AD-A258 692

ATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	Aug. 31, 1992	FINAL 7/89 - 7/92

4. TITLE AND SUBTITLE

Coincident Pulse Techniques for Hybrid Electronic Electronic Optical Computer Systems

5. FUNDING NUMBERS

G AFOSR-89-0469

(2)

6. AUTHOR(S)

D.M. Chiarulli, R.G. Melhem & S.P. Levitan

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Departments of Computer Science and
Electrical Engineering
University of Pittsburgh
Pittsburgh, PA 15260

8. PERFORMING ORGANIZATION REPORT NUMBER

AFOSR-TR- 10 25

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Office of Scientific Research
Electronic & Material Sciences Dir.
AFOSR/NE Building 410 Dr. Alan Craig
Bolling Air Force Base, Washington DC 20332

Craig

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

2305 B1

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Unrestricted

12b. DISTRIBUTION CODE

DTIC ELECTE DEC 2 9 1992 S A D

13. ABSTRACT *(Maximum 200 words)*

This research is an investigation of the application of coincident pulse techniques to multiprocessor interconnection networks. The research focuses on three main areas: an examination of the applicability of coincident pulse techniques and required hardware to multiprocessor applications, an investigation of the limits of scalability, and an exploration of various interconnection structures which can be created using these techniques.

14. SUBJECT TERMS

Electro/optical Systems
Optical Computing

15. NUMBER OF PAGES

120

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	

Final Technical Report
For:

The Air Force Office of Scientific Research
Electronic and Material Sciences Directorate

Coincident Pulse Techniques for Hybrid Electronic Optical Computer Systems

Grant Number: AFOSR-89-0469

by:

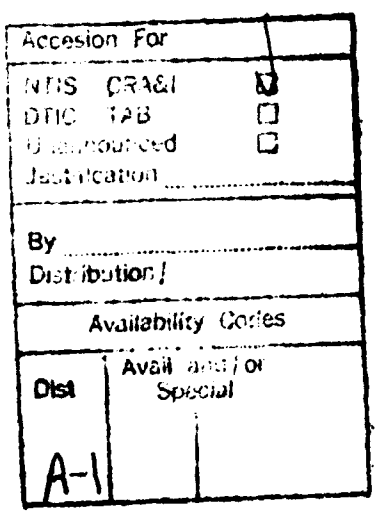
Donald M. Chiarulli
Department of
Computer Science
(412) 624-8839
don@cs.pitt.edu

Rami G. Melhem
Department of
Computer Science
(412) 624-8426
melhem@cs.pitt.edu

Steven P. Levitan
Department of
Electrical Engineering
(412) 648-9663
levitan@ee.pitt.edu

University of Pittsburgh
Pittsburgh, PA 15260

Contents



1 Project Summary

This research is an investigation of the application of coincident pulse techniques to multiprocessor interconnection networks. The research focuses on three main areas: an examination of the applicability of coincident pulse techniques and required hardware to multiprocessor applications, an investigation of the limits of scalability, and an exploration of various interconnection structures which can be created using these techniques.

2 Project Objectives

Specific objectives of this research are:

- The determination of the limits placed by current technology on implementations of coincident pulse structures. These limits include, pulse width, detection limits, degree of overlap for coincidence, power distribution and pulse synchronization.
- The study of specific structures which are capable of supporting simulcasting and multicasting communications and the comparison of these structures with functionally comparable electronic systems.
- The resolution of specific configuration issues related to clock distribution mechanisms for latching data in the simulcasting structure.
- The characterization of the tradeoff between complexity and latency for choosing the number of dimensions appropriate for a particular coincident structure.
- The study of techniques for error detection and recovery in two dimensional structures.

3 Project Status

As noted in the following sections, the specific objectives listed above were met, in particular:

- An investigation of the limits placed by current technology on the coincident pulse technique.
- Power and distribution issues for specific bus configurations.
- Investigations of specific configurations for multicasting and simulcasting.
- The identification of the limits to scalability for these systems.
- The quantification and resolution of the "shadow problem" in linear and multi-dimensional structures.

- The generalization of the coincident structures to pipelined bus structures and the analysis of inherent advantages of pipelined communication structures for both optical and electronic interconnections.

Further, Several of the specific objectives have been modified based on our research. In particular we pursued research on the following:

- Active amplification for tapped bus structures. Therefore we investigated the use of non-linear (Erbium doped) fiber to create a "lossless" tapped bus structure.
- The application of the signal pipelining results to reconfigurable time/space division multiplexed structures
- The generalization of of our work in TDM structures to more general reconfigurable optical interconnection networks
- The identification of bandwidth as a virtual resource which can be allocated dynamically on a variety of networks.
- The use of locality in source-destination address pairs to provide a mechanism of providing a dynamic reconfiguration mechanism which provides channels at optical message speed, while optimizing resources at computer algorithm speeds.

The next sections summarize the contributions of this project. The most recent results of this work has not yet appeared in print. The results of earlier work is reported in the papers given in the Appendix of this report.

3.1 Demonstration of an All Optical Addressing Circuit

A demonstration is presented of both single and parallel selection in a one of four addressing circuit using coincident pulse addressing. Scalability issues of synchronization and power distribution are also addressed.

Introduction

This experiment is based on two properties of optical signals, unidirectional propagation and predictable path delay. Using these properties, logic systems can be devised in which information is encoded as the relative timing of two optical signals. Coincident pulse addressing is an example of such a system. In this case, the address of a detector is encoded as the delay between two optical pulses which traverse independent optical paths to a 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, see both pulses independently, separated in time.

Stated more formally, consider a fiber of length L with two optical pulse sources, P_1 and P_2 coupled to each end. Each source generates pulses of width τ and height h. Define $l = \tau c_g$ 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 2×2 passive couplers, n detectors, labeled D_0 through D_n, are placed in the fiber with the two tap fibers from each coupler cut to equal length and joined at the detector site. The location of each coupler/detector is carefully measured so that the kth detector is located at $(L-nl)/2+(k-1)l$. The optical bus in the center of figure 1 shows such an arrangement for $n = 4$. To uniquely address any detector, a specific delay between the pulses generated by P_1 and P_2 is chosen. If this delay corresponds to $t_1 - t_2$, then when $t_1 - t_2 = [n - 1 - 2(k-1)]\tau$ the two pulses will be coincident at detector D_k

The same technique can be generalized to support parallel selections. If one of the sources is allowed to generate a series of pulses with each t_k timed relative to t_1 to select a specific detector k, then according to the addressing equation t_k will be in the range $-(n-1)\tau \leq t_1 - t_k \leq (n-1)\tau$, *for* $k = 1..n$. In other words, any or all of the k 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 in separated by 2τ it follows that the system latency, $\sigma = 2n\tau$. Since up to n locations may be selected in parallel within a single latency period, the system throughput is thus $\nu = 1/2\tau$. For a more complete discussion of the general application of coincident pulse techniques see, [CML87] and [LCM90].

Experimental Results

Figure 1 is a diagram of the prototype structure. The fiber bus consists of a length of multimode fiber tapped four 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.

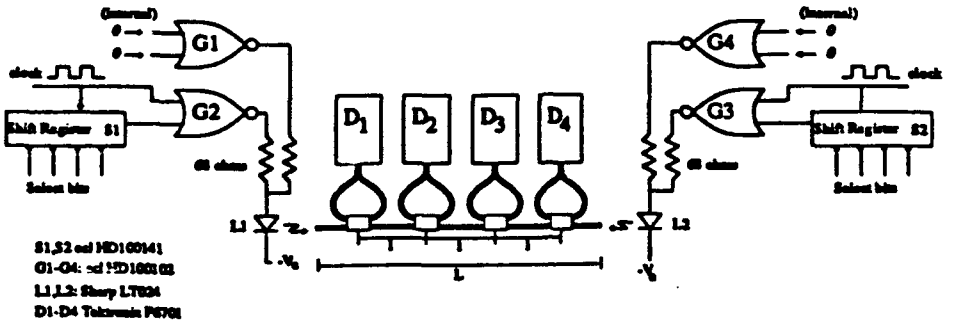


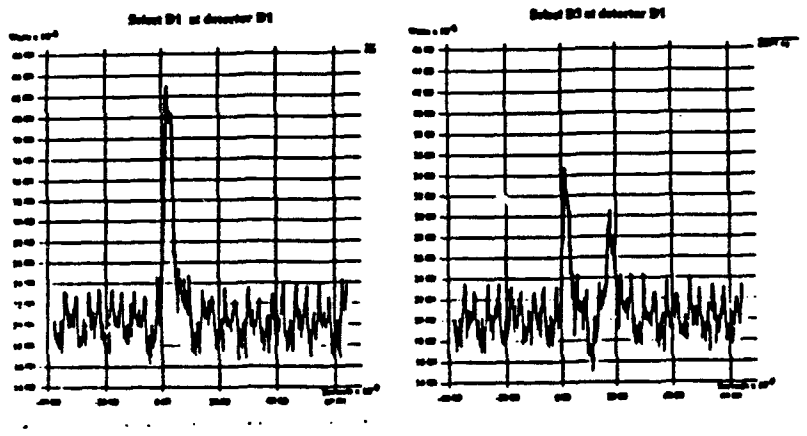
Figure 1: Experimental Setup

Figures 2 and 3 show the output waveforms for detectors D1 and D3 for various selection patterns. Figures 2a and 2b show coincident and non-coincident waveforms at detectors D1 and D3 respectively. Note that in both cases, the non-coincident waveforms shown on the right are of unequal power. This is due to the fact that each pulse has passed through a different number of couplers and has hence become attenuated to different levels. Thus the relative power between coincident and non-coincident pulses is a function of the detector location. The amount of additional power in the coincident pulse relative to the largest non-coincident pulse is called the power margin, m, and is defined as a fraction of the maximum non-coincident pulse power by $m = [p_1 + p_2 - max(p_1, p_2)]/max(p_1, p_2)$. For both of the single selection experiments shown in figures 2a and 2b, the power margin is in excess of 0.5. This is true even for D_1 which is leftmost on the bus.

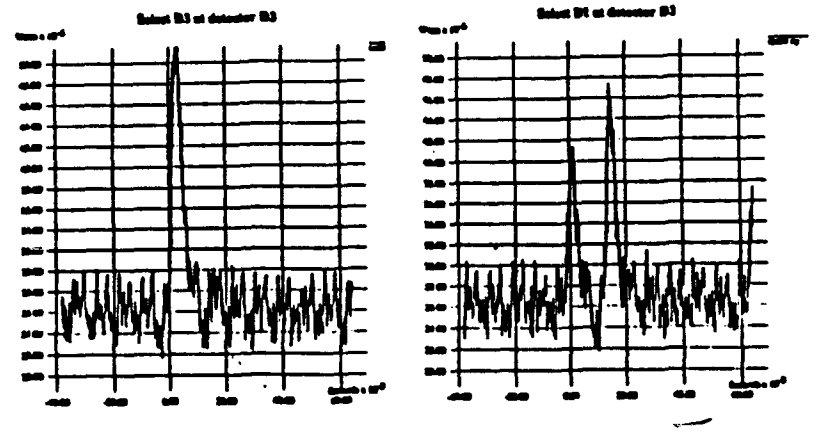
Figures 3a and 3b are examples of parallel selections. The left waveform in figure 3a shows a parallel selection waveform at detector site D_3 for the selection of three detectors, including D_3. This coincident waveform peak compares to the non-coincident waveform on the right in which D_3 has been removed from the set of selected locations. Similarly figure 3b shows parallel selection of all four detectors at sites D_1 and D_3.

Pulse Synchronization

In a second experiment, measurements were made to characterize the effect of synchronization error between the reference and select pulses on the power margin of the coincident pulse. Since clearly



(a)



(b)

Figure 2: Single Select Waveforms

this error is 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. The apparatus used was identical to the previous experiment except that the number of detectors was reduced from four to three. This allowed detector D_2 to be located in the center of the bus resulting in exactly equal non-coincident pulse heights as shown in figure 4a. 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 bus. Length was added first on the reference pulse end of the bus, and then on the select pulse end of the bus.

