Terabit optical local area networks for multiprocessing systems

Ted H. Szymanski, Albert Au, Myriam Lafrenière-Roula, Victor Tyan, Boonchuay Supmonchai, James Wong, Belkacem Zerrouk, and Stefan Thomas Obenaus

The design of a scalable optical local area network for multiprocessing systems is described. Each workstation has a parallel-fiber-ribbon optical link to a centralized complementary metal-oxide silicon (CMOS) switch core, implemented on a single compact printed circuit board (PCB). When the Motorola Optobus fiber technology is used, each workstation has a data bandwidth of 6.4 Gbits/s to the core. A centralized switch core interconnecting 32 workstations supports a 204-Gbit/s aggregate data bandwidth. The switch core is based on a conventional broadcast-and-select architecture, implemented with parallel CMOS integrated circuits (IC's). The switch core scales well; by incorporation of the CMOS optoelectronic IC's with optical input–output, the electrical core can be reduced to a single-chip optoelectronic IC with terabit capacities. A prototype of an optoelectronic switch core has been fabricated and is described. The appeal of the architecture includes its reliance on commercially available parallel-fiber technology, its reliance on the well-developed markets of local area networks and networks of workstations, and its smooth scalability from the electrical to optical domains as technology matures.

OCIS codes: 060.0060, 060.2310, 060.2330, 060.4250, 060.4510.

1. Introduction

The performance of microprocessors and communication networks has been growing exponentially over the past decade, and this trend is expected to continue well into the future.[1,2] Current industry requirements are for computing machines with teraFLOP and petaFLOP performances, and the U.S. Accelerated Strategic Computing Initiative has undertaken a program to accelerate further the growth in computing performance, aiming to achieve 100 teraFLOP machines by the year 2003.[3] Such machines will occupy several rows of cabinets and will require a complex high-performance interprocessor communication network. For sustaining these increases the interconnection networks must also exhibit exponential growth in their capacity, and sometime after the year 2010 communication networks are expected to have capacities of several tens of terabits per second.[2] Hence new network architectures that exploit optical technologies seem necessary.

In this paper we propose a scalable optical network architecture for high-bandwidth interconnects. The network can be used to connect printed circuit boards (PCB's) in a multigigabit fiber backplane, to connect backplanes in a cabinet, or to connect cabinets in a massively parallel processing system. In this paper we focus on interconnecting a network of workstations,[4] as shown in Fig. 1. The network architecture can be called an optical star on the basis of its centralized starlike topology. Each workstation has a dedicated parallel-fiber ribbon connection to the switch core, providing typically 6.4 Gbits/s per link.[5,6] A centralized electronic switch core interconnecting 32 workstations, implemented with complementary metal-oxide silicon (CMOS) integrated circuits (IC's) can be implemented on a single high-speed PCB supporting a bandwidth of 204.8 Gbits/s. The switching is performed with parallel CMOS IC's, either field-programmable gate arrays (FPGA's) or custom application-specific integrated circuits (ASIC's).

The authors are with the Department of Electrical Engineering and School of Computer Science, Microelectronics and Computer Systems Laboratory, McGill University, Montreal, Canada H3A 2A7.

Received 11 April 1997; revised manuscript received 11 August 1997.

0003-6935/98/020264-12$10.00/0

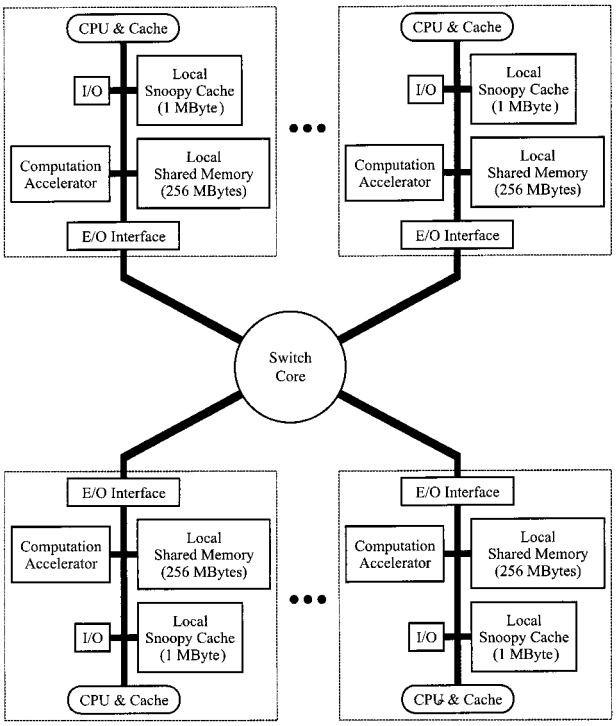


Fig. 1. Distributed shared memory architecture. I/O, input–output; E/O, electrical–optical.

The bandwidth of parallel-fiber ribbons will likely exhibit rapid growth rates after their use becomes widespread. Parallel-fiber ribbons with 32 or 64 fibers/row, with clock rates in the tens of gigabits-per-second range, are plausible within a decade. Fiber ribbons with 32 fibers and 500-MHz clock rates already exist.[7] However, the construction of an electrical switch core with terabit capacity on a single PCB will become difficult to achieve. In this paper we illustrate a smooth scalability from the electrical to optoelectronic domains, i.e., how the single PCB switch can be reduced to a single-chip optoelectronic IC, by use of a relatively inexpensive free-space optical imaging system. We also describe a prototype CMOS–self-electro-optic-device (SEED) optoelectronic IC that implements the switching functions. This optoelectronic IC can scale to several terabits capacity as the smart-pixel technology matures.[8]

High-speed switching and networking are being explored at various locations (see, e.g., Refs. 9–24), and many of these projects share common attributes: Many rely extensively on buffering within the switch core; shared buffering is described in Refs. 11 and 12, input buffering is described in Refs. 13–15, and output or virtual output buffering is described in Refs. 14 and 16–19. Many rely on complex (hence relatively slow) electronic circuits to perform the arbitration among many buffered packets contending for the few output ports; iterative arbitration algorithms that require several passes in time are described in Refs. 14, 18, and 20. Many rely on the use of phase-locked loops (PLL's) or other complex techniques to maintain clock synchronization among all the distributed components.[14] In these projects the impact that parallel-fiber technology may have on localized multigigabit networks has not been considered.

Our research project is exploring and unifying several novel approaches to achieve multigigabit networking, with particular emphasis on the following: (1) parallel-fiber technology, which can alter many fundamental design decisions; (2) the broadcasting of timing, synchronization, and control from a central master timekeeper; (3) a deemphasis of buffering, in which most packet buffers are eliminated and replaced with pipelined virtual circuits (VC's) over parallel fibers; (4) an emphasis on reduced-complexity pipelined CMOS switch-core IC's with very high internal clock rates (≥400 MHz); (5) an emphasis on very fast (i.e., of the order of nanoseconds) one-pass concentration (or arbitration) with optimal logarithmic latency; (6) an emphasis on error control codes (ECC's) to lower bit-error rates (BER's) to levels comparable with main memory, thereby eliminating software-based error control; (7) a deemphasis on asynchronous transfer mode (ATM) for multiprocessing systems; and finally (8) a cost-effective optical imaging system. We now elaborate on these aspects.

Parallel-fiber technology can support tens of fibers and hundreds of gigabits of bandwidth and can simplify many fundamental network design decisions. Given the abundance of high-bandwidth low-skew parallel fibers (1-ps skew/m), they are attractive to broadcast timing, synchronization, and control from the central switch core that serves as a master timekeeper. This approach can eliminate the PLL's usually needed to maintain clocking in distributed systems, a significant simplification on its own. It is also attractive for deemphasizing buffering within the switch core, as it pushes many of the buffers back into the workstations where there is plenty of memory already. The majority of an integrated CMOS switch is usually occupied by buffer memory and controllers; hence this design philosophy can reduce the complexity and the VLSI area of the core significantly. It then becomes attractive to pipeline the reduced-complexity core and operate at very high clock rates, which further reduces the number of logic gates required. Our analysis indicates that CMOS cores with narrow fast 400-MHz data paths are possible with 0.8-μm technology. In contrast, all the buffered IC's described above are restricted to wide and slow data paths, typically with 100-MHz clocks.

