A Gbyte/s Parallel Fiber-Optic Network Interface for Multimedia Applications

B. Raghavan, Y.-G. Kim, T.-Y. Chuang, B. Madhavan, and A. F. J. Levi

Department of Electrical Engineering
University of Southern California
Los Angeles, CA 90089-1111

Abstract

We explore parallel fiber-optics and single-chip CMOS interface solutions to support broadband multimedia applications in a small workgroup environment. Initial implementation of a network interface card used in our testbed of Intel Pentium-based PCs is described. Feasibility of a CMOS bridge to parallel optics is demonstrated using an adapter chip implemented in 0.5 μm CMOS technology for a point-to-point configuration and providing a measured peak bandwidth of 4.8 Gbit/s on independent transmit and receive directions. In addition, we describe the design of a Gbyte/s parallel fiber-optic network interface for a slotted ring network currently being developed. We present a medium access control to be implemented on a single CMOS chip. The proposed network is a potentially cost-effective solution for small workgroup clusters.

ACKNOWLEDGEMENTS

We thank the PONI Team at Hewlett-Packard Laboratories, Palo Alto, California for their support and providing us with HP-POLO-2 optoelectronic modules. We thank P. Wijetunga for his assistance in simulations of the P2P chip, the contributions of B. Sano and Y. Hong to initial network designs, and the discussions with S. Zeadally and W. Cui. This work is supported by DARPA under agreement #MDA972-94-3-0038 and #MDA972-97-3-0008.

1. Introduction

The convergence of computing and communications in recent times has enabled a new class of applications with increasing bandwidth requirements and more stringent delivery-time constraints. In the case of high-performance workgroup environments in small campus networks, several bandwidth intensive multimedia applications such as multicast videoconferencing, remote graphics visualization, and distributed computing have emerged. Conventional network interfaces and network designs are severely strained to meet these demands at an affordable cost.

The continuing advances in Complementary Metal Oxide Semiconductor (CMOS) technology and shrinking of minimum feature sizes in accordance with "Moore's Law" have enabled the development of high-performance processors and highly integrated systems on a chip. Inexpensive Personal Computers (PC) such as Intel Pentium-based PCs have been consistently improving in performance and commercialization of processors with clock rates in excess of 1 GHz is fast approaching [1]. They thus offer a cost-effective alternative to conventional workstations. By using a cluster of such machines in a high-bandwidth networked environment significant reconfigurable computing resources can be achieved at reasonable cost.

High-performance CMOS-based integrated circuits (ICs) can be designed to interface between a host computer and network [2] providing bandwidth capabilities that were only available previously using expensive technologies such as bipolar Emitter-Coupled Logic (ECL). CMOS-based IC design is attractive because it leverages the cost benefits of an inherently simpler process compared to silicon bipolar and leverages the infrastructure of high-volume commodity ICs for high-performance circuit applications. It also offers a higher level of integration and the potential to replace older multi-chip circuitry with single-chip CMOS-based solutions.

The network physical medium needed to meet the requirements of a professional high-performance workgroup environment could be either electrical or optical [3]. Current small area networks are constructed using electrical links. Copper cables are however bulky and restricted in density, bandwidth and interconnect distance capabilities. For example, the electrical link described in [4] provides a 10 Gbit/s serial link over distances less than 20 m, using thick coaxial cable. In contrast, optical fiber-based links offer immense benefits including essentially unlimited bandwidth, immunity from electromagnetic interference, and a significantly higher

edge connection density (form-factor). The fiber medium used in optical links could be either single-mode or multimode. While single-mode fiber is capable of achieving a higher bandwidth-distance product than multimode fiber, the cost of the transceiver module, packaging and fiber connectors is higher rendering it inappropriate for workgroup environments. A parallel fiber-ribbon constructed using multimode fibers is a natural and low-cost solution to provide the needed total interconnect-bandwidth while interfacing with the wide data buses of CMOS-based systems in a less complex, and power-efficient way compared to conventional serial link approaches. Multimode fiber-ribbon based links can be used at Gbit/s/line signal rates over distances greater than a kilometer [5]. Fiber Distributed Data Interface (FDDI) [6] and the Scalable Coherent Interface standard (SCI) [7] are examples of existing link standards that use serial fiber-optic technology. The High Performance Parallel Interface HIPPI-6400 is an emerging standard for parallel fiber-optic links [8].

The actual data bits are transmitted on fiber using light from laser diodes which is detected using photodiodes. Advances in Vertical Cavity Surface Emitting Lasers (VCSEL) [9][10] and new optoelectronic packaging technologies make possible the construction of low-cost optoelectronic components so that there is now the potential for constructing inexpensive high-performance network interfaces.

In this paper, we explore the use of parallel fiber-optics and single-chip CMOS interface solutions to construct broadband network interconnects. We propose to construct a shared-medium ring network targeted at supporting high-performance multimedia applications in a small workgroup environment where total cost of the network interface is particularly significant. As an initial step towards this goal, we have constructed a prototype testbed for a point-to-point communication link. We first present a description of and results obtained with this testbed. Subsequently, we describe the architecture of a ring network. This ongoing experimental work reflects our participation in the Parallel Optical Network Interconnect (PONI) program, which is a DARPA-funded consortium of industry and academic partners [11] developing components and exploring the use of parallel fiber-optic interconnects in systems.

In Sections 2-4, we present a description of an implemented testbed connecting two PCs on a point-to-point link using a network interface card (NIC) that we have constructed. We describe relevant aspects of a CMOS interface chip known as the P2P, a PCI host interface card and an experimental optoelectronic transceiver that comprise the NIC. In the remaining sections, we

describe our proposal for a network based on a slotted-ring topology known as the PONI network. We present details of a proposed medium access control (MAC) protocol to be implemented using a CMOS link adapter chip known as LAC. We conclude with a comparison of PONI with other LANs and a discussion on the important lessons learned from our experiences.

2. Experimental link testbed

In this section, we describe a prototype experimental testbed that implements a point-to-point link connecting two PCs. We demonstrate the potential for constructing CMOS-based interfaces with parallel fiber-optics. The testbed also enables us to develop multimedia applications and obtain initial system performance measurements. The architecture and photograph of the link testbed are shown in Figure 1.