Figure 4b shows the reduction factor, f, of the power margin as a function of percent synchronization error. Percent synchronization error is the error, in time units, 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 is, by definition, the power margin. Synchronization error in either the select pulse, shown as positive error, or the reference pulse, shown as negative error, reduces the power margin by the factors shown. The solid line in figure 4b is the experimental result. The dotted line is a simulated result generated from the coincidence of

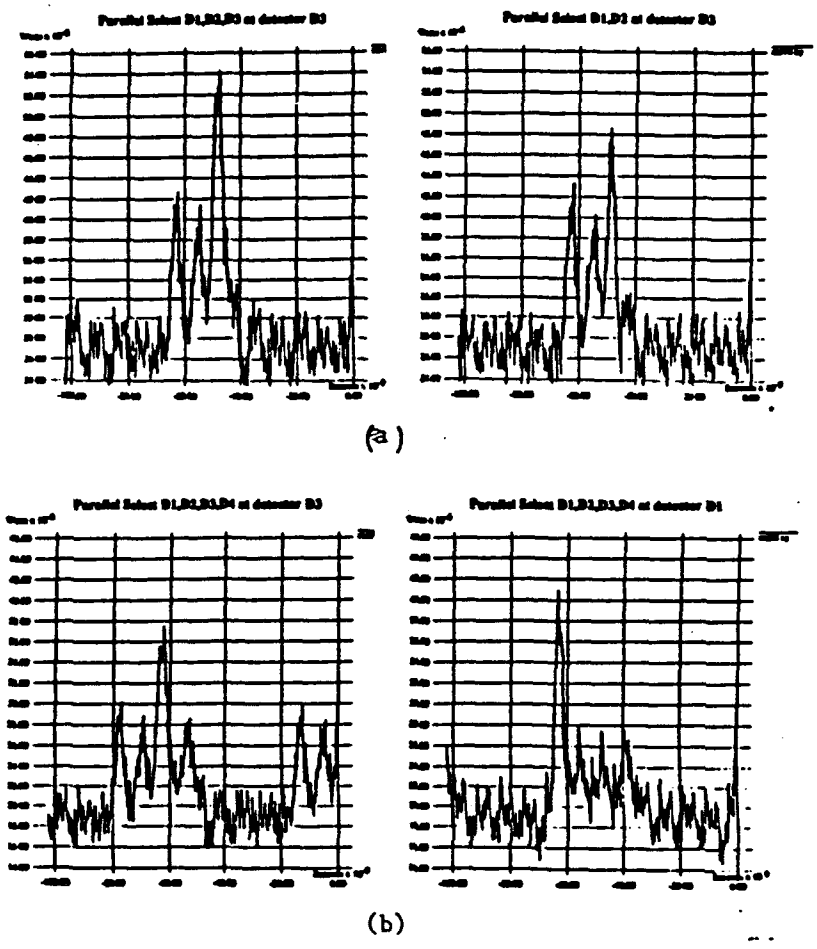


Figure 3: **Parallel Selection Waveforms**

sinusoidal pulse waveforms. In both cases power margin falls off in roughly the shape of the coincident waveforms. Thus the "flatness" of the experimental pulses results in a flattening of the power margin curve, while the sinusoids fall off somewhat more smoothly. These waveforms characterize the temporal limits on scalability. That is to say, the limit on pulse width, latency, and throughput.

Power Distribution

Since the bus configuration chosen for this experiment requires bidirectional propagation, we are constrained to use a single tapping ratio, r, for all couplers. Therefore, assuming a unit height pulse from each direction, the optical power p_1 and p_2 at detector D_k are given by the equations

$$p_1 = r^{(k-1)}(1-r), \quad p_2 = r^{(n-k)}(1-r)$$

Since the absolute power falls off geometrically with increasing n, and power margin essentially bounds scalability, the size of the system is highly sensitive to the value of r. In figure 5 we have

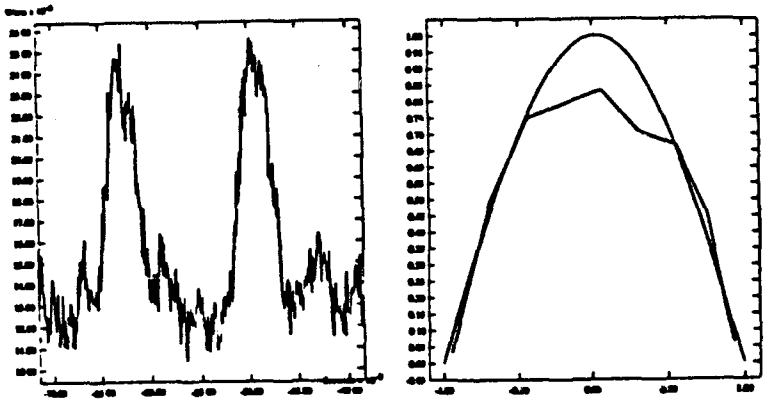


Figure 4: Pulse Synchronization Experiment

plotted worst case power margin versus coupling ratio for various bus sizes n. To determine an overall bound on system scale, the effects of synchronization error and power distribution limits must be considered jointly. The following procedure can be used. First, a minimum power margin m_d is selected such that a reasonable threshold can be established based on signal to noise ratio. Next, synchronization error, based on the pulse width and the accuracy of the fiber lengths, is used to determine worst case reduction in power margin, f. The actual power is calculated as m_d/f. Finally, the maximum number of detector sites can be determined based on figure 4 and the power equations above.

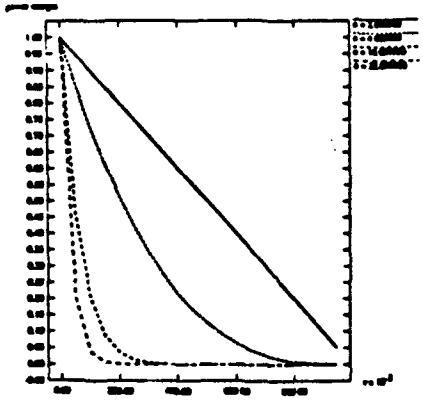


Figure 5: Power Margin vs Coupling Ratio

Discussion

Clearly, three factors, threshold power margin, synchronization error, and coupling ratio determine system scale. Based on current and near term technology, our experiments show that synchronization error does not contribute significantly to the bounds calculated above. Rather, power distribution effects dominate. However, we believe that near term technologies such as fiber amplifiers as well as alternate bus structures will alleviate this problem, as discussed below. The fact that temporal scalability limits show that significantly shorter pulses can be supported, is very encouraging for the long term application of this technique.

3.2 An Analysis of Power Distribution in Optical Buses

In this section, we present an analysis of power distribution in two tapped fiber network structures.

Introduction

Fiber optic interconnections can be generally classified as either point to point or tapped fiber bus systems. Point to point systems, those with exactly one transmitter and one receiver, are widely used in commercial telecommunications systems. Tapped fiber busses, those with one or more transmitter and multiple receivers, have been less widely adopted primarily because of scalability limits based on power distribution [NTM85]. However, the recent development of low ratio passive couplers [Gou] and the prospect of fiber based optical amplifiers [GDT+89, LFR+89] suggest a closer examination of the power distribution problem. In this paper we present an analysis of power distribution in each of two tapped fiber network structures. The first is a simple linear structure with a single backbone and a series of passive coupler taps. The second is a dual level structure which consists of a backbone fiber and a series of secondary distribution fibers from which power is tapped. The result of this analysis is a simple relationship for the number of taps supportable given a minimum detector power and coupling ratios for each tap.

In this analysis, we assume passive, bidirectional, 2x2, symmetric fiber couplers [Gou, All90]. Since the couplers are bidirectional, we arbitrarily let A, B be the inputs ports and α, β be the output ports. Equation (1) shows power distribution from the input to the output

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

where r is the coupling ratio. To find a upper bound on the maximum number of couplers, we assume the couplers are ideal (i.e., the performance is not degraded by excess loss).

The Linear Model

As seen in figure 6, a linear bus consists of n detectors (and n couplers). For this example, the signal originates on the left from a single transmitter and propagates to the right. We will assume that we are given only one type of coupler with a ratio of r.

For this model, the power available at any detector i is given by

$$p_i = r^{i-1}(1-r) \tag{2}$$

where p_i is the power at site i and q is the coupling ratio. The upper bound on the number of detectors ,n, is determined by the sensitivity of the last detector on the bus. If the last detector has a sensitivity $Pmin$, then the number of detectors supportable is

$$N = \frac{log(\frac{Pmin}{1-r})}{log(r)} + 1 \tag{3}$$

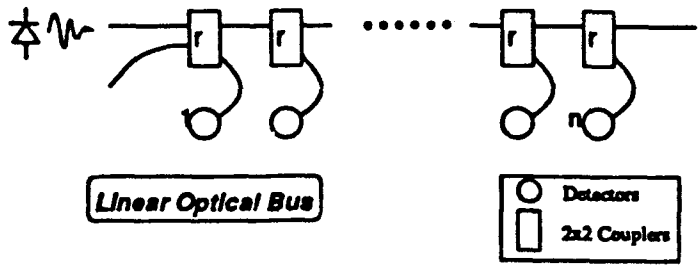


Figure 6: Linear Optical Bus

where N is the maximum number of detectors, $Pmin$ is a normalize minimum detectable power, and r is the coupling ratio. Equation (3) is shown graphically in figure 7 for a set of coupling ratios r. 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 sites.

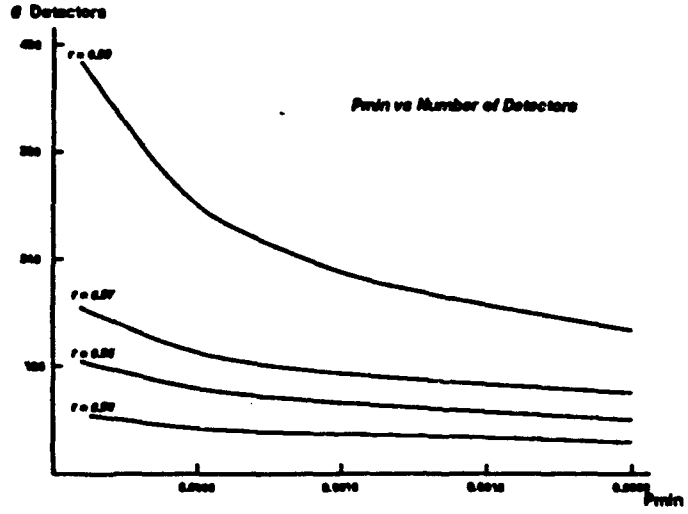


Figure 7: Pmin vs. Number of Detectors

The Dual-level Model

The problem with the linear model is that as the signal propagates down the bus, it becomes multiplictively weaker as each coupler diverts a percentage of the power. Therefore, detectors at the start of the bus use more power than needed and latter detectors are starved. One way to try to correct this problem would be to relax the requirement of fixed ratio taps in favor of varying the coupling ratios. However to achieve any significant gains, the number of distinct, precisely tuned

couplers needed must approach the number of detector sites. No couplers exist which would allow tuning to a precision of more than one or two percent. In addition, the use of tuned couplers forces the network to be uni-directional since coupling ratios must decrease in the direction of propagation.

An alternative method that does not require multiple coupling ratios is to adopt a dual-level bus structure. As shown in figure 8, we split the bus into a main fiber and a sub-level to create a section of the bus, labeled m. The sub-level contains m detectors in a linear arrangement except for the last detector which feeds back the remaining power into the main fiber and next section. In the main fiber, care must be taken to assure that the distance propagated is the same as the sub-section so that the signal arrives together at the next section. The dual-level bus consists of a series of these sections. The input is from the left (into the upper leg of the first coupler) and propagates to the right. The detectors are numbered linearly in the direction of propagation.

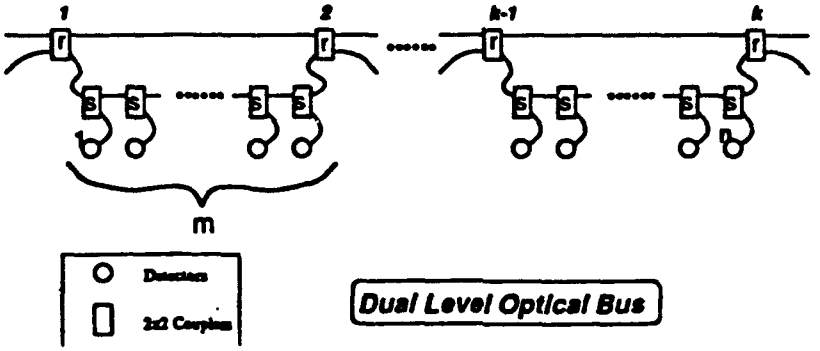


Figure 8: **Dual Level Optical Bus**

We assume that we have two types of couplers with splitting ratios of r and s for the main level and sub-level, respectively. The power at any given detector site in figure 8 is given by

$$p_i = (1-r \quad r)A^k \begin{pmatrix} 1 \\ 0 \end{pmatrix} s^l(1-s) \tag{4}$$

where p_i is the power at site i, r and s are coupling ratios, k is $i \; div \; m$, l is $i \; mod \; m$, m is the number of detectors in a sub-level, and

$$A = \begin{pmatrix} r & 1-r \\ (1-r)s^m & rs^m \end{pmatrix}. \tag{5}$$

From linear algebra we know that an equations of the form $u_k = A^k u_o$ can be rewritten as $u_k = c_i \lambda_i{}^k x_i$ where λ_i represent the eigenvalues of matrix A, the x_i's are the associated eigenvectors and c_i is determined from the initial condition u_o. For our analysis, we rewrite the matrix of equation (4) in the form

$$A^k \begin{pmatrix} 1 \\ 0 \end{pmatrix} = c_1 \lambda_1{}^k x_1 + c_2 \lambda_2{}^k x_2 \tag{6}$$

A more detailed analysis of our model shows as k increases the $c_1 \lambda_1{}^k x_1$ term in equation (6) quickly dominates. Therefore a good approximation is given by

$$p_i = (1-r \quad r)c_1 \lambda_1{}^k x_1 s^l(1-s) \tag{7}$$

In the linear model, we examined the bounds for the minimum power needed at the last detector. For the dual level model, we will examine the minimum power seen at the last detector of the last complete section. This minimum power is given by the equation

$$Pmin = {\lambda_1}^k(x_1(1-r)+r)c_1 s^{m-1}(1-s) \tag{8}$$

where $c_1 = \frac{1}{x_1 - x_2}$. As with the linear case, the ability to support as many detectors as possible is largely dependent upon maximizing the values of r and s. In order to support as many detectors as possible λ_1 must be maximized. λ_1 is a function of r,s and m. As with the linear case, maximizing the coupling ratios r and s will increase the maximum number of detectors. However, for a given r and s the optimal number of detectors for the system is determined by the selection of m. This relationship is shown in figure 9.