We are also exploring the use of ECC's to improve the reliability of distributed networking. Many previous projects relied on software protocols to handle errors, arguing that current links have BER's as low as 10^{-10} or lower. However, if a network is moving one terabit per second, this low BER still results in several hundred bit errors per second on average, and these bit errors necessitate complex software to perform error recovery. If more reliable links with hardware-based error control are used, it may be possible to achieve networks that are as reliable as main memory (which also uses hardware-based error con-

trol), potentially eliminating the need for software-based error control. In many of the above-mentioned projects ATM was selected as the underlying packet format. Although our proposed system can use ATM, we have selected a reconfigurable multiprocessor packet format instead for our prototype development, allowing us to explore novel hardware-based communication protocols specifically designed for multiprocessing.

In this paper we focus on the design issues of a scalable 32-port optical local area network (LAN), gained from our experience in developing prototypes of all the major components addressed in this paper. A first-generation prototype PCB switch core that links Optobus fiber ribbons was fabricated and will become operational in 1998, with an objective of a 32-Gbit/s bandwidth. A first-generation single-chip optoelectronic switch-core IC has been fabricated and tested and should be integrated into an optical system in 1998. A second-generation device, which interfaces to eight Optobus fiber ribbons, was designed as part of the 1997 Lucent Technologies/Consortium for Optical and Optoelectronic Technologies in Computing (CO-OP) Workshop and has been submitted for fabrication.

This paper is organized as follows. Section 2 provides an architectural overview of the proposed optical LAN system. Section 3 describes the design of a single PCB switch core with 204-Gbits/s capacity. Section 4 describes the design of a single-chip optoelectronic switch core, which scales to terabit capacities. Section 5 describes the error detection schemes and reliability schemes. Finally, Section 6 provides concluding remarks.

2. Architectural Overview

Conventional distributed shared-memory (DSM) systems are limited to small systems, typically 32 processors, as scalability of electrical broadcast busses is a major limitation.[1] However, an optical LAN network can overcome this obstacle and extend the DSM concept to hundreds of processors by providing large-scale broadcasting capability with terabits of bandwidth. To support DSM, a workstation on the optical LAN typically requires (1) a local shared-memory module, (2) a local snoopy or directory-based cache module, and (3) an electrical–optical (EO) interface module to interface to the optical LAN. These modules interact with the workstation CPU and the input–output (I/O) module by means of the CPU bus, as shown in Fig. 1. FPGA's can outperform workstations by a factor of 10 to 100 for scientific computations. In order to utilize the gigabits of bandwidth of an optical LAN, each conventional workstation should be equipped with a computation-accelerator card with several FPGA's, along with several megabytes of high-speed memory and a CPU interface.[25]

The Motorola Optobus is described in detail in Ref. 5. Currently the parallel-fiber ribbon provides 10 fibers with optical clock rates of 800 MHz. The electronic buses within the workstations will run much more slowly, typically at 100 MHz. The EO interface module shown in Fig. 2 interfaces between these two domains with different clock rates. It consists of four basic modules, the message-processor (MP) clock-multiplier, clock-divider, and Optobus transceiver modules. The MP implements the communication protocols required for the LAN in FPGA hardware and provides a 64-bit-wide data path for data to be transmitted. The clock-multiplier module takes the slow wide data path and generates a fast narrow data path clocked at 800 MHz to be transmitted. The Optobus transceiver module includes the Motorola transceiver IC as well as peripheral IC's which perform electrical signal translation. The clock-divider module takes a fast narrow data path from the Optobus module and generates a slow wide data path in CMOS, which is fed to the MP.

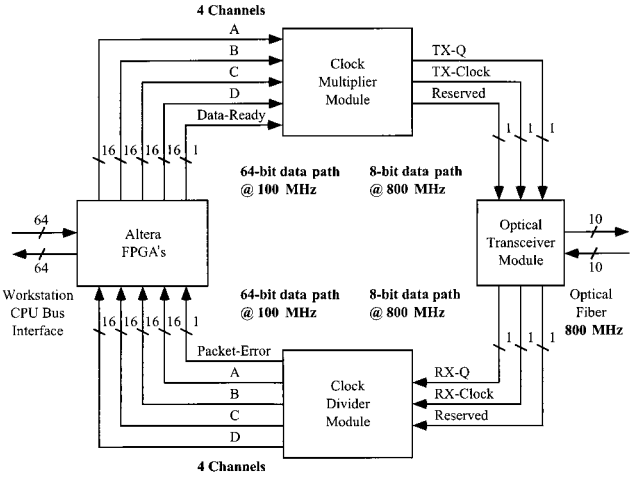


Fig. 2. EO interface module. TX, transmitter; RX, receiver.

Our design reserves two optical fibers in each ribbon for broadcasting of timing, synchronization, and control from the core. One is used for the bit clock, i.e., the 800-MHz clock used by the clock-divider module. The second bit is the frame signal used to denote the start of a packet frame. These signals allow for a self-timed design, in which each receiver uses the bit clock of the sender to sample the data. This approach eliminates the need for PLL's throughout the network, as all workstations receive timing information from the low-skew fibers, which are all 10 m long in our localized network. (The frame signal is not necessary and may be replaced by a dc balance bit.)

The CMOS MP handles the LAN communication protocols. Our protocols are lean and will span two traditional networking functions: the data link control protocol for reliable bit-stream communications and automatic repeat request protocols for error and flow control and for message fragmentation and reassembly.[26] The data link control protocol performs error detection and optional error correction over a single Optobus link. The hardware-based automatic repeat request protocol performs end-to-end functions, such as optional error detection, error cor-

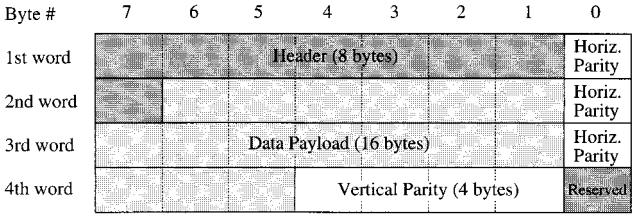


Fig. 3. Typical 256-bit packet format (format reconfigurable).

rection, sliding window flow control, message fragmentation, and message reassembly between two communicating workstations. Messages are supplied to the MP's, where they are fragmented into fixed-size packets, assigned source identification numbers and sequence numbers for error control, queued in the workstations, and then transmitted over parallel-fiber links to the centralized switch core to their destinations. The MP's also perform the receiving protocols.

In a DSM model, typical messages include (1) reads and writes to a shared-memory word (8 bytes), and (2) reads and writes of shared-memory cache blocks (16 . . . 128 bytes). Our optical LAN design uses a fixed-size 32-byte packet format, as shown in Fig. 3. However, it is possible to use variable-length packets or ATM cells, although this may increase the complexity of the core.

3. Electrical Switch-Core Design

The design of a CMOS switch core with a bandwidth of 204.8 Gbits/s that can be implemented on a single PCB is described. The CMOS switch core is based on a conventional multichannel broadcast-and-select architecture, as shown in Fig. 4. (The knockout switch and dilated crossbar are examples of this classic architecture, i.e., see Refs. 19 and 27.) The switch core requires $N = 32$ I/O ports, and each port interfaces to an Optobus link. To communicate, a sender simply broadcasts the packet over its own reserved contention-free channel. All output ports

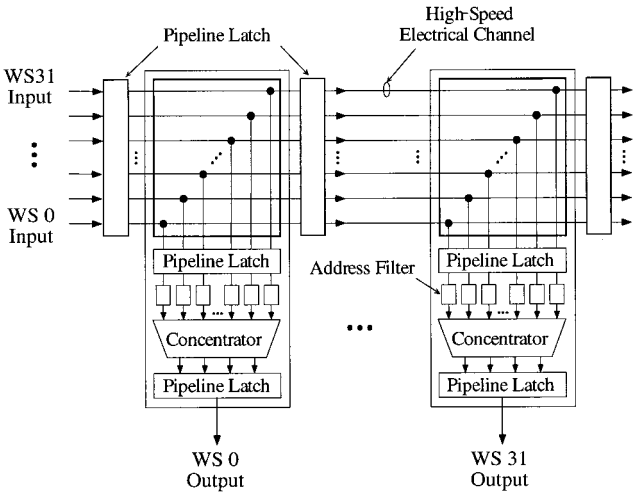


Fig. 4. Broadcast-and-select switch core. WS, workstation.

