# Optoelectronic Buses for High Performance Computing*

D.M. Chiarulli, S.P. Levitan, R.G. Melhem
M. Bidnurkar, R. Ditmore, G. Gravenstreter, Z. Guo, C. Qao, J. Teza
University of Pittsburgh
Pittsburgh, PA 15260

January 15, 1996

### Abstract

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.

# Contents

# 1 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, multi-chip-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 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 networks require both addressing and arbitration of accesses. However, 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. For example, our optical address bus provides two paths by which signals may reach a node. Optical addressing is achieved by encoding an address as a 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 of an optical feedback wavefront in lieu of a clocking signal in an optical priority chain.

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 two, outlines previous research on both networks and optical bus implementations. Section three describes the data bus. Section four introduces our solution to all optical address generation and decoding. Section five deals with the access arbitration problem and presents an electro-optical distritbuted solution. Section six shows the combined implementation for the bus, in Section seven we report on various experiments designed to determine the technological limits on the implementation.

## 2 Background

Optical interconnections offer the potential for gigahertz transfer rates in an environment free from capacitive bus loading, and electro-magnetic interference. The effectiveness of optical interconnections has been examined from both theoretic [4, 5] and practical perspectives [6, 7, 8, 9, 10].

Over the past decade, much of the research in optical communications networks has focused on applications to wide area networks (WANS) [11, 12] and metropolitan area networks (MANS) [13, 14, 15, 16, 17]. More recently, specialized high speed local area networks (HSLNs) for computer interconnections have been studied [18, 19, 20] and commercial standards have emerged [21, 22]. Other research groups have investigated the implementation of parallel computers using optical interconnections in multiprocessor applications [23, 24, 25, 18, 26, 27].

Device technology in electro-optics has also matured to a point where small, low power, and low cost devices exist which are suitable for use in bus level implementations[28]. Initial efforts have focused on direct technology substitution [29] in board level [30] and chassis to chassis [31] links. However, there are 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. Even though optical pulses as short as few femto-seconds may be generated and detected[32, 33, 34, 35], such short pulses may not be useful to transmit data on an optical bus without an electro-optic interface of comparable speed. In other words, the speed of electronics bounds the transmission speed of optical buses.

In switched networks, time division switched (TDS) [36, 37, 38], space division switched (SDS) [39, 40], and wavelength division switched (WDS) [41, 42] 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[43, 44, 45, 46, 47].

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 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 [48]. Examples of this class are Fibernet [49, 50] and Fibernet II [51]. 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

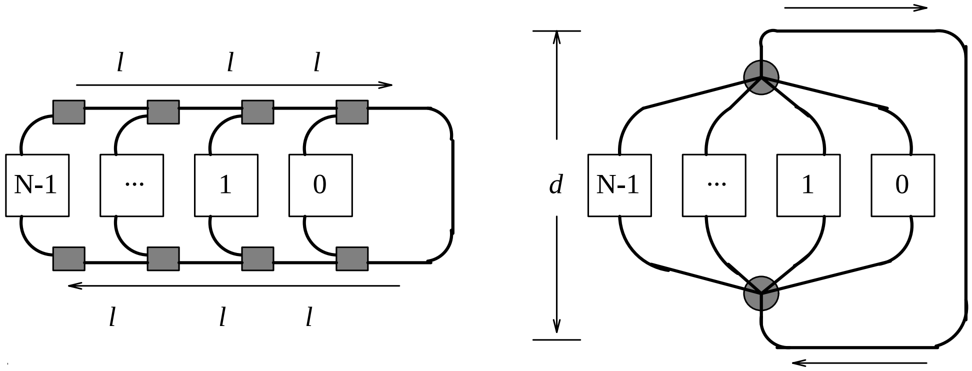

Figure 1: **Tapped Fiber and Star Coupled Bus Designs**

the speed and complexity of centralized electronic control[52, 53]. The second class is the Demand assignment multiple access (DAMA) protocol. Examples include EXPRESSNET and FASTNET [51, 54, 55, 56]. 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 [57] is a fair, bounded delay, random access protocol which also uses this technique. Other DAMA based protocols have also been recently suggested by Joen et. al. [58] and Kovacevic et. al. [59]. The third class of control strategies are based on token rings. The 80 Mbps fiber ring network from Proteon [60], and the 100Mbps Fiber Distributed Data Interface (FDDI) ring [61], are examples of this control structure. Other designs have been recently proposed be Schaffa et. al. [62], and Gerla et. al. [63].

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 substantial overhead in controller complexity. In addition, depending on the control structure adopted, a significant time overhead is paid for either collision handling or control information appended to message packets.

On the other hand, bus applications are characterized 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.

## 3 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 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 Figure 1. Both or these structures are functionally equivalent. However, their temporal and power distribution characteristics differ significantly. For example, the time of flight, $T$, for a message to traverse a star coupled system is $T = dc_g$ where $d$ the total length of fiber connected to 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.

In its power distribution characteristics, the star coupler has a significant advantage over a tapped fiber. 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 $\epsilon$ 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, multi-level taps can be used to reduce losses[2], or fiber amplifier segments can be introduced to restore power [64] respectively. An analysis of the limits on power distribution in tapped fiber is presented in section 7.

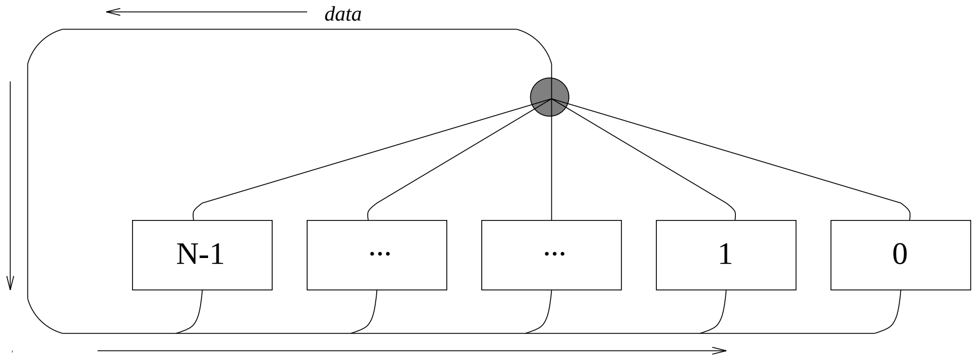

Figure 2: **Data bus interconnection**

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.

Figure 2 shows the structure of data bus portion of the system bus implementation. The temporal differences between the input side and the output side of the bus are intentional in order to match the temporal characteristics of the address bus. As we will show, data and addresses concurrently traverse identical path lengths and arrive in a fixed temporal relationship at each of the receiver sites.

## 4 Optical Addressing using Coincident Pulse Logic

The address bus implementation exploits two properties of optical signals, unidirectional propagation and predictable path delay, which make it possible to encode an address as the relative timing of two optical signals. In this technique, called coincident pulse addressing, 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.