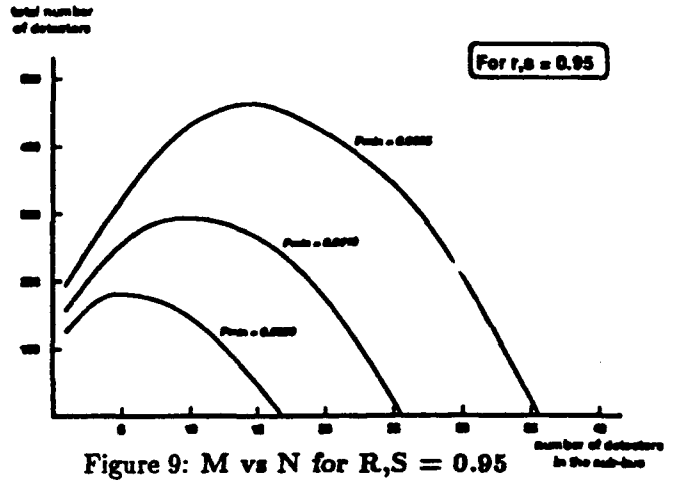


Figure 9: M vs N for R,S = 0.95

Having chosen the values for r,s and m, we can now rewrite equation (8) to compute the number of detectors supportable as a function of $Pmin$.

$$N = m\frac{log(\frac{Pmin}{[(1-r)x_1+r]c_1 s^{m-1}(1-s)})}{log(\lambda)} \tag{9}$$

A plot of equation (9) gives figure 10.

Equation (9) allows a direct comparison of the dual-level bus performance shown in figure 10 with linear bus performance derived in equation (3) and plotted in figure 7.

Figure 11 is a plot of both equations for equal coupling ratios and with m optimized in the dual level case. From this comparison, we can see that the optimized dual level bus gives approximately a factor of 10 improvement over the simple linear configuration.

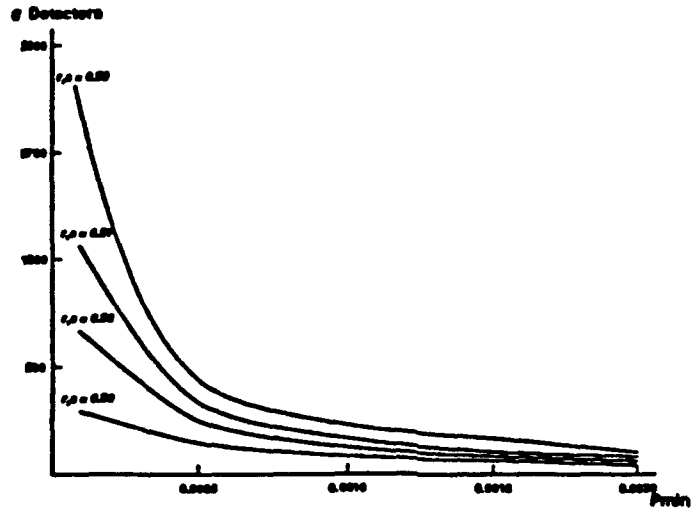


Figure 10: Pmin vs N for Different Values of R,S

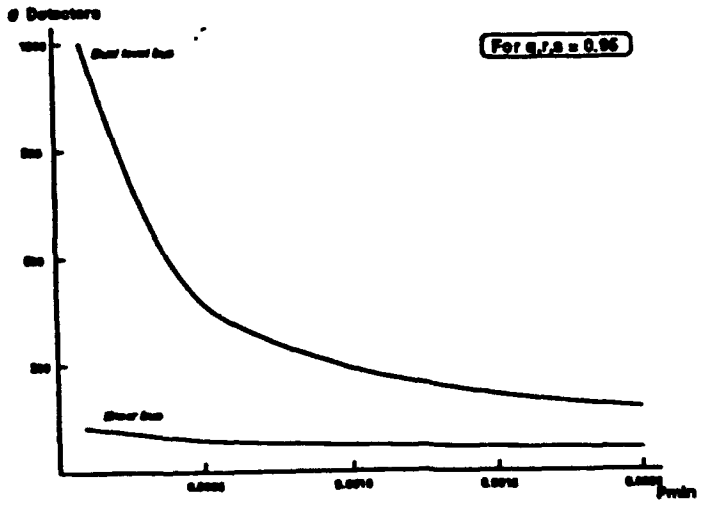


Figure 11: Pmin vs N for Linear and Dual Bus Structures

3.3 Routing Messages and Mapping Communication Structures in Array Processors with Pipelined Optical Busses

Array Processors with Pipelined Busses are a new hybrid optical-electronic parallel computer architecture utilizing optical bus interconnections. We present results for efficient communications on this architecture.

One of the most significant roles that the state-of-the-art optics may play in high speed computation is to provide high communication bandwidth in parallel computers. A new hybrid optical-electronic parallel computer architecture, called *Array Processors with Pipelined Busses* (APPB), has recently been proposed which utilizes message-pipelined optical busses for interprocessor communications and is shown to achieve an asymptotically linear (in number of processors on the bus) improvement in communication bandwidth over conventional multiprocessor architectures with nearest neighbor or exclusive access bus interconnections [GMH+90b]

Figure 12(a) illustrates an array of N processors connected by an optical bus (waveguide), where each processor is coupled to the bus with two couplers, one for writing signals and the other for receiving signals. For analytical convenience, we let D_o be the optical distance between each pair of consecutive processors and τ be the time taken for an optical signal to traverse the optical distance D_o. To transfer a message from a processor j to processor i, ij, the sender j writes its message on the bus. After a time $(i-j)\tau$ the message will arrive at the receiver i, which then reads the message from the bus.

To facilitate our discussion, for the system in figure 12(a) we define $N\tau$ as a *bus cycle*, and correspondingly τ as a *petit cycle*. Assume that the system is synchronized such that every processor writes its message on the bus at the beginning of a bus cycle and the optical distance D_o is larger than the length of each message, then all the processors can send their messages on the bus *simultaneously*, and the messages will travel from left to right on the bus in a pipelined fashion without collision. Here by collision we mean that two messages sent by two distinct processors arrive at some point on the bus simultaneously. This is in contrast to an electronic bus, where writing access to the bus is *exclusive*.

Several addressing mechanisms can be used for transferring messages on the optical bus. In cases where the communication pattern is known to the receiver, a *wait* register in each processor may be programmed such that it indicates the time at which the processor should read its message from the bus. Similarly, a *skip* register may be used to count the number of messages, real or dummy, to be skipped before reading the right message, which drops the requirements for timing accuracy and equal distance between each pair of consecutive processors. In cases where the routing pattern is unknown, the destination address can be put in each message or otherwise use the coincident pulse techniques [LCM90] such that an addressing pulse and a reference pulse coincide at the destination processor, thereby addressing it.

Connecting all processors in a system with a linear optical bus, as was done in figure 12(a), has the disadvantage that a message transfer incurs $O(N)$ time delay in an $N-$processor system.

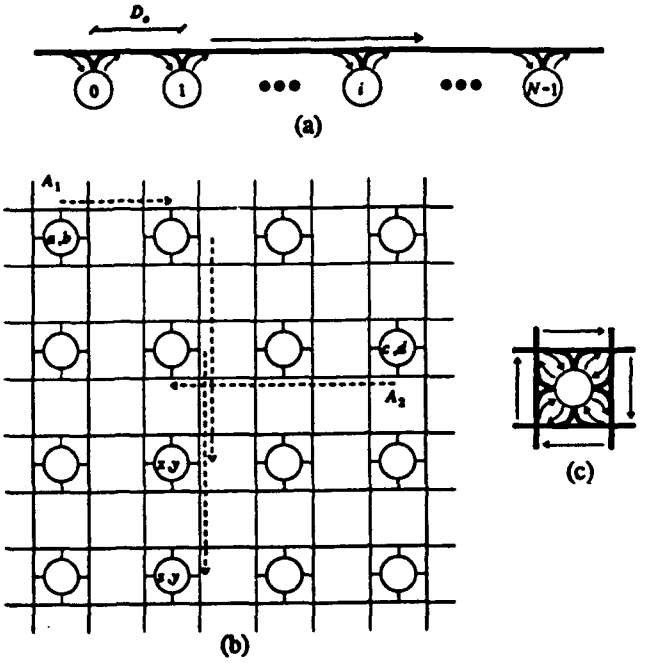


Figure 12: (a) A processor array connected with a single optical bus. (b) A schematic of the basic two-dimensional architecture of Array Processors with Pipelined Busses (APPB) where each processor is coupled to four busses as shown in (c).

Assuming $N = n \times n$, then this delay can be reduced to $O(\sqrt{N})$ by using the two-dimensional $n \times n$ APPB in figure 12(b), where each processor is connected to four busses as shown in figure 12(c).

Although this basic two-dimensional APPB architecture achieves a significant reduction in the length of a bus cycle, in general it takes at least two bus cycles, a row bus cycle ($n\tau$) and a column bus cycle ($n\tau$), for two processors to communicate with each other, while only 1 bus cycle is necessary for the same communication in the linear case. Such a two-cycle communication requires a message relay by an intermediate processor. As an example, in figure 12(b) the message A_1, which is traveling from (a, b) to (x, y), has to be sent to and buffered at processor (a, y) in the row bus cycle and then relayed to its destination in the column bus cycle. The same is true for the message A_2, which is being sent from (c, d) to (x, y) and thus has be relayed by processor (c, y). (Note that the two messages do not collide since they are buffered at the end of the row cycle and then simultaneously written on the column bus at the beginning of the column cycle.) In the worst case, e.g., when all the messages from the same row are to be sent to the same column, up to n message relays by a single processor will be needed in an $n \times n$ APPB. Further, each message relay involves an optical-electronic-optical information conversion, which reduces the communication efficiency.

There are two approaches to dealing with this disadvantage. The first is a "software" approach

which relies on designing algorithms that require only communications between two processors on the same bus. This approach has been used to obtain mappings of many well known communication structures, e.g., binary trees, hypercubes, and pyramids, onto the basic two-dimensional APPB such that any two neighboring processors in the source structure are mapped to the same (row or column) bus in the two-dimensional APPB, thus allowing them to communicate with each other using a single bus cycle without a message relay [GM90] [Guo91]. Such mappings improve the efficiency in emulating these communication structures on the two-dimensional APPB. However, they have the disadvantage that processors in the APPB may not be fully utilized, resulting in a higher expansion cost, defined as the ratio of the number of processors in the two-dimensional APPB to that in the source structure. Therefore, an efficient mapping algorithm for the basic two-dimensional APPB should use few or no relays, and have a low expansion cost. These are conflicting requirements and trade offs often have to be made between them.

The second is a hardware approach in which we propose an architectural variation of the basic two-dimensional APPB, called APPB with switches. In APPB with switches, some optical switches, e.g., Ti:$LiNbO_3$ switches, are used at each processor, as shown in figure 13(a), to switch an optical signal from a row (column) bus to a column (row) bus, thus eliminating the optical-electronic-optical conversion in the basic APPB. Each switch may assume one of the two states *straight* and *cross* as defined in figure 13(b). Initially all the switches are in state *straight*, which corresponds to the case where there is no message switching. That is, the APPB with straight switches is identical to the basic APPB architecture. When message switching is desired at some processor, a switch at that processor must be set to the *cross* state.

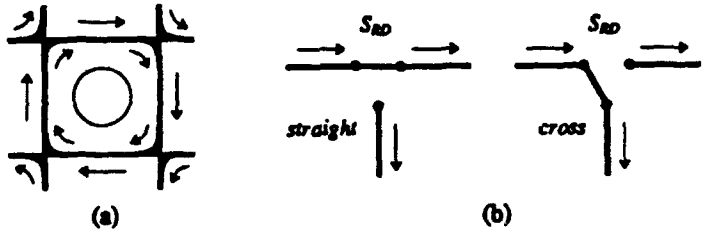


Figure 13: (a) Switch connections at each processor in APPB with switches. (b) Definition of switch states.

Note that at each intersection of a row bus and a column bus, there may be four modes for switch settings, as shown in figure 14(a). Only the first two modes are useful for message transfer since the other two may cause message collisions. As a result, one Ti:$LiNbO_3$ switch is sufficient for the switching implementation at each intersection, as shown in figure 14(b). In an $m \times n$ APPB with switches, a bus cycle is defined as $(m + n)\tau$.

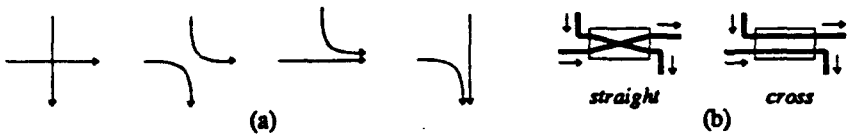


Figure 14: (a) Four possible switching modes at a intersection of a row bus and a column bus in APPB with switches; Only the first two are useful for message transfer since the last two may cause message collisions. (b) Implementation of the two useful switching modes using one Ti:$LiNbO_3$ switch.

In the basic APPB architecture, messages will not collide as long as every processor starts sending its message simultaneously and D_o is larger than the message length. This, however, is not sufficient for APPB with switches. A sufficient and necessary condition is established in the following Lemma:

> *Lemma*: Assume that the optical distance D_o between two consecutive processors is larger than the message length and that all processors start sending their messages simultaneously. Then, two messages A_1 and A_2 sent by two distinct processors (a, b) and (c, d), respectively, passing processor (x, y) on the same bus will not collide iff
>
> $$|a - x| + |b - y| \,!= |c - x| + |d - y|$$

For example, if switches are provided in figure 12(b), then the two messages A_1 and A_2 will collide at the downward bus at processor (x, y). Note that, as mentioned previously, the two messages do not collide in the basic APPB. Thus when designing algorithms for APPB with switches, extra care must be taken such that the message collision does not occur [Guo91][GMH+90b].