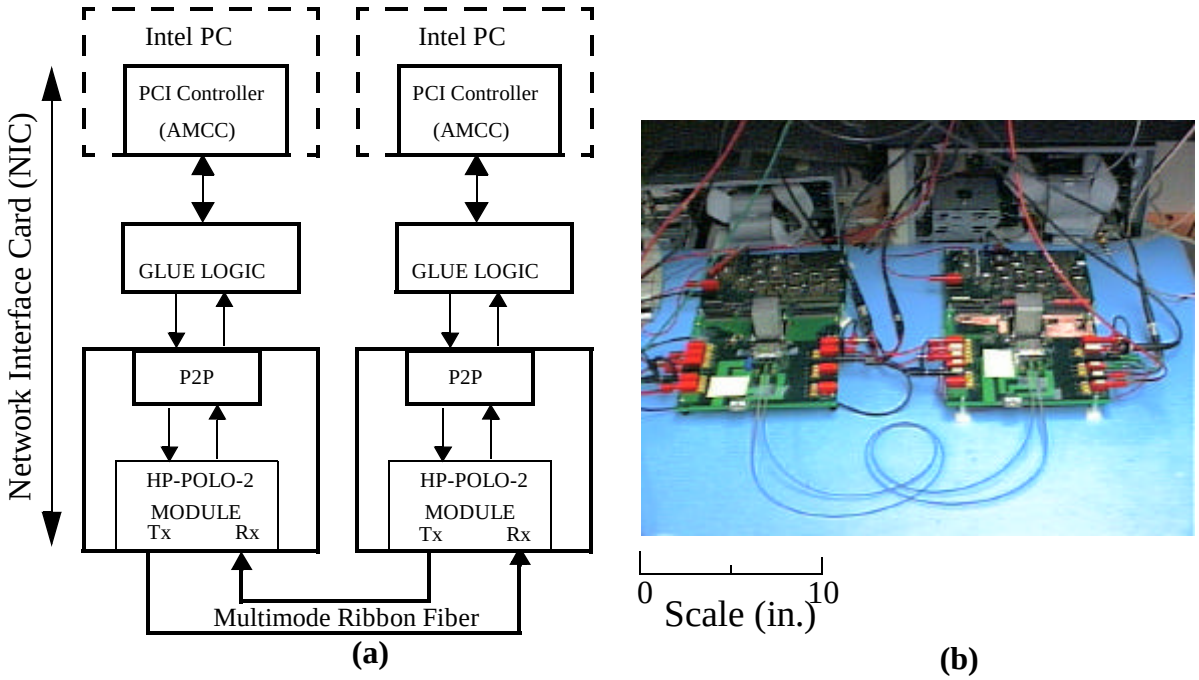

Figure 1: (a) Illustrates the experimental arrangement for a point-to-point PCI-based link. (b) Photograph of the experimental arrangement. In the foreground is the looped multimode fiber-ribbon which connects two HP-POLO-2 transceivers. Intel Pentium-based PCs in the background constitute the host computers. A commercially available PCI controller (AMCC) is installed inside the PC and connected to an external glue logic board using an electrical ribbon connector.

The initial link testbed interconnects two PCs over a point-to-point link using a network interface card (NIC) that we constructed. The schematic of the NIC is shown in Figure 1(a). The

NIC was implemented using a combination of novel specialized hardware solutions, commercially available components, and experimental components. Intel Pentium-based PCs running the Windows NT 4.0 operating system constitute the host. A commercially available AMCC Matchmaker developer's board [12] installed in the host provides a 32-bit 33 MHz half-duplex interface with the internal PCI [13] bus. A commercially available device driver [14] was modified by us for performing data transfers over the AMCC board.

A photograph of the custom-designed portion of the NIC is shown in Figure 2. The board on the left is called the glue logic board. The glue logic board performs signaling functions necessary to bridge the AMCC board with a custom-designed CMOS chip known as the P2P. Currently, the P2P chip that has been designed and tested substitutes for the final link adapter chip (LAC) of our proposed ring network. It is described in detail in the following section. The glue logic and P2P are located on separate boards for ease of testing in our experimental implementation. By handling the glue logic functions external to the P2P, its use is not restricted to any particular I/O bus interface. By constructing a new glue logic board, the same P2P can be connected with any other existing or emerging bus standards such as the Accelerated Graphics Port (AGP). The glue logic board consists of independent 4 kbyte buffers on its transmit and receive directions. These buffers temporarily store and synchronize data transfers between the PCI controller and the P2P. The glue logic interface with P2P is a full-duplex 32-bit interface designed to operate at a maximum frequency of 50 MHz providing a peak bisection bandwidth of 3.2 Gbit/s. The glue logic interface with the PCI controller is a 32-bit half-duplex interface at a clock frequency of 33 MHz.

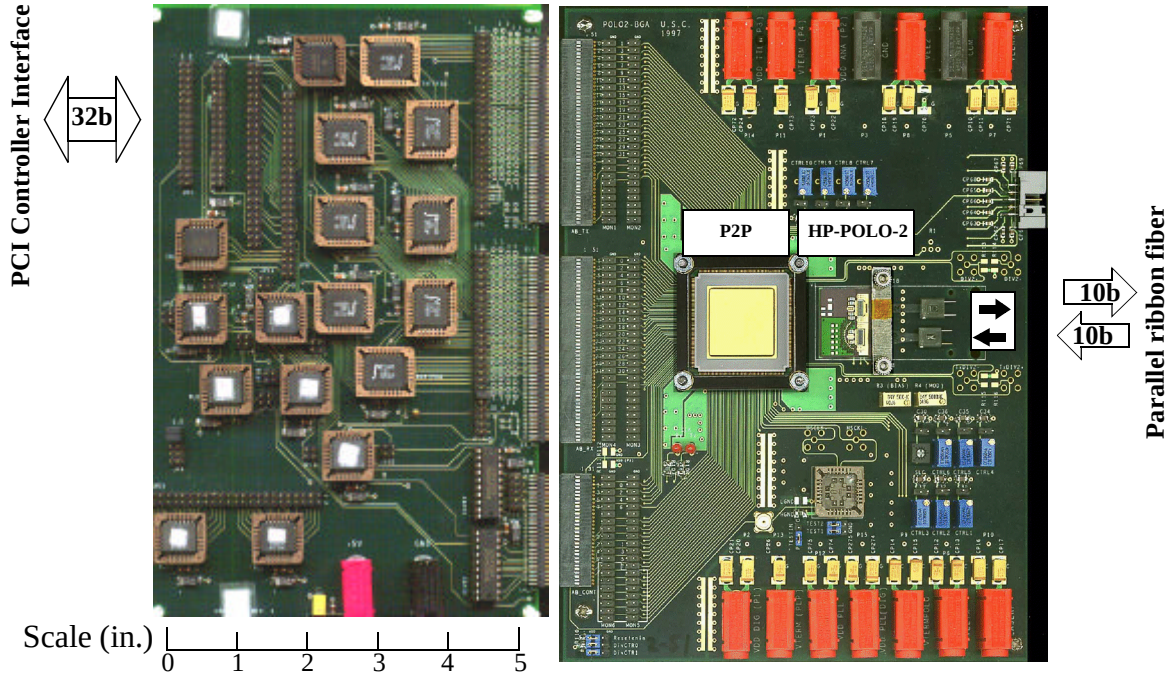

Figure 2: Glue logic board (left) with FIFOs for data storage and PLDs for control logic. The P2P chip that was designed for point-to-point interconnections (which will later be replaced by the LAC in a ring network currently being designed) and HP-POLO-2 parts are mounted on the board shown on the right. The fiber connector I/Os are indicated by the two solid black arrows on the right and demonstrate the high edge-connection density offered by optics.

The board on the right in Figure 2 has two main components - the P2P and a HP-POLO-2 optoelectronic module. The HP-POLO-2 module is an 850 nm VCSEL/PIN detector array-based fiber transmit/receive interface [11] designed by Hewlett-Packard (HP) Research Laboratories for parallel fiber-optic links. Each POLO-2 module consists of 10 transmitters and 10 receivers in a compact Ball Grid Array (BGA) package. A newer version of the HP optoelectronic module called the PONI module (shown in Figure 3) has separate 12-wide transmitters and receivers. The modules can handle data rates of 2.5 Gbit/s per multimode fiber. More details on the PONI modules can be found in [15].