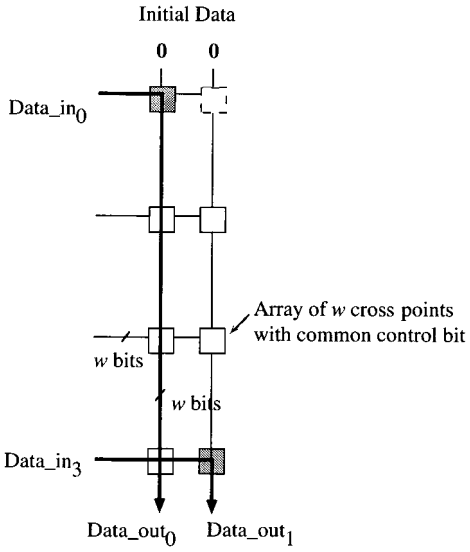


Fig. 5. Data plane of a 4-to-2 daisy-chain concentrator (dashed cell not needed).

listen to all channels, performing address filtering functions on the packets passing by and copying and extracting packets addressed to them through the concentrator circuits. We can improve the fault tolerance of the system by operating $M > N$ channels, using the extra channels as spares (see Section 5).

Each workstation output port has M packet address filters, one filter per channel, and an M-to-n concentrator, a circuit that can extract a packet from any of the M channels onto n output ports (n is typically 1 . . . 8). The concentrators (or arbiters) largely determine the speed and the complexity of the switch. In the field of communications the design of efficient concentrators is a well-studied problem without an optimal solution (see Ref. 24 and the references therein). An optimal M-input concentrator will determine its state in logarithmic time, i.e., $O[\log(M)]$. The knockout concentrators[19] can be used; however, these are complicated circuits and they lack systematic VLSI layouts. The arbiters in Ref. 20 require linear time and are thus slow. Our design uses a logarithmic time variation of the daisy-chain concentrator[27] shown in Fig. 5 for fast one-pass arbitration. An M-to-n concentrator consists of an $M \times n$ array of controller cells. Each column arbitrates access to one output port. For an expanded discussion of this concentrator see Refs. 27 and 28.

Blocking will occur at an output port when greater than n packets simultaneously attempt to reach the same destination workstation. We can reduce the blocking probability PB by exploiting statistical multiplexing of more VC's onto each Optobus parallel link, i.e., we may space-multiplex $n = 8$ VC's onto each Optobus link, with one VC per fiber. Each VC transports one packet. The fraction of time a workstation is transmitting a packet is given by α. By generalizing the analysis in Ref. 19, we can show that

the probability a packet is blocked, PB, at a switch-core output port is given by[29]

$$\mathrm{PB} = \frac{N}{\alpha}\left[\sum_{j=n+1}^{\infty}(j-n)\,\frac{\exp(-\alpha/N)(\alpha/N)^j}{j!}\right], \quad (1)$$

where n is the degree of space multiplexing, N is the size of the switch core, and α is the normalized load. For $n = 8$, PB = 0.005 at a load of $\alpha = 0.5$.[27] In other words, 99.5% of all packets will be successful in their first transmission attempt. The remaining packets will not be acknowledged and must be retransmitted. The acknowledgments and the retransmissions are handled by the MP's in the hardware. For these parameters, the average number of time slots required for transferring a packet is $1/(1 - \mathrm{PB}) = 1.005$.

It should be noted that a buffered switch core, when, for example, virtual output queueing is used, would require a moderate amount of packet buffers and complex and slow arbitrators and queue controllers to achieve a throughput of 99.5%. An unbuffered core exploiting multiple VC's per link can deliver most packets in their first transmission attempt and can operate at very high clock rates.[27]

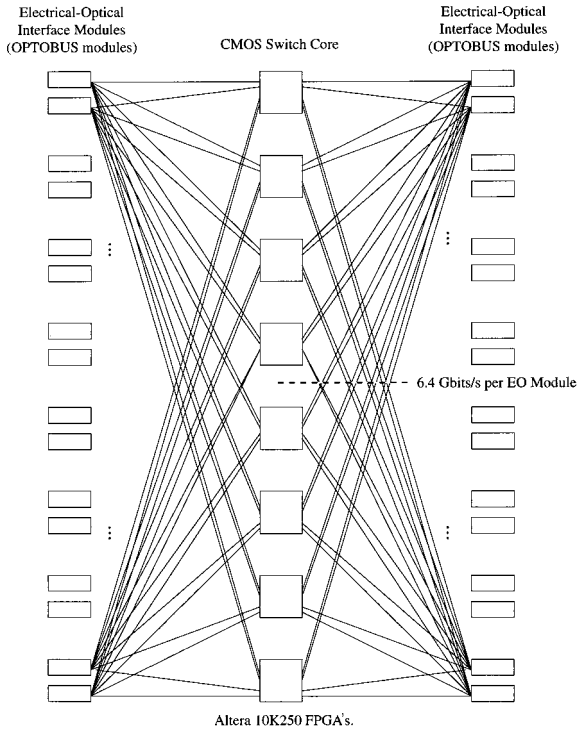


Fig. 6. Packaging of the FPGA-based switch core.

A. Complementary Metal-Oxide Silicon Implementation of the 204.8-Gbit/s Switch Core

The design of a 32 × 32 CMOS switch core implemented with parallel FPGA's on a single PCB is first described. Within the current technology FPGA's can maintain clock rates of up to 100 MHz. To support the Optobus data rate, each channel must be 64 bits wide. It can be verified that a dilated 32 × 32 crossbar with 64-bit-wide data paths, in which each I/O port supports eight VC's, will require roughly 2×10^6 logic gates[27,28] and 4096 I/O pins.

FPGA's allow for the dynamic reconfiguration of the switching functions to suit the application, providing an unprecedented flexibility. The core can be reconfigured to perform computing as well as switching, i.e., fast Fourier transforms, sorting, or any other parallel computation. Furthermore, if the digital functions remain static in a large-volume application, the same hardware functional specifications can be recompiled to generate the VLSI masks for a semicustom ASIC by use of standard cell technology.

Because of gate limitations of the FPGA's, it is not possible to implement the entire switch core on a single FPGA. Hence a standard bit-slice approach can be used. To support the full 64-bit-wide switch core, a bank of 16 FPGA's can be operated in parallel. Each FPGA will implement a 4-bit slice of the original switch, requiring 125,000 logic gates. The Altera 10K250 FPGA's support typically 250,000 programmable logic gates and can easily implement the required functions. The bit-slice approach also results in improved fault tolerance, as spare FPGA's can be hardwired into the core initially.

B. Packaging the Switch Core onto a Single Printed Circuit Board

The packaging of the electrical switch core onto a single PCB is illustrated in Fig. 6. The EO interface modules are on the sides and the CMOS IC's are arranged in the center. (In Fig. 6, the center array illustrates only one column of eight IC's.) Each EO module has a 4-bit data path to each IC, resulting in perfect-shuffle interconnections. The entire PCB requires 16 FPGA's and 32 EO modules. Each FPGA occupies 2 in. × 2 in. (5.08 cm × 5.08 cm) of PCB real estate; hence the PCB will be 16 in. high if the FPGA's are arranged in two vertical columns of eight FPGA's each, as shown in Fig. 6.

The PCB must employ the high-speed board design techniques described in Ref. 30 to reduce cross talk and skew. A multilayer PCB can be used with pairs of power–ground planes interleaved between pairs of horizontal and vertical signal layers to lower power supply noise and cross talk. The use of solid power–ground planes between signal planes reduces cross talk between traces significantly. To reduce cross talk further, guard traces that are grounded at both ends can be inserted between signal traces. For maximizing performance the trace geometries must be controlled to maintain 50-Ω impedance lines, i.e., traces are 0.010 in. wide and 0.0014 in. thick (1-oz. copper is used) and are spaced 0.010 in. apart (i.e., on a 0.020-in. pitch). With guard traces, each signal layer supports 25 signal traces per inch of PCB, and this figure can be used to estimate the number of signal layers and the width of the PCB. For eliminating reflections and maintaining signal integrity,

all high-speed signal traces are end terminated at the receiver. For minimizing skews, all traces in a data path should have comparable length. In addition, the switching IC's can use relatively simple configurable deskewing circuits.

