Optoelectronic Buses for High-Performance Computing

DONALD M. CHIARULLI, STEPHEN P. LEVITAN, RAMI G. MELHEM, MANOJ BIDNURKAR, ROBERT DITMORE, GREGORY GRAVENSTRETER, ZICHENG GUO, CHUNGMING QIAO, MEMBER, IEEE, MAJD F. SAKR, MEMBER, IEEE, AND JAMES P. TEZA

Invited Paper

Modern computer buses are typically organized by the three functions of data transfer, addressing, and arbitration/control. In this paper we present a fiber-based bus design which provides optical solutions for each of these functions. The design includes an all-optical addressing system, based on coincident pulse addressing, which eliminates the latency contribution and bandwidth limitation associated with electronic address decoding. The control system uses time-of-flight relationships between a priority chain and a feedback waveguide to implement fully distributed asynchronous and self-timed bus arbitration.

I. INTRODUCTION

Buses are by far the most commonly implemented communication structure within a modern computer system. As optical technology moves from the realm of local-area and wide-area networks between computer systems, to board level, multichip module, or even chip-level communications within computer systems, an optical solution to the fundamental issues in bus design must be devised. In this paper we present a design for a multiple-access optical bus. The design includes optical solutions to the problems of data transfer, bus arbitration, and device addressing. The problem domain we have chosen is the backplane of a closely coupled multiprocessor system. In a closely coupled multiprocessor system the resources can be accessed via a single bus level operation without I/O transfers, and in a manner which is transparent to both systems and application software. Although the design is presented in this domain it is also applicable to a variety of high-speed bus applications.

There are a number of defining characteristics for bus applications which distinguish them from other types of optical communications networks. Buses are multiple-access links implemented either with tapped fibers or optical star couplers. The end-to-end length and propagation delay are relatively small. Bus-level transactions consist of short messages which occur with a volume of distinct messages per source which is higher than is typically experienced in network environments.

Multiple-access buses require both addressing and arbitration of accesses. With short messages, and low end-to-end latency, the overhead for access arbitration and address decoding dominate the total message latency time. At short distances the bandwidth required to be competitive with electronic implementations is substantially higher than other optical network applications. Thus in order to support the low latency and high bandwidth requirements of this application it is imperative that the optical links provide more than simply a communication channel. A substantial portion of the bus control logic must also be implemented in optics.

Two unique properties of optical signals, unidirectional propagation, and predictable path delay, make it possible to base a logic system on the time of flight and relative delay between two signals. We use these properties heavily in our implementation of addressing and control. Our optical address bus provides two paths by which signals may reach a node. Optical addressing is achieved by encoding an address as difference in path length between the two paths and using the time of flight and relative signal delays as the address. Arbitration is similarly achieved by using the time of flight of an optical-feedback wavefront in lieu of a clocking signal in an optical priority chain.

Manuscript received January 1, 1994; revised July 1, 1994. This work was supported in part by AFOSR F-49620-92-1-0023.

D. M. Chiarulli, R. G. Melhem, R. Ditmore, and G. Gravenstrater are with the Departments of Computer Science, Mineral Industries Building, University of Pittsburgh, Pittsburgh, PA 15260 USA.

S. P. Levitan, M. Bidnurkar, and J. P. Teza are with the Department of Electrical Engineering, 348 Benedum Engineering Hall, University of Pittsburgh, Pittsburgh, Pittsburgh, PA 15261 USA.

Z. Guo is with the Department of Electrical Engineering, Louisiana Tech University, Ruston, LA 71272 USA.

C. Qiao is with the ECE Department, State University of New York at Buffalo, Amherst, NY 14260 USA.

M. F. Sakr is with NEC Research Institute, Princeton, NJ 08540 USA.

IEEE Log Number 9405124.

This paper presents a synthesis of several investigations into the key issues of optical bus design. Separate solutions have been previously devised for the problems of data transfer [1], device addressing [2], and arbitration/control [3] of an optical bus. We present here a complete and operational bus design, which verifies the compatibility of these techniques and analyzes their combined performance.

The presentation is organized as follows: Section II outlines previous research on both networks and optical bus implementations. Section III describes the data bus. Section IV introduces our solution to all optical address generation and decoding. Section V deals with the access arbitration problem and presents an electrooptical distributed solution. Section VI shows the combined implementation for the bus, in Section VII we report on various experiments designed to determine the technological limits on the implementation.

II. Background

Optical interconnections offer the potential for gigahertz transfer rates in an environment free from capacitive bus loading and electromagnetic interference. The effectiveness of optical interconnections has been examined from both theoretic [4] and practical [5]–[7] perspectives. Over the past decade, much of the research in optical communications networks has focused on applications to wide-area networks (WAN's) [8] and metropolitan-area networks (MAN's) [9]. More recently, specialized high-speed local-area networks (HSLN's) for computer interconnections have been studied [10] and commercial standards have emerged [11]. Other research groups have investigated the implementation of parallel computers using optical interconnections in multiprocessor applications [12]–[14].

Device technology in electrooptics has also matured to a point where small, low-power, and low-cost devices exist which are suitable for use in bus-level implementations. Initial efforts have focused on direct technology substitution [15] in board level [16] and chassis-to-chassis [17] links. However, there are obvious limitations to such substitution. For example, any interface between electronics and optics limits the speed of that interface to the speed of electronics.

In switched networks, time division switched (TDS) [18], space division switched (SDS) [19], and wavelength division switched (WDS) [20], [21] implementations have been used to perform message routing in both HSLN and multiprocessor applications. However, since switching device technology has developed more slowly than technology for other components, many recent designs have implemented low-latency "single-hop networks." These networks are composed of groups of processors linked by multiple passive star couplers which efficiently use optical power, and have simple control structures [22].