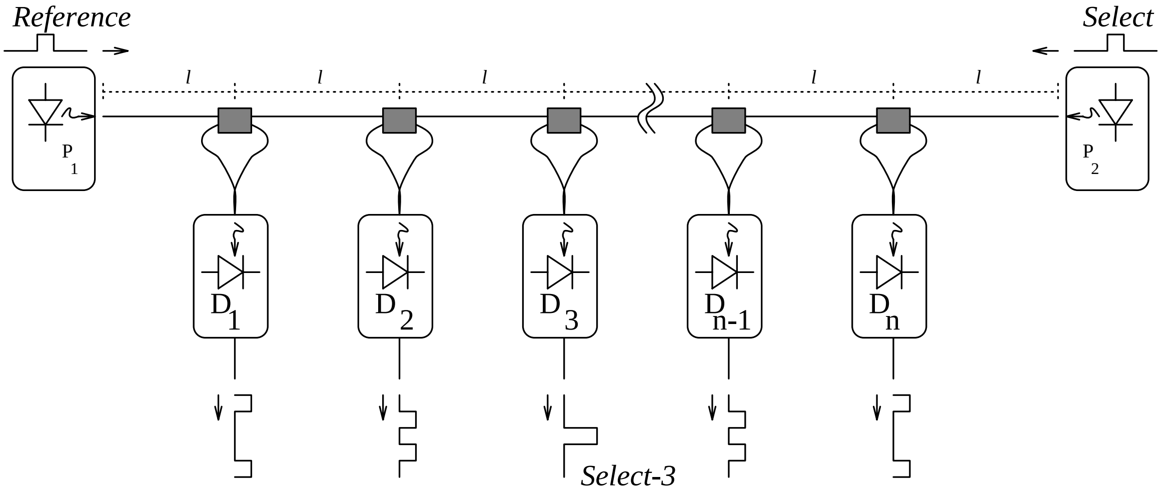

Figure 3: **Coincident pulse addressing structure**

Consider the optical addressing structure shown in figure 3. 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 $\tau$ 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 pulse width. Using $2 \times 2$ 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 $t_i, i \in \{1..n\}$ 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 $2\tau$, it follows that the system latency, $\sigma = 2n\tau$. Further, up to $n$ locations may be selected in parallel within a single latency period. Therefore the system throughput is $\nu = 1/2\tau$.

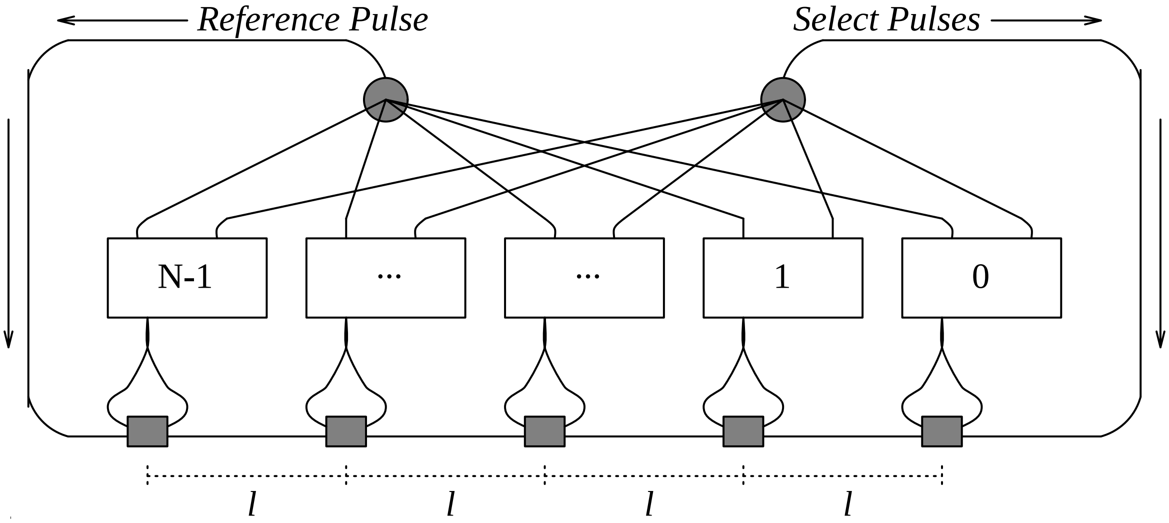

Figure 4: **Coincident Pulse Address bus**

This simple technique can be used as the basis for a practical addressing mechanism for a system bus. Figure 4 shows a design in which the select and reference pulse generators in Figure 3, 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 figure 2. 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.

## 5 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 5, shows the external interconnections for a typical control node. There are two electronic signals, *BusRequest*

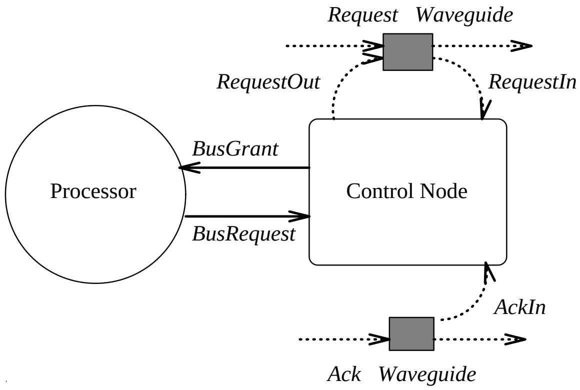

Figure 5: **Control Node External Connections**

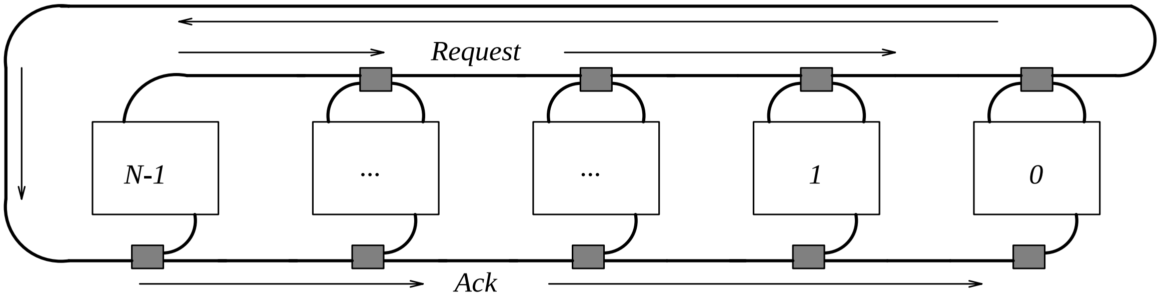

Figure 6: **Control Bus, Request/Ack Waveguide feedback structure**

and *BusGrant* which connect the processor to the control node. Three optical signals, one output and two input, 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* waveguides shown in figure 6.