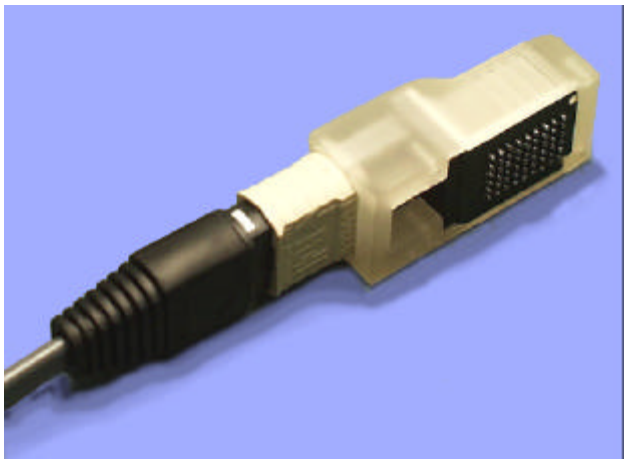

Figure 3: PONI module for separate transmitter or receiver functions. The figure shows a single 12-wide transmitter module that can support data rates of 2.5 Gbit/s per multimode fiber with low bit error ratio (BER<10^{-14}). The module is shown on its side to expose the Ball Grid Array (BGA) on the underside of the package. The BGA provides electrical I/O and is surface mounted onto a printed circuit board. The optical I/O to the module is via an MT push/ pull fiber-ribbon connector seen on the left in the figure.

3. P2P

A CMOS-based chip known as the P2P, containing the minimum functionality necessary to perform transmit and receive functions on point-to-point connections, has been implemented. The P2P was designed to demonstrate the feasibility of constructing a CMOS-based chip for multi-Gbit/s data rates. It is important because it shows that CMOS can be used to bridge a conventional CMOS electrical interface to a high-performance parallel fiber-optic module. It retains compatibility with the glue logic interface enabling integration into a system. Aspects of P2P design will be reused in the LAC.

The P2P provides a measured link data rate of 622 Mbit/s per signal line on each of its 8 parallel data lines thus giving an aggregate bisection bandwidth of 9.6 Gbit/s on its independent transmit and receive ports. The P2P CMOS die, constructed in 0.5 μm CMOS technology via the MOSIS service, measures 10.35 mm x 4.4 mm. The architecture of the P2P is shown in Figure 4(a). The host interface transmit and receive buffers, TxFIFO and RxFIFO, are two 1 kbyte on-chip buffers constructed using dual-ported static random access memory (SRAM). Cells are written into the TxFIFO and a special signal line is used by the host interface to indicate the end of a cell. Cells are transmitted as soon as a complete cell has been loaded onto the on-chip buffer. The 32-bit wide data read out of the TxFIFO is multiplexed onto eight parallel lines by the serial-

izer. The eight data lines, clock and frame can be transmitted onto the network either on optical fiber or connected directly to 50-ohm characteristic impedance coaxial cables. The logic of the FIFOs and their interface with the high-speed serializer and deserializer circuitry runs at a maximum frequency of 155 MHz. The TTL host interface has a programmable maximum clock speed of 39 MHz, which is within the 50 MHz design specification of the glue logic interface described earlier.

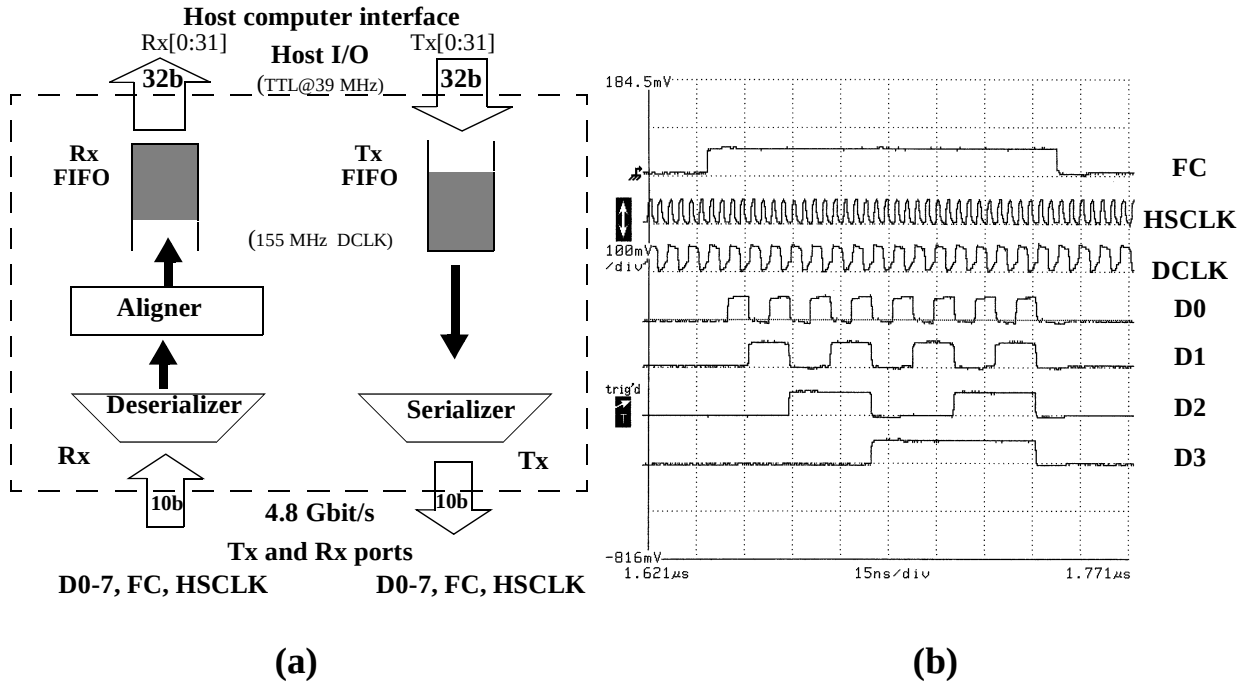

Figure 4: (a) Architecture of the P2P chip built in 0.5 µm CMOS with 4.8 Gbit/s peak transfer rates on independent transmit and receive directions. The TxFIFO and RxFIFO are 1kbyte dual-ported SRAM buffer memories. The P2P chip is a precursor to the ring network LAC. (b) Tx high-speed port output signal voltages displayed on a digital sampling oscilloscope (from top): frame control line(FC), high-speed interface clock at 311 MHz (HSCLK), digital logic clock at 155 MHz (DCLK), and four data lines (D0-D3) at 622 Mbit/s per signal line in the P2P chip. The horizontal scale is 15 ns/division while the vertical scale is 1 V/division for each signal (which has been attenuated by 20 dB).

At the receiver, the eight high-speed parallel lines are demultiplexed to again produce 32-bit wide data. This data is word aligned in hardware by the aligner module. Alignment is required when the demultiplexed signals at the receive port are not phase aligned with the multiplexed signals of the transmit port. The waveforms shown in Figure 4(b) demonstrate the P2P chip transmit lines functioning at a data rate of 622 Mbit/s. Valid data is when the frame control line shows a logic high value. Data is clocked on both edges of the clock to increase throughput for the same clock frequency, thus enabling low-power operation. The average power consumption is 8 W.

Based on our simulations, we estimate that the ring network LAC will consume less than 10 W of power.

