15] with a bandwidth of 1.2 Gbit/s per channel. The peak bandwidth of the switch is 19.2 Gbit/s. The switch contains 16 fully independent 16:1 mulitplexers and allows each output port to accept data from any of the 16 input ports. The switch is statically configured by sequentially loading each output port's program latch with the desired input port's 4-bit address, which is performed by the controller workstation. The input port address is then used as the multiplexer select lines for that output port.

Our current demonstrator implements the switching with high speed electrical PECL switches. However, as described in [5] the switching functions can be implemented using other schemes. We have also designed and fabricated a custom all-electrical VLSI self-routing switch core which can scale to hundreds of Gbit/s of I/O bandwidth [8]. We have also developed a single-chip optoelectronic CMOS ASIC, integrated with SEED-based optical I/O [7]. The optical signals from the parallel fiber-ribbon can be fed directly into the optoelectronic IC through an optical imaging system [5]. As described in [5], this technology would allow the fiber optic LAN to scale to Terabits of bandwidth, and represents the future step in this project.

5. Experimental Results

The fiber optic LAN demonstration system, consisting of the central switch core board, two parallel fiber optic datalinks and one NIC, is shown in Fig. 5.

Fig. 6 shows the transmission of serial data from one NIC card, through 10 meters of parallel fiber-ribbon, through the centralized switch core, through 10 meters of parallel fiber-ribbon, and back to the NIC. The output signal from the Transmitter (TX) module and the input signal into the Receiver (RX) module are observed for 2 frame clock frequencies. In Fig. 6(a), the 20-bit parallel data word "10100010..0" is clocked at 8 Mhz into the Transmitter module. The 4-bit control field "1101" is appended. Thus the observed serial data rate is (20+4) * 8 = 192 Mbit/s. In Fig. 6(b), the same 20-bit parallel data word is clocked at 12 Mhz into the Transmitter module. The same 4-bit control field is appended. Thus the observed serial data rate is (20+4) * 12 = 288 Mbit/s. These tests demonstrate successful transmission of data through the LAN.

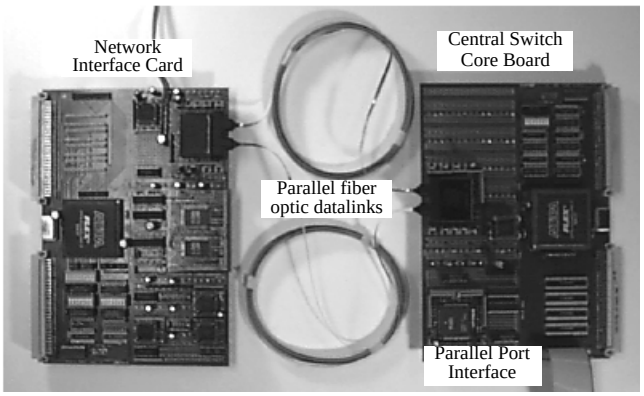


Fig. 5. Key components of the fiber optic LAN demonstrator.

6. Conclusions

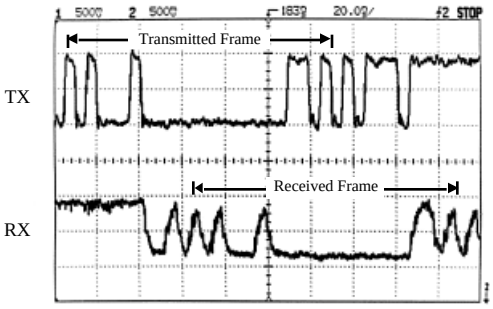
The design of a fiber optic LAN demonstrator was described. A complete system would connect 16 PCs, each having an NIC and a point-to-point parallel optical datalink to a centralized switch core. The switch core switches the data generated by the 16 NICs, for an aggregate bandwidth of 128 Gbit/s. A subset of the complete demonstrator has been constructed. The design and testing of our first prototype system is summarized. We have successfully shown the transmission of 288 Mbit/s serial data through the LAN, from one NIC card, through 10 meters of parallel fiber-ribbon and the centralized switch core, and back to the NIC.

In principle, this LAN architecture, using off-the-shelf technology, can supply interconnect requirements of up to a few Terabits per second. Each LAN could switch

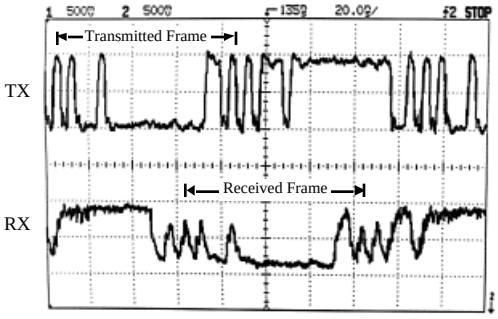
approximately 100 Gbit/s of data, and several switch cores could be interconnected with fiber ribbons using various standard topologies, such as 3-stage Clos networks or Fat-Trees. The architecture also provides a smooth growth path to much higher capacities, in the 100 Tbit/s range, when the switch core is implemented onto a single optoelectronic IC [2][5][6]. We hope to demonstrate LAN switching over such an IC in the near future.