The functions of the *Request* and *Ack* waveguides 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 grouped more closely on the bus. We show this effect in section 7.2, where we present a simulation analysis.

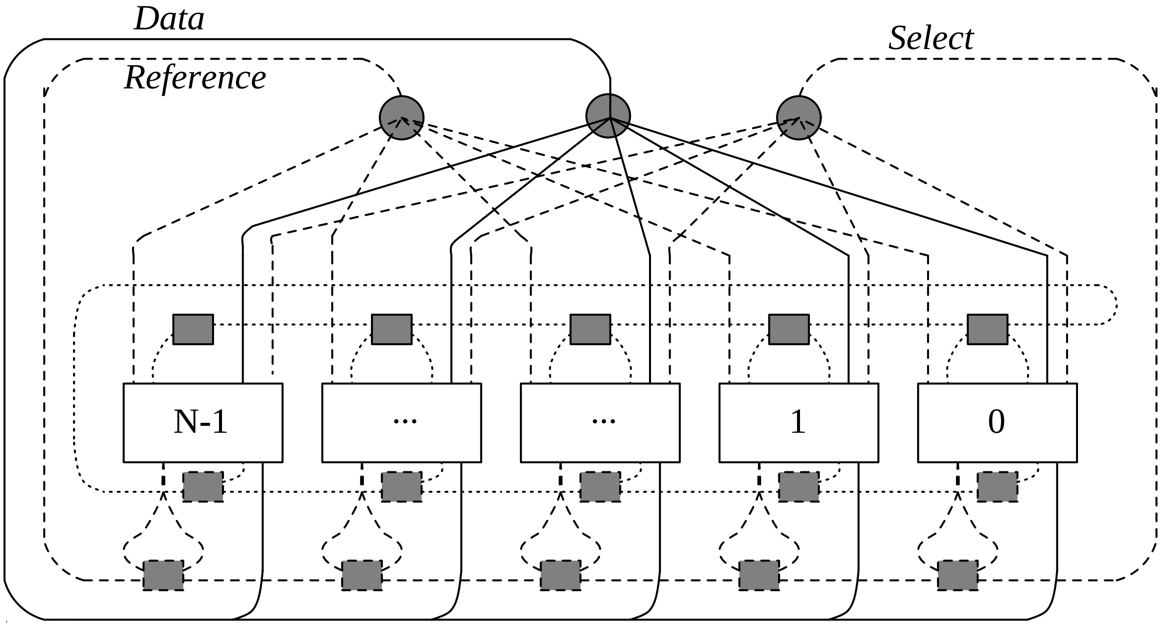

Figure 7: **Complete System Bus**

# 6 Combined Implementation

In this section we combine the data, address and control buses from the previous sections into a combined implementation. The design is shown in Figure 7. The data, address, and control buses operate as described in the previous sections. In this section we focus on the combined operation, specifically on the timing of each bus cycle and the latency of bus transfers.

In this description it is important to note a specific difference between the operation of optical and electronic bus implementations. In electronics, signal validity is implicitly assumed to be spatially invariant along the length of the bus. In other words, a bus signal is only considered valid when it is stable at all inputs, along the entire bus length. Variations in transport time are attributed to signal skew, and the worst case skew must be accounted for in the timing analysis. On an optical bus the transport time, or time of flight, of a signal is considered as part of the signal state. Thus, signal validity is both temporally and spatially variant. As a practical matter, this implies that a timing diagram representing the state of a bus line cannot be drawn for an optical bus. Timing diagrams of bus signals are only meaningful when confined to a specific location on the bus. In other words at the inputs or outputs of a specific node.

The accuracy and reproducibility of of the transport time between the transitions of a bus signal at each node input is a key requirement in the design of the addressing and control structures. Optical signal propagation time is a function of the wavelength of the signal, the length of fiber, signal mode, and the refractive index of the fiber medium. Although these parameters are subject to environmental effects and manufacturing tolerances, variation effects are on the order of picoseconds/kilometer. Bus lengths are typically on the order of meters. Thus, even for bandwidths of hundreds of gigahertz, time of flight variations are three orders of magnitude smaller than a bit time.

The timing diagram in figure 8 represents the states of bus signals at two arbitrary control nodes, $node_i$ and 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 8 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.

As stated above, bus transfers consist of interleaved control and data transfer cycles. 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 occur at any time during the data transfer or short 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