4. PCI throughput measurements

We performed some measurements to determine the maximum sustainable send throughput over the AMCC PCI bus interface board using a file-based I/O transfer over Windows NT 4.0. Theoretically, the highest possible send throughput achievable through a PCI bus operating at 33 MHz is 1.06 Gbit/s. However, in a real system such as a typical Pentium-based PC, this throughput value cannot be sustained. We performed some throughput measurements on the PCI bus of a Triton (82430 VX PCI chipset) motherboard in a 166 MHz Intel Pentium-based PC using Windows NT 4.0. The sustained send throughput as shown in Figure 5 saturates at a value of 163 Mbit/s, while that obtained using a file-based I/O transfer over DOS resulted in a sustained throughput of 300 Mbit/s. Throughput degradation is due to the scatter/gather operations on physically discontiguous memory data being performed in software, since the current AMCC card does not have hardware scatter/gather capabilities. The send throughput obtained using actual applications will be further limited by protocol and operating system overheads [16]-[18] and is beyond the scope of discussion of this paper.

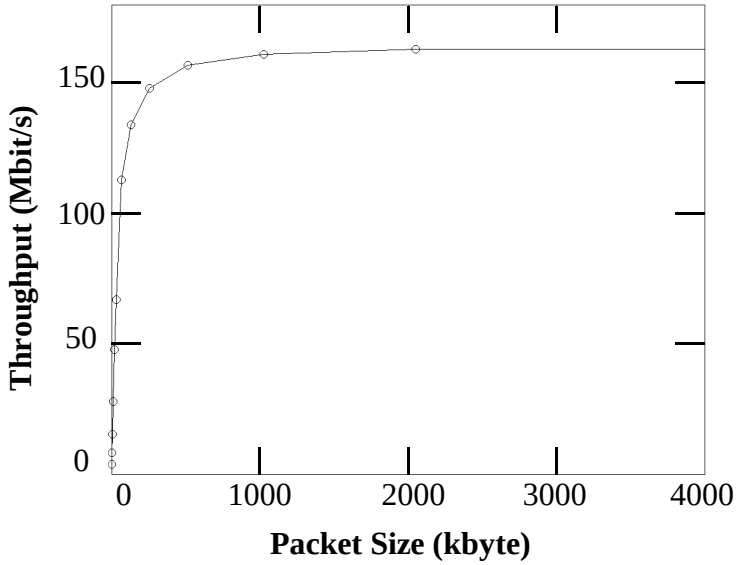

Figure 5: Measured sustained send throughput using file-based I/O for a 166 MHz Intel Pentium-based Triton motherboard with 82430 VX PCI chipset (33 MHz) motherboard using Windows NT 4.0 operating system. Due to file I/O transfer overheads, the throughput saturates at 163 Mbit/s.

5. Need for a slotted ring network

The point-to-point link described thus far is a testbed implemented to verify the feasibility of constructing high-performance CMOS-based interfaces to parallel fiber-optics. Based on this, we are presently designing an NIC for a network of PCs based on the slotted ring architecture. In this section, we present reasons motivating our choice of network topology. Many earlier high-speed commercial and research networks were shared medium bus and ring networks [6][19]-[25]. The advantages of shared medium ring networks include the simplicity of the topology which leads to less complex hardware and allows the possibility of performance optimization at low cost. Importantly, deterministic bandwidth and delay bounds may be achieved without penalty of a significant drop in overall network performance. In addition, the broadcast and multicast modes needed for multimedia applications are natural implementations. Congestion control is implemented by traffic sources monitoring the current network traffic of the shared medium before initiating transmission, avoiding the need for backward propagation of congestion information to the sender. These factors justify further exploration of rings for their bandwidth capabilities and cost-effectiveness relative to alternative schemes such as those based on ATM switches.

Three media access control protocols for a multi-Gbit/s ring network are token, slotted and buffer insertion. The increases in system I/O bus bandwidth have not kept pace with those in the network bandwidth. A single host cannot constantly transmit data to utilize the peak network bandwidth or receive at the same rate unless it has very large buffers on chip. However, on-chip area for buffers is limited. A token ring protocol is hence inappropriate for a multi-Gbit/s ring network since it leads to an under-utilization of network bandwidth. Thus some form of multiplexing is needed for media access to the ring to ensure efficient bandwidth utilization. A slotted ring, where the available network bandwidth is sectioned into slots, is the simplest alternative and has the potential for lowest node latency. This is because the alternative buffer insertion ring requires considerable hardware to arbitrate access to the ring and buffer incoming cells while a host transmits. The simpler slotted ring protocol does not have to handle this complexity and simply waits for a free transmission slot.

In subsequent sections, we describe the architecture and proposed implementation of an experimental low-cost Gbyte/s slotted ring network called the PONI network. Some of the goals

guiding our network design are:

- Construction of the entire medium access control (MAC), support logic and high-speed circuitry using a single CMOS link adapter chip (LAC) to provide a link data rate of a Gbit/s/signal line
- Provide fair access to hosts along with bandwidth guarantees for multimedia applications
- Low node latency and low delay variation (jitter)
- Support for broadcast and multicast modes
- Interoperability with ATM networks in the B-ISDN domain
- Interface to PCs through a high-performance PCI bus interface

6. PONI Network Architecture

The PONI network design is a low latency, ultra-high bandwidth, unidirectional slotted ring for use in a scalable cluster of host computers which could be PCs or workstations (see Figure 6). The physical medium is a 10-wide multimode fiber-ribbon. The physical layer portion of the PONI network design is optimized for parallel fiber-optic link technologies with an 8 Gbit/s data transfer rate. The end systems in a typical network are envisioned to be, though not restricted to, low-cost PCs.

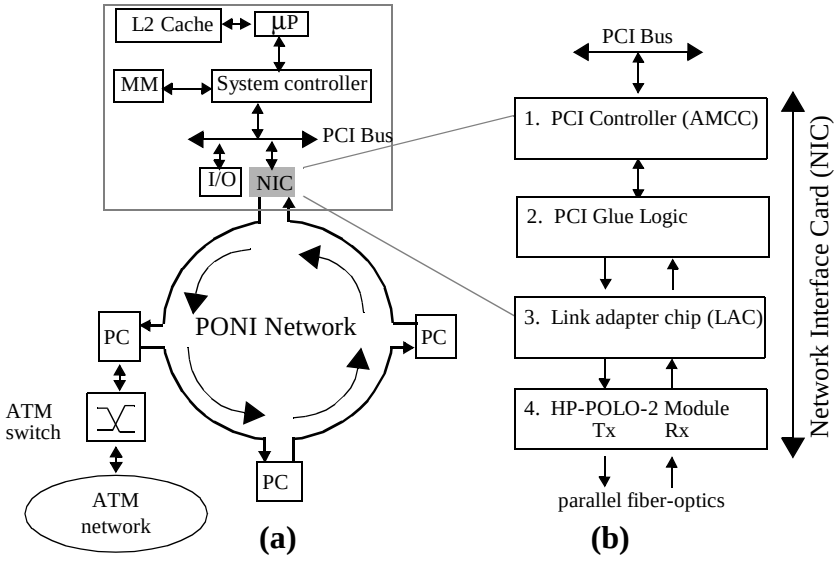

Figure 6: (a) Schematic of unidirectional PONI ring network showing interconnected host PCs (b) Components of the NIC in a host PC that interfaces between the PCI bus and the parallel multimode fiber-ribbon network medium. A commercially available PCI controller interfaces with the host PCI bus. The link adapter chip (LAC) will implement the medium access control (MAC) and our glue logic design bridges PCI and LAC. The HP-POLO-2 is an experimental fiber transceiver designed by Hewlett-Packard Research Laboratories.