7. Acknowledgments

This research was funded by NSERC Canada Grant OGP011211601, and in part by SPAR Space Systems. Computing equipment, CAD tools and device fabrication were supplied by the Canadian Microelectronics Corporation (CMC).



(a) 192 Mbit/s Serial Data.



(b) 288 Mbit/s Serial Data.

Fig. 6. Observed serial data at (a) 192 Mbit/s, and (b) 288 Mbit/s. The upper waveform is the output from the Transmitter (TX) Module. The lower waveform is the input to the Receiver (RX) Module.

8. References

[1] T. E. Anderson, D. E. Culler, and D. A. Patterson, "A Case for NOW (Networks of Workstations)", *IEEE Micro*, Vol. 16, No. 1, Feb. 1995, pp. 54-64.

[2] A. Benner, E. Schenfeld, J. Sauer, L. Rudolph, T. Sterling, and T. H. Szymanski, "Interconnect Design Options for Delivering a 100 TFlop/sec Parallel Supercomputer in 2003", Paper and Panel Discussion, *Fifth International Conference on Massively Parallel Processing using Optical Interconnects*, Las Vegas, Nevada, June 15-17, 1998.

[3] T. Chaney, J. A. Fingerhut, M. Flucke, and J. S. Turner, "Design of a Gigabit ATM Switch," *Proc. IEEE INFOCOM*, vol. 1 , 1997, pp. 2-11.

[4] J. W. Lockwood, H. Duan, J. J. Morikuni, S. M. Kang, S. Akkineni and R. H. Campbell, "Scalable Optoelectronic ATM Networks: The iPOINT Fully Functional Testbed," *IEEE Journal of Lightwave Technology*, vol. 13, no. 6, Jun. 1995, pp. 1093-1103.

[5] T. H. Szymanski, A. Au, M. Lafrenière-Roula, V.Tyan, B. Supmonchai, J. Wong, B. Zerrouk, and S. T. Obenaus, "Terabit Optical Local Area Networks for Multiprocessing Systems", *Applied Optics, Special Issue on Massively Parallel Optical Interconnects for Multiprocessor Systems*, vol. 37, no. 2, Jan. 1998, pp. 264-275.

[6] T. H. Szymanski, "Parallel Computing with Intelligent Optical Networks", in *Parallel Computing using Optical Interconnections*, eds. K. Li, Y. Pan and S. Q. Zheng, Kluwer Academic Publishing, 1998, pp. 24-46. (http://www.mcs.newpaltz.edu/~li/pcuoi.html).

[7] T. H. Szymanski, M. Saint-Laurent, V. Tyan, A. Au, and B. Supmonchai, "A Field Programmable Gate Array with Optical I/O", *Proc. Optics in Computing (OC'99)*, Snowmass, Colorado, USA, Apr. 1999.

[8] B. Supmonchai and T. H. Szymanski, "High Speed VLSI Concentrators for Terabit Intelligent Optical Backplanes", *Proc. Optics in Computing (OC'98)*, Brugges, Belgium, Jun. 1998, pp. 306-310.

[9] D. R. Engebretsen, D. M. Kuchta, R. C. Booth , J. D. Crow, and W. G. Nation, "Parallel Fiber-Optic SCI Links," *IEEE Micro*, vol. 16, no. 1, Feb. 1996, pp. 20-26.

[10] L. J. Norton, F. Carney, N. Choi, C. K. Y. Chun, R.K. Denton, Jr., D. Diaz, J. Knapp, M. Meyering, C. Ngo, S. Planer, G. Raskin, E. Reyes, J. Sauvageau, D. B. Schwartz, S. G. Shook, J. Yoder and Y. Wen, "OPTOBUS I: A Production Parallel Fiber Optical Interconnect", *Proc. 47th Electronic Components and Technology Conference*, IEEE, New York, NY, USA, 1997, pp. 204-209.

[11] D. Bertsekas, and R. Gallager, *Data Networks*, 2nd Ed., Prentice Hall, 1992.

[12] *Altera FPGA Data Book 1996*, Altera Corporation, San Jose, CA, USA.

[13] H. W. Johnson and M. Graham, *High-speed Digital Design: A Handbook of Black Magic*, Prentice-Hall, 1993.

[14] Chu-Sen Yen, R. C. Walker, P. T. Petruno, C. Stout, B. W. H. Lai and W. J. McFarland, "G-Link: A Chipset for Gigabit-Rate Data Communication", *Hewlett-Packard Journal*, Vol. 43, No. 5, Oct. 1992, pp. 103-116.

[15] TriQuint Semiconductor, *TQ8017 - 1.2 Gigabit/sec 16x16 Digital PECL Crosspoint Switch*, TriQuint Semiconductor, Beaverton, OR, USA.