The work presented here shares the "single hop" concept of the multiple passive star networks but is specifically adapted to single backbone designs intended to compete with electronic buses and crossbar switches for parallel processors. Given that on a bus each message is broadcast, access arbitration, rather than switching, becomes the control

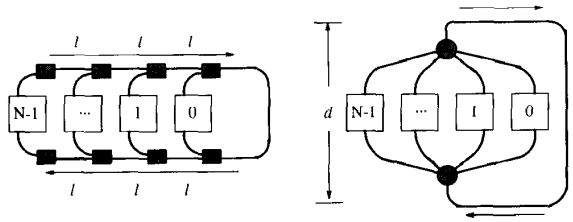


Fig. 1. Tapped fiber and star-coupled bus designs.

problem. All such networks use control algorithms which are generically called multiple access and whose implementation falls into one of the following three classes. The first class is the carrier sense multiple access (CSMA/CD) control protocol [23]. Examples of this class are Fibernet [24] and Fibernet II [25]. The primary motivation for optical CSMA/CD is compatibility with electronic ethernet systems. Thus most of the proposed designs resort to electronic collision detection and their performance is bound by the speed and complexity of electronic control [26]. The second class is the demand assignment multiple access (DAMA) protocol. Examples include EXPRESSNET and FASTNET [27]–[29]. Also known as access-and-defer systems, these networks have controllers which monitor their inputs and outputs on unidirectional fiber-optic waveguides to simultaneously transmit and detect network activity. If during a transmission, activity is sensed from the upstream direction of the waveguide, the controller aborts the transmitted message in deference to the upstream controller. RATO-net [30] is a fair, bounded-delay, random-access protocol which also uses this technique. Other DAMA-based protocols have also been recently suggested in [31]. The third class of control strategies are based on token rings. The 80 Mb/s fiber ring network from Proteon [32], and the 100-Mb/s Fiber Distributed Data Interface (FDDI) ring [33], are examples of this control structure. Since all of these systems are designed for use in HSLN applications, where relatively long packets of information are sent for each transaction, they share the common characteristic of being relatively insensitive to message latency.

On the other hand, bus applications are characterized by short messages, with a high volume of messages per source. Also, bus lengths are on the order of meters. Given these two characteristics, we are specifically motivated by two corresponding design requirements. The first is to minimize control latency since, in this application, control time dominates overall message latency. The second is to eliminate the additional latency imposed by electronic address decoding. The unique contribution of the work presented in this paper is that a substantial percentage of the overhead required for a bus implementation is processed in the optical domain.

III. The Data Bus

By definition, a bus has multiple senders and multiple receivers. In an optical implementation, the light output from any sender must be seen by the input detectors of

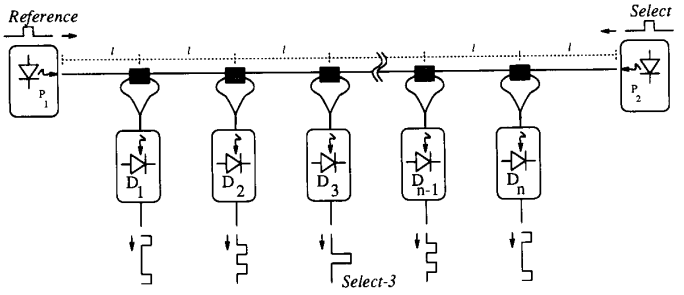


Fig. 2. Coincident pulse addressing structure.

all other devices on the bus. The most common fiber-based designs are based on either tapped fibers, or optical star couplers as shown in Fig. 1. Both of these structures are functionally equivalent. However, their temporal and power distribution characteristics differ significantly. The time of flight T for a message to traverse a star coupled system is $T = dc_g$ where d the total length of a path through both sides of the coupler and c_g is the speed of light in the fiber. T is thus independent of the number of transmitting and receiving nodes. The time of flight in a tapped fiber system is $T = (n_t + n_r)lc_g$ where n_t and n_r are node numbers, counted from the end of the bus, for the transmitting and receiving processors respectively, and l is the length of fiber between each node on the bus. Thus in a star-coupled system, each message arrives at all receivers simultaneously while on a tapped fiber each message arrives at successive time intervals given by the difference in optical path length between the receivers. For this reason tapped fibers are often referred to as tapped-delay lines.

The star coupler has a significant advantage over a tapped fiber in its power distribution characteristics. Each of the outputs from the star coupler sees an equal percentage of the optical power injected into the coupler by a transmitting node. If a star coupler has a fanout of N fibers, the optical power in each of the output fibers is $P = (p - \epsilon)/N$ where p is the input optical power from any source fiber, and ϵ is the excess loss in the coupler. This compares with the power characteristics of a tapped fiber, in which $P = p(1 - k)^{n_r+n_t}$ where k is the percentage of power removed at each tap.

It is possible to emulate the temporal characteristics of a tapped fiber in a star-coupler design by merely trimming the lengths of each fiber of to differ by length l. This retains the favorable power distribution characteristics of the star coupler. When such an implementation is not practical, multilevel taps can be used to reduce losses [2], or fiber amplifier segments can be introduced to restore power [34], respectively.

Since temporal characteristics are a key attribute in our bus design, the remainder of this discussion uses both tapped-fiber and star-coupler structures as a representation of specific temporal characteristics but not as implementation requirements. Thus where a tapped fiber is shown it represents the fact that an optical path length difference between the transmitters or the receivers must exist for proper operation. Similarly, connections shown through star couplers assume equal path lengths. However, in either case, temporally equivalent implementations could be substituted.

IV. Optical Address Bus

The address bus implementation uses a technique called coincident pulse addressing where the address of a detector site is encoded as the delay between two optical pulses which traverse independent optical paths to the detector. The delay is encoded to correspond exactly to the difference between the two optical path lengths. Thus pulse coincidence, a single pulse with power equal to the sum of the two addressing pulses, is seen at the selected detector site. Other detectors along the two optical paths (for which the delay did not equal the difference in path length) detect both pulses independently, separated in time.

Consider the optical addressing structure shown in Fig. 2. It consists of an optical fiber with two optical pulse sources, P_1 and P_2, coupled to each end. Each source generates pulses of width τ and height h. Assume $l = \tau c_g$ where c_g is the speed of light in the fiber. In other words l is the length of fiber corresponding to the pulsewidth. Using 2×2 passive couplers, n detectors, labeled D_1 through D_n, are placed in the fiber with the two tap fibers from each coupler cut to equal lengths and joined at the detector site. The location of each coupler/detector is carefully measured so that the ith detector is located at il. To uniquely address any detector, a specific delay between the pulses generated by P_1 and P_2 is chosen. If this delay is $(n - 2i + 1)\tau$ the two pulses will be coincident at detector D_i.