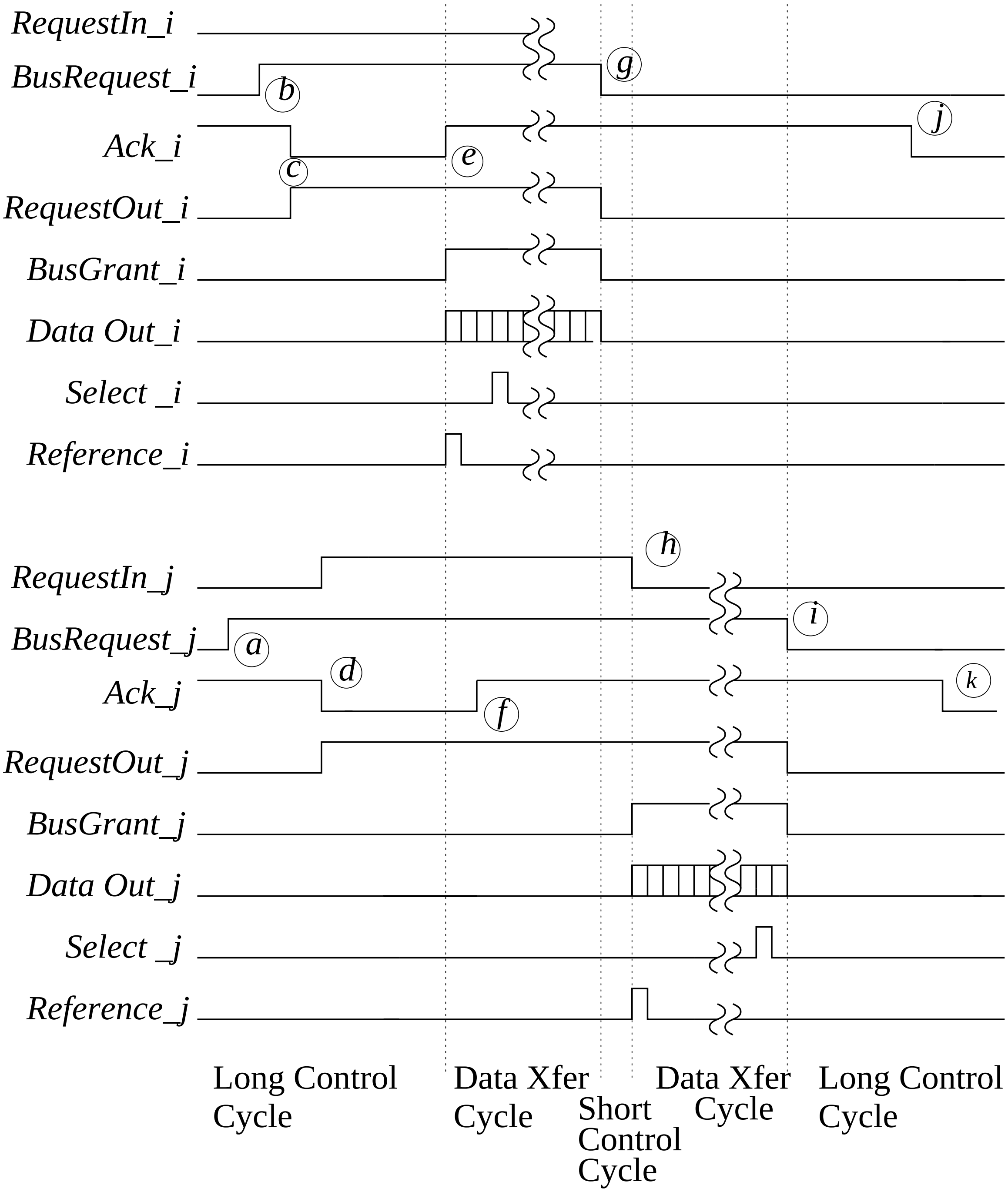

Figure 8: **Control Timing**

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, $node_i$ lower $RequestOut_i$ 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 $RequestOut_j$ 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 of the system by experimental and simulation analysis and discuss the scalability limitations for such a design.

# 7 Performance and Scalability Limitations

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 pulse width, 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. Latency in the control bus design was characterized by simulation analysis for various synthetic traffic loads. Power distribution was characterized analytically for linear tapped fiber structures.

## 7.1 Temporal Limits

In this section, we present results from a laboratory experiment on a prototype of the address bus 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 non-coincident waveforms before we discus the experiment itself.

Figure 9 is a diagram of the prototype structure. The fiber bus consists of a length of multimode fiber tapped three times using Gould 10 dB fiber couplers. Select and reference bit patterns are generated by modulating the 4ns 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 outputs 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.

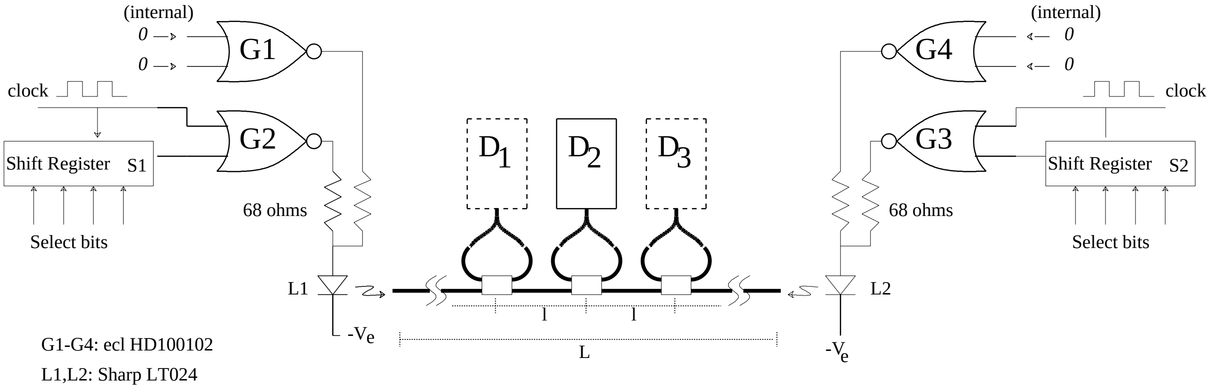

Figure 9: **Synchronization Experiment**

Figure 10 shows waveforms for both coincidence and non-coincidence measured at detector 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 non-coincident 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 non-coincident pulses is a function of the detector location as discussed in section 7.3.

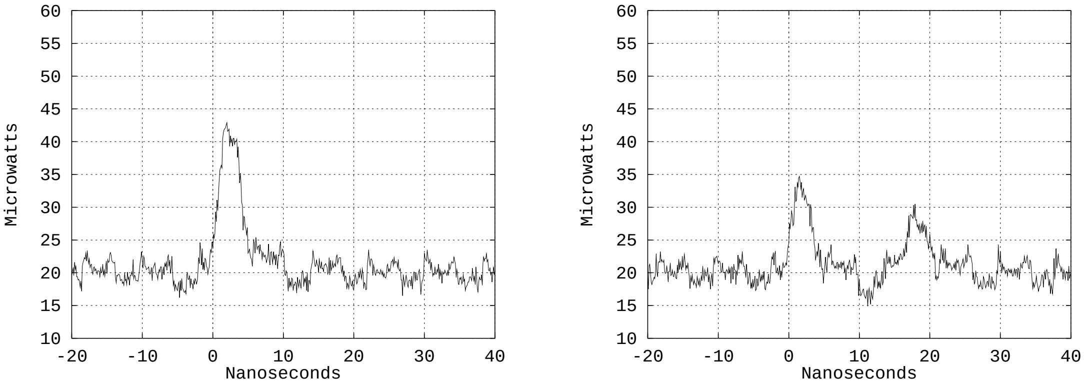

Figure 10: **Select D1 measured at D1, and Select D3 measured at D1**

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 pulse width, synchronization precision has a direct bearing on the absolute width and height of an addressing pulse that can be effectively detected.

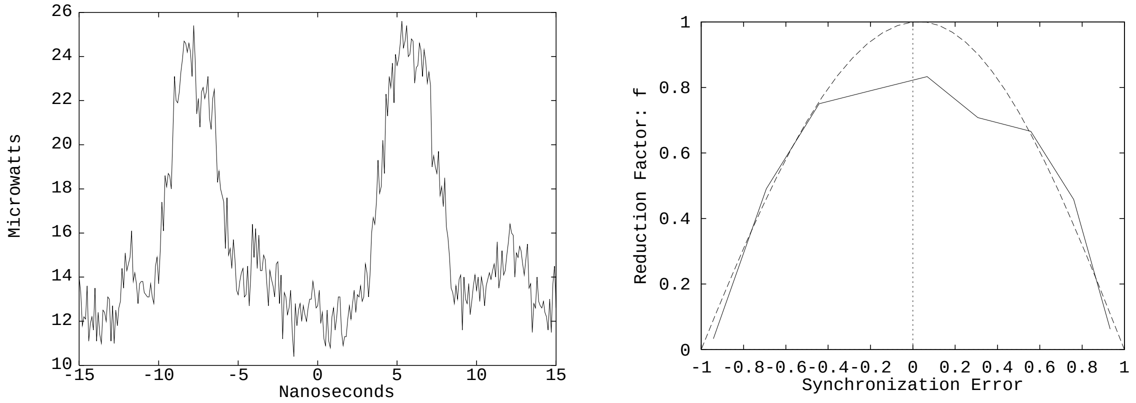

Figure 11: **Synchronization Results**

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 figure 11 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 figure 11 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 pulse width. 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 the reference pulse, shown as negative error, reduces this power by the factors shown. The solid line in figure 11 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.