Many efficient interprocessor communication schemes have been obtained for the basic two-dimensional APPB architecture and APPB with switches. In Table 1, these results are summarized into two categories: fundamental message routing tasks and mappings of well known communication structures. The fundamental message routing tasks include one-to-one, broadcast, semigroup communication [CCCS90, Lev87], permutations, sorting, matrix transpose and rotation. All these routings require a constant number of bus cycles and a constant number of registers (memory) at each processor, except for the semigroup communication which uses $O(\log n)$ bus cycles (but still a constant number of registers). These message routing tasks are fundamental since they are used in many parallel computations. Their efficient accomplishment will significantly improve the efficiency of these parallel computation tasks. Mappings of other communication structures include binary trees, pyramids, hypercubes, shuffle-exchange networks, X-binary-trees [DP78], and X-quad-trees (an extension of X-binary-trees). Such efficient mappings of these well known communication structures onto the APPB architectures will allow all algorithms designed for these structures to be efficiently executed on the APPB architectures. They also allow the APPB to be logically reconfigured as the architecture which is more suitable for a given computation task.

Fundamental message routings

One-to-one -- $O(1)$ time and memory

Broadcast -- $O(1)$ time and memory

Semigroup Communication -- $O(\log n)$ time and $O(1)$ memory

Permutations -- $O(1)$ time and memory

Sorting -- $O(1)$ time and memory (to sort n data)

Matrix Transpose and Rotation -- $O(1)$ time and memory

Mappings of Communication Structures

On basic APPB, without relay:

Binary tree -- expansion cost = 1.12

Pyramid -- expansion cost < 1.23

Hypercube -- expansion cost = 1

On basic APPB, with one relay:

Binary tree, X-binary-tree -- expansion cost $\approx$ 1

Pyramid, X-quad-tree -- expansion cost $\approx$ 1

On APPB with switches, without collision:

Binary tree, X-binary-tree -- expansion cost $\approx$ 1

Pyramid, X-quad-trees -- expansion cost $\approx$ 1

Shuffle-exchange -- expansion cost = 1

Table 1. Results for routing messages and mapping communication structures in an $n \times n$ Array Processor with Pipelined Busses. Time is measured in number of bus cycles.

3.4. Reconfiguration with TDM in Multistage Interconnection Networks

In multiprocessor systems, a processor may communicate with other processors from time to time, but not all of the time. Therefore, it may be neither feasible, nor efficient to establish all connections at all times. Instead, establishments of required connections may be interleaved such that each subset of connections is alternately established for a fixed period of time called a time slot. That is, the available bandwidth of the interconnection network may be shared among these connections in a time division multiplexed way.

3.4.1. RTDM in multistage interconnection networks (MINs)

Let the set of the input ports and the set of the output ports of a $N \times N$ MIN be I and O respectively, where $I = O = \{0, 1, ..., N-1\}$. A path in the MIN between $i \in I$ and $j \in O$ is denoted by $p_{i \rightarrow j} = (i, j) \in I \times O$. Define a mapping, M, to be a set of paths that can be established at the same time without conflicts in the MIN. More specifically,

$M = \{p_{i \rightarrow j} \mid \text{all } p_{i \rightarrow j} \text{ can be established at the same time without conflicts, where } 0 \leq i, j < N-1\}$

Note that, an *admissible* (or permissible) permutation is a mapping that contains N paths. We will refer to mappings that contain less than N paths as *partial* mappings. Since establishment of two paths at the same time may cause conflicts, not every set of paths is a mapping. We refer to the establishment of all the paths in a mapping as the *realization* of the mapping.

Given a mapping, there is a way to set switches in the MIN to realize the mapping. Let $m = \frac{N}{2}$ be the number of switches per stage, and let $n = \log N$ be the number of stages in the MIN. Define a *switch setting* array to be an $m \times n$ array, whose i-th element at j-th column corresponds to the i-th switch at j-th stage in the MIN. Denote the switch setting array of a mapping M by $SS(M)$ and its elements by $SS(M)<i, j>$ where $1 \leq i \leq m$ and $1 \leq j \leq n$. The value of an element in $SS(M)$ is "0" or "1" if the corresponding switch has to be set to "straight" or "cross" to realize mapping M. The value of an element is "×" if the corresponding switch can be in any state without affecting the realization of the mapping, in which case, the mapping must be a partial mapping. Two mappings M_1 and M_2, are said to be not *compatible* with each other, if there are some i and j such that the two elements, $SS(M_1)<i, j>$ and $SS(M_2)<i, j>$, are either "0" or "1" but not equal. That is, M_1 and M_2 are not compatible if the realization of both mappings at the same time will cause conflicts in switch setting. Otherwise, M_1 and M_2 are said to be *compatible* with each other, in which case, the two mappings can be merged into one mapping, namely $M = M_1 \bigcup M_2$.

MINs under consideration in this report are generalized cube networks, which are topologically equivalent to many blocking MINs. An $N \times N$ generalized cube network has n (where $N = 2^n$) stages of 2×2 switches with cube-type connections. Figure 1 shows an example of such a MIN with $N = 8$ in which stages are numbered 1 to $n = 3$ from left to right. Each switch is assumed to have two states : straight or cross, as shown in the figure. Three switch control strategies are possible for this type of MINs. Individual switch control assumes one control signal per switch. Individual stage control assumes one control signal per stage and partial stage control assumes $i+1$ control signals in stage i. In general, individual switch control is used since it yields more powerful connectivities in a MIN.

Given a set of paths $P \subseteq I \times O$, it may not be possible to establish all paths in P at the same time without conflicts. However, P can be partitioned into several mappings, $P = M_1 \bigcup M_2 \bigcup \cdots \bigcup M_t$. Each mapping M_i, $i = 1, 2, ..., t$, may be realized for a fixed length of time, which we call a time slot. By doing so, every path in P is established once in a time slot and P is said to be realized through time-division multiplexing. Note that, the switch setting arrays for different mappings are usually different. Therefore, the MIN has to change its switch setting after each time slot.

We call a MIN a *Time-Division Multiplexed MIN* (TDM-MIN) if it repeatedly realizes a sequence of mappings in a time-division multiplexed way. More specifically, a t-way TDM-MIN changes its switch setting after each time slot to realize one of t mappings $M_1, M_2, ..., M_t$ in a round-robin fashion. Without loss of generality, we

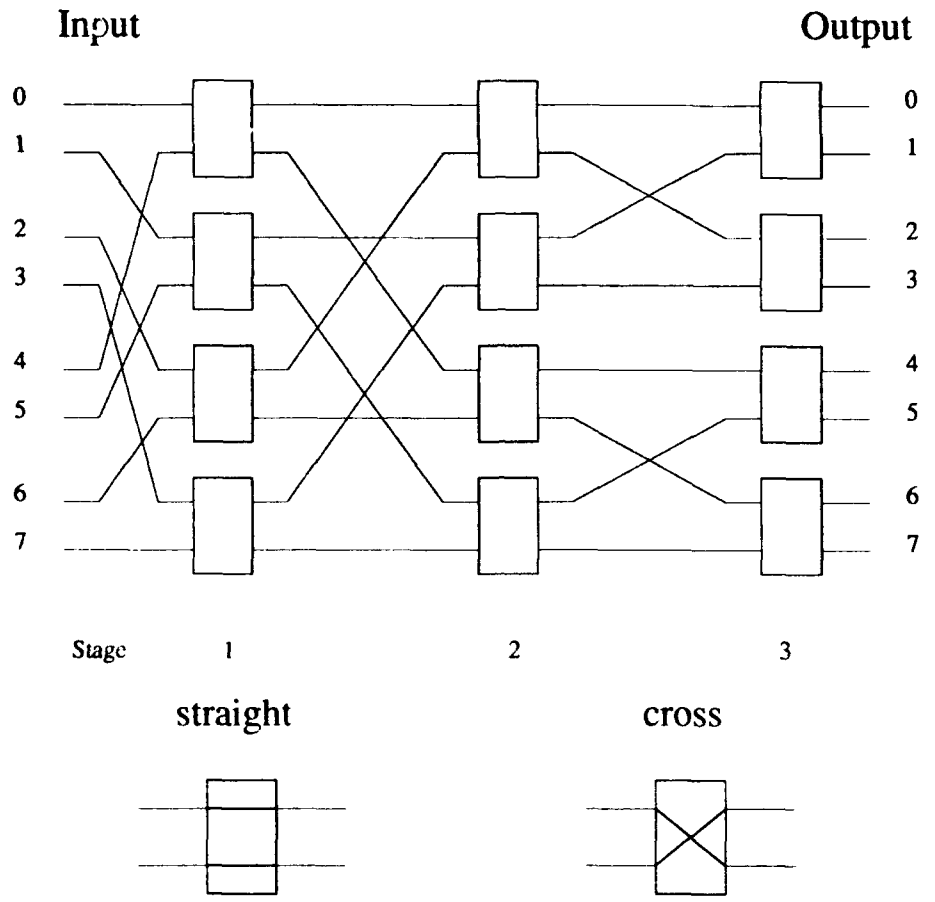


Figure 1. An 8×8 generalized cube network

assume that M_i is realized during the i-th time slot ($1 \le i \le t$). We call the ordered sequence $[M_1, M_2, ..., M_t]$ a *configuration* of the t-way TDM-MIN and the number of mappings in the sequence, t, the *Multiplexing Cycle Length* (MCL) of the configuration. Note that, this definition of configuration of a MIN is different from the conventional one that defines a configuration of a MIN as a set of numberings of input and output ports within a given MIN topology.

Once a configuration of a TDM-MIN has been determined, each mapping and its corresponding switch setting array in each time slot are determined. That is, in each time slot, the output port to which any given input port is connected is known. So is the state to which any given switch should be set. A global clock may be used to synchronize all input ports and switches in the MIN at the beginning of each time slot. Each input port maintains a list of output ports to which it is connected during different time slots of a multiplexing cycle. More specifically, the k-th entry in the list of source node i is j if $p_{i \to j} \in M_k$. Each switch is assumed to have a shift register whose size is no less than the multiplexing cycle length, t, of a configuration. The sequence of t states that a switch should be set to is stored in the shift register. The k-th bit of the shift register of the i-th switch at stage j is either "0" or "1" if the corresponding element of $SS(M_k)$ is "×". Otherwise, it should be equal to the value of its corresponding element of $SS(M_k)$.

At the beginning of each time slot, every switch is set to the state specified by the content of its shift register. After switches are set properly, an input port can transmit a message to the output port to which it is connected in

this time slot. Note that, if individual stage control is used, only one shift register per stage is required.

In MIMD environments, a MIN is commonly either circuit-switched or packet-switched or a combination of both. In circuit-switching, only a limited number of circuits can be established without conflicts. Dynamically setting up or releasing a circuit involves run time overheads. In packet switching, having one routing tag for each packet also introduces run time overheads. In addition, switches are more complex since they need to do buffering and arbitrations based on routing tags of incoming packets.

In a TDM-MIN with multiplexing cycle length t, up to tN different connections can be established. If a set of connections required by an application is known, a MIN can be set to a TDM configuration statically. This means that after execution begins, the time slot in which each connection is established is predetermined and routing decisions are as simple as waiting for the appropriate time slots. As a result, messages do not have to contain routing information such as destination addresses, nor do they need to be buffered at any intermediate switch. Overheads due to arbitration and path conflicts in circuit or packet switchings are eliminated at run time.

Even in applications that require dynamically changes of connections, the overall communication pattern is expected to change slowly during execution. Dynamic reconfiguration may affect one or more mappings but most of the other mappings will remain intact. As a result, overheads due to dynamic establishments or releases of connections are relatively lower than using circuit switching. In essence, the RTDM connection paradigm takes advantage of the relative stability of communication patterns to simplify control and to reduce overheads. On the other hand, however, the multiplexing degree in a TDM-MIN affects the latency of a connection, which must be kept low to achieve high communication efficiency.

3.4.2. Static Reconfiguration

Connection Request Graphs

Communication requirements of an application can often be obtained as a result of compile time analysis. After data allocation and processor assignment are done, memory access patterns or inter-processor communication patterns can be represented by a bipartite graph, which we call *Connection Request* (or CR) graph.

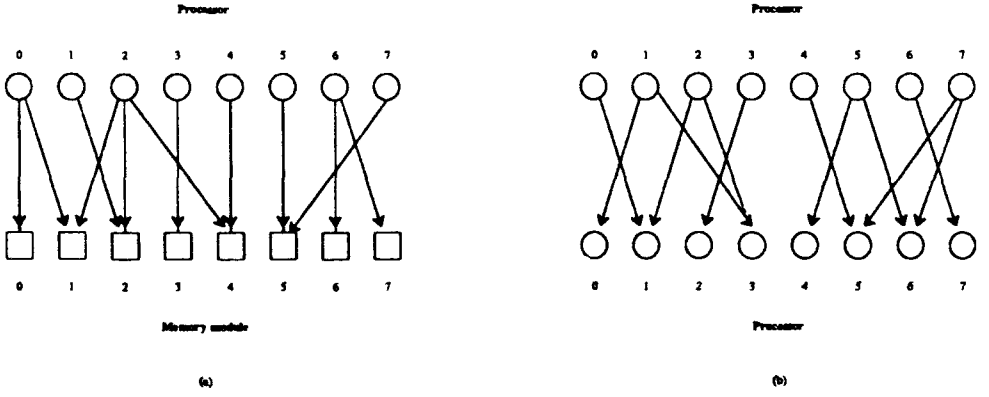


Figure 2. Examples of Connection Request (CR) Graphs

Figure 2 shows two examples of CR graphs, one for shared memory systems and another for message passing systems. Figure 2(a) shows a CR graph based on processor to memory connections. A directed edge from processor i at the top to memory module j at the bottom means that processor i may address memory module j. On the other hand, Figure 2(b) shows a CR graph based on inter-processor connections. A directed edge from a source processor i at the top to a destination processor j at the bottom means that processor i may send messages to processor j. An edge from processor i to itself is meaningless in CR graphs for interprocessor connections. Note that, due to

the dynamic nature of memory access requests in many applications, it is relatively difficult to construct CR graphs for shared memory systems.