The same technique can be generalized to support parallel selections. If the P_1 source generates a single pulse at time t_1 and the source P_2 generates a series of pulses at times $t_i, i \in \{1 \cdots n\}$ with each t_i timed relative to t_1. Then, according to the addressing equation above, to select a specific detector i each t_i will be in the range $-(n-1)\tau \leq t_1 - t_i \leq (n-1)\tau$. Therefore, any or all of the i detectors can be uniquely addressed by a positionally distinguishable pulse from source P_2. For convenience, this pulse train is referred to as the select pulse train and the single pulse emanating from P_1 is called the reference pulse. Since the length of the select pulse train is n, and each pulse in the return to zero encoding is separated by

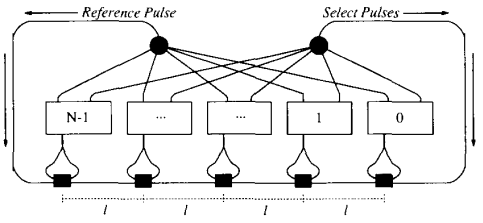


Fig. 3. Coincident pulse address bus.

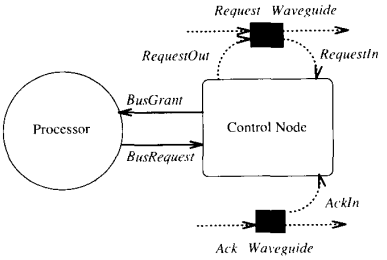


Fig. 4. Control node external connections.

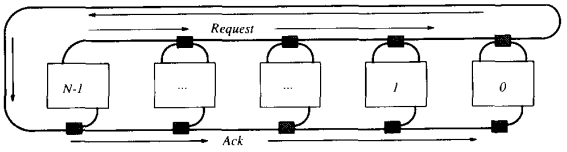


Fig. 5. Control bus, *Request*/*Ack* waveguide feedback structure.

2τ, it follows that the system latency $\sigma = 2n\tau$. Further, up to n locations may be selected in parallel within a single latency period.

This simple technique is the basis for a practical addressing mechanism for a system bus. Figure 3 shows a design in which the select and reference pulse generators in Fig. 2 are replaced by star couplers. In this structure, each processor, when granted bus access, can independently generate select and reference pulses. Addresses are encoded at each node as relative delays between the reference and select pulses using the coincidence equations above. Coincidences resulting between the select pulses and the reference pulse may select one or more destination nodes for each message. Once selected, messages are read by the node from a separate data bus as shown in Fig. 6. Since the design uses multiple sources both for the reference and select pulse trains, only one node at a time may transmit on the bus. The arbitration of bus access is the subject of the next section.

V. Control and Arbitration

Bus control and arbitration is fully distributed among the nodes and no central bus arbiter is required. Each processor accesses the bus via an electronic control node. Figure 4 shows the external interconnections for a typical control node. There are two electronic signals, *BusRequest* and *BusGrant* which connect the processor to the control node. Three optical signals, one output and two inputs, connect the control node to the optical control bus. The optical output signal *RequestOut* indicates a pending request at the corresponding control node. The *RequestIn* optical input signal reflects the state of the *RequestOut* signal of all higher priority control nodes. The third optical input *AckIn* is used as a feedback mechanism to trigger state transitions in the control node circuitry. A processor may request access to the bus by asserting its electronic request signal, *BusRequest*. Similarly, when the bus is made available to the processor, the corresponding control node electronically asserts a bus grant signal *BusGrant* to the processor. Both *BusRequest* and *BusGrant* are held active for the duration of the bus transfer cycle.

The design is asynchronous. Two waveguides, *Request* and *Ack*, form the optical control bus. At each control node, *RequestIn* and *RequestOut*, respectively, sample the upstream *Request* waveguide and drive the *Request* waveguide downstream. The *AckIn* input at each control node, reads the state of the *Ack* waveguide. The substitution of the *Ack* waveguide for a global clock signal is accomplished by the feedback structure between the *Request* and *Ack* waveguide shown in Fig. 5.

The functions of the *Request* and *Ack* waveguide are as follows. The *Ack* waveguide defines two operating states for the control bus. When there is no light in the *Ack* waveguide, the control bus state is in the batch-formation state. In this state, one or more control nodes make requests by injecting light into the *Request* waveguide, the feedback mechanism between the *Request* and *Ack* waveguides causes a transition from dark to light in the *Ack* waveguide.

With light in the *Ack* waveguide the bus enters the batch-service state. In this state, the *Request* waveguide acts as a priority chain. Each control node with a pending request defers from bus access so long as there is light upstream in the *Request* waveguide. When there is no upstream signal, the control node grants the bus access to the attached processor and on completion removes the optical output from its *RequestOut* waveguide. Note that no control node may assert *RequestOut* during the servicing state. Thus any new requests must be held pending by the control node until there is no light in the *Ack* waveguide. This organization has the effect of creating a *batch* from all pending requests at the time of the transition on the *Ack* waveguide. Batching eliminates the starvation problems which characterize other priority chain arbitration systems. Once a request enters a batch during the batch-formation state, it is guaranteed bus access in the next batch-service state.

Operating the control nodes in this fashion has a desirable side effect. Specifically, the control delay for arbitration of requests is now proportional to the optical path length between the two asserting control nodes. Only in the worst case, that is, for a batch size of one, will this delay equal the round-trip delay time of the *Request* and *Ack* waveguides. For any combination of multiple requests, the delay is always less than the round-trip delay. In addition, for a high-contention environment, where the number of pending requests and thus batch size, is large, the average control overhead per message will decrease, as the requests are

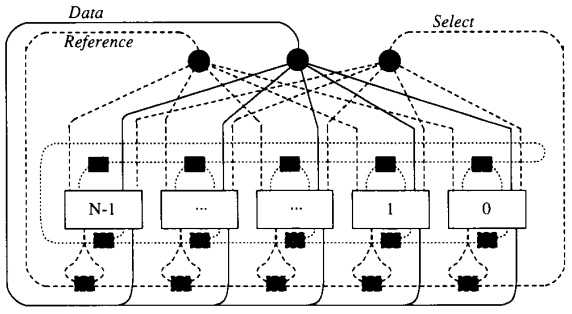


Fig. 6. Complete system bus.

grouped more closely on the bus. We show this effect in section VII-C, where we present a simulation analysis.