In order to analyze this result, we must consider the sources of synchronization error. Assuming that tolerances for electronic components and errors in fiber length measurements can be compensated for by tuning the system, the primary sources of synchronization error will be thermal variations in both the optical characteristics of the fiber and in the performance of electronic components. For the former, studies [65] have shown that the variability of the index of refraction of the fiber versus temperature is on the order of 40ps/kilometer-degree C, and that this is the dominant temperature effect. This represents a very minor variation in effective time delay. Obviously, the electronic variations with temperature will be the predominant source of synchronization error. However, from figure 11 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 pulse width) in the electronics can be tolerated without significant degradation of the coincident signal.

The experiment shows that the important system issues of latency and throughput which are related to pulse width limits are highly scalable. Based on current and near term technology, we have shown that synchronization error does not contribute significantly to the bounds calculated above. On the other hand, physical scalability issues such as the size of the bus and the number of detectors that can be supported are more severely restricted due to power distribution in a system built from passive couplers. In section 7.3 we discuss the implications of power distribution on scalability issues. First we consider the latency imposed by the control structure itself.

## 7.2 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, $\tau_{nrd}$, is the period that any processor will wait before issuing its next bus request after completion of a bus transfer cycle. Average transfer length, $\tau_{trans}$, 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 $\tau_{trans}$. Thus the actual length of each simulated transfer was randomly generated within a small range bounded by $\tau_{trans}/2$. To simulate various levels of bus contention, $\tau_{nrd}$ was varied in each simulation. We began with a relatively low demand environment and incrementally increased demand, by a proportional decrease in $\tau_{nrd}$, 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 12 shows clearly the reduction in overhead with increased contention. In this figure we identify three possible bus 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 figure 12 the percentage of total time the bus spends in each of the three states versus increasing bus demand. The uppermost plot, *busy* = □, increases as

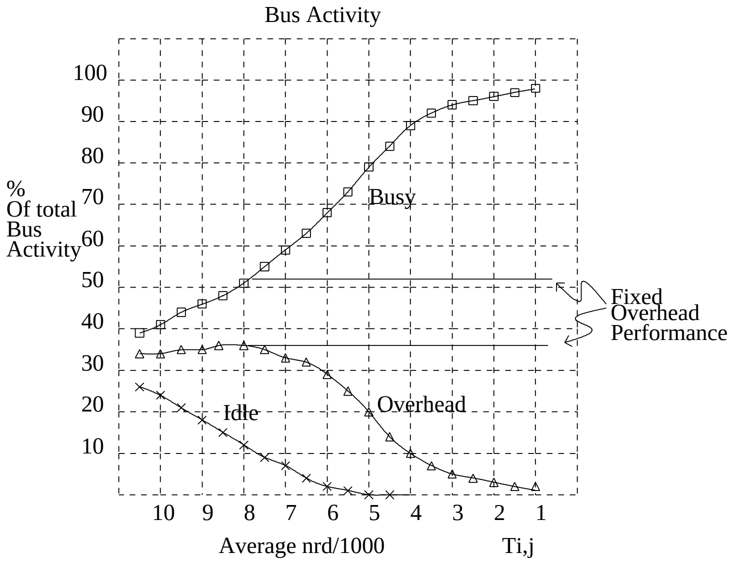

Figure 12: **Simulation Results**

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 $\tau_{nrd}/processors < \tau_{trans}$, 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 figure 12. 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 $\tau_{nrd} < \tau_{trans}$. 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, $\tau_{i,j}$.

### 7.3 Power Distribution

In this section we address the third limitation to the proposed bus organization, power distribution. We present an analysis of power distribution in a bi-directional tapped fiber as used in the experiment above and discussed in section 6. In this analysis, we use passive, bidirectional, $2 \times 2$, symmetric fiber couplers [66, 67] and assume no excess loss in the couplers.

Since the couplers are symmetric and bidirectional, either side can be considered the input or the output. The power distribution from the input to the output is:

$$\begin{pmatrix} A' \\ B' \end{pmatrix} = \begin{pmatrix} r & (1-r) \\ (1-r) & r \end{pmatrix} \begin{pmatrix} A \\ B \end{pmatrix}$$

Where $A$ and $B$ are the input ports, $A', B'$ are the output ports, and $r$ is the coupling ratio.

As is shown in figure 13, a linear bus built from a tapped fiber consists of $n$ detectors (and $n$ couplers). Assuming two, unit height, pulses starting at opposite ends of the bus, and one type of coupler with a ratio of $r$, the optical power from each pulse $p_i^1$ and $p_i^2$ at detector $D_i$ is given by the equations:

$$p_i^1 = r^{(i-1)}(1-r), \qquad p_i^2 = r^{(n-i)}(1-r). \tag{3}$$

Because the bus is symmetrical, we can analyze one signal which originates on the left from a single transmitter and propagates to the right as shown in figure 13.

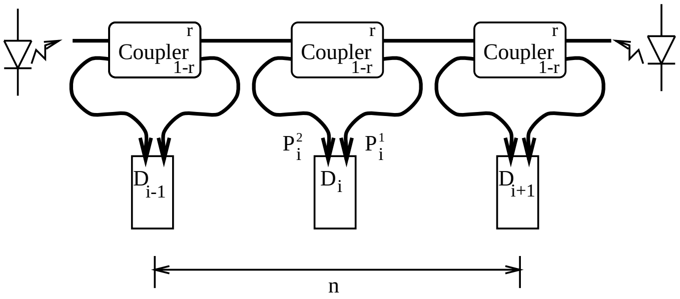

Figure 13: **Linear Optical Bus**

Figure 14 is a plot of $p_i^1$ versus $i$ for various values of $r$. Note that the values of $i$ are plotted on a logarithmic scale. The topmost curve is for a bus with $r = 90\%$ where the power at the first detector is 10% of the initial power. The lowest curve is for a bus with $r = 99\%$ where power at the first detector is 1% of the initial pulse power. For all the curves, the absolute power falls off geometrically.

A bound on the number of detectors, $n$, is determined by the sensitivity of the last detector on the bus. In other words, it is the bound for a detector to discriminate between "no pulse" and "pulse". If the last detector has a sensitivity $Pmin$, then the maximum number of detectors supportable is:

$$n = \frac{log(\frac{Pmin}{1-r})}{log(r)} + 1 \tag{4}$$

Equation (4) is shown graphically in figure 15 for a set of coupling ratios $r = 90\%$, 95%, 97%, 98%, 99% and $.01\% \leq Pmin \leq 1\%$ of the input power on a logarithmic scale. This graph confirms the intuition that by improving either the coupling ratio $r$, or the sensitivity of the detectors $Pmin$ we will be able to support more detectors on the bus. We also note the sharp drop in $n$ for high values of $Pmin$ and $r$ which reflects the situation where much of the available power flows off the end of the bus and is wasted.