The PONI network design is cost-scalable (same incremental cost with each additional node) as it will be formed from identical nodes. The terms “host” and “node” are used interchangeably in this paper and may be considered equivalent. The address field in the prototype PONI network packet for transmitting sources is 5 bits long and hence the network is scalable currently to a maximum of 32 nodes each capable of multi-Gbit/s throughput access to the ring. It would be fairly straightforward to increase the allowed number of network nodes. The interface hardware of the network is designed to provide ease of connection to the PCI bus standard along with ATM Adaptation Layer (AAL) support which simplifies bridging functions to WAN/ATM networks. Figure 6(a) shows the topology of the PONI network design and Figure 6(b) shows the various components that constitute the PONI network interface. In a typical network, high-performance PCs constitute the host computers. These PCs are connected to the high-bandwidth network using the NIC. The maximum total network physical layer length in a 32-node network is 9.6 km (due to the link length limitation of 300 m of the HP-POLO-2 module), corresponding to a total fiber latency of less than 50 μs.

The custom-designed link adapter chip (LAC) will implement a MAC protocol. The LAC will replace the earlier described P2P on our network interface card shown in Figure 2. The PONI network has been designed to support the high bandwidth and delay sensitive requirements of multimedia traffic such as uncompressed live video and voice, while at the same time providing adequate support for other traditional data traffic such as file transfer and e-mail. The MAC protocol used in the PONI network design is discussed in Section 7.

7. PONI network protocol design

The following section describes the MAC for a slotted ring network that we designed, simulated, and propose to implement on the LAC. The physical layer of the PONI slotted ring network consists of 10 signal lines - eight parallel Gbit/s data lines for the serialized data streams resulting from a 32-bit host computer interface, one line for the clock and one line for a frame control signal. The clock is transmitted along with the control and data lines, thereby avoiding the need for clock extraction which is a necessity in conventional serial links. The clocking mechanism in the PONI network is a distributed scheme that is based on the one used in FDDI [26]. Slot boundaries are indicated using the frame control line which is enabled to a logic high value for the entire

duration of a slot as shown in Figure 7(a). Short idle gaps (a minimum of 8 bytes) are inserted between slots. The elasticity buffer located in the input receiving stage of the LAC uses the idle gaps to synchronize between the clock received from the network and the local chip clock. The differences in received network clock and local clock frequencies may lead to expansion or shrinkage of the idle gap. A smoother module located after the elasticity buffer in the datapath preserves a minimum idle separation between slots. The slot busy/free status is indicated by a single bit in the header. Access to a free slot is negotiated using the MAC.

Ring initialization is performed by a ring master which also performs error monitoring functions during normal ring operation. While the ring master could be selected actively by the network through an election procedure, we select the master by setting a control register on the LAC of one of the hosts on the network, through the device driver. Since the same LAC is used at all nodes, any host is potentially capable of being a ring master. The addresses for the various nodes in the ring are allocated during the ring initialization process initiated by the master. This is achieved using a hop-count field which is incremented by all nodes during initialization along the ring propagation direction. Subsequently, the master loads the ring with slots to enable normal ring operation.

The ring master can optimize the network for a variety of traffic patterns by adjusting the size and number of slots on the ring. These parameters are configured onto the control registers of the master via the device driver during ring initialization. A packet in the host is broken into smaller cells. A data cell fits within slots of the PONI ring and encapsulates ATM adaptation layer (AAL) protocol data units. The slot size can be as small as 16 bytes and as large as 1 kbyte, which is also the size of the transmit and receive buffers on the chip. The PONI network design does not rely on the physical layer diameter to accommodate the required number of slots. Instead, the smoother located on each chip provides buffer space adequate to hold the desired number of slots. The result is a flexible, ultra-high bandwidth network capable of supporting a variety of cluster

applications from scalable servers to the delivery of multimedia data in a workgroup environment.

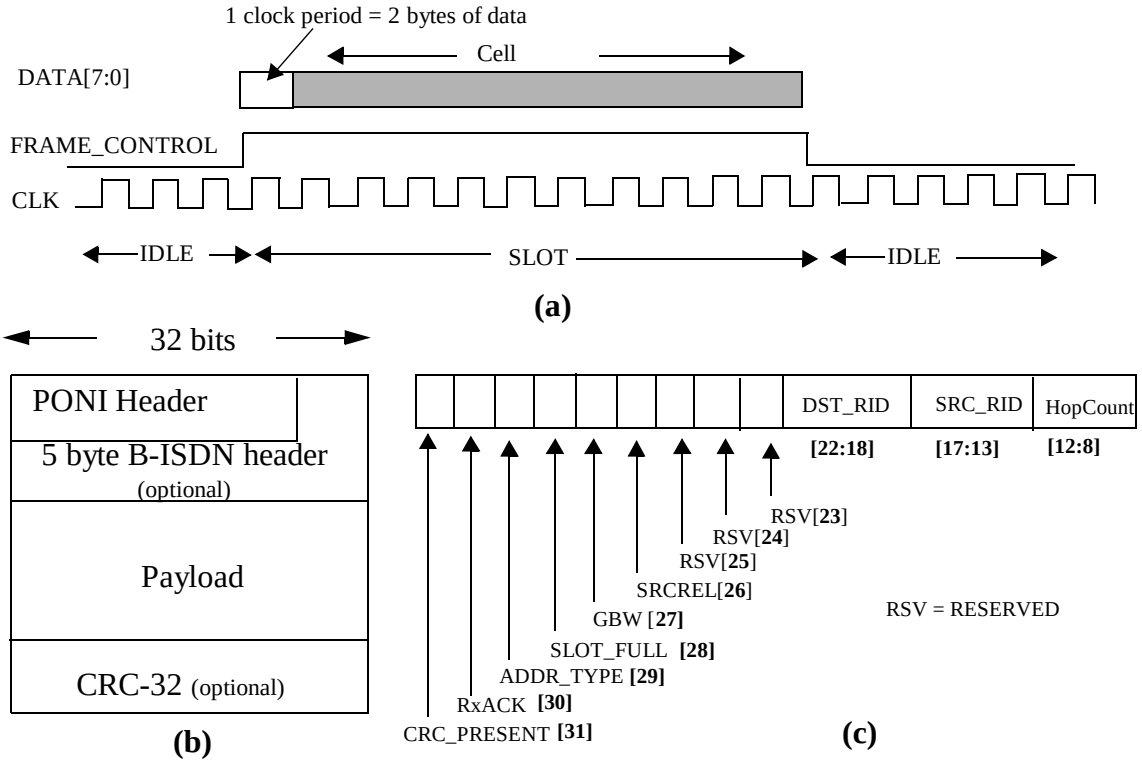

Figure 7: (a) Clock, frame control and data line format on the high-speed lines. Slot duration is marked by the frame control line. There are two bits of data on each of the 8 parallel data lines in every clock cycle. (b) Cell format with a 3-byte PONI header and an optional 5-byte B-ISDN header. (c) PONI header format. The SLOT_FULL bit indicates if a slot is busy or free. Transmission access rights are negotiated based on the GBW and SRCREL bits.