VI. Combined Implementation

In this section the data, address, and control buses from the previous sections are combined into a complete implementation. The design is shown in Fig. 6. The timing diagram in Fig. 7 represents the states of bus signals at two arbitrary control nodes, $nodes_i$ and $node_j$, such that $node_i$ is physically upstream on the *Request* waveguide and thus has a higher priority. Bus transfers consists of interleaved control and data transfer cycles in which the control cycle may be one of two types: a long-control cycle, or a short-control cycle. Long-control cycles correspond to the batch formation state of the control bus, short cycles are control operations between nodes within a batch during the batch servicing state. If the optical path length between each node on the *Ack* and *Request* waveguides is l, then the latency of a long-control cycle is equal to 1.5 round-trip propagation delay times on the control bus. In other words, $3Nlc_g$ for an N-node bus. Short cycles vary in length from lc_g to $(N-1)lc_g$ depending on the relative position of the nodes in a batch. Assuming all nodes are equally active, the average latency of a short cycle is $(N-1)lc_g/2$.

Figure 7 shows the timing for two bus transfers, one each from $node_i$ and $node_j$, assuming that both transfers take place within a single batch. The top set of waveforms show control, address, and data bus connections for $node_i$ and the lower set for $node_j$. The time axis is in units of lc_g. The bus is assumed to connect five nodes with $node_i$ and $node_j$ separated by an optical path length of $2lc_g$ on the control bus. For simplicity, electronic delays within the control node circuitry are not represented. This is reasonable since optical delays in the design are only measured against other optical delays. While electronic delays add to the total latency, they do not invalidate the asynchronous handshaking based on the relative time of flight of the optical signals.

The bus activity represented here begins with transitions on the $BusRequest_i$ input lines for $node_i$ and $node_j$. These transitions, marked a and b, respectively, in the timing diagram, are shown to occur when the *Ack* input is high. In fact, the two requests would be placed in the same batch if they occurred at any time during the data transfer or short

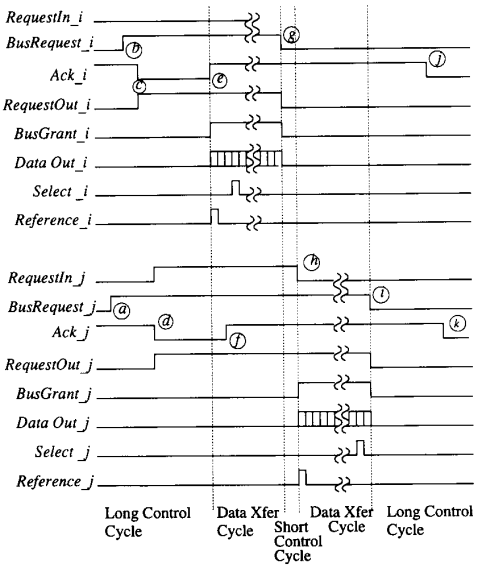


Fig. 7. Control timing.

control cycles of the previous batch or during the long-control cycle of the current batch. Since at time a the Ack_i input for $node_i$ is high, the control node takes no action until the falling edge of the *Ack* input. Similarly, $node_j$ sees Ack_j high at time b and also holds $BusRequest_j$ pending. The falling edge of Ack_i at time c marks the beginning of the long control cycle at $node_i$. In response to the low level, $node_i$ asserts $RequestOut_i$. At $2lc_g$ time units later, $node_j$ sees the same low going transition on its Ack_j input and similarly asserts $RequestOut_j$. Both *RequestOut* signals traverse the feedback path into the *Ack* waveguide. The resulting rising edge at each *Ack* input ends the long control cycle at each node.

At time e, $node_i$ sees the rising edge on Ack_i. With no upstream control nodes asserting *RequestOut*, $RequestIn_i$ is dark. A data transfer cycle thus begins at $node_i$ which asserts *BusGrant* to the corresponding processor. $Node_j$ sees this same transition at time f as the edge traverses the *Ack* waveguide, but with $RequestIn_j$ held high by the output from $node_i$ it defers bus access to $node_i$. During the data-transfer cycle, both the address and data outputs from $node_i$ are active. The data output is a serial bit stream containing the message. The two address outputs generate reference and select pulses with the reference pulse aligned relative to the first bit of the data message and the select pulse delayed by $(N-2i+l)\tau$ time units to address the ith out the N receiving nodes. The end of the $node_i$ data transfer cycle happens at time g, when the processor lowers the $BusRequest_i$ input. In response, $node_i$ lowers $RequestOut_i$ and $2lc_g$ time units later, at time h, $RequestIn_j$ becomes dark. This period, between time g and h is a short control cycle in which access is arbitrated along the priority chain. Time h begins the data transfer cycle for $node_j$ which continues until the falling edge of $BusRequest_j$ at time i. With no other control nodes in the batch, the lowering of

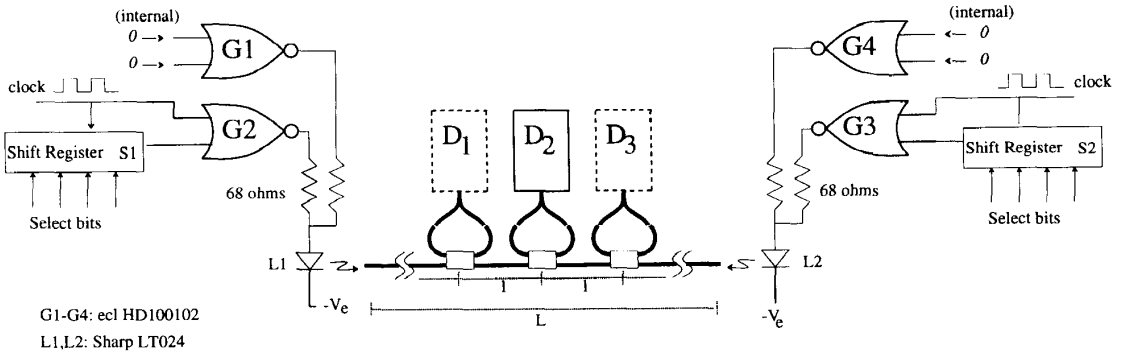


Fig. 8. Synchronization experiment.

RequestOut$_j$ creates a low going transition in the feedback fiber to the *Ack* waveguide. This initiates a long control cycle for the next batch.

Thus the three operations of control, addressing, and data transfer are supported. We turn now to a validation of the system by experimental and simulation analysis and discuss the scalability limitations for such a design.