A node in a CR graph is called either a *source node* or a *destination node*. Let source node i, $0 \le i \le N-1$, use input port i of an $N \times N$ MIN and let destination node j, $0 \le j \le N-1$, use output port j of the same MIN. Therefore, an edge from source node i to destination node j in a CR graph requires the establishment of a path $p_{i \to j}$ in the MIN. We will use the same notation for a path to denote an edge and use the terms "edge" and "path" interchangeably.

Denote the set of all edges in the CR graph by E, and the number of edges in the set by $|E|$. As an example, the set E in the CR graph in Figure 2(b), with $|E| = 12$, is given in Eq. 3.1 below.

$$E = \{(0,1), (1,0), (1,3), (2,1) (2,3), (3,2), (4,5), (5,4), (5,6), (6,7), (7,5), (7,6)\} \tag{3.1}$$

Note that, any directed or undirected communication graph can be converted into a corresponding bipartite CR graph. The number of edges going out from a node is called the *out-degree* of the node and the number of edges coming into a node is called the *in-degree* of the node. We will refer the maximum of the out-degree and the in-degree of a node as the *degree* of a node.

Given a CR graph, we call a configuration $[M_1, M_2, ..., M_t]$ a *Minimal Connection* (or MC) configuration for the CR graph if it satisfies the following two conditions.

(1). $E \subseteq \bigcup_{i=1}^{t} M_i$.

(2). for any $i, j \in \{1, 2, ..., t\}$, M_i and M_j are not compatible.

The first condition states that any edge $(i, j) \in E$ is established in a mapping M_i and the second condition states that any two mappings in the configuration can not be merged together. We call a configuration for a CR graph *optimal* if it is an MC configuration for the graph and it has the least multiplexing cycle length among all other MC configurations for the same graph. Note that, if t is equal to the maximum degrees of nodes in a CR graph, then the MC configuration is optimal.

Embeddings of Regular Communication Structures

When the communication structure of an application is regular, finding a configuration for its CR graph is often called *embedding*. Note that, the ability to embed regular communication structures efficiently is important since there are many existing applications designed for them. The multiplexing cycle length t of a configuration is a measure of the efficiency of the embedding. This measure is, in some sense, similar to the dilation cost in conventional embeddings. We also define *path utilization*(PU) to be the ratio of the number of connections required versus the number of connections that can be established in one multiplexing cycle. That is, $PU = \frac{|E|}{Nt}$.

Since there are N^2 paths between every input port and every output port in a completely connected CR graph and at most N paths can be established in each mapping, an MC configuration that embeds a completely connected network has at least N different mappings. We call a configuration $[M_1, M_2, ..., M_N]$ such that

$$\bigcup_{i=1}^{N} M_i = I \times O \tag{3.2}$$

a *completely connected* (CC) configuration since every path in a completely connected network is established in one of the N mappings of the configuration. Such an embedding is clearly an optimal one with its multiplexing cycle length $t = N$ and path utilization $PU = 1$. There are more than one CC configurations. As one example, define

$$M_{s(k)} = \{p_{i \to j} \mid j = (i + k) \bmod N \quad \text{for } i = 0, 1, ..., N-1\} \tag{3.3}$$

and call it a *shift-k* mapping. Therefore, the configuration $[M_{s(0)}, M_{s(1)}, ..., M_{s(N-1)}]$ establishes paths from any input port to all N output ports and, thus, is a CC configuration. As another example, define

$$M_{f(k)} = \{p_{i \to j} \mid j = i \text{ xor } k \quad \text{for } i = 0, 1, ..., N-1\} \tag{3.4}$$

and call it a *flip-k* mapping where *xor* is the bit-wise *Exclusive-OR* operation. The $flip-k$ mapping can be realized by individual stage control. Therefore the configuration $[M_{f(0)}, M_{f(1)}, ..., M_{f(N-1)}]$ is also a CC configuration. Note that, CC configurations are functionally equivalent in terms of their multiplexing cycle lengthes and path utilizations. However, the CC configuration with $flip-k$ mappings may be chosen for the purpose of embedding a completely connected network in the time domain due to its control simplicity. It is also worth noting that in the case of processor-to-processor interconnection, the *identity* mappings, such as $M_{s(0)}$ or $M_{f(0)}$, that establish no paths other than those from a node to itself can be deleted from the CC configurations.

Since any CR graph is a subgraph of a completely connected graph, it can be embedded in any CC configuration of a TDM-MIN. This, however, requires an N-way TDM-MIN and thus may be inefficient in terms of both the multiplexing cycle length and the path utilization. An alternative is to find an MC configuration of length $t < N$ which embeds the CR graph. In other words, a t-way TDM-MIN instead of an N-way TDM-MIN, for some $t < N$, can be used to increase the embedding efficiency. The smaller the ratio of $\frac{t}{N}$ is, the more efficient it is to use such an MC configuration. Table 1 summarizes embedding results of several regular communication structures.

Structure	Nodes	MCL	Optimal
Ring	N	2	yes
Mesh	$N=m^2$	4	yes
Hypercube	$N=2^n$	n	yes
Cube-Connected Cycle	$N=2^n$	3	yes
Complete Binary Tree	$N=2^n-1$	4	?

Table 1. Embedding Regular Structures in TDM-MINs

Static Reconfigurations Based On Arbitrary CR Graphs

For non-regular CR graphs, an MC configuration can always be obtained by selecting a subset of mappings from a CC configuration. Such an MC configuration can often improve performance of applications by reducing the multiplexing cycle length to t time slots, for some $t < N$. Note that, given a specific path, there is only one mapping in a CC configuration that establishes that path. The mapping can usually be determined by either a simple arithmetic operation or a table look-up. For example, if the CC configuration consists of flip-k mappings, the mapping that will establish the path $p_{i\rightarrow j}$ is $M_{f(k)}$ where $k = i \; xor \; j$.

Given a CR graph containing a set of edges E and a CC configuration $[M_1, M_2, ..., M_N]$, an MC configuration that establishes the edges in the CR graph can be found by using the selection algorithm below. We use the symbol [ ] to denote an empty MC configuration with no mappings and the operation | | to denote the addition of a mapping to an MC configuration.

Selection Algorithm

1. Set initially MC = [ ]
2. For each edge $p_{i\rightarrow j} \in E$ repeat
   2.1. Determine the mapping M_k such that $p_{i\rightarrow j} \in M_k$
   2.2. If $M_k \notin$ MC then MC = MC | | M_k

For example, consider the CR graph in Figure 2(b) and the CC configuration consisting of *flip-k* mappings. The MC configuration selected by the algorithm is $[M_{f(1)}, M_{f(2)}, M_{f(3)}]$. Since $t = 3$ and $N = 8$, a 3-way rather than an 8-way TDM-MIN may be used for the application to improve the efficiency. Note that, the maximum number in the set of in-degrees and out-degrees of nodes in the graph is 2 and, thus, this configuration may not be

optimal.

In fact, an optimal MC configuration with $t = 2$ will be obtained in the next section.

Probabilistic analysis of the average multiplexing cycle length of MC configurations for random CR graphs can be carried out as follows. Given a CR graph in which S out of N source nodes are each connected randomly to D out of N destination nodes, define $s = \frac{S}{N}$ and $d = \frac{D}{N}$. The probability that an edge is established in any mapping of a given CC configuration is $\frac{1}{N}$. Since edges that go out from the same source node must be established in different mappings, exactly D mappings are needed to establish paths from a source node to D destination nodes. Therefore, after selecting D mappings from the given CC configuration, the probability that any mapping has *not* been selected is $p = 1 - d$. Since each source node randomly selects D mappings independent of others, the probability that a mapping in the CC configuration remains un-selected after all S source nodes have selected their mappings is $P = p^S$. The probability that exactly i mappings have been selected for an MC configuration is thus

$$Prob(i) = \binom{N}{i} P^{N-i}(1-P)^i \tag{3.5}$$

Therefore, the average (expected) number of mappings selected, that is, the expected multiplexing cycle length of a MC configuration is

$$t_{av} = \sum_{i=d}^{N} i \times Prob(i) \tag{3.6}$$

Clearly, $t_{av} \geq S$, since S is the out-degree of a source node. The percentage of communication load of an application can be approximated by the ratio of $|E|$ versus N^2. In the above analysis, the load percentage is proportional to $s \times d$. Figure 3 shows calculated values of $\frac{t_{av}}{N}$ for different system size N under different load conditions assuming $s = d$. It can be seen that with reasonably small system size and under low load condition, the selection algorithm can generate an MC configuration that improves over a CC configuration. Note that, by using the CC configuration with $flip-k$ mappings, the selection algorithm can be simple and so does the resulting MC configuration because of individual stage control. The time complexity of the algorithm is linear in the number of connections requested, that is $O(|E|)$.

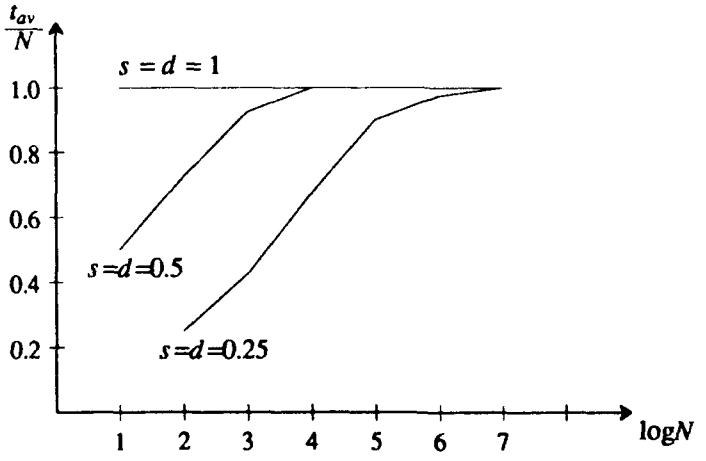


Figure 3. Percentage of mappings selected

The problem with the selection algorithm is that it restricts itself to the set of N mappings in a chosen CC configuration. With individual switch control, however, a total of $2^{mn} = N\sqrt{2^N}$ mappings (excluding partial map-

pings) are possible in the MIN. Given a CR graph with a set of edges E, a MC configuration can be obtained by composing mappings based on the set E, which may be different from any mapping in a chosen CC configuration. More specifically, assuming individual switch control, the following *composition* algorithm composes each mapping in a greedy fashion. That is, starting with an empty set of paths, a mapping is composed by including as many required paths as possible provided that they do no conflict.

The Composition Algorithm

Set MC = [ ] and $k = 1$. Repeat until E is empty

1. Reset mapping $M_k = \phi$ and all elements of $SS(M_k)$ to "×"

2. For each edge $p_{i \to j} \in E$

   If $\{p_{i \to j}\}$ is compatible with M_k

   2.1. $M_k = M_k \cup \{p_{i \to j}\}$ and update $SS(M_k)$ accordingly

   2.2. delete $p_{i \to j}$ from the set E.

3. MC = MC || M_k and $k = k + 1$

For example, an MC configuration that is composed by the algorithm for the CR graph in Figure 2(b) is $[M_1, M_2]$. For easy verifications by the readers, we will show each mapping in the MC configuration with its corresponding switch setting array in Eq. 3.7.

$$M_1 = \{(0,1), (1,0), (2,3), (3,2), (4,5), (5,4), (6,7), (7,6)\}$$
$$M_2 = \{(1,3), (2,1), (5,6), (7,5)\} \tag{3.7a}$$

Their switching setting arrays are respectively:

$$SS(M_1) = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \qquad SS(M_2) = \begin{bmatrix} 0 & 1 & 1 \\ 0 & \times & 1 \\ 1 & \times & 0 \\ 0 & 1 & 1 \end{bmatrix} \tag{3.7b}$$

In this example, the MC configuration composed by the algorithm is optimal with $t = 2$, which improves over the MC configuration generated by the selection algorithm. Simulations have been done to determine the average multiplexing cycle length of MC configurations composed by the composition algorithm. A random number generator is used to generate D distinct destination nodes for each of the S source nodes. Figure 4 shows simulation results of t_{av} for different system sizes under different load conditions where $s = \frac{S}{N}$ is equal to $d = \frac{D}{N}$. It can be seen that under low or medium load conditions, the composition algorithm improves over the selection algorithm as expected. However, when the load is extremely high, the multiplexing cycle length of an MC configuration could exceed N. That is, configurations under high load condition using this algorithm may be worse than simply using a CC configuration. Note that, $\frac{t_{av}}{N}$ does not vary much with the system size N. Figure 5 shows simulation results for a system with $N = 32$ with various s and d. Given that it takes $O(logN)$ time to compute switch settings for a path, the composition algorithm has the time complexity of $O(|E|^2 logN)$.