The size of the PCB is easily computed from Fig. 6. Each EO module sends and receives 6.4 Gbits/s of bandwidth to–from the CMOS IC's, and half of this bandwidth crosses the dotted horizontal bisector in Fig. 6. Because the data paths operate at 100 MHz, the number of traces from each EO module crossing the bisector is 64. On either half of the PCB the total number of traces crossing the bisector is 1024. Hence for implementing these traces the shuffles are 8 in. wide with five vertical signal layers (more layers can be used to reduce the width). Therefore an 18-layer PCB will suffice, which is well within current PCB capability. The PCB also needs 8 in. of width for the EO cards. Hence the entire PCB is 16 in. high and 26 in. wide, which is well within the limits of current PCB technology. Furthermore, the design has been conservative and has used relatively wide traces. High-performance PCB's can use much finer metal trace widths (0.004 in.) and up to 40 layers of metal,[30] resulting in much smaller PCB's.

C. High-Speed Complementary Metal-Oxide Silicon Application-Specific Integrated Circuit Design

The single PCB solution can also exploit custom CMOS ASIC's to perform the switching. Bull S.A. in Europe has commercialized a high-speed low-power CMOS I/O technology, called the high-speed link technology, which operates at up to 1-GHz data rates.[31] These I/O pads sense the characteristics of the electrical PCB traces and are then self-configured to support a 1-GHz clock rate. Similar results are reported in Ref. 32. By following our design approaches, one can design a fast reduced-complexity switch-core IC around this high-speed CMOS I/O technology. The simplified core can be scaled upward to 32 I/O ports and pipelined to operate the internal data paths at 400 MHz or higher without exceeding the density of the CMOS process. However, the new high-speed CMOS I/O pads are quite large, measuring roughly 1 mm $\times$ 1 mm each.[31] Hence it is not feasible to implement more than one bit-serial 32 $\times$ 32 crossbar on a single IC because of limitations on the number of I/O pads around the perimeter of the die. Hence once again a bit-slice approach can be used in which eight bit-serial crossbars are operated in parallel.

Eight bit-serial 32 $\times$ 32 high-speed CMOS IC's can be packaged in the center of the PCB in Fig. 6. In this design, the PCB will be smaller, as the slow wide external data paths to the IC's (64 bits at 100 MHz) will be replaced by narrow fast external data paths (8 bits at 800 MHz). As follows from the previous analysis, 6.4 Gbits/s from each EO module will cross the bisector. At an 800-MHz data rate per trace this bandwidth will require eight vertical traces from each EO module (the EO modules are also simplified, as the external electrical data paths between IC's operate at the same clock as the Optobus). In the Bull S.A., high-speed link technology eight vertical traces must cross the bisector for each EO module. In total, 128 vertical traces cross the bisector. These 128 traces can be implemented on two vertical layers that are 3 in. wide. Hence the entire PCB requires six layers and two pairs of vertical–horizontal signal layers, separated by one pair of power–ground layers. The entire PCB is 16 in. high and 16 in. wide. Therefore the use of high-speed CMOS and fast narrow datapaths between IC's has reduced the PCB complexity considerably.

4. Single-Chip Optoelectronic Switch Core that Uses Fiber to Silicon

In this section a smooth scalability from the electronic PCB of Section 3 to a single-chip optoelectronic IC is described. Several optoelectronic switches that use a fiber-to-CMOS–SEED approach have been proposed.[9,21–23] Other technologies have also been used.[10] The single-chip optoelectronic solution results in several optimizations over the single PCB solution. Our general approach is to implement the functionality of the PCB shown in Fig. 6 directly on the optoelectronic IC. The switching IC's shown in Fig. 6 are replaced by integrated crossbars on the silicon substrate; the PCB signal traces that form the external data paths between crossbars and optical I/O (the shuffles in Fig. 6) are replaced by microelectronic traces a few micrometers wide on the IC. Hence the entire PCB can be collapsed onto a single optoelectronic IC. The optical signals from the Motorola Optobus fiber ribbons are fed directly into the optoelectronic IC through an optical imaging system, eliminating 32 Motorola Optobus transceivers in the EO modules shown in Fig. 6. Each Optobus transceiver has a commercial value in the neighborhood of $1000; hence this approach can save several thousands of dollars in a 32-port switch (likely offsetting the cost of the optical imaging system, which is described below.)

Another significant optimization is the hardware savings owing to the use of fast narrow electrical data paths within the crossbars and between the crossbars and the I/O. The Synopsys Hardware Design System was used to synthesize our very-high-speed IC hardware description language design of the switch core, targeted for a semicustom CMOS with a 0.8-μm standard cell library (the Canadian Microelectronics Corporation K-cell Library[33]). Timing analysis of the pipelined layout indicates that the electrical channels within the crossbar switches can be clocked at 400 MHz. Hence, for interfacing to the bytewide Optobus channels clocked at 800 MHz, the internal electrical channels within the crossbars need be only 16 bits wide. Therefore the internal data path width can be reduced from 64 bits in the FPGA's to 16 bits in a CMOS ASIC with a 400-MHz internal clock rate. The use of fast narrow data paths will cause a significant reduction in the logic gate complexity, and it can be verified that the entire switch core will require $\sim 10^6$ logic gates.

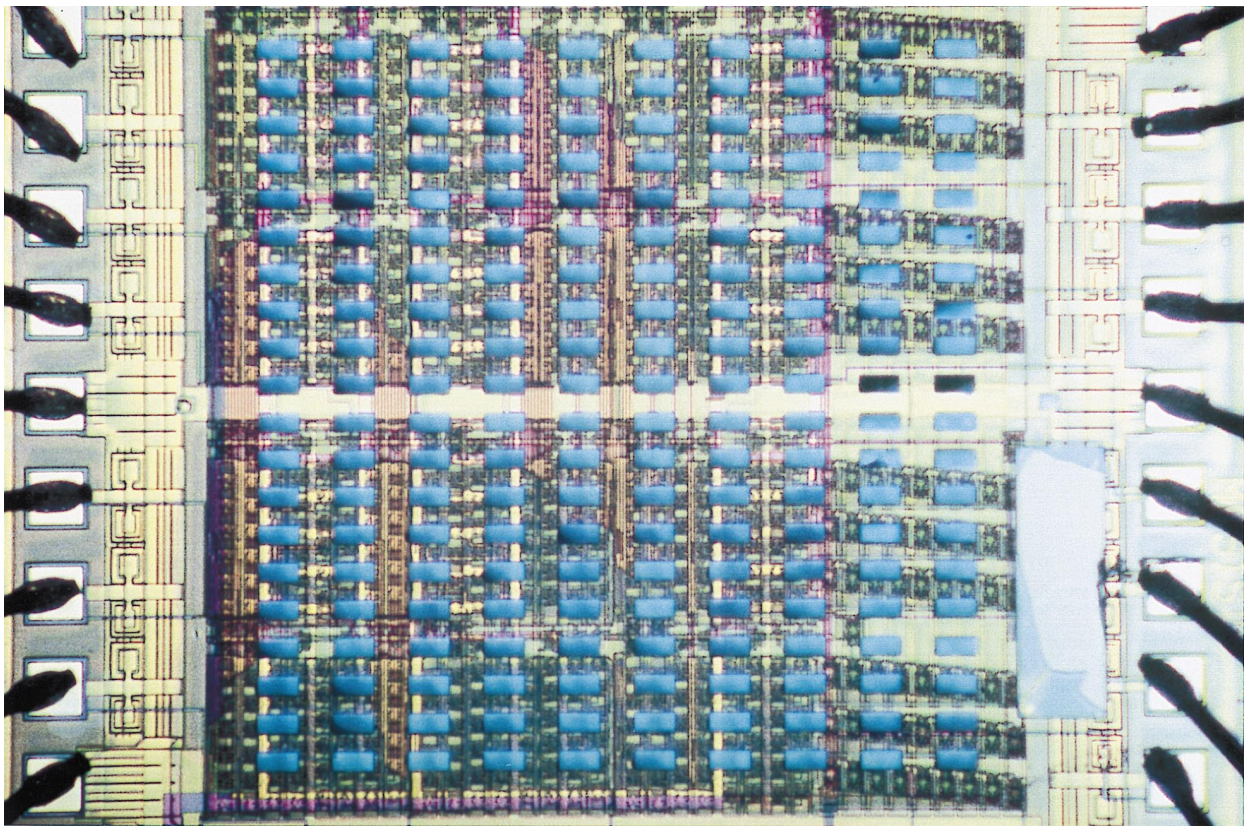


Fig. 7. Optical diodes on a CMOS–SEED IC, which act as I/O's for the 2-D array of optical bits. (See Ref. 39.)