However, it is clear that it is not the absolute power but rather the *power margin* which imposes a bound on the size of the system. We define the power margin $Pm$ to be the amount of additional

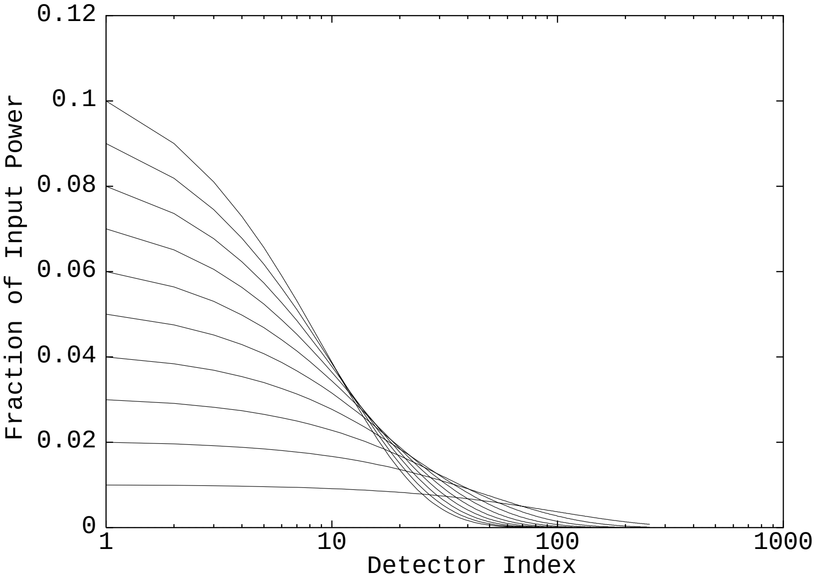

Figure 14: **Power $p_i^1$ at detector $D_i$ for** $90\% \leq r \leq 99\%$

power in a coincident pulse relative to the largest non-coincident pulse seen by a detector. This is given as a fraction of the maximum non-coincident pulse power:

$$Pm = (p_1 + p_2 - max(p_1, p_2))/max(p_1, p_2) = min(p_1, p_2)/max(p_1, p_2)$$

$Pm$ indicates the threshold level needed for a detector to discriminate between coincident and non-coincident pulses. That is, for each detector on the bus the threshold should be set to be at:

$$((Pm + 1) \times max(p1, p2))/2$$

For any linear structure, $Pm$ has its maximum value, $Pm = 1$, at the center of the bus, where each pulse is at equal power, and coincidence is reflected as a doubling of power seen by the detector. It is at its minimum value at the ends of the bus. Another constraint is that the configuration of the coincident address bus requires bidirectional propagation. Therefore, we are constrained to use a single tapping ratio, $r$, for all couplers.

Based on these two constraints of power margin and a single coupling ratio, the graph shown in figure 16 which is a plot of worst case power margin $Pm$ versus $1 - r$ for various bus lengths, confirms that the power margin for the coincident structure bounds scalability more strongly than absolute power. We can see from figure 15 that using 90 percent couplers, and assuming we can tolerate a $Pmin$ of .001 of input power we could achieve bus lengths of about 50 detectors. However, figure 16 shows that for a power margin of $Pm = 20\%$ we could only reach lengths of 16 detectors. Therefore, due to both minimum power constraints but more strongly by power margin issues the system scale is highly sensitive to the fixed value of $r$.

There are three solutions to this problem. First, is the use of more sophisticated taps that use cladding modes to give tapping ratios as small as 36db[68]. Second is the use of non-linear fiber

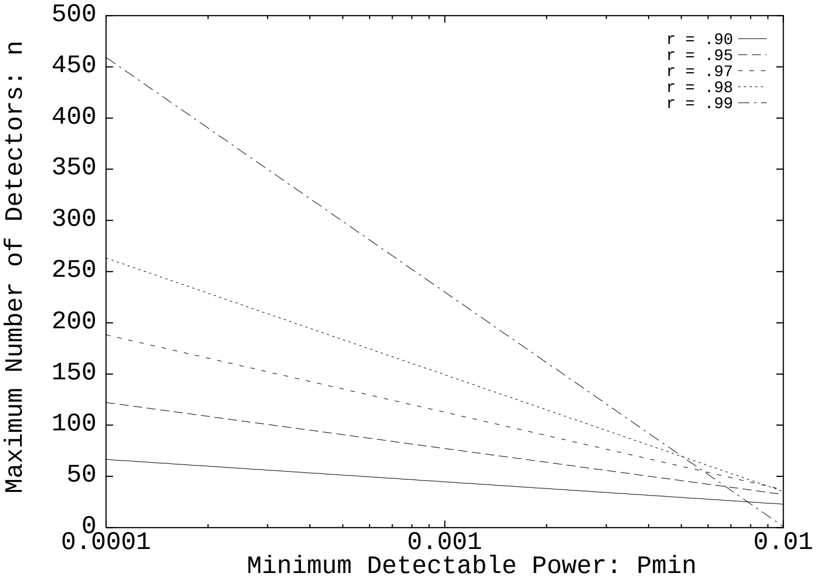

Figure 15: **Number of Detectors versus Pmin for various values of** $r$

amplifiers to restore power on the bus[64]. Third, is the use use of multi-level structures, proposed by Nassehi[56].

## 8 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. *Journal of Parallel and Distributed Computing*, 12(3):269–282, 1991.

[2] Donald M. Chiarulli, Robert M. Ditmore, Steven P. Levitan, and Rami G. Melhem. An all optical addressing circuit: Experimental results and scalability analysis. *IEEE Journal of Lightwave Technology*, 9(12), December 1991.