VII. Scalability and Performance

As discussed above, there are three limitations to the large-scale implementation of the proposed bus architecture. These are bandwidth limits, which determine the minimum pulsewidth, detector spacing, and temporal limits on coincidence; latency limits, which bound the acceptable delay for a bus transfer and are determined by the speed and complexity of the bus arbitration and control algorithms; and power budget; which sets the minimum amount of power required at each detector to provide acceptable bit-error rates and noise margins.

Each of these limits have been separately characterized for our bus design. Temporal limits have been established experimentally by testing the tolerances for pulse overlap when detecting coincidence [35]. Power distribution was characterized analytically for linear tapped fiber structures [2]. Latency in the control bus design was characterized by simulation analysis for various synthetic traffic loads [3]. These results are summarized below.

A. Temporal Limits

In this section, we present results from an experimental prototype of the address bus used to investigate the relationship of coincident pulse power as a function of the synchronization of the arriving pulses. We first discuss the experimental structure, then we show typical coincident and noncoincident waveforms before we discuss the experiment itself.

Figure 8 is a diagram of the prototype structure. The fiber bus consists of a length of multimode fiber tapped three times using 10-dB fiber couplers. Select and reference bit patterns are generated by modulating the 4-ns pulse output of a Tektronix PG502 pulse generator, shown in the diagram as clock, with the output of two ECL shift registers, one for select, one for reference, at gates $G2$ and $G3$. Gates $G1$ and $G4$ simultaneously hold the diode current for laser diodes $P1$ and $P2$, respectively, at threshold while the output of $G2$ and $G4$ generate modulation current. The result is two, 4-bit, return-to-zero bit streams which encode the information in each of the shift registers. As explained above, this allows us to select any subset of the three detectors. The use of two shift registers allows us flexibility in the positioning of the reference pulse relative to the select pulse train.

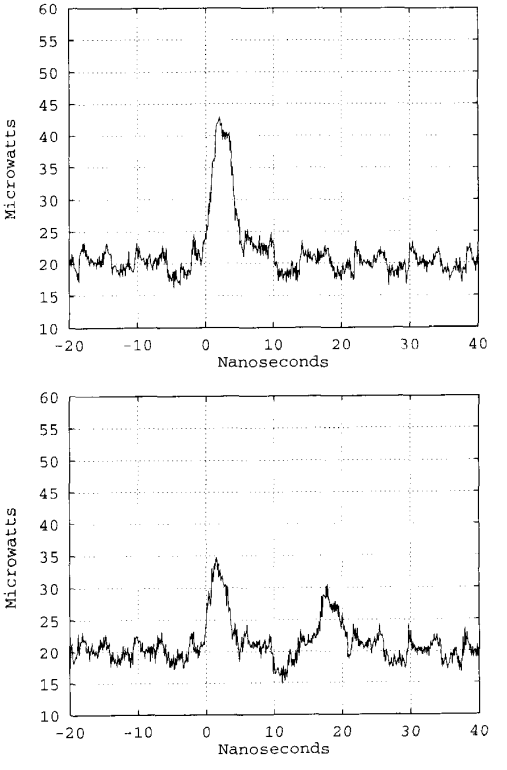


Fig. 9. Select $D1$ measured at $D1$, and select $D3$ measured at $D1$.

Figure 9 shows waveforms for both coincidence and noncoincidence measured at detectors $D1$. The left waveform shows a single double-height pulse as seen at the detector for the case that the reference and the select pulse arrive at the detector simultaneously. The right waveform shows two pulses, each of lower amplitude and separated in time, at the detector for the case that a different detector ($D3$) was chosen. Note that in this case, the noncoincident pulses are of unequal power. This is due to the fact that each pulse has passed through a different number of couplers and, hence, has become attenuated to different levels. This shows that the relative power between coincident and noncoincident pulses is a function of the detector location as discussed below.

One of the limits to the bandwidth which can be supported on the bus comes from the synchronization error which can be tolerated, while still detecting coincidence. Therefore, measurements were made to characterize the effect of synchronization error between the reference and select pulses on the power of the coincident pulse. Since clearly this error can be characterized as a percentage of the pulsewidth, synchronization precision has a direct bearing on the absolute width and height of an addressing pulse that can be effectively detected.

In this experiment, the reference and select pulse trains were configured to select D_2. In each step of the experiment synchronization error was introduced by adding successively longer lengths of fiber to the ends of the bus. Length was added first on the reference pulse end of the bus, and then on the select pulse end of the bus. The two pulses shown on the left of Fig. 10 show the pulse waveform at the end of the experiment, after sufficient delays were added to the fiber to bring the pulses completely apart.

The right half of Fig. 10 shows the reduction factor f of the coincident pulse power as a function of percent synchronization error. Percent synchronization error is the error, in time, introduced by each length of fiber divided by the pulsewidth. In other words, pulses at perfect coincidence (synchronization error $= 0$) yield a reduction factor of $f = 1.0$ which implies a coincident power equal to twice the single pulse power.

Synchronization error in either the select pulse, shown as positive error, or reference pulse, shown as negative error, reduces this power by the factors shown. The solid line in Fig. 10 is the experimental result. The dotted line is an analytical result generated from the coincidence of two sinusoidal pulse waveforms. In both cases, the power falls off in roughly the shape of the coincident waveforms themselves. We can see that a timing error of up to 50% only decreases the coincident pulse power to about 70% of its ideal value. Therefore, large variations (on the order of one half of a pulsewidth) can be tolerated without significant degradation of the coincident signal.