A. Imaging System

We summarize our approach to the imaging system. We have successfully imaged several Optobus fibers onto the CMOS–SEED device by using this approach. An interesting optical fan-out system has been described in Ref. 34. An advantage of this system is the reliance on commercial camera lenses; a disadvantage is that the large-scale fan-out of optical signals lowers their power (by 9.88 dB for 1-to-9 splitting), such that only slow broadcasting is allowed (in the megahertz range). Our approach retains the advantages of the system described in Ref. 34 while avoiding the disadvantages of low power and speed.

In Fig. 7 the pitches of the optical I/O on the CMOS–SEED IC are 62.5 μm and 125 μm. The CMOS–SEED technology is optimized for a wavelength of 850 nm, which matches the nominal wavelength of the Motorola Optobus.[5] The power of each Optobus fiber is 0.8 mW at the source, a very high initial power level. Our measurements indicate that each Motorola parallel-fiber ribbon is 3125 μm wide and 450 μm high (3.125 mm × 0.45 mm), with a small variance. Within any ribbon the pitch between fibers is 250 μm. An aluminum frame with an inner open window that is 8 mm × 3.125 mm is used to contain physically the polished ends of 32 one-dimensional (1-D) fiber ribbons, which are stacked on top of each other. When each ribbon is separated with a Teflon sheet ~50 μm thick, the 32 ribbons occupy 8 mm and fit tightly within the frame. The spacing between fibers in neighboring ribbons along one axis is 500 μm, and the spacing between fibers within any ribbon is 250 μm. (The use of Teflon sheets to maintain a 500-μm pitch is important, as otherwise the images would not align with the IC.) This simple mechanical frame supports 32 fiber ribbons and a two-dimensional (2-D) 32 × 10 array of optical beams. A conventional camera lens is used to focus these beams onto the optoelectronic IC. With an image compression of 4, these beams are focused down to a smaller 2-D array that matches exactly the pitch of the optical receivers on the CMOS–SEED device (125 and 62.5 μm). A representative output of such an imaging system is shown in Fig. 8.

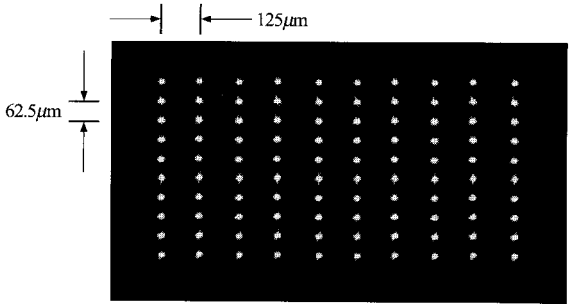


Fig. 8. Representation of the optical output of a 1-D array of 1-D parallel fiber ribbons after 4-to-1 compression.

Table 1. Projections for CMOS–SEED Technology[8,38]

Year	Feature Size (μm)	Gates[38] (× 10^6)	Area (mm^2)	Maximum Number of Optical I/O's[8]	Optical Clock[8]	Maximum Optical I/O BW^a (Tbps)
1995	0.35	0.8	400	6000	200 Mhz	0.6
1998	0.25	2	600	12,000	350 Mhz	2.1
2001	0.18	5	800	24,000	500 Mhz	6
2004	0.12	10	1000	40,000	700 Mhz	14
2007	0.10	20	1,250	50,000	1 Ghz	25

[a]*BW* is the product of the optical I/O times the optical clock divided by 2.

A similar system is used for imaging the 2-D optical output from the IC onto the fiber ribbons (with an image expansion of 4), which carry optical data back to the workstations. If the optoelectronic IC uses vertical-cavity surface-emitting laser arrays to generate optical outputs (e.g., see Refs. 35 and 36), then no additional imaging apparatus is needed. The total cost of the imaging assembly comprises the metal frames (negligible) and the camera lens (with a cost of less than $2000). If the optoelectronic IC uses SEED modulators, then an additional optical power imaging system for supplying the SEED modulators with laser power is required. Such optical power systems are used in all SEED devices; for example, see Ref. 22 and the references therein. The optical power system can be placed on either side of the camera lens; the misalignment tolerance is larger if the power supply is used before the image compression, making system alignment easier. The cost of the laser power supply can be currently estimated at $5000, with approximately $4000 for the remaining optical components. Overall, the cost of this optoelectronic solution seems to compare favorably with the conventional PCB solution; the optoelectronic solution can potentially realize a net savings because 32 Optobus transceivers are eliminated from the core.

The multimode fibers in the Optobus are 62.5 μm wide, and their beams diverge after leaving the fiber.[5] Assume conservatively that each beam expands to an 80-μm waist before focusing. After a 4-to-1 compression the beam is focused to a 20-μm waist, which is imaged onto the SEED diodes (which have dimensions of approximately 25 μm × 55 μm) on the IC. We have demonstrated excellent focusing of fibers onto SEED's by using this approach, with spot sizes as small as 5 or 10 μm, without requiring any microlenses. In the reverse direction the optical outputs from the camera lens will have wide waists, resulting in only moderately efficient coupling into the multimode fibers. Refractive microlens arrays with pitches of 500 and 250 μm can be used to focus the individual beams for efficient coupling. In this application inexpensive polymer microlenses such as those described in Ref. 37 can be used.

B. Scalability

Table 1 illustrates the scalability of the CMOS technology[38] and the CMOS–SEED technology[8] over the next decade. With 1998 technology a high-performance smart-pixel array can easily meet the requirements of our single-chip 204-Gbit/s core, i.e., 640 optical I/O and 2 × 10^6 logic gates. A CMOS–SEED device that can be programmed to implement a simplified single-chip optoelectronic switch core has been developed with the 1995 AT&T–U.S. Advanced Research Project Agency–CO-OP smart-pixel technology (0.8-μm CMOS) and became operational in 1996. The device is shown in Fig. 7 and is described in Ref. 39. The device demonstrates the feasibility of dense electronics with optical I/O and supports approximately 1000 logic gates per optical I/O.

5. Error Control

The eye diagrams of a Motorola Optobus link at various clock rates are shown in Fig. 9. Various direct output traces are also shown in Fig. 9. These figures indicate that the BER increases with frequency as the eye opening narrows.

With the growing use of high-bandwidth optical data links, hardware-based error detection is necessary, and optional error correction may be advantageous. The more advanced error correction codes employ spectral domain methods, e.g., fast Fourier transforms.[40,41] Popular ECC schemes such as linear block codes, Reed–Solomon codes, and convolution codes[40] offer excellent correction capability in the presence of random as well as burst errors. However, these techniques are not well suited for high-bandwidth optical networks, as they require a significant amount of hardware for implementation. For example, it was shown in Ref. 41 that, given an input BER of 10^{-4} and a required output BER of 10^{-12}, a state-of-the-art parallel spectral Reed–Solomon decoder to process a 77-Gbit/s data link will require a chip area of 1 cm × 1 cm with a 1-μm CMOS process. A decoder for a 1-Tbit/s optical data link will require 13 cm × 13 cm of CMOS area, which is unrealistically large. Hence, even as CMOS technology improves, traditional ECC codes for high-bandwidth optical data links will be infeasible.

Our optical LAN uses a different approach to error control. Our LAN has a great deal of bandwidth, and it is reasonable to trade optical bandwidth for ECC simplicity. Simpler codes will require more redundant bits and have a lower coding efficiency but will also require much less processing. Because we have ample bandwidth, the trade-off for hardware simplicity is worthwhile. Hence we are exploring the use of a simple hardware-based 2-D parity check to detect and optically correct errors on an Optobus data link. The basic 2-D scheme is not as robust as the Reed–Solomon scheme reported in Ref. 41 in the presence of burst errors, but our experimental results suggest that burst errors are extremely uncommon.