The header portion of a network cell as shown in Figure 7(b) is 3 bytes wide and precedes the 5 byte B-ISDN header (which is optional) and the AAL protocol data unit. The PONI network header shown in Figure 7(c) contains a minimal set of address and control information for slotted ring operation. The first field consists of eight bits of control information. The slot busy/free status is indicated by the SLOT_FULL bit. The next two fields contain the 5-bit source and destination short node addresses. The last field contains the hop-count which is used by the master during normal ring operation for monitoring purposes. The PONI network design supports two forms of addressing - one that uses the short node addresses and can be used for rapid decoding, and a second that uses the B-ISDN virtual channel identifier (VCI) address. The VCI address table entries are configured and can be subsequently modified using the device driver. The ADDR_TYPE bit indicates which of the two is used. The LAC can currently recognize up to 1024 addresses with provisions to expand by a further 2048 addresses. It is intended that the VCI address space be used to provide broadcast and multicast capabilities.

Packet removal is a critical issue for any slotted ring. The designed network should balance the competing goals of maximizing bandwidth available to requesting nodes and ensuring fairness of access. A cell could be removed from the network either at the source node or at the destination node, with the source or destination node being allowed to reuse the slots for another transmission. Both schemes, if not correctly monitored, could result in a few nodes utilizing all the network resources and depriving other nodes of transmission opportunities.

The PONI network design implements a source removal scheme with or without source reuse. The alternative destination removal scheme with spatial reuse provides higher overall network utilization and bandwidth. Various fairness schemes for such networks have been proposed, however the hardware implementation is more complex. Our goal was to implement the MAC protocol on a single chip that could be accommodated within a high-performance general-purpose package that we had designed. The source removal scheme we use results in simpler hardware implementation which makes it easier to optimize for higher speeds, and simplifies the task of testing. In a small workgroup cluster, applications still have access to very high network bandwidths.

The PONI network design provides four priority classes using two bits - the GBW bit and the SRCREL bit. Currently, only two of the four possible priority classes are used. The SRCREL bit being set indicates that slot access is controlled under the source release protocol, while the GBW bit being set indicates that slot access is negotiated under the higher priority guaranteed bandwidth protocol. The network operation under the two schemes is as follows. The source release protocol guarantees fair network access. Under this scheme, after the cell is successfully received by the destination, the source resets the slot full bit. This slot cannot be reused by the source immediately and is passed on to the next downstream neighbor. The guaranteed bandwidth protocol provides applications with a guaranteed upper bound on access latency, since in a network with n nodes, the source release protocol can cause a worst case access latency that is n times greater. It emulates a constant bit rate connection. Under this scheme, the ring is configured with some of the total number of slots being allocated to specific nodes. The number of such slots can be flexibly controlled by the ring master. A node can transmit only on those slots whose header source address field matches its own ring identifier. A returning slot whose contents were received successfully can immediately be reused by the source in this mode.

The remaining PONI header bits are used to provide hardware assistance for realizing appli-

cation performance gains. The RxACK bit is set by the destination on successful reception of a cell. The master monitors the ring for cells recirculating indefinitely, either due to receiver problems or due to corrupted header bits. The payload of short cells is padded in hardware to accommodate cells smaller than the configured slot size. There is hardware provision to calculate Cyclic Redundancy Check (CRC) on a per slot basis. This is under the direction of the CRC_PRESENT bit which indicates whether a CRC has been attached by the host in software or is to be computed in hardware. The CRC algorithm adopted is the AAL5 CRC-32 polynomial. Off-loading the CRC calculations to hardware reduces the amount of data processing to be performed by the host and could result in substantial gains in end-to-end performance. At the destination node, the calculated CRC is appended to the cell transferred to the host. Only cells with correct CRC's can be used in the calculation of higher level checksum operations (e.g., TCP/IP checksum). A cell with an erroneous CRC is passed on to the host in any case. The decision to use or discard this cell is made at the higher levels in the host and not in the hardware since there may be applications where occasional errors are tolerated. There are three unused header bits to accommodate future additions.

The modules comprising the datapath that implements the digital logic have been constructed from a 0.5 μm CMOS standard-cell library that we developed. They have been individually and exhaustively tested using a Verilog-based switch-level simulator. A switch-level simulation of a ring consisting of three nodes has been performed using key functional vectors that verify ring initialization, protocol correctness under normal ring operation and error-handling. The estimated size of the PONI LAC as obtained from preliminary builds is 10 mm x 7 mm. The node latency is less than 150 ns, with additional latency, if any, resulting from that deliberately inserted using the adjustable smoother buffer memory.

8. Worst-case throughput analysis of the PONI network protocol design

In the following section, we present a simple, deterministic analysis of the worst-case throughput of a node in the PONI network protocol design. The purpose of this analysis is to establish a lower bound on network bandwidth available for applications at a node. Traffic at transmitting nodes is assumed to be generated continuously in a deterministic pattern so that nodes transmit data on every accessible free slot. A more detailed analysis of the network proto-

col performance would include a description of the dynamic behavior of the network under random traffic, and is outside the scope of discussion of this paper.

In a ring consisting of n nodes, let the total number of slots be N_{total} which consists of N_{sr} number of slots accessed using the source release protocol, and the remaining slots N_{gbw} accessed using the guaranteed bandwidth protocol. The round-trip latency of the network is T_L. The maximum available system bandwidth at a node is equal to

$$BW_{max} = \frac{Number\ of\ data\ bits}{Number\ of\ data\ bits + Number\ of\ idle\ bits} \times Link\ data\ rate$$

The total number of bits on the ring depends on the link data rate and the network latency including the adjustable node latency. The number of idle bits depends on the minimum required for correct operation of the elasticity buffer to accommodate a pre-specified clock variation between adjacent nodes. The worst case access latency for a node in a network in a steady state with k ring nodes that are currently active using the source release protocol (with no source reuse) is $(k+1)T_L$. If BW_{sr} is the total system bandwidth available under the source release protocol from N_{sr} slots, the bandwidth $BW(i)_{sr}$ available to node i averaged over an interval equal to this latency is given by

$$BW(i)_{sr} = \frac{BW_{sr}}{k+1}$$

The latency for a node to access a slot under the guaranteed bandwidth protocol is exactly one round-trip latency, T_L. The bandwidth $BW(i)_{gbw}$ available to node i for every slot that it has access to under the guaranteed bandwidth protocol is given by

$$BW(i)_{gbw} = \frac{BW_{max}}{N_{total}}$$

If $n(i)_{gbw}$ represents the number of slots available to the node i under the guaranteed bandwidth protocol, the total bandwidth available to node i is equal to

$$BW(i) = \frac{n(i)_{gbw}}{N_{total}} \bullet BW_{max} + \frac{N_{total} - N_{gbw}}{N_{total}} \bullet \frac{1}{k+1} \bullet BW_{max},\ k \le n$$