It is desirable, not only to do better than the selection algorithm under low or medium load conditions, but also to bound the multiplexing cycle length of any MC configuration by N under high load conditions. One way is to use the selection algorithm first to determine a set of up to N $flip$-k mappings needed. Then each mapping is examined to see if it can be deleted from the configuration by *migrating* paths established in it to other mappings in the configuration. Given a CR graph with a set E, we can use the following merge algorithm to achieve the above objective.

The Merge Algorithm

1. Run the selection algorithm to determine an MC configuration.

2. For each mapping $M_k \in$ MC, repeat step 3

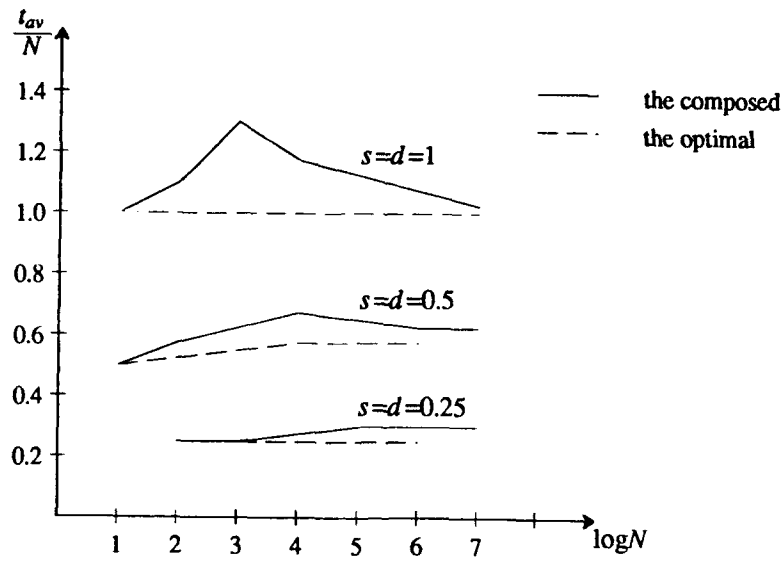


Figure 4. Percentage of mappings composed

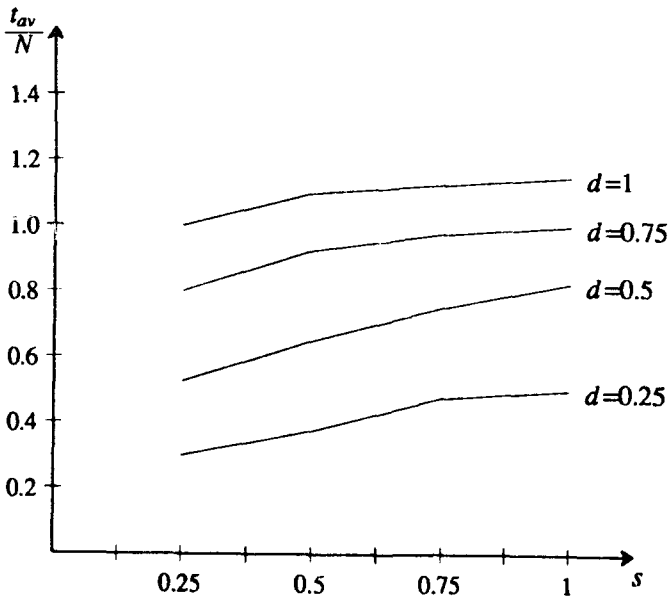


Figure 5. Different load conditions when $N = 32$

3. If every $p_{i \to j} \in M_k$ is such that $\{p_{i \to j}\}$ is compatible with a M_l currently in MC where $l \neq k$

3.1. For every $p_{i \to j} \in M_k$

$M_l = M_l \cup \{p_{i \to j}\}$ if $\{p_{i \to j}\}$ is compatible with M_l

3.2 Remove M_k from MC

Simulations have been done under similar assumptions to those used for the composition algorithm. Figure 6 shows the results of the merge algorithm for a system with $N = 32$. It can be seen that the merge algorithm

performs as good as the composition algorithm under low or medium load conditions and converges to the selection algorithm under high load conditions. The complexity of this algorithm can be shown to be $O(N \mid E \mid^2 \log N)$.

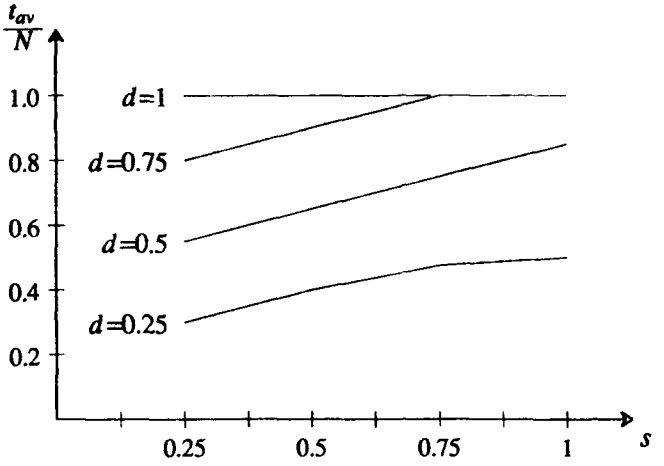


Figure 6. Performance of the merge algorithm

3.4.3. Dynamic reconfiguration

Static reconfigurations work well if all paths are required to be established from the beginning to the end of executions of an application. For applications such as those in telecommunication, connection requests are usually generated at run time. Even if a CR graph that contains all edges needed during the execution can be constructed at compile time, it may be inefficient to perform static reconfigurations based on such graph since some paths are used only for a certain duration of time and are wasted for the remaining time during execution. It is possible to achieve more efficient communication in such an application by reconfiguring a TDM-MIN dynamically based on run time requests. This means, mappings realized in a time slot may be different from time to time. We call such reconfigurations *dynamic reconfigurations*.

Centralized Reconfigurations

Run time requests may include requests for establishing new paths and releasing existing ones. Dynamic reconfigurations can be done incrementally based on an existing configuration. If a request is to establish a path $p_{i\rightarrow j}$, the current configuration is examined to find a mapping M_k that is compatible with $\{p_{i\rightarrow j}\}$. If successful, path $p_{i\rightarrow j}$ is added to mapping M_k by updating $SS(M_k)$. In the array, only elements that correspond to switches along the new path and have value of "×" are updated to either "0" or "1". Consequently, the k-th bit of each shift register of those switches whose corresponding elements have changed values may need to be updated. The source node i also sets its k-th entry in the list of output ports to j. At this time, reconfiguration based on the request is completed.

If, however, the current configuration does not contain any mapping that is compatible with $\{p_{i\rightarrow j}\}$, a new mapping that establishes $p_{i\rightarrow j}$ can be added to the configuration. This requires that all source nodes be informed of the additional time slot in the multiplexing cycle. Shift registers of all switches have to be updated accordingly. Before adding the new mapping, one can migrate existing paths in a mapping into other mappings so that it may become compatible with $\{p_{i\rightarrow j}\}$. This way, the new mapping may be avoided. Note that there are tradeoffs between overheads of migrating paths and overheads of adding a new mapping.

If a run time request is to release a path, the mapping that currently establishes the path may be deleted if all remaining paths in the mapping can be migrated into other mappings in the configuration. Such explicit release requests may not be necessary if replacement algorithms or garbage collection algorithms are used by the central

controller based on the usage of existing connections. All these involves tradeoffs. Note that, dynamic reconfiguration can also be done by buffering run time requests and periodically executing a static reconfiguration algorithm. At each selected instance, a snapshot of the CR graph is constructed based on all current paths that need to be established.

Distributed Reconfiguration

Assume that nodes connected to a MIN has distributed control but global synchronization is still applicable. In this case, the multiplexing cycle should always consist of a fixed number, k, of time slots. This is because each node will not be aware of either increment or decrement of the number of time slots in the multiplexing cycle (in a timely way) without being informed by a centralized control mechanism.

Each source node that wants to establish a path reserves a time slot in which the path may be established by routing a reservation packet to the destination on the path. These reservation packets use links and switches, called *reservation links* and *reservation switches* respectively, separate from those used by data packets. Note that the separation can be either *logical* or *physical*. An example of logical separation could be a MIN with links and switched used by reservation packets and data packets in a time-multiplexed way. In the following discussion, terms "link" and "switches" are used to refer to reservation links and reservation switches respectively.

Let an input port of a switch s be denoted by $l(s)$ and let an output port of the switch be denoted by $r(s)$. Let a path be represented by a sequence of $n = \log N$ pairs of left and right ports of switches at each stage. That is, $p_{i \to j}$ can be represented by $(<l(s_1), r(s_1)>, <l(s_2), r(s_2), ..., <l(s_n), r(s_n)>)$. Note that, this implies that $r(s_i)$ is connected to $l(s_{i+1})$. Let every output port $r(X)$ maintain a set of time slots that is not used by any paths. Denote that set by $AVAL(r(X))$. Assume that each source node also maintains an $AVAL(l(Y))$ list for an input port $l(Y)$ to which it is connected. Let "lock" and "unlock" be mutual exclusive operations on a switch port. Only the reservation packet that can successfully "lock" the port can update its $AVAL$ list while other reservation packets are buffered at the switch until the port is "unlocked". When the TDM-MIN system is started, all ports are unlocked and for any port Z, $AVAL(Z) = \{1, 2, ..., k\}$.

Each reservation packet maintains a set of time slots that are available for possibly establishing the corresponding path. Denote by $AVAL(R)$ the set of available time slots maintained by a reservation packet R. When a reservation packet R is generated, its $AVAL(R)$ is set to $AVAL(l(s_1))$. As a reservation packet goes through each switch, it locks the corresponding ports and updates its own $AVAL(R)$ to the set of time slots that are available at all switch ports visited so far. If the reservation packet reaches the destination, it chooses a time slot, namely $ts \in AVAL(R)$, and returns to the source along the same path in reverse order. As it passes each switch, it deletes the ts from the $AVAL$ lists of each port visited and unlocks these ports. At the same time, the ts-th bit of the shift register of the switch is loaded with a proper state. When it comes back to the source node, the destination to which it is sent to is recorded in the ts-th entry of the list of output ports by the source node.

Before the control packet reaches its destination, if $AVAL(R)$ would become empty at a switch, the reservation packet may be blocked. Two strategies similar to "holding" and "dropping" in circuit switching can be used when a packet is blocked. If holding is used, the packet stays in the buffer of the switch. An advantage is that whenever some paths using the same switch port are released, the reservation packet can continue its routing without repeating from the source up to that switch. However, a disadvantage is that switch ports that have been locked by the packet can not be used by other reservation packets while the packet is blocked. An alternative is to use dropping, in which the reservation packet reverses its way, unlocking switch ports and undoing changes to $AVAL$ sets of switches. The source node may queue the packet and try to send the packet again after a random interval. Note that, a combination of these two strategies, which drops a packet after holding it for a certain period, can also be used.

If a source node wants to release a path, it sends a cancellation packet $\bar{R}$ with $AVAL(\bar{R})$ containing the time slot in which the path is established. The cancellation packet can add the time slot in $AVAL(\bar{R})$ into $AVAL$ sets of every switch ports visited on the way to its destination.

3.4.4. Summary

To summarize, reconfiguration with TDM is a connection paradigm that can be applied to multiprocessor systems using multistage interconnection networks. It provides more architectural flexibilities and could achieve potentially higher communication bandwidthes than conventional switching methods. It is especially promising for optical interconnection networks because, first, high optical communication bandwidths make time-division multiplexing feasible and more attractive; Important properties of optical signal propagation, namely unidirectional propagation and predictable path delay, enable pipelined transmissions over optical waveguides. The way in which switches are set in TDM-MIN can further simplify pipelinings between stages. Second, partitioning connection requests and establishing subsets in a time division multiplexed way can simplify controls and eliminate the needs for message relaying, optical delay loops (optical time-slot interchangings) and costly conversions between optical and electronic signals. As a result, current photonic switching technology, can be readily adopted for TDM-MINs implementations.

3.5. Optical Multicasting In Linear Arrays

In this research, we use coincident pulse techniques to implement multicasting among processors connected by optical buses. To reduce addressing latency and overcome system size limits, we propose a two level addressing implementation in which multicasting introduces the problem of possibly addressing unintended processors (called *shadows*). We show how additional addressing pulses can be used to reduce these *shadows*.

3.5.1. Coincident pulse addressing

Coincident pulse techniques are based on two properties of optical pulse transmission, namely unidirectional propagation and predictable propagation delay per unit length. The technique was first introduced in the context of parallel memory addressing but was also applied to multiprocessor interconnection structures. In this research, coincident pulse techniques will be applied as an addressing mechanism for multicasting among optical bus connected processors.

As an introduction to using coincident pulse techniques as an addressing mechanism, we discuss two models of multiprocessor systems in which processors are connected by optical buses. In the first model, called the *Folded Bus* model, each processor transmits on the lower half segment of a bus, while receiving from the upper half segment. In the second model, called the *Dual Bus* model, each processor is connected to two buses, one for downstream transmitting and upstream receiving and the other for upstream transmitting and downstream receiving.

An optical bus consists of three waveguides, one for carrying messages, one for carrying *reference pulses* and one for carrying *select pulses*, which we call the *message waveguide*, the *reference waveguide* and the *select waveguide* respectively. Messages are organized as *message frames*, which have a certain fixed length. The propagation delay on the reference waveguide is the same as that on the message waveguide but not the same as that on the select waveguide. A fixed amount of additional delay, which we show as loops (see Figure 2), are inserted onto the reference waveguide and the message waveguide.