In our 2-D parity check each 256-bit packet is divided into b blocks, in which each block is arranged in an $n \times m$ array of bits and each row and column contains a parity bit. In conventional terminology this yields an $[n \times m + (n + m), nxm]$ code, with a

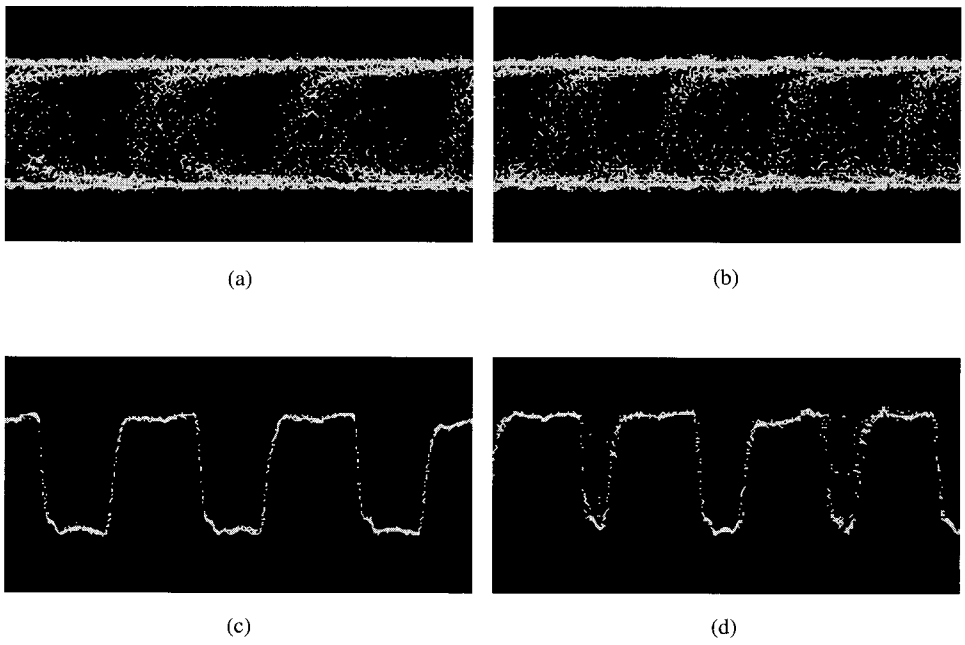


Fig. 9. (a) Eye diagram of Motorola Optobus fibers at 600 Mbits/s. (b) Eye diagram of Motorola Optobus fibers at 800 Mbits/s. (c) Waveform of Motorola Optobus fibers at 600 Mbits/s. (d) Waveform of Motorola Optobus fibers at 800 Mbits/s.

code-word length of $[n \times m + (n + m)]$ bits and with $(n + m)$ check-sum bits. Given that bit errors are modeled as independent events, it can be verified that the probability of an undetected error in a packet is given by

$$\Pr(\text{undetected error in packet}) \cong b\binom{n+1}{2}\binom{m+1}{2}p^4(1-p)^{(n+1)(m+1)-4}, \quad (2)$$

where p is the BER on the parallel-fiber link.

Given the fixed packet format shown in Fig. 3, the effects of varying the number of blocks are investigated (assuming a BER of 10^{-8} at 800 MHz). Figure 10(a) plots the probability of an undetected error within a packet versus the BER on a fiber ribbon for one, two, four, and six blocks. The assumed operating point of our system is indicated in Fig. 10(a), and it shows a very low probability of undetected error in the 10^{-29} range, after hardware-based error control.

A. Mean Time to Undetected Error, after Error Control

The mean time to undetected error (MTTE) is defined as the average time between the occurrence of two consecutive undetected errors, after our hardware error control scheme, and it can be computed, given the data transmission bandwidth and the BER, as

$$\text{MTTE}_{\text{system}} = [\Pr(\text{undetected error per bit}) \times \text{transmission bandwidth}]^{-1}. \quad (3)$$

This equation is useful for determining whether any additional error checking with software is required. The bandwidth of our optical LAN is 204.8 Gbits/s, and we assume a BER on the fiber of 10^{-8} at 800 Mbits/s. After the hardware-based error control scheme, our MTTE is computed as 5.8×10^{19} s [see Fig. 10(b)]. It is estimated that there are only 10^{17} s left in the life of our universe. Hence undetected errors in the optical LAN would never occur over the life of any machine we can construct (if the assumption on the frequency and distribution of burst errors is correct). Hence it is plausible that additional error checking in software may not be necessary.

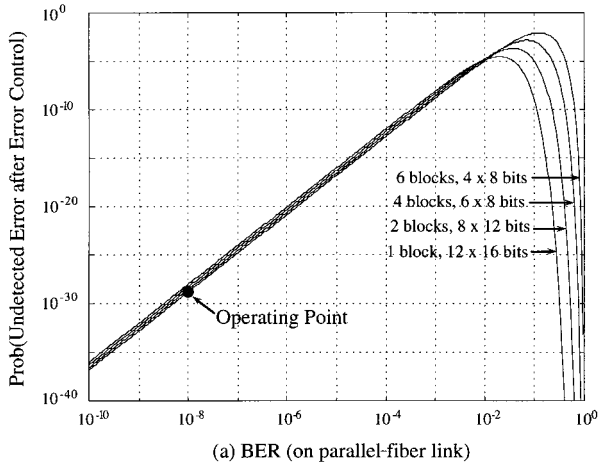
Without hardware-based error control, bit errors would occur in the multiprocessor system once every 4.8×10^{-4} s or, equivalently, approximately 10^4 errors/s. These errors could quickly contaminate the computation of the entire multiprocessor system. Without hardware detection and automatic correction these errors would have to be handled by software protocols.

B. Reliability

If the reliability of a component is exponentially distributed, the reliability of a series system with n components, in which all components must be functioning, can be expressed as[42]

$$R_{\text{series-system}}(t) = \exp\left(-t\sum_{i=1}^{n}\lambda_i\right). \quad (4)$$

Reliability can be improved by the addition of redundant components, yielding an m-out-of-n system.[42] Such a system contains a total of n components and requires that only m of them be functioning for the entire system to function. If we assume that the reliability of a component is expo-



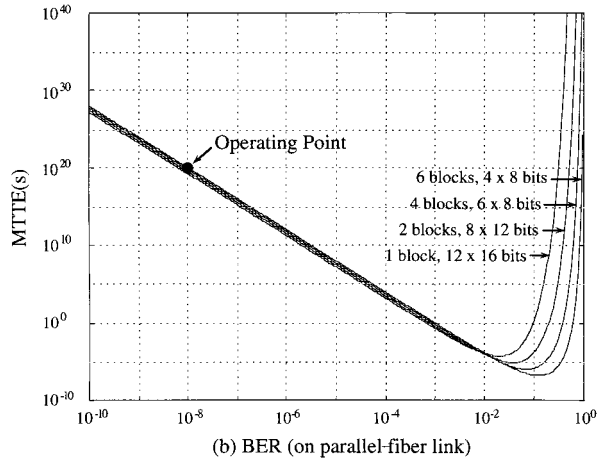


Fig. 10. (a) Probability of undetected error, after error control, versus BER in fiber. (b) Mean time to undetected error, after error control, versus BER in fiber.

nentially distributed, the reliability of an m-out-of-n system is given by

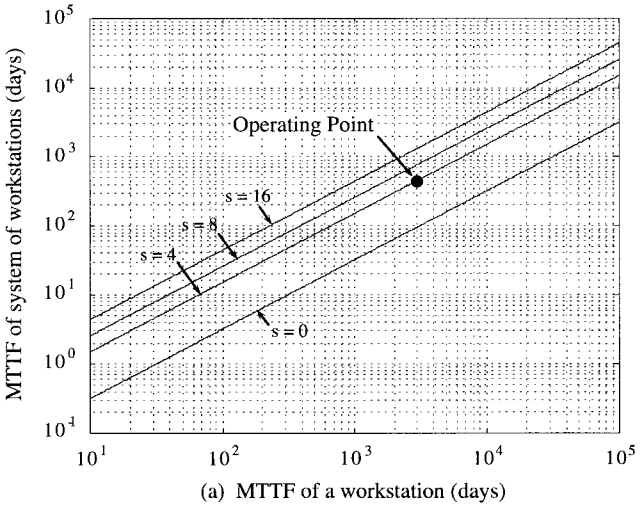
$$R_{m\text{-out-of-}n}(t) = \sum_{i=m}^{n} \binom{n}{i} \binom{i-1}{m-1} (-1)^{i-m} \exp(-i\lambda t), \quad (5)$$

and the mean time to failure (MTTF) is expressed as

$$\mathrm{MTTF}_{m\text{-out-of-}n} = \sum_{i=m}^{n} \frac{1}{i\lambda}. \quad (6)$$

Our electronic switch core is implemented with a bank of parallel FPGA's, as shown in Fig. 6. If there are two spare FPGA's in the core, there are a total of 18 hardwired FPGA's and the switch core can be modeled as a 16-out-of-18 system. Referring to Fig. 1, if we have four spare workstations and require 32 functioning workstations, the workstations form a 32-out-of-36 system.