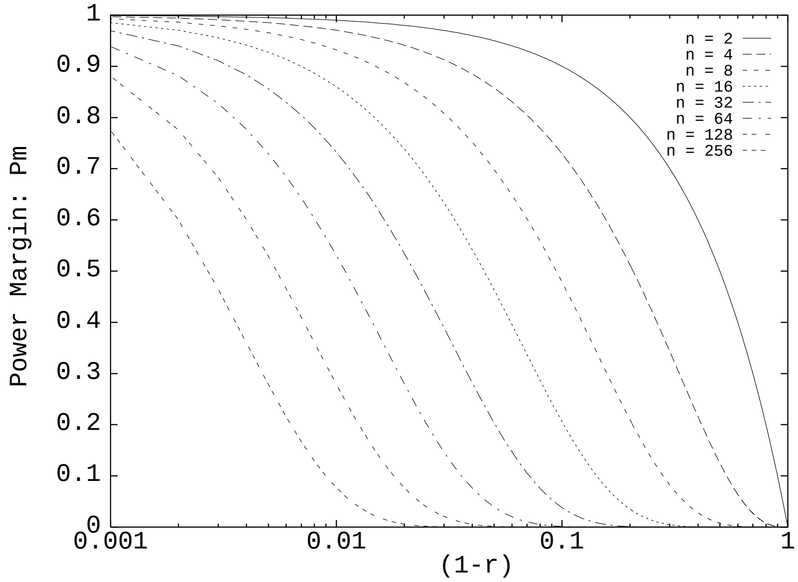

Figure 16: **Power margin** $Pm$ **versus** $0.001 \geq (1-r) \geq 1$ **for various bus sizes**

[3] D.M. Chiarulli, S.P. Levitan, and R.G. Melhem. Asynchronous control of optical busses for distributed multiprocessors. *Journal of Parallel and Distributed Computing*, 10:45–54, 1990.

[4] H.M. Ozaktas and J.W. Goodman. Lower bound for the communication volume required for an optically interconnected array of points. *Journal of the Optical Society of America A (Optics and Image Science)*, 7(11):2100–6, November 1990.

[5] H.M. Ozaktas and J.W. Goodman. Implications of interconnection theory for optical digital computing. *Applied Optics*, 31(26):5559–67, 10 Sept. 1992.

[6] J. Goodman, F. Loenberger, S. Kung, and R. Athale. Optical interconnections for VLSI systems. *Proceedings of the IEEE*, 72(7):850–866, July 1984.

[7] A. A. Sawchuck and B. K. Jenkins. Dynamic optical interconections for parallel processors. *Applied Optics*, 25:143–153, 1986.

[8] S. Redfield. Opportunities for optics in computing. In *Twenty-Third Asilomar Conference on Signals, Systems, and Computers*, volume 1, pages 84–8. IEEE, Oct-Nov 1989.

[9] L. D. Hutcheson and P. Haugen. Optical interconnects replace hardware. *IEEE Spectrum*, pages 30–35, March 1987.

[10] Stuart K. Tweksbury, editor. *Microelectronic System Interconnections Performance and Modeling*. IEEE Press, New York, 1994.

[11] J. Hecht. Bell labs transmits 8 gbits/s over 68km. *Lasers and Applications*, May, 1986.

[12] W.R. Franta and J.P. Hughes. Extended high speed networks employing HIPPI switches, high speed WANS, and FDDI rings. *Journal of High Speed Networks*, 1(2):167–92, 1992.

[13] M.-S. Chen, N.R. Dono, and R. Ramaswami. A media-access protocol for packet-switched wavelength division multiaccess metropolitan area networks. *IEEE Journal on Selected Areas in Communications*, 8(2):78–88, Aug. 1990.

[14] M.J. Karol and R.D. Gitlin. High-performance optical local and metropolitan area networks: enhancement of FDDI and IEEE 802.6 DQDB. *IEEE Journal on Selected Areas in Communications*, 8(5):490–2, Oct. 1990.

[15] J. Bannister, M. Gerla, and M. Kovaucevic. An all-optical multifiber tree network. *Journal of Lightwave Technology*, 11(5-6):997–1008, May-June 1993.

[16] F.J. Janniello, R. Ramaswami, and D.G. Steinberg. A prototype circuit-switched multi-wavelength optical metropolitan-area network. *Journal of Lightwave Technology*, 11(5-6):777–82, May-June 1993.

[17] P.E. Green, L.A. Coldren, K.M. Johnson, J.G. Lewis, , et al. All-optical packet-switched metropolitan-area network proposal. *Journal of Lightwave Technology*, 11(5-6):754–63, May-June 1993.

[18] K.A. Falcone and O.K. Tonguz. Access methods for fiber-optic interconnection of LAN's. *Journal of Lightwave Technology*, 11(5-6):1113–24, May-June 1993.

[19] R. Ramaswami. Multiwavelength lightwave networks for computer communication. *IEEE Communications Magazine*, 31(2):78–88, Feb. 1993.

[20] H.T. Kung. High-speed networks for high-performance computing. In *COMPCON Spring '90*, pages 68–72. IEEE, Feb-March 1990.

[21] F.E. Ross and J.R. Hamstra. Forging FDDI. *IEEE Journal on Selected Areas in Communications*, 11(2):167–92, Feb. 1993.

[22] M. Gerla and J.A. Bannister. A guide to data communications: high-speed local-area networks. *IEEE Spectrum*, 28(8):26–31, Aug. 1991.

[23] N.R. Dono, P.E. Green, K. Liu, R. Ramaswami, et al. A wavelength division multiple access network for computer communication. *IEEE Journal on Selected Areas in Communications*, 8(2):78–88, August 1990.

[24] P.W. Dowd. Optical bus and star-coupled parallel interconnection. In *Proceedings of the Fourth Annual Parallel Processing Symposium.*, volume 2, pages 824–38. IEEE, April 1990.

[25] A. Hartmann and S. Redfield. Design sketches for optical cross bar switches intended for large-scale parallel processing applications. *Optical Engineering*, 28(4):315–27, April 1989.

[26] R. K. Kostuk, Yang-Tung Huang, and M. Kato. Multiprocessor optical bus. In *International Symposium on Advances in Interconnection and Packaging*, volume 1389, pages 515–22. SPIE, November 1991.

[27] J.R. Sauer, D.J. Blumenthal, and A.V. Ramanan. Photonic interconnects for gigabit multicomputer communications. *IEEE LTS*, 3(3):12–19, Aug. 1992.

[28] R.G. Hunsperger. *Integrated optics: theory and technology. 3rd edition.* Springer-Verlag, Berlin, Germany, 1991.

[29] W.A. Crossland, P.A. Kirkby, J.W. Parker, and R.J. Westmore. Some applications of optical networks in the architecture of electronic computers. *Optical Computing & Processing*, 1(3):199–207, July-Sept. 1991.

[30] 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*, volume 994, pages 57–64. SPIE, September 1989.

[31] J.W. Parker, P.J. Ayliffe, T.V. Clapp, M.C. Geear, et al. Multifibre bus for rack-to-rack interconnects based on opto-hybrid transmitter/receiver array pair. *Electronics Letters*, 28(8):801–3, 9 April 1992.

[32] C. V. Shank. *The Role of Ultrafast Optical Pulses in High Speed Electronics.* Springer Verlag, 1985.

[33] M.J. LaGasse, K.K. Anderson, H.A. Haus, and J.G. Fujimoto. Femtosecond all-optical switching in AlGaAs waveguides using a time division interferometer. *Applied Physics Letters*, 54(21):2068–70, 22 May 1989.

[34] J.G. Fujimoto. New technologies for ultrashort pulse generation in solid state lasers. *Optics & Photonics News*, 2(3):8–13, March 1991.