The basic idea of using coincident pulse techniques as an addressing mechanism is as follows. Addressing of a destination processor is done by the source processor which sends a reference pulse and a select pulse with appropriate delays, so that after these two pulses propagate through their corresponding waveguides, a coincidence of the two occurs at the desired destination. The source processor also sends a message frame which propagates synchronously with the reference pulse. Whenever a processor detects a coincidence of a reference pulse and a select pulse, it reads the message frame. In essence, the address of a destination processor is unary encoded by the source processor using the relative transmission time of a reference pulse and a select pulse.

3.5.2. Two level memory addressing

Using unary addressing, an address frame is N units long. There are two reasons why we want to reduce the length of address frames by using two level addressing. One has to do with efficiency. Unary addressing could be very inefficient in a large multiprocessing system where the address frame is longer than the message frame. The other reason has to do with the physical limitation of optical path length between two adjacent processors. One way to ensure that the frames sent by one processor do not collide with other frames sent by other processors is to arbitrate the bus to allow exclusive access by one processor at a time. Another way is to pipeline the bus. That is, to synchronize all processors such that they will send messages at the beginning of each cycle. The propagation delays between two adjacent processors should be large enough to prevent frames from overlapping. If unary addressing is used, it is necessary for the optical path between any two adjacent processors to be long enough to prevent overlapping of the address frames. Although the required minimum optical path length can be reduced by shortening the pulse width w, the address frame length, which is linear in the system size, becomes a limiting factor.

A two level addressing implementation divides the whole system into logical clusters. Addressing of a single destination is accomplished by using one level of unary addressing to select a particular cluster and another level of

unary addressing to select an individual processor within the selected cluster. Two trains of select pulses are used, one for each level of addressing and their pulse trains are sent in parallel. Therefore, the length of address frames can be reduced as neither the number of clusters nor the size of any cluster is larger than the system size.

Assume that $N = n^2$ processors are linearly connected. If every n consecutive processors constitute one logical cluster, two level addressing in this linear system is logically equivalent to addressing a two dimensional array. More specifically, we can view the linear system as the result of embedding an $n \times n$ array in row major fashion. Each row of processors of the array is embedded into n consecutive processors in the linear system. Hence, selecting a logical cluster is equivalent to selecting a row while selecting an individual processor within a cluster is equivalent to selecting a column processor within a row.

3.5.3. Shadow Reduction and avoidance

Shadows are created because of the unintended couplings of a $W1$ pulse with a $W2$ pulse. One way to reduce shadows is to further identify the intended pairs by using additional select waveguides, called *check* waveguides for carrying select pulses called *check* pulses. Check pulses are arranged such that they do not coincide with the reference pulse at places where shadows were created. Only processors at which coincidences of a reference pulse and all select pulses occur are addressed. Note that having an additional check waveguide in two level addressing is different from having three level addressing. The latter would be logically equivalent to addressing a three dimensional array. That is, addressing a single destination would required three select pulses. The address frame length would be further reduced while more shadows would be likely when multicasting with three level addressing. One such set of check pulses are 45° diagonal select pulses and another are –45° diagonal select pulses.

One way to avoid shadows when multicasting to a group of processors is to partition the whole group into several subgroups such that each subgroup is multicasted within one cycle without creating any shadows. A number of subgroups is called a *maximal SF* partition if it is a *SF* partition and if multicasting to more than one of the subgroups within one cycle will create a shadow. Therefore the number of subgroups of a maximal partition is the number of cycles needed to complete the multicasting to the whole group G.

In some applications, such as finite element analyses and image processing, multicasting patterns can be quite regular. For example, a convolution of an $n \times n$ array involves multicastings of an element to its $w \times w$ neighbors, where w is the current window size. A group to be multicasted could also be all processors of a row, or of a column or of a diagonal line. By embedding a physical 2-D structure into our linear structure in the row-major fashion, these regular 2-D patterns can be characterized by a group of four parameters. More formally, in an embedded $n \times n$ system, we consider a group G of m processors starting with the processor numbered as k (called offset) with increment of d (called stride). We call a group a *dense* group if d is less than n, a *sparse* group otherwise.

While we can make tradeoffs between the number of select waveguides used and the number of cycles needed to multicast to a group of processors, we only analyze here simple cases in which only two select waveguides are used. The results may be extended to cases in which four select waveguides are used.

Definition 1. A row of processors in a logical two dimensional array is *incomplete* with regard to a group $G(k, d, m, n)$ if and only if the row contains two processors i and j such that $i \in G$, $j \notin G$ and $|j - i| = b \times d$ for some integer $b > 0$. A row is *complete* if and only if the row contains at least one processor of the group G and is not an *incomplete* one.

Definition 2. Define $I(k, d, m, n)$ to be the number of *incomplete* rows with regard to the group G.

Let the first processor of the group be $k = r_k \times n + c_k$ and the last processor of the group $l = k + (m-1) \times d = r_l \times n + c_l$ for some integers $0 \le r_k, c_k, r_l, c_l < n$. And let condition 1 be that $c_k > d$, condition 2 be that $n - c_l > d$ and condition 3 be that $r_k \neq r_l$. There will be two *incomplete* rows, namely row r_k and row r_l if and only if all three conditions are true. There will be no *incomplete* rows if and only if neither condition 1 nor condition 2 is true. Otherwise, there will be only one *incomplete* row. Therefore, I has an upper bound of 2. Noting that

for a *sparse* group $G(k, d, m, n)$ where $d \geq n$, neither condition 1 nor condition 2 is true, therefore $I = 0$.

Lemma 1. Two processors of a group $G(k, d, m, n)$ numbered as i and $j, i < j$, will be in the same column if and only if $j - i = b \times LCM(n, d)$ for some integer $b > 0$†.

Proof. Let $i = r_i \times n + c_i$, and $j = r_j \times n + c_j$ as before. On the one hand, if $c_j = c_i$, then $j - i = (r_j - r_i) \times n$ and $r_j > r_i$. Since both processors i and j are in the same group, $j - i$ must be a multiplier of d, therefore $j - i$ should be a common multiplier of both n and d.

3.5.4. Summary

Coincident pulse techniques have been applied as an efficient addressing mechanism for multicasting among multiprocessors connected by optical buses. Two basic models of a unary addressing implementation have been discussed, and a two level addressing implementation has been proposed to reduce the address frame length. Two approaches to deal with the shadow problem have been presented. One approach reduces the number of shadows by using check pulses. Another approach avoids possible shadows by constructing *SF* partitions. It has been shown that for regular multicasting patterns, *SF* partitions can be constructed systematically and processors can multicast to their communicating processors within one cycle in many applications. A partitioning algorithm has also been presented for arbitrary multicasting patterns. The overall results of the two level addressing implementation are higher efficiency, lower minimum optical path requirements and potential speed ups. This reinforces our belief that coincident pulse techniques are a promising addressing mechanism which can be applied in both parallel memory structures and multiprocessor systems.

†*LCM* stands for Least Common Multiplier and *GCD* stands for Greatest Common Divider.

3.6 Model of Lossless Bus Structure Using Erbium Fiber Amplifiers Pumped near 820nm

Fiber amplifiers based on erbium or other rare earth doped fibers have proven to be highly efficient in giving high gains and low noise power. In this section we present a model of a bus structure which adapts this technology to multiprocessor interconnections. Specfically, we model a tapped fiber bus using distibuted 820nm laser diode pump sources.

Introduction

The tapped fiber bus architecuture is an important design for fiber interconnected multiprocessors. In this architecture, all the devices communicate through a central fiber optic link, called the bus. A bus is made up of many fibers connected to each other with couplers. The light signal on the bus can be tapped out at any coupler and read by a detector. Thus each coupler along with coupling light from one fiber to the other, also results in a certain amount of loss of the light power. This is a major drawback of the system. Due to this the power level of the light pulse being transmitted, falls after each coupler. Consequently the power level at each detector also falls. And thus, the threshold power of the detector imposes a constraint on the maximum number of couplers we can have in the system.

Therefore some mechanism has to be provided, to restore power when it becomes too low. This would make it feasible to have a large number of couplers in the system. This has given the required impetus to the development of Er-doped fiber amplifiers. When these erbium amplifiers are introduced in an optical bus, the number of couplers the system can support is increased[Raj:90].

In this section the modelling of such an erbium amplifier is discussed. The models developed earlier were by Giles, et. al.[Gie:91] and by Sunak, et. al.[Sunak]. The model of Sunak, et. al., has the drawback of being complicated and also difficult to simulate. The model used in this research was the one developed by Giles, et. al.[Gie:91]. The mathematical equations for the amplifier are the ones developed in this model. The values of the various parameters used, were taken from papers published earlier.

The function of an erbium amplifier is to amplify the signal power on the bus. The inputs to the amplifier are lights at two wavelengths, pump and the signal. The pump results in the amplification of the signal. The working of the amplifier, is due to the lasing action of the Er-doped fiber. The model makes use of a simple two-level laser system[Gie:91]. In this system, the pump is used to excite Er-ions to a higher energy state. These excited ions, then fall back to the ground state, upon stimulation by the signal, resulting in amplification.

The light shining into the fiber, is assumed to be in the form of distinct optical beams. Each beam is centered at a certain frequency ν_k, and has a frequency spread $\Delta\nu_k$. In the model, only two such beams, one for the pump and one for the signal are considered. The signal and the pump wavelengths used are 1550 nm and 820 nm, respectively. The model essentially breaks up the fiber into many small pieces. It then solves a pair of coupled differential equations, for each piece, till

the end of the fiber is reached.

Another model is built for the optical bus. A bus essentially consists of couplers, referred to as 'taps' and optic fibers. Couplers are necessary to tap out or introduce light at any point on the bus. The model uses passive, bi-directional, 2x2, symmetric couplers.

The bus and amplifier models are then integrated. Thus, between any two couplers in the bus, we can introduce an Er amplifier. Then the transmission of light through such a bus is studied, using the coupling ratios of the couplers, the placement strategy for the Er amplifiers and the amplifier characteristics, as variables. The results of the experiments show that, with a proper choice of the amplifier characteristics and the placement strategy, it is possible to build optical buses using Er amplifiers. In such buses, the logical one (higher signal power) can be maintained upto a certain number of couplers. While the logical zero (lower signal power) can be attenuated. This implies that, beyond a certain number of couplers only the logical one remains in the bus. The logical zero dies out.

2. Amplifier Model

The amplification properties of the erbium amplifer are attributed to the lasing action of the fiber. In the model, a two-level laser system is assumed. The light at the pump wavelength($820nm$), excites Er-ions from the ground state to a higher energy state. These ions then fall back to the ground state upon stimulation by light at the signal wavelength($1550nm$). This leads to the amplification of the signal.

The light travelling in the fiber is assumed to be composed of a number of optical beams [Gie:91]. Each beam has a central frequency ν_k, and a frequency spread $\Delta\nu_k$ around the central frequency. The variable k is a dummy index which is summed over the total number of optical beams. Each of these beams, affects the populations of the Er-ions in the various energy levels.

The analysis is done in cylindrical co-ordinates r, ϕ and z. Here r is the distance in the direction perpendicular to the fiber axis, z is the distance along the fiber axis and ϕ is the azimuthal angle. The distance z is measured from the point at which the signal is introduced into the fiber. In the model the pump and the signal, are assumed to be co-directional.

The light intensity of the k^{th} optical beam at any point (r, ϕ, z) in the fiber is given by $I_k(r, \phi, z)$. Since, the intensity is power per unit area, the total power at a cross-sectional plane $P_k(z)$, is given by the integration of $I_k(r, \phi, z)$ over the cross-section as in Equation 1.

$$P_k(z) = \int_0^{2\pi} \int_0^{\infty} I_k(r, \phi, z) r dr d\phi \tag{1}$$

Here we have implicitly assumed that, $I_k(r, \phi, z)$ is zero outside the fiber core. Hence the limits of integration for r, are from 0 to ∞. Now the normalized optical intensity $i_k(r, \phi, z)$, is defined as the ratio of the light intensity $I_k(r, \phi, z)$ and the total power $P_k(z)$, at that cross-section of the

fiber. This is given by Equation 2.

$$i_k(r,\phi) = \frac{I_k(r,\phi,z)}{P_k(z)} \tag{2}$$

Let n_1 and n_2 be the densities of Er-ions, in the ground state and first excited state. Since the intensities of light beams are varying with time, n_1 and n_2 are also varying with time. The variation of n_2 with time is given by Equation 3. The total number of Er-ions in the two states combined, remains constant and is given by Equation 4.

$$\frac{dn_2}{dt} = \sum_k \frac{P_k i_k \sigma_{ak}}{h\nu_k} n_1(r,\phi,z) - \sum_k \frac{P_k i_k \sigma_{ek}}{h\nu_k} n_2(r,\phi,z) - \frac{n_2(r,\phi,z)}{\tau} \tag{3}$$

where σ_{ak} and σ_{ek} are the absorption and emission cross-sections of the fiber, for the k^{th} optical beam and τ is the lifetime of the Er-ions in the excited state.

$$n_t(r,\phi,z) = n_1(r,\phi,z) + n_2(r,\phi,z) \tag{4}$$

In Equation 3 the first term on the right hand side, is for the absorption. This is proportional to the density of the Er-ions in the ground state. The second term takes into account the stimulated emission, which is proportional to the density in the excited state. And the third term is for spontaneous emission. From Equation 3, it is obvious that absorption tends to increase n_2, while the spontaneous and stimulated emissions tend to decrease n_2.