Our two modules (switch core and the collection of workstations) form a series system. Let $\lambda_{\text{switch-core}}$ be the failure rate of the m-out-of-n switch core and



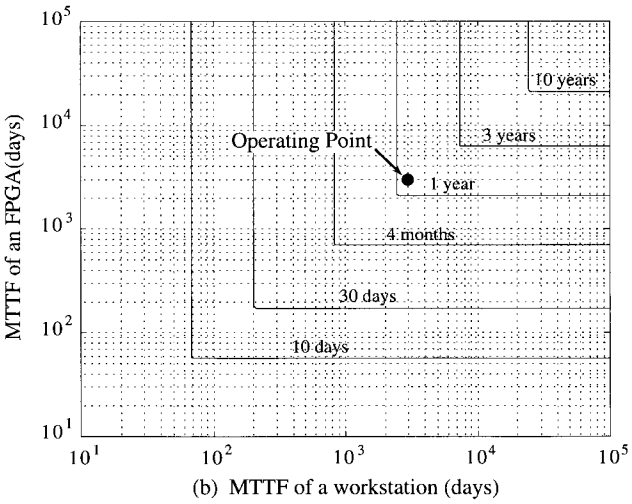


Fig. 11. (a) MTTF of the system of workstations with a varying number of spares (32 out of 32 + s workstations). (b) MTTF of the overall system composed of the switch core (16-out-of-18 FPGA's) and the workstations (32-out-of-36 workstations).

$\lambda_{\text{workstations}}$ be the failure rate of the m-out-of-n collection of workstations. However, the lifetime of each module is not exponentially distributed. In this case the reliability of the system is determined by the reliability of the weakest component in the system, and we can evaluate the overall MTTF as[42]

$$\mathrm{MTTF}_{\text{overall}} = \min(\mathrm{MTTF}_{\text{switch-core}}, \mathrm{MTTF}_{\text{workstations}}). \quad (7)$$

According to Altera reliability reports,[43] each FPGA has a typical failure rate of 1.35×10^{-8} failures/h (140 °C, 6-V operation), which corresponds to less than a single failure in 8000 years of use. This figure neglects other factors such as electronic pin failures, which are much more common. In our analysis we assume a more realistic failure rate of one FPGA failure in 10 years, i.e., a MTTF of 3000

days. We assume a similar failure rate for the workstations.

Figure 11(a) illustrates the effects of adding spare workstations to the system. The MTTF of the m-out-of-n collection of workstations is computed with Eq. (6). At our assumed operating point, the MTTF of the collection of workstations increases to approximately 450 days from 100 days by the addition of only four spare workstations. Therefore there is a sizable improvement in the MTTF of the collection of workstations with relatively low overhead. The addition of spare FPGA's leads to a similar improvement in the switch-core reliability.

Figure 11(b) illustrates the MTTF of the entire multiprocessing system as computed with Eq. (7). Our assumed operating point yields an overall MTTF of greater than 1 year. Without the addition of spares the overall MTTF was 2 months. Thus the addition of redundancy can significantly improve the system's reliability.

Contrary to the conventional view that single-chip devices can be unreliable, the reliability of the single-chip optoelectronic IC can be made very high (i.e., much more reliable than several parallel electronic IC's) with a careful design. The switch core can use the bit-sliced approach shown in Fig. 6, with several parallel crossbars implemented on a single IC. Each crossbar can use independent electrical I/O pins for power and ground, and independent devices for optical I/O. Hence each crossbar will fail independently of any other. Electrical I/O pin failure is the dominant failure mode of an IC. These crossbars require only a small number of electrical I/O pins (for power and ground) as all data are sent and received through optical I/O's. Hence each crossbar will be much more reliable than the FPGA's modeled above. Therefore our proposed single-chip optoelectronic IC will be much more reliable than the purely electronic switch core or previously proposed optoelectronic switches. The optoelectronic core can also implement spare crossbars on the same IC substrate, so that the above analyses are still applicable.

6. Conclusions

The design of a multigigabit optical network for multiprocessing systems has been proposed. The network can be used to connect PCB's in a multigigabit backplane, to connect backplanes in a cabinet, or to connect cabinets in a massively parallel processing system. We have illustrated smooth scalability from the electronic to optoelectronic domains. In the electronic domain, the switch core can be implemented on a single PCB with parallel CMOS IC's. In the optoelectronic domain, the switch core can be implemented with a single-chip optoelectronic IC that scales to terabits of bandwidth. In both cases, parallel fibers interconnect the workstations to the switch core.

Finally, we are exploring several relatively novel concepts in optical networking. We show how the use of parallel fiber can simplify many design decisions in an optical network. The availability of low-skew fiber makes the broadcasting of timing information from the switch core attractive. This approach can be used to coordinate packet transmissions and thereby reduce packet buffering in the core and can eliminate the need for PLL's to recover clocking information. Our design favors the space-multiplexing of multiple pipelined VC's over fiber ribbons, resulting in relatively simple switch-core IC's, which can then be pipelined and operated at very fast clock rates, further reducing their logic complexity. We are also exploring simple and fast hardware-based error control schemes, with the objective of achieving an optical LAN with communication that is as error free as the main memory, thereby eliminating the need of software-based error control schemes. Our results thus far indicate that these approaches should prove feasible.

This research was funded by Natural Sciences and Engineering Research Council of Canada grant OGP011211601 and in part by Spar Aerospace Limited. The fabrication of the CMOS–SEED chip was funded by the U.S. Advanced Research Project Agency. Computing equipment, computer-aided-design tools, and device fabrication were supplied by the Canadian Microelectronics Corporation. A large number of graduate students in the Microelectronics and Computer System Laboratory at McGill University have contributed to our research project, including Manoj Verghese, Razvan Buhescu, Andrew Olum, Palash Desai, Stephane Gagnon, Sabeen Randhawa, Sherif Sherif, Michael Kim, and Winnie Ho. Our thanks to Sherif Sherif for the VLSI layout of the IC in Fig. 7.

References and Note

1. J. L. Hennessey and D. A. Patterson, *Computer Architecture, A Quantatative Approach*, 2nd ed. (Morgan-Kauffman, San Francisco, 1995).
2. T. Lewis, "The next $10{,}000_2$ years: parts 1," IEEE Comput. **29**(4), 64–70 (1996); "Part 2," IEEE Comput. **29**(5), 78–86 (1996).
3. D. Clark, "Breaking the teraflops barrier," IEEE Comput. **30**(2), 12–14 (1997).
4. T. E. Anderson, D. E. Culler, and D. A. Patterson, "A case for NOW (networks of workstations)," IEEE Micro. **16,** 54–64 (1995).
5. *OPTOBUS Data Sheet*, Logic Integrated Circuits Division, Motorola Inc., Chandler, Ariz. 85248, 1995.
6. D. B. Schwartz, C. K. Y. Chun, B. M. Foley, D. H. Hartman, M. Lebby, H. C. Lee, C. L. Shieh, S. M. Kuo, S. G. Shook, and B. Webb, "A low-cost, high performance optical interconnect," IEEE Trans. Components, Packag. Manuf. Technol. B **19,** 532–539 (1996).
7. D. R. Engebretsen, D. M. Kuchta, R. C. Booth, J. D. Crow, and W. G. Nation, "Parallel fiber-optic SCI links," IEEE Micro. **16,** 20–26 (1996).
8. A. V. Krishnamoorthy and D. A. B. Miller, "Scaling optoelectronic-VLSI circuits into the 21st century: a technology roadmap," IEEE J. Sel. Topics Quantum Electron. **2,** 55–76 (1996).
9. A. V. Krishnamoorthy, J. E. Ford, K. W. Goossen, J. A. Walker, B. Tseng, S. P. Hui, J. E. Cunningham, W. Y. Jan, T. K. Woodward, M. C. Nuss, R. G. Rozier, F. E. Kiamilev, and D. A. B. Miller, "The AMOEBA chip: an optoelectronic switch