[35] P.J. Delfyett, D.H. Hartman, and S.Z. Ahmad. Optical clock distribution using a mode-locked semiconductor laser diode system. *Journal of Lightwave Technology*, 9(12):1646–9, Dec. 1991.

[36] H.F. Jordan and J.R. Sauer. A multi-gb/s optoelectronic packet switching network. In *LEOS Summer Topical on Optical Multiple Access Networks*, pages 59–60. IEEE, July 1990.

[37] M.A. Santoro and M.J. Karol. Experimental and theoretical performance of ring-shaped passive-bus optical networks. *IEEE Photonics Technology Letters*, 3(5):490–2, May 1991.

[38] J.R. Sauer. Multi-gb/s optical computer interconnect. In *Advanced Fiber Communications Technologies*, volume 1579, pages 49–61. SPIE, September 1991.

[39] D.J. Blumenthal, K.Y. Chen, J. Ma, R.J. Feuerstein, and others. Demonstration of a deflection routing 2*2 photonic switch for computer interconnects. *IEEE Photonics Technology Letters*, 4(2):169–73, Feb. 1992.

[40] P. Healey. Minimising crosspoints and spatial highways in multidimensional bus networks. *Electronics Letters*, 25(22):1515–17, 26 Oct. 1989.

[41] D.J. Blumenthal and J.R. Sauer. Multiwavelength information processing in gigabit photonic switching networks. In *Multigigabit Fiber Communications*, volume 1787, pages 43–54. SPIE, September 1992.

[42] S. Banerjee and B. Mukherjee. FairNet: a WDM-based multiple channel lightwave network with adaptive and fair scheduling policy. *Journal of Lightwave Technology*, 11(5-6):1104–12, May-June 1993.

[43] U. Krackhardt, F. Sauer, W. Stork, and N. Streibl. Concept for an optical bus-type interconnection network. *Applied Optics*, 31(11):1730–4, 10 April 1992.

[44] Y. Birk. Fiber-optic bus-oriented single-hop interconnections among multi-transceiver stations. *Journal of Lightwave Technology*, 9(12):1657–64, Dec. 1991.

[45] Y. Birk. Fiber-optic bus-oriented single-hop interconnections among multi-transceiver stations. In *IEEE INFOCOM '92: Conference on Computer Communications*, volume 3, pages 2358–67. IEEE, May 1992.

[46] O. Stucky, R.L. Shoemaker, R. Manner, and P.H. Bartels. Optical interconnection for multiprocessor computer bus systems. *Optical Engineering*, 28(11):1185–92, Nov. 1989.

[47] W.G. Briscoe. Fiber optic data bus networks. In *SOUTHEASTCON '89 Energy and Information Technologies in the Southeast*, volume 3, pages 1140–4. IEEE, April 1989.

[48] A J. Reedy and J. R. Jones. Methods of collision detection in fiber optic csma/cd networks. *IEEE Journal on Selected Areas of Communications*, SAC-3(6):890–896, November 1985.

[49] E. G. Rawson and R. M Metcalfe. Fibernet: Multimode optical fibers for local comptuter networks. *IEEE Transactions on Communications*, July, 1978.

[50] R. Kelly, J. Jones, V. Bhatt, and P. Pate. Transceiver design and implmentation experience in an ethernet-compatable fiber optic local area network. In *INFOCOM 84*, 1984.

[51] R. Schmidt, E. G. Rawson, R. Norton, S. Jackson, and M. Bailey. Fibernet II: A fiber optic ethernet. *IEEE Journal on Selected Areas in Communications*, November, 1983.

[52] J.A. Lageman and J.A. Pence. Performance analysis of a CSMA/CD bus based multiprocessor system. In *MILCOM 89*, volume 3, pages 799–805. IEEE, October 1989.

[53] Chong Ho Yoon and Chong Kwan Un. Unslotted CSMA-CD protocols with combined retransmission strategy for fiber optic bus and ring networks. *Computer Networks and ISDN Systems*, 21(5):381–97, July 1991.

[54] F. Tobagi, F. Borgonovo, and L. Fratta. Expressnet: A high performance integrated services local area network. *IEEE Journal on Selected Areas of Communications*, SAC1(5), November 1983.

[55] F. Tobagi and M. Fine. Performance of unidirectional broadcast local area networks: Expressnet and fastnet. *IEEE Journal on Selected Areas in Communications*, SAC-1(5), November 1983.

[56] M. Nassehi, F. Tobaji, and M. Marhic. Fiber optic configurations for loacal area networks. *IEEE Journal on Selected Areas in Communications*, November, 1985.

[57] C. Yeh, M. Lin, M. Gerla, and P. Rodrigues. Rato-net: a random-access protocol for unidirectional ultra-high-speed optical fiber network. *Journal of Lightwave Technology*, 8(1):78–89, Jan. 1990.

[58] H.B. Jeon, B.C. Shin, and C.K. Un. Probabilistic reservation protocol for high-speed unidirectional bus networks. *Computer Communications*, 16(3):140–6, March 1993.

[59] M. Kovacevic and M. Gerla. Rooted routing in linear lightwave networks. In *IEEE INFOCOM '92*, volume 1, pages 39–48, May 1992.

[60] E. K. Thurber. *The LOCALNetter Designer's Handbook*. Architecture Technology Corporation, November 1985.

[61] S. Joshi. High performance networks: A focus on the fiber distributed data interface (FDDI) standard. *IEEE Micro*, June, 1986.

[62] F. Schaffa, M. Willebeek-LeMair, B. Patel, and M. Gerla. A demand driven access protocol for high speed networks. In *Proceedings of the Third Workshop on Future Trends of Distributed Computing Systems*, pages 158–64. IEEE, April 1992.

[63] M. Gerla, P. Camarda, and G. Chiaretti. Fault tolerant pon topologies. In *IEEE INFOCOM '92*, volume 1, pages 49–56. IEEE, May 1992.

[64] M. M. Bidnurkar, S. P. Levitan, R. Melhem, and D. M. Chiarulli. Model of lossless bus structure using erbium fiber amplifiers pumped near 820nm. In *Optical Computing Technical Digest*. Optical Society of America, March 15-19 1993. poster.

[65] D. Sarrazin, H. Jordan, and V. Heuring. Digital fiber optic delay line memory. In *Digital Optical Computing II*, volume 1215, Los Angles, CA, January 1990. SPIE.

[66] Gould Electronics, Glenn Burnie,MD. *Gould Fiber Optics Technical Notes*.

[67] F.C. Allard. *Fiber Optics Handbook For Engineers and Scientists*. McGraw–Hill, 1990.

[68] P.R. Prucnal, E.E. Harstead, and S.D. Elby. Low-loss, high-impedance integrated fiber-optic tap. *Optical Engineering*, 29(9):1136–42, Sept. 1990.