B. Power Distribution

The second limitation to the proposed bus organization is power distribution. In [2] we present an analysis of power distribution in a bi-directional tapped fiber as used in the

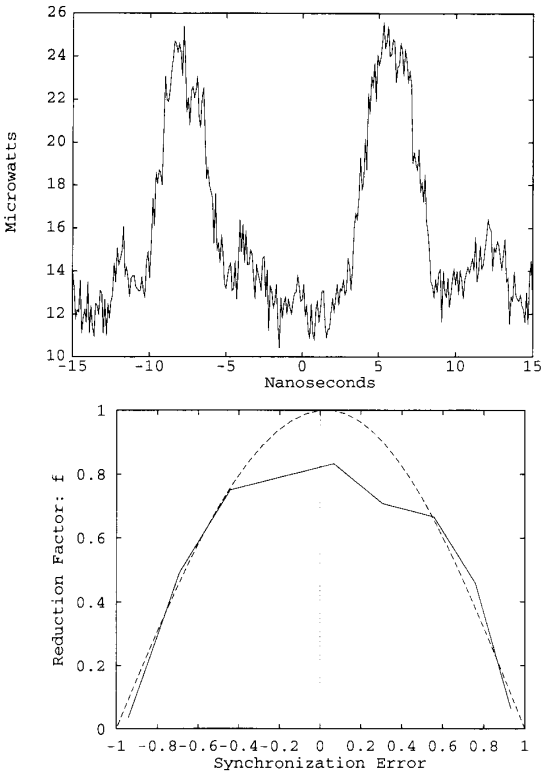


Fig. 10. Synchronization results.

experiment above. The results of this analysis are that the use of passive, bidirectional, 2×2, symmetric fiber couplers in a tapped structure limit the practical bus length in two ways. First, the absolute power delivered to each node on the bus decreases geometrically. While this could be accommodated by amplification and variable thresholding, the second restriction is more severe. The issue is that the relative power between the two coincident pulses is a function of the detector location on the bus. At the ends of the bus, the ratio of coincident to noncoincident power is severely reduced. Together these problems would limit the practical length of a bus to under 100 nodes.

There are three solutions to this problem. First, we can use fiber amplifiers to restore power [34]. Second, we can use multilevel taps to minimize the power lost at each detector. Multilevel taps act as taps with very high coupling ratios. Third, where appropriate, we can star couplers with fiber lengths trimmed to introduce the delays required for the logical operation of the bus.

C. Control Latency

In this section we present a simulation study of the bus performance under various load conditions. Minimizing control overhead is one of our primary motivations in the design. Thus we focus on an analysis of the time spent in control operations versus data transfers. One of the side effects of batching as it has been implemented here is that the average control time per-message required to manage

bus access decreases with increasing traffic. This is because the ratio of long control cycles to short control cycles becomes more favorable.

To analyze bus performance, we conducted a discrete event simulation study on an eight processor model. For simplicity we assume in the model that the processors are arranged in a spiral in order to minimize the feedback path length. The physical separation of each processor, and hence the delay between processors $\tau_{i,j}$, is the same for all pairs of adjacent processors. Further, the round-trip delay is equal to $\tau_{i,j} \times$ *the number of processors.* While this topology is more restrictive than can be supported in general, it provides a convenient time unit for performance measurements which is independent of other parameters such as the number of processors.

Two parameters in the model determine the level of bus contention: average next request delay and average transfer length. Average next request delay, τ_{nrd}, is the period that any processor will wait before issuing its next bus request after completion of a bus is transfer cycle. Average transfer length, τ_{trans}, is the period a processor will hold the bus once a bus grant is issued. For this simulation we have chosen a fixed value for τ_{trans}. Thus the actual length of each simulated transfer was randomly generated within a small range bounded by $\tau_{\mathrm{trans}}/2$. To simulate various levels of bus contention, τ_{nrd} was varied in each simulation. We began with a relatively low-demand environment and incrementally increased demand, by a proportional decrease in τ_{nrd}, until bus saturation. In the final saturated test, new requests arrive at each processor more often than the average transfer length. This assured that in the final simulation each new batch included all other processors.

Figure 11 shows clearly the reduction in overhead with increased contention. In this figure we identify three possible but states: *idle*, *busy*, and *overhead* during any time unit. The bus is *idle* when no bus requests are pending and no transfers are in progress. The *busy* state is defined to be the period when the bus has been granted to a processor and the requested bus transfer is in progress. *Overhead* occurs between the termination of a busy state and the next grant, if a request is currently pending, or the time from request to grant if the bus is currently idle. We have accounted for and plotted in Fig. 11 the percentage of total time the bus spends in each of the three states versus increasing bus demand. The uppermost plot, *busy* $= \square$, increases as expected for larger numbers of bus requests being serviced. The lower plot, *idle* $= \times$, shows the corresponding decrease in bus idle time with demand. *Overhead* $= \triangle$, initially increases with increasing bus traffic until all idle bus cycles have been exhausted. At this level of demand, where $\tau_{\mathrm{nrd}}/\mathit{processors} < \tau_{\mathrm{trans}}$, one or more new requests will always arrive during each bus transfer. In a fixed overhead system this would be defined as the point of bus saturation. The expected behavior of the *busy* and *overhead* plots would be as shown by the solid horizontal lines of Fig. 11. In the protocol we have proposed, it is at this point that batching becomes a dominate effect and control overhead begins a decrease proportional to further increases in the level of demand. The decreasing overhead trace in this region corresponds to additional bus capacity provided by overhead reduction. It continues to decrease to actual bus saturation, where $\tau_{\mathrm{nrd}} < \tau_{\mathrm{trans}}$. At this point there is always a pending request at the neighboring processor upon completion of a bus transfer cycle. Thus control overhead reduces to its minimum, $\tau_{i,j}$.

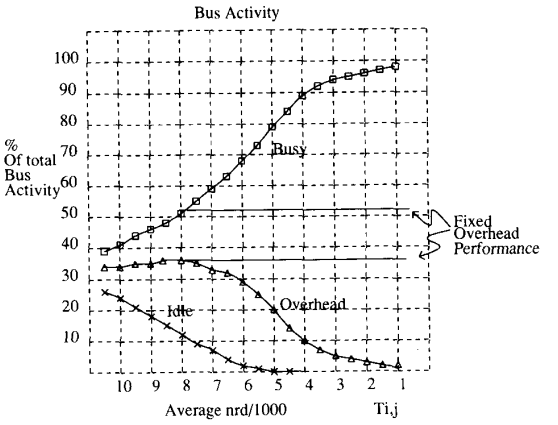


Fig. 11. Simulation results.

VIII. Summary

In this paper we have presented a complete design for an optical fiber bus suitable for applications such as multiprocessor backplanes or other systems applications. The design incorporates optical processing as well as data transfer into the communication links. The resulting system includes an all optical addressing system which eliminates the latency contribution and bandwidth limitation associated with electronic address decoding. The control system uses time of flight relationships between a priority chain and feedback waveguide to implement fully distributed asynchronous and self-timed bus arbitration.