for multiprocessor networking using dense-WDM," in *Proceedings of the Third International Conference on Massively Parallel Processing Using Optical Interconnections* (*MPPOI'96*) (Institute of Electrical and Electronics Engineers Computer Society, Los Alamitos, Calif., 1996), pp. 94–100.
10. W. A. Crossland and T. D. Wilkinson, "Optically transparent switching in telecommunications using ferroelectric liquid crystals over silicon VLSI circuits," in *Proceedings of the 1996 IEEE/LEOS Summer Topical Meeting* (Institute of Electrical and Electronics Engineers Service Center, Piscataway, N.J., 1996), pp. 22–23.
11. T. Chaney, J. A. Fingerhut, M. Flucke, and J. S. Turner, "Design of a gigabit ATM switch," Tech. Rep. WUSC-96-07 (Applied Research Laboratory, Department of Computer Science, Washington University, St. Louis, Mo., 1996).
12. C. B. Stunkel, D. G. Shea, B. Abali, M. G. Atkins, C. A. Bender, D. G. Grice, P. Hochschild, D. J. Joseph, B. J. Nathanson, R. A. Swetz, R. F. Stucke, M. Tsao, and P. R. Varker, "The SP2 high-performance switch," IBM Syst. J. **34,** 185–204 (1995).
13. 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," J. Lightwave Technol. **13,** 1093–1103 (1995).
14. N. McKeown, M. Izzard, A. Mekkittikul, W. Ellersick, and M. Horowitz, "Tiny Tera: a packet switch core," IEEE Micro. **17,** 26–33 (1997).
15. M. J. Karol, M. G. Hluchyj, and S. P. Morgan, "Input vs. output queueing on a space division packet switch," IEEE Trans. Commun. **COM-35,** 1247–1356 (1987).
16. S. Scott, "The gigaring channel," IEEE Micro. **16,** 27–34 (1996).
17. M. Galles, "Spider: a high-speed network interconnect," IEEE Micro. **17,** 34–39 (1997).
18. T. E. Anderson, S. S. Owicki, J. B. Saxe, and C. P. Thacker, "High-speed switch scheduling for local-area networks," ACM Trans. Comput. Syst. **11,** 319–352 (1993).
19. Y. S. Yeh, M. G. Hluchyj, and A. S. Acampora, "The knockout switch: a simple modular architecture for high performance packet switching," IEEE J. Sel. Areas Commun. **5,** 1274–1283 (1987).
20. Y. Tamir and H. C. Chi, "Symmetric crossbar arbiters for VLSI communication switches," IEEE Trans. Parallel Distribut. Syst. **4,** 13–27 (1993).
21. S. J. Hinterlong, A. L. Lentine, D. J. Reiley, J. M. Sasian, R. L. Morrison, R. A. Novotny, M. G. Beckman, D. B. Buchholz, T. J. Cloonan, and G. W. Richards, "An ATM switching system demonstration using a 40 Gb/s throughput smart pixel optoelectronic VLSI chip," in *Proceedings of the 1996 IEEE/LEOS Summer Topical Meeting* (Institute of Electrical and Electronics Engineers Service Center, Piscataway, N.J., 1996), pp. 47–48.
22. T. J. Cloonan, "Comparative study of optical and electronic interconnection technologies for large asynchronous transfer mode packet switching applications," Opt. Eng. **33,** 1512–1523 (1994).
23. A. L. Lentine, K. W. Goossen, J. A. Walker, L. M. F. Chirovsky, L. A. D'Asaro, S. P. Hui, B. T. Tseng, R. E. Leibenguth, J. E. Cunningham, W. Y. Jan, J. M. Kuo, D. Dahringer, D. Kossives, D. D. Bacon, G. Livescu, R. L. Morrison, R. A. Novotny, and D. B. Buchholz, "Optoelectronic VLSI switching chip with greater than 4,000 optical flip-chip-bonded GaAs/AlGaAS MQW modulators and detectors on silicon CMOS circuitry," in *Conference on Lasers and Electro-Optics*, Vol. 9 of 1996 OSA Technical Digest Series (Optical Society of America, Washington, D.C., 1996), pp. 517–518.
24. T. Robertazzi, ed., *Performance Evaluation of High Speed Switching Fabrics and Networks: ATM, Broadband ISDN, and MAN Technology* (Institute of Electrical and Electronics Engineers, New York, 1993).
25. T. H. Szymanski and B. Supmonchai, "Reconfigurable computing with optical backplanes—an economic argument for optical interconnects," in *Proceedings of the Third International Conference on Massively Parallel Processing Using Optical Interconnections* (*MPPOI'96*) (Institute of Electrical and Electronics Engineers Computer Society, Los Alamitos, Calif., 1996), pp. 321–328.
26. D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed. (Prentice-Hall, Englewood Cliffs, N.J., 1992), Chap. 2.
27. T. H. Szymanski, "Design principles for practical self-routing nonblocking switching networks with $O(N\log N)$ bit complexity," IEEE Trans. Comput. **46**, 1057–1069 (1997).
28. Further information will be available in a future paper entitled "Fast self-routing concentrators for optoelectronic IC's," by T. H. Szymanski and B. Supmonchai. B. Supmonchai is with the Microelectronics and Computer Systems Laboratory, Department of Electrical Engineering, McGill University, 3480 University Street, Montreal, Quebec, Canada H3A 2A7.
29. T. H. Szymanski and H. S. Hinton, "Reconfigurable intelligent optical backplane for parallel computing and communications," Appl. Opt. **35,** 1253–1268 (1996).
30. H. W. Johnson and M. Graham, *High-Speed Digital Design: A Handbook of Black Magic* (Prentice-Hall, Englewood Cliffs, N.J., 1993), Chaps. 4–6.
31. B. Zerrouk, A. Greiner, V. Reibaldi, F. Potter, A. Derieux, R. Marbot, and R. Nezamzadeh, "The HIC high speed link technology and associated router," Real-Time Mag. **3,** 73–77 (1996).
32. W. J. Dally and J. Poulton, "Transmitter equalization for 4-Gbps signaling," IEEE Micro. **17,** 48–56 (1997).
33. BiCMOS Design Kit V2.0 for Synopsys, Canadian Microelectronics Corporation, Carruthers Hall, Queen's University, Kingston, Ontario, K7L 3N6, Canada, 1996.
34. F. E. E. Frietman, *Opto-Electronic Processing and Networking: A Design Study* (Delft University of Technology, Faculty of Applied Physics, Lorentzweg 1, NL-2628 CJ Delft, The Netherlands, 1995).
35. E. M. Hayes, R. D. Snyder, R. Jurrat, S. A. Feld, C. W. Wilmsen, K. D. Choquette, K. M. Geib, and H. Q. Hou, "8×8 array of smart pixels fabricated through the Vitesse foundry integrating MESFET, MSM, and VCSEL elements," in *Proceedings of the 1996 IEEE/LEOS Summer Topical Meeting* (Institute of Electrical and Electronics Engineers Service Center, Piscataway, N.J., 1996), pp. 103–104.
36. S. Matsuo, T. Nakahara, Y. Kohama, Y. Ohiso, S. Fukushima, and T. Kurokawa, "Monolithically integrated photonic switching device using an MSM PD, MESFET's, and a VCSEL," IEEE Photonics Technol. Lett. **7,** 1165–1167 (1995).
37. B. P. Keyworth, D. J. Corazza, and J. N. McMullin, "Single-step fabrication of refractive microlens arrays," Appl. Opt. **36,** 2198–2202 (1997).
38. "The national technology roadmap for semiconductors," Semiconductor Industry Association, San Jose, Calif., 1994.
39. S. Sherif, T. H. Szymanski, and H. S. Hinton, "Design and implementation of a field programmable smart pixel array," in *Proceedings of the 1996 IEEE/LEOS Summer Topical Meeting* (Institute of Electrical and Electronics Engineers Service Center, Piscataway, N.J., 1996), pp. 78–79.
40. R. E. Blahut, *Theory and Practice of Error Control Codes* (Addison-Wesley, Reading, Mass., 1983).
41. M. A. Neifeld and S. K. Sridharan, "Parallel error correction using spectral Reed–Solomon code," J. Opt. Commun. **18**, 144–150 (1997).
42. K. S. Trivedi, *Probability & Statistics with Reliability, Queueing, and Computer Science Applications* (Prentice-Hall, Englewood Cliffs, N.J., 1982), Chaps. 3 and 4.
43. "FPGA data book," Reliability Rep. No. 26 (Altera Corp., San Jose, Calif., 1996).