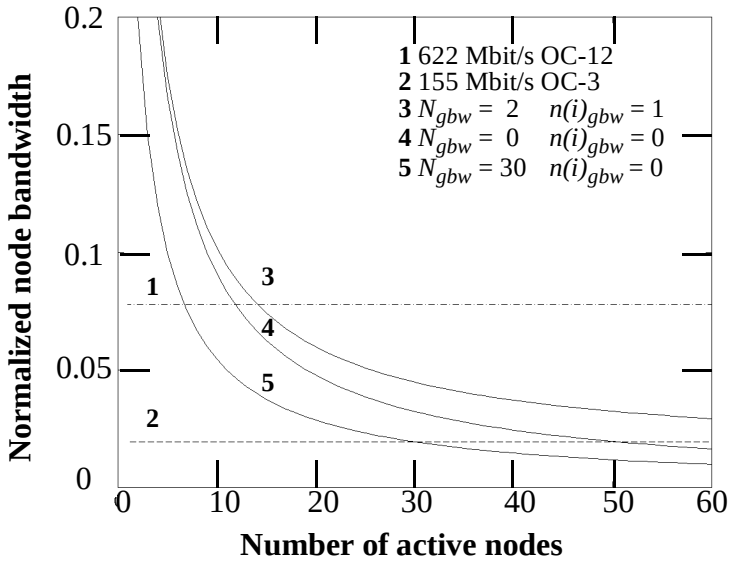

Figure 8: Calculated node bandwidth normalized to 8 Gbit/s total bandwidth with increase in number of active nodes for N_{total} = 75, and (1) ATM using 622 Mbit/s OC-12 link (2) ATM using 155 Mbit/s OC-3 link (3) N_{gbw} = 2 $n(i)_{gbw}$ = 1 (4) N_{gbw} = 0 $n(i)_{gbw}$ = 0 (5) N_{gbw} = 30 $n(i)_{gbw}$ = 0. The bandwidth allocated to a node varies depending on ring configuration. It could be greater than (curve '3') or less than (curve '5') the bandwidth available under a purely source release scheme (curve '4') thereby providing adaptability to workgroup needs.

Figure 8 shows the bandwidth that is available to a node as a fraction of the total available bandwidth, BW_{max} as the number of active nodes in the network increases. Each node for ATM is assumed to be connected to a centralized switch box using a full-duplex link. As can be seen, an increase in the number of active ring nodes results in a decrease in the bandwidth available to a node. By allocating some slots in the guaranteed bandwidth mode, a node can access more bandwidth than it otherwise would have had access to under a purely source release protocol, so long as the following relationship is observed in a ring with *k* active nodes:

$$\frac{n(i)_{gbw}}{N_{gbw}} > \frac{1}{k+1}$$

If each of the *k* active nodes in the guaranteed bandwidth mode use equal proportions of the guaranteed bandwidth, the above relationship is always satisfied. A node that uses a larger share will result in some nodes obtaining less bandwidth than available in a purely source release scheme. The aggregate system throughput, obtained by summing individual node bandwidths, is

given by the following expression:

$$System\ throughput\ =\ BW_{max} \bullet \left(1 - \frac{N_{total} - N_{gbw}}{N_{total}} \bullet \frac{1}{k+1}\right), k \le n$$

Higher system throughputs are obtained when using some slots in the guaranteed bandwidth protocol than when all slots are accessed using the source release protocol.

9. A brief comparison of PONI with a few LANs

A previous network whose access control has similarities to that used in the PONI network design was the Cambridge Fast Ring (CFR) [20] implemented in 1986. This was a 100 Mbit/s slotted ring with two modes of operation - the normal mode and the channel mode. Due to the lower link speed, the CFR had restrictions on the slot size (data field of 32 bytes) as well as the total number of slots. In the CFR network protocol, a node was allowed to use only one slot at a time while in the PONI network design, a node is allowed to use multiple slots. In the initial implementation of CFR, channel mode was not implemented. The node latency in CFR was in the microseconds range. The later Cambridge Backbone Network (CBN) [27] was operational at 512 Mbit/s and relaxed the one slot per revolution restriction. The CFR and CBN used relatively expensive ECL technology to achieve high line speeds and single-mode optical fiber for the transmission medium.

Examples of current fiber-based local area networks are Gigabit Ethernet and ATM. Gigabit Ethernet [28] is specified to run at a maximum data rate of 1 Gbit/s over 550 m when using multimode fiber. The network can operate in half-duplex or full-duplex modes. Collisions due to contention in the half-duplex mode which uses the CSMA/CD MAC protocol degrade the available 1 Gbit/s bandwidth. Collisions can be avoided using the full-duplex mode. A multiported Gigabit Ethernet switch is however necessary to fully avoid collisions in addition to the NICs in each host.

In contrast, the PONI network uses a single NIC installed in each host connected to the network medium. This is sufficient to implement a full-duplex network protocol that incorporates admission control features and simultaneous collision-free access by multiple network nodes. The LAC chip in PONI is designed to provide a higher net link data rate of 8 Gbit/s. The PONI network is cost-scalable. The total cost of the PONI network when amortized over the high available bandwidth makes it a potentially attractive alternative to provide gigabit access to the desk-

top in small to medium-sized clusters.

ATM has a complex Quality of Service (QoS) allocation scheme to handle multiple traffic classes with widely varying requirements. While such complex bandwidth reservation schemes are essential in a WAN environment, less sophisticated protocols are sufficient and cost-effective in a high-bandwidth small LAN environment that PONI targets. Broadcast and multicast modes are easier to support in a shared-medium ring network such as PONI as opposed to switch-based networks such as ATM or full-duplex collision-free Gigabit Ethernet. The PONI header format provides WAN compatibility through support for interoperability with ATM.

The PONI ring network design has potentially lower latency than centralized switches making it an attractive choice for cluster applications. The measured latency for a 64-byte cell through an Ethernet switch between two switch ports in the same chassis and under minimum load is 17 μs [29]. The latency for a 32-node PONI network under the guaranteed bandwidth protocol scheme, irrespective of total network load, has the following components:- (a) worst-case slot access latency of 6.4 μs (b) the time taken to load or unload a 64 byte packet from FIFOs at a TTL speed of 62.5 MHz is 256 ns (c) internodal fiber latency of 50 ns over a distance of 10 m per node (d) a node's input receive port to output transmit port latency of 150 ns. The node latencies for the source and destination nodes aggregate to that of an entire node. In a 32-node network, this gives a worst-case latency of 13.1 μs for a destination that is immediately upstream to the source. Thus, by using small networks with less than 32 nodes the round-trip network latency can be kept low. Larger networks can be constructed using a multiple hierarchy of interconnected rings as described in earlier work on ring networks [20].

10. Conclusions

We have described the design and architecture of a Gbyte/s slotted ring network interface for multimode fiber-based parallel fiber-optic links targeted at high-performance workgroup cluster multimedia environments. We propose a potentially low-cost scheme based on multimode fiber technologies and CMOS-based single-chip implementation of a medium access control and high-speed interface. Feasibility has been verified by system integration of parallel fiber-optic components with CMOS technologies to provide a preliminary multi-Gbit/s, low-latency network that utilizes high-performance PCI bus interfaces to interconnect Intel Pentium-based PCs.

Currently, workgroup cluster performance limitations arise from protocol and operating system overheads and limited I/O bus interface bandwidths. High-bandwidth shared medium slotted ring networks such as PONI may be more cost-effective here than centralized switch-based solutions. Particularly appealing is the fact that multimedia applications such as broadcast video, multicast video, and remote graphics visualization exploit the natural strengths of shared medium ring networks.

By shrinking the feature size used to implement the CMOS link adapter chip, more functionality can be incorporated on a single chip. Since 0.1 μm CMOS will potentially allow signaling rates per signal line to increase to 10 Gbit/s, PONI network data rates can increase to 80 Gbit/s allowing simple ring networks to remain an attractive and competitive solution for workgroup cluster needs in the future. By designing separate I/O bus specific bridge ICs which integrate the glue logic, it will be possible to construct a very low-cost set of components for network interfaces based on the slotted ring network.