References

[1] Z. Guo, R. G. Melhem, R. W. Hall, D. M. Chiarulli, and S. P. Levitan, "Pipelined communications in optically interconnected arrays," *J. Parallel and Distributed Comput.*, vol. 12, no. 3, pp. 269–282, 1991.

[2] D. M. Chiarulli, R. M. Ditmore, S. P. Levitan, and R. G. Melhem, "An all optical addressing circuit: Experimental results and scalability analysis," *J. Lightware Technol.*, vol. 9, no. 12, Dec. 1991.

[3] D. M. Chiarulli, S. P. Levitan, and R. G. Melhem, "Asynchronous control of optical busses for distributed multiprocessors," *J. Parallel and Distributed Comput.*, vol. 10, pp. 45–54, 1990.

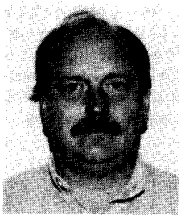
[4] H. M. Ozaktas and J. W. Goodman, "Implications of interconnection theory for optical digital computing," *Appl. Opt.*, vol. 31, no. 26, pp. 5559–5567, Sept. 10, 1992.

[5] J. Goodman, F. Loenberger, S. Kung, and R. Athale, "Optical interconnections for VLSI systems," *Proc. IEEE*, vol. 72, no. 7, pp. 850–866, July 1984.

[6] A. A. Sawchuck and B. K. Jenkins, "Dynamic optical interconnections for parallel processors," *Appl. Opt.*, vol. 2, pp. 143–153, 1986.

[7] S. K. Tweksbury, Ed., *Microelectronic System Interconnections Performance and Modeling.* New York: IEEE Press, 1994.

[8] W. R. Franta and J. P. Hughes, "Extended high speed networks employing HIPPI switches, high speed WANS, and FDDI rings," *J. High Speed Net.*, vol. 1, no. 2, pp. 167–92, 1992.
[9] M. J. Karol and R. D. Gitlin, "High-performance optical local and metropolitan area networks: enhancement of FDDI and IEEE 802.6 DQDB," *IEEE J. Selected Areas Commun.*, vol. 8, no. 5, pp. 490–492, 1990.
[10] H. T. Kung, "High-speed networks for high-performance computing," in *COMPCON Spring '90*, IEEE, Feb.–Mar. 1990, pp. 68–72.
[11] F. E. Ross and J. R. Hamstra, "Forging FDDI," *IEEE J. Selected Areas Commun.*, vol. 11, no. 2, pp. 167–192, Feb. 1993.
[12] N. R. Dono, P. E. Green, K. Liu, R. Ramaswami *et al.*, "A wavelength division multiple access network for computer communication," *IEEE J. Selected Areas Commun.*, vol. 8, no. 2, pp. 78–88, Aug. 1990.
[13] P. W. Dowd, "Optical bus and star-coupled parallel interconnection," in *Proc. 4th Annual Parallel Processing Symp.*, vol. 2, IEEE, Apr. 1990, pp. 824–38.
[14] J. R. Sauer, D. J. Blumenthal, and A. V. Ramanan, "Photonic interconnects for gigabit multicomputer communications," *IEEE LTS*, vol. 3, no. 3, pp. 12–19, Aug. 1992.
[15] W. A. Crossland, P. A. Kirkby, J. W. Parker, and R. J. Westmore, "Some applications of optical networks in the architecture of electronic computers," *Opt. Comput. & Process.*, vol. , no. 3, pp. 199–207, July–Sept. 1991.
[16] D. H. Hartman, G. R. Lalk, T. C. Banwell, and I. Ladany, "Board level high speed photonic interconnections: recent technology developments," in *Optoelectronic Materials, Devices, Packaging and Interconnects II*, SPIE, vol. 994, pp. 57–64, Sept. 1989.
[17] J. W. Parker, P. J. Ayliffe, T. V. Clapp, M. C. Geear *et al.*, "Multifibre bus for rack-to-rack interconnects based on optohybrid transmitter/receiver array pair," *Electron. Lett.*, vol. 28, no. 8, pp. 801–803, Apr. 9, 1992.
[18] H. F. Jordan and J. R. Sauer, "A multi-gb/s optoelectronic packet switching network," in *LEOS Summer Topical on Optical Multiple Access Networks*, IEEE, July 1990, pp. 59–60.
[19] P. Healey, "Minimizing crosspoints and spatial highways in multidimensional bus networks," *Electron. Lett.*, vol. 25, no. 22, pp. 1515–1517, Oct. 26, 1989.
[20] D. J. Blumenthal and J. R. Sauer, "Multiwavelength information processing in gigabit photonic switching networks," in *Multigigabit Fiber Communications*, SPIE, vol. 1787, Sept. 1992, pp. 43–54.
[21] S. Banerjee and B. Mukherjee, "FairNet: A WDM-based multiple channel lightwave network with adaptive and fair scheduling policy," *J. Lightwave Technol.*, vol. 11, nos. 5–6, pp. 1104–1112, May–June 1993.
[22] Y. Birk, "Fiber-optic bus-oriented single-hop interconnections among multi-transceiver stations," *J. Lightwave Technol.*, vol. 9, no. 12, pp. 1657–1664, Dec. 1991.
[23] A. J. Reedy and J. R. Jones, "Methods of collision detection in fiber optic csma/cd networks," *IEEE J. Selected Areas Commun.*, vol. SAC-3, no. 6, pp. 890–896, Nov. 1985.
[24] E. G. Rawson and R. M. Metcalfe, "Fibernet: Multimode optical fibers for local comptuter networks," *IEEE Trans. Commun.*, July 1978.
[25] R. Schmidt, E. G. Rawson, R. Norton, S. Jackson, and M. Bailey, "Fibernet II: A fiber optic ethernet," *IEEE J. Selected Areas Commun.*, Nov. 1983.
[26] C. H. Yoon and C. K. Un, "Unslotted CSMA-CD protocols with combined retransmission strategy for fiber optic bus and ring networks," *Comput. Networks ISDN Syst.*, vol. 21, no. 5, pp. 381–397, July 1991.
[27] F. Tobagi, F. Borgonovo, and L. Fratta, "Expressnet: A high performance integrated services local area network," *IEEE J. Selected Areas Commun.*, vol. SAC-1, no. 5, Nov. 1983.
[28] F. Tobagi and M. Fine, "Performance of unidirectional broadcast local area networks: Expressnet and fastnet," *IEEE J. Selected Areas Commun.*, vol. SAC-1, no. 5, Nov. 1983.
[29] M. Nassehi, F. Tobaji, and M. Marhic, "Fiber optic configurations for local area networks," *IEEE J. Selected Areas Commun.*, Nov. 1985.
[30] C. Yeh, M. Lin, M. Gerla, and P. Rodrigues, "Rato-net: A random-access protocol for unidirectional ultra-high-speed optical fiber network," *J. Lightwave Technol.*, vol. 8, no. 1, pp. 78–79, Jan. 1990.
[31] H. B. Jeon, B. C. Shin, and C. K. Un, "Probabilistic reservation protocol for high-speed unidirectional bus networks," *Comput. Commun.*, vol. 16, no. 3, pp. 140–146, Mar. 1993.
[32] E. K. Thurber, *The LOCALNetter Designer's Handbook*. Architecture Technology Corp., Nov. 1985.
[33] S. Joshi, "High performance networks: A focus on the fiber distributed data interface (FDDI) standard," *IEEE Micro*, June 1986.
[34] M. M. Bidnurkar, S. P. Levitan, R. Melhem, and D. M. Chiarulli, "Model of lossless bus structure using erbium fiber amplifiers pumped near 820 nm," in *Optical Computing Technical Digest*. Optical Soc. America, Mar. 15–19 1993, poster.
[35] D. M. Chiarulli, S. P. Levitan, and R. G. Melhem, "Demonstration of an all optical addressing circuit," in *Optical Computing*. Salt Lake City, UT: Optical Soc. of America, Mar. 1991.