The important issue of providing fault tolerance due to ring breakdown has not been addressed in this paper. FDDI has a well-established technology to perform optical bypassing of faulty nodes. Similar technologies have yet to be established for parallel fiber-optic links.

References

[1] J. Silberman, N. Aoki, D. Boerstler, J. Burns, S. Dhong, A. Essbaum, U. Ghoshal, D. Heidel, P. Hofstee, K. Lee, D. Meltzer, H. Ngo, K. Nowka, S. Posluszny, O. Takahashi, I. Vo, and B. Zoric, "A 1.0GHz single-issue 64b PowerPC integer processor," in *1998 IEEE International Solid-State Circuits Conference*, pp. 230-231, February 1998 (IEEE cat# 98CH36156).

[2] B. Sano, B. Madhavan, and A. F. J. Levi, "8 Gbps CMOS interface for parallel fiber-optic links," *Electron. Lett.*, Vol. 32, No. 24, pp. 2262-2263, November 1996.

[3] B. Sano, and A. F. J. Levi, "Networks for the professional campus environment," *Multimedia Technology for Applications*, pp. 413-427, IEEE Press, Piscataway, NJ, 1998.

[4] R. C. Walker, K.-C. Hsieh, T. A. Knotts, and C.-S. Yen, "A 10 Gb/s Si-Bipolar TX/RX Chipset for Computer Data Transmission," in *1998 IEEE International Solid-State Circuits Conference*, pp. 302-303, February 1998 (IEEE cat# 98CH36156).

[5] A. P. Kanjamala and A. F. J. Levi, "Subpicosecond skew in multimode fibre ribbon for synchronous data transmission," *Electron. Lett.*, Vol. 31, No. 16, pp. 1376-1377, 1995.

[6] ANSI X3.148-1988, "Fiber Distributed Data Interface (FDDI) - Token Ring Physical Layer," *American National Standards Institute*, Nov. 1988.

[7] IEEE Std 1596-1992, "IEEE Standard for Scalable Coherent Interface (SCI)," *Institute of Electrical and Electronics Engineers*, August 1993.

[8] HIPPI-6400-OPT Working Draft T11-1, Project 1249-D, http://www.cic-5.lanl.gov, Rev 0.7, *American National Standards Institute*, August 1998.

[9] G. M. Yang, M. H. MacDougal, and P. D. Dapkus, "Ultra-low threshold vertical cavity surface emitting lasers obtained with selective oxidation," *Electron. Lett.*, Vol. 31, No. 11, pp. 886-888, 1995.

[10] D. G. Deppe, D. L. Huffaker, H. Y. Deng, Q. Deng, and T. H. Oh, "Ultra-low threshold current vertical cavity surface emitting lasers for photonic integrated circuits," *IEICE Transactions on Electronics*, Vol. E80-C, No. 5, pp. 664-674, May 1997.

[11] K. Hahn, K. S. Giboney, R. E. Wilson, J. Straznicky, E. G. Wong, M. R. Tan, K. T. Kaneshiro, D. W. Dolfi, E. H. Mueller, A. E. Plotts, D. D. Murray, J. E. Marchegiano, B. L. Booth, B. J. Sano, B. Madhavan, B. Raghavan, and A. F. J. Levi, "Gigabyte/s Data Communications with POLO Parallel Optical Link," in *The 46th Electronics Components and Technology Conference*, Orlando, Florida, pp. 301-307, May 28-31, 1996 (IEEE cat# 96CH35931).

[12] Applied Micro Circuits Corporation, http://www.amcc.com, "S5933 PCI controller data book," San Diego, CA, Spring 1996.

[13] T. Shanley, D. Anderson, "PCI System Architecture," Third Edition, Addison-Wesley Publishing Company, July 1995.

[14] BlueWater Systems, Inc., http://www.bluewatersystems.com, "WinDK User's Manual," Edmonds, WA, September 1996.

[15] L. Buckman, A. Yuen, K. Giboney, P. Rosenberg, J. Straznicky, K. Wu, and D. Dolfi, "Parallel Optical Interconnects," in *Hot Interconnects 6 Symposium*, Stanford University, Palo Alto, California, August 1998.

[16] J. Brustoloni and B. Bershad, http://reports-archive.adm.cs.cmu.edu/cs.html, "Simple Protocol Processing for High-Bandwidth low-latency networking," CMU-CS-93-132, School of Computer Science, CMU, March 1992.

[17] D. Clark, V. Jacobson, J. Romkey, and H. Salwen, "An Analysis of TCP processing overhead," *IEEE Comm. Mag.*, Vol. 27, No. 6, pp. 23-29, June 1989.

[18] J. Kay and J. Pasquale, "The Importance of Non-Data Touching Processing Overheads in TCP/IP," in *Proc. ACM SIGCOMM '93*, pp. 259-267, 1993.

[19] K. Imai, T. Ito, H. Kasahara, and N. Morita, "ATMR: Asynchronous transfer mode protocol," *Computer Networks and ISDN Systems*, Vol. 26, Nos. 6-8, pp. 785-798, March 1994.

[20] A. Hopper, and R. Needham, "The Cambridge Fast Ring Networking System," *IEEE Transactions on Computers*, Vol. 37, No. 10, pp. 1214-1223, Oct. 1988.

[21] J. L. Adams, "Orwell," *Computer Networks and ISDN Systems*, Vol. 26, Nos. 6-8, pp. 771-784, March 1994.

[22] H. R. van As, W. W. Lemppenau, H. R. Schindler, and P. Zafiropulo, "CRMA-II: A MAC protocol for ring-based Gb/s LANs and MANs," *Computer Networks and ISDN Systems*, Vol. 26, Nos. 6-8, pp. 831-840, March 1994.

[23] IEEE Standards for Local and Metropolitan Area Networks: Distributed Queue Dual Bus (DQDB) Subnetwork of Metropolitan Area Network (MAN), 902.6-1990.

[24] G. Watson, D. Banks, C. Calamvokis, C. Dalton, A. Edwards and J. Lumley, "AAL5 at a gigabit for a kilobuck," *Journal of High Speed Networks*," Vol. 3, No. 2, pp. 127-145, 1994.

[25] Y. Ofek, "Overview of the MetaRing architecture," *Computer Networks and ISDN Systems*, Vol. 26, Nos. 6-8, pp. 817-830, March 1994.

[26] J.D. Hutchison, C. Baldwin, and B.W. Thompson, "Development of the FDDI Physical Layer," Digital Technical Journal, Vol. 3, No. 2, pp. 1-13, Spring 1991.

[27] D. J. Greaves, K. Zielinski, "The Cambridge Network: an overview and preliminary performance," *Computer Networks and ISDN Systems*, Vol. 25, pp. 1127-1133, 1993.

[28] IEEE P802.3z Gigabit Task Force, http://grouper.ieee.org/groups/802/3/z/index.html, *Institute of Electrical and Electronics Engineers*, 1998.

[29] The Tolly Group, http://www.tolly.com, "MultiSwitch 900/VNswitch 900 Fast Ethernet and Multi-topology switching performance," Report #7303, pp. 1-6, Nov. 1997.