Donald M. Chiarulli received the B.S. degree in physics in 1976, from Louisiana State University, Baton Rouge, the M.S. degree in computer science in 1979 from Virginia Polytechnic Institute, Blacksburg, and the Ph.D. degree in 1986 also from Louisiana State University.

In 1986 he joined the Department of Computer Science at the University of Pittsburgh, Pittsburgh, PA, where he is currently an Associate Professor.

Dr. Chiarulli is a member of the Association for Computing Machinery, ACM SIGARCH, SIGMICRO, the IEEE Computer Society, and the Optical Society of America.

Steven P. Levitan received the B.S. degree from Case Western Reserve University, Cleveland, OH, in 1972 and the M.S. (1979) and Ph.D. (1984) degrees, both in computer science, from the University of Massachusetts, Amherst.

He was an Assistant Professor from 1984 to 1986 in the Electrical and Computer Engineering Department at the University of Massachusetts. In 1987 he joined the Electrical Engineering faculty at the University of Pittsburgh where he is the Wellington C. Carl Associate Professor of Electrical Engineering.

Dr. Levitan is a member of the IEEE Computer Society, , ACM, SPIE, and OSA.

Rami G. Melhem was born in Cairo, Egypt, in 1954. He received the B.E. degree in electrical engineering from Cairo University in 1976, the M.A. degree in mathematics (1981), the M.S. degree in computer science (1981), and the Ph.D. degree in computer science (1983) all from the University of Pittsburgh, Pittsburgh, PA.

Since 1989, he has been an Associate Professor of Computer Science at the University of Pittsburgh. Previously, he was an Assistant Professor at Purdue University, West Lafayette, IN, and at the University of Pittsburgh.

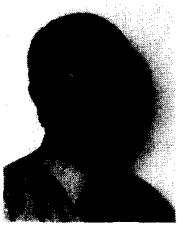
Dr. Melhem is a member of the IEEE Computer Society, the Association for Computing Machinery, and the International Society for Optical Engineering.

Manoj Bidnurkar, photograph and biography not available at the time of publication.

Robert Ditmore, photograph and biography not available at the time of publication.

Gregory Gravenstreter received the B.S. degree in mathematics and computer science from the University of Pittsburgh, Pittsburgh, PA, and is currently a graduate student in the Computer Science Department there.

From 1971 to 1985 he was a researcher in the Chemical Physics and Computer Science Departments at Westinghouse Electric Corporation's Research and Development Center, Pittsburgh. From 1988 to 1992, he worked as Director of Systems and Operations for Guidance Technologies, also in Pittsburgh. His interests include architectural and systems issues for parallel and distributed systems and massively parallel processing.



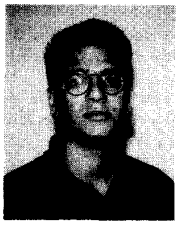
Zicheng Guo received the B.S. degree form Xi'an Jiaotong University, China, in 1982. He received the M.S. and Ph.D. degrees in electrical engineering from the University of Pittsburgh, Pittsburgh, PA, in 1986 and 1991, respectively.

He was on the faculty at Xi'an Jiaotong University in the Electrical Engineering Department for three years. He has been an Assistant Professor in the Electrical Engineering Department at Louisiana Tech University, Ruston, since 1991.

Chunming Qiao (Member, IEEE) was born in Suzhou, P. R. China, in 1965. He received the B.S. degree in computer science and engineering from the University of Science and Technology of China in 1985, and the M.S. and Ph.D.degrees, both in computer science, from the University of Pittsburgh, Pittsburgh, PA, in 1990 and 1993, respectively.

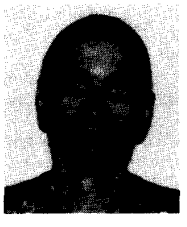
He joined the Department of Electrical and Computer Engineering, State University of New York at Buffalo as an Assistant Professor in 1993.

Dr. Qiao is a member of the International Society for Optical Engineering and a member of the IEEE Computer Society.



Majd F. Sakr (Member, IEEE) was born in Beirut, Lebanon, on January 20, 1969. He received the B.S. degree in 1992 and the M.S. degree in 1994, both in electrical engineering , from the University of Pittsburgh, Pittsburgh, PA.

He is currently pursuing the Ph.D. degree in electrical engineering at the University of Pittsburgh. His research interests include optical interconnection networks, parallel computer architecture, neural networks, and VLSI design. He is an intern at NEC Research Institute, Princeton, NJ, performing his Ph.D. research there.



James P. Teza received the B.S. degree in physics from the University of Pittsburgh, Pittsburgh, PA, where he is currently completing the requirements for the M.S. degree in electrical engineering.

His research interests include optical interconnection networks for multiprocessor systems.

Mr. Teza is a student member of the IEEE Computer Society.