The variation of light power through the fiber for the k^{th} optical beam, is given by the variation of power $P_k(z)$, with respect to the distance z along the fiber axis. This is given by Equation 5.

$$\begin{aligned}\frac{dP_k}{dz} &= \sigma_{ek} \int_0^{2\pi} \int_0^{\infty} i_k(r,\phi) n_2(r,\phi,z)(P_k(z) + h\nu_k \Delta\nu_k)\ r\ dr\ d\phi \\ &\quad - \sigma_{ak} \int_0^{2\pi} \int_0^{\infty} i_k(r,\phi) n_1(r,\phi,z) P_k(z)\ r\ dr\ d\phi\end{aligned} \tag{5}$$

Here the first term on the right hand side takes into account the increase in power due to stimulated emission. The second term accounts for the decrease in power due to absorption. These equations need to be expressed in terms of the absorption (α_k) and the emission (g_k^*) coefficients for the fiber. These coefficients are given by Equations 6 and 7.

$$\alpha_k = \sigma_{ak} \int_0^{2\pi} \int_0^{\infty} i_k(r,\phi) n_t(r,\phi,z) r\ dr\ d\phi \tag{6}$$

$$g_k^* = \sigma_{ek} \int_0^{2\pi} \int_0^{\infty} i_k(r,\phi) n_t(r,\phi,z) r\ dr\ d\phi \tag{7}$$

If we assume that Er-ions are uniformly concentrated in a disk of radius b around the fiber axis, the equations become:

$$\alpha_k = \sigma_{ak}\Gamma_k n_t \tag{8}$$

$$g_k^* = \sigma_{ek}\Gamma_k n_t \tag{9}$$

where

$$\Gamma_k = \int_0^{2\pi}\int_0^b i_k(r,\phi)\, r\, dr\, d\phi \tag{10}$$

This Γ_k, is called the overlap integral, between the dopant and the optical mode of the beam. Substituting these new parameters into the fiber Equation 5, we get:

$$\begin{aligned}\frac{dP_k}{dz} &= \frac{\alpha_k + g_k^*}{\Gamma_k} P_k \int_0^{2\pi}\int_0^b \frac{n_2(r,\phi,z)}{n_t} i_k r\, dr\, d\phi - (\alpha_k + l_k)P_k \\ &+ \frac{g_k^*}{\Gamma_k} h\nu_k \Delta\nu_k \int_0^{2\pi}\int_0^b \frac{n_2(r,\phi,z)}{n_t} i_k r\, dr\, d\phi\end{aligned} \tag{11}$$

Here term l_k has been added to take into account the excess loss in the fiber. In steady state, the density of Er-ions in the various energy levels is going to remain constant with time. Consequently, we can equate the left hand side of Equation 3 to zero. After rearranging the terms we get:

$$n_2(r,\phi,z) = n_t \frac{\sum_k \tau P_k i_k(\tau\sigma_{ak})/(h\nu_k)}{1 + \sum_k \tau P_k i_k(\sigma_{ak} + \sigma_{ek})/(h\nu_k)} \tag{12}$$

From Equation 12 it is apparent that the value of n_2 depends on the powers of all the light beams. Equations 11 and 12, are then solved for the required number of optical beams. The thing to be noted is that, Equation 11 is a *set* of differential equations, for $k = 1$ to m, where m is the total number of beams. The presence of n_2 in Equation 11, couples the equations together.

The model makes certain assumptions regarding the light propagation in the fiber. Firstly, only two optical beams are taken into consideration. One for the pump $k = 1$ and the other for the signal $k = 2$. This means that we have only two coupled differential equations to be solved. The second assumption is that, the fiber is assumed to be a single mode fiber. This implies that only the zeroth modes are dominant for the pump and the signal. The expression for the light intensity for the zeroth mode is given by Equation 13[Lee:86].

$$i_k(r) = C_k e^{-r^2/r_k^2} \tag{13}$$

where C_k and r_k are constants.

Table 1: Nomenclature

Symbol	Explanation
ν_k	Frequency of the k^{th} optical beam.
$\Delta\nu_k$	Frequency spread of the k^{th} optical beam.
P_k	Power of the k^{th} optical beam.
I_k	Intensity of the k^{th} optical beam.
i_k	Normalized intensity of the k^{th} optical beam.
σ_{ak}	Absorption cross-section of the k^{th} optical beam.
σ_{ek}	Emission cross-section of the k^{th} optical beam.
α_k	Absorption spectrum of the k^{th} optical beam.
g_k^*	Emission spectrum of the k^{th} optical beam.
Γ_k	Overlap integral for the k^{th} optical beam.
l_k	Excess loss in the fiber for the k^{th} optical beam.
n_1	Density of Er-ions in the ground state.
n_2	Density of Er-ions in the excited state.
n_t	Total denstiy of Er-ions, in the two states combined.
b	Radius of the disk in which Er-ions are concentrated.

The explanation of the variables used in the model is given in Table 1. The values for the variables used by the model are given in Table 2.

The model breaks the length of the fiber into a large number of small pieces, connected end to end. Then for each piece, the pair of coupled differential equations is solved, till the end of the fiber is reached.

Let P_{k1}, P_{k2}, P_{k3},..., be the powers of the k^{th} optical beam, at the inputs of pieces 1, 2, 3,..., respectively. Since the pieces are end to end, P_{k2}, P_{k3},..., are also the powers at the outputs of pieces 1, 2,..., respectively. Let $z_1(=0)$, z_2, z_3,..., be the distances of the inputs of the pieces from the beginning of the fiber. For a small piece of fiber, Equation 11 can be approximated by:

$$\begin{aligned} dP_k &= \left(\frac{\alpha_k + g_k^*}{\Gamma_k} P_k \int_0^{2\pi}\int_0^b \frac{n_2(r,\phi,z)}{n_t} i_k r\, dr\, d\phi - (\alpha_k + l_k)P_k \right. \\ &\quad \left. + \frac{g_k^*}{\Gamma_k} h\nu_k \Delta\nu_k \int_0^{2\pi}\int_0^b \frac{n_2(r,\phi,z)}{n_t} i_k r\, dr\, d\phi\right) dz \end{aligned} \tag{14}$$

where dP_k is the difference in the powers at the input and the output of the piece and dz is the length of the piece. Thus for the first piece we can write

$$P_{k2} - P_{k1} = \left(\frac{\alpha_k + g_k^*}{\Gamma_k} P_{k1} \int_0^{2\pi}\int_0^b \frac{n_2(r,\phi,z)}{n_t} i_k r\, dr\, d\phi - (\alpha_k + l_k)P_{k1}\right.$$

Table 2: Symbols and Values

Symbol	Value
σ_{a1}	$0.18 * 10^{-25} m^2$[Min:91]
σ_{a2}	$5.1 * 10^{-25} m^2$[Bar:91]
σ_{e1}	$0\ m^2$
σ_{e2}	$4.4 * 10^{-25} m^2$[Bar:91]
Γ_1	0.6[Bar:91]
Γ_2	0.6[Bar:91]
n_t	$10^{24} - 10^{26}$ per m^3[Min:91, Agg:91]
ν_1	$3.66 * 10^{14}$ per sec
ν_2	$1.94 * 10^{14}$ per sec
$\Delta\nu_1$	$4.462 * 10^{12}$ per sec
$\Delta\nu_2$	$1.249 * 10^{12}$ per sec
τ	10 ms[Bar:91]
b	6.2 $microns$

$$+ \frac{g_k^*}{\Gamma_k} h\nu_k \Delta\nu_k \int_0^{2\pi} \int_0^b \frac{n_2(r,\phi,z)}{n_t} i_k r\ dr\ d\phi)(z_2 - z_1) \qquad (15)$$

Since we know the pump and the signal powers at the input i.e. P_{11} and P_{21}, we can substitute them into Equation 15, to get the signal and pump powers at the end of the first piece. These can then be used as inputs for the second piece. Repeating the same calculations for the second piece, we get the inputs to the third and so on. This is repeated till the end of the fiber is reached. Thus finally, we get the pump and the signal powers at the end of the fiber.

3. Linear Bus Model

The linear bus consists of a single unbranched fiber running from the input to the output, as shown in Figure 15. The signal can be tapped or introduced at any coupler in the bus. The tapped signal is fed to a detector. Each coupler also results in a certain amount of loss of the signal power. The couplers are 2x2, passive, bi-directional couplers. In the figure, q is the coupling ratio of the couplers.

4. Integration of the Amplifier and the Bus Models

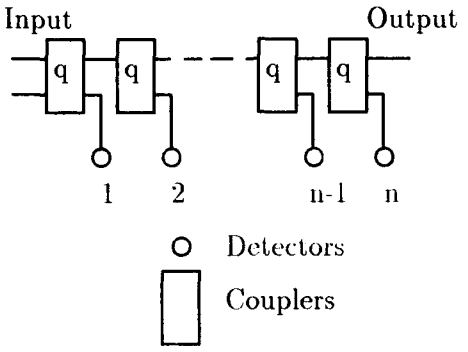


Figure 15: Linear Bus

The models for the linear bus and the erbium amplifier are integrated. In this bus, an erbium amplifier can be introduced between any two couplers in the bus. Otherwise the two couplers can be connected by an ordinary fiber. The characteristics of the erbium amplifier, the placement strategy and the coupling ratios of the couplers, can be varied. Thus it is possible to configure the system according to our requirements. The integrated model is shown in Figure 16.

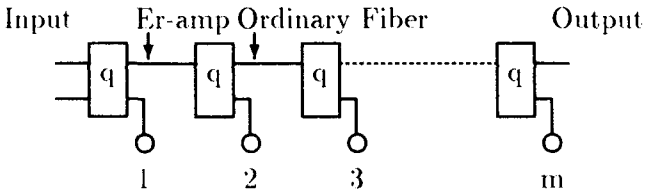


Figure 16: Linear Bus With Er Amplifiers

5. Experiments

Two sets of experiments were carried out. The first set was aimed at validating the amplifier model and getting the data which would be required, when introducing the amplifiers in the optical bus. The second set of experiments was on the linear bus, with the erbium amplifiers. Many different configurations of the system were taken into consideration. While doing so, the characteristics of the erbium amplifiers, the placement strategy and the coupling ratios, were used as variables.

5.1 Study of Amplifier Characteristics

The first experiment carried out, was to find the variation of the dB gain as a function of length for the amplifier. The aim of this experiment was to see if the results of the model show the known

characteristics for Er amplifiers. One of the most important characteristic, is the presence of an *optimum length*, for which the amplification is maximum. The results of this experiment could then be compared with those in earlier papers.

The input signal power was held constant at $50\mu W$. The length of the amplifier was varied from $0m$ to $4m$, in steps of $0.5m$. This was done for different values of pump powers. The density of the Er-ions was $10^{24}/m^3$. The pump powers used were from $10mW$ to $100mW$, in steps of $10mW$. The results obtained are shown in Figure 17.

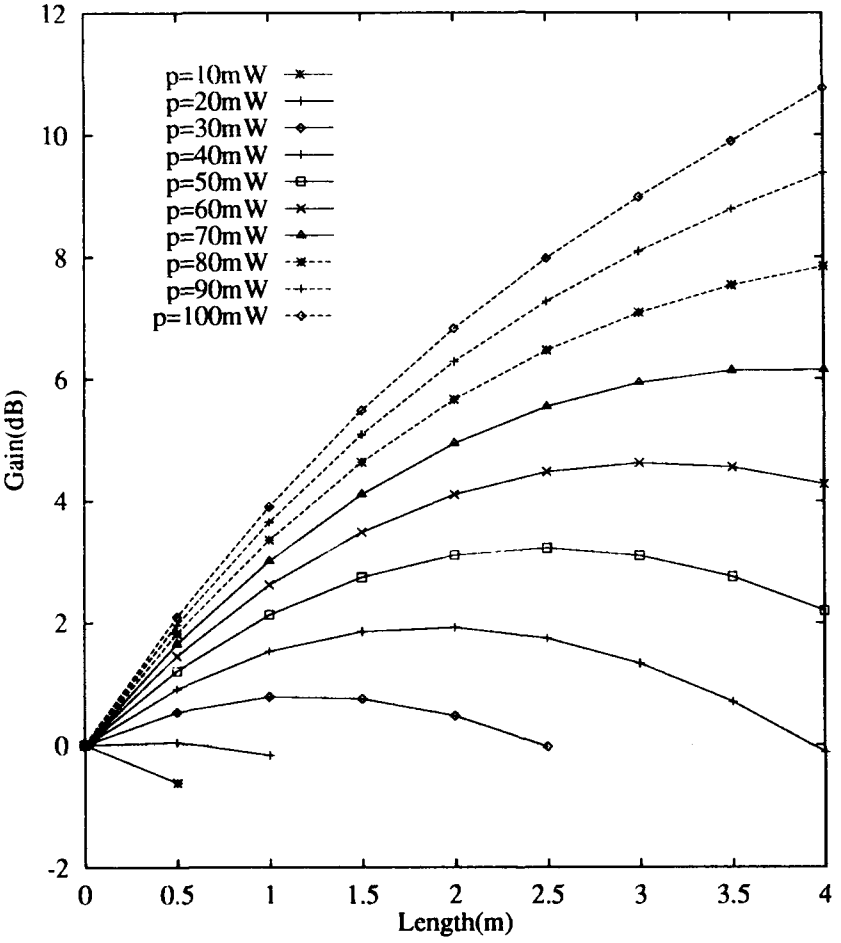


Figure 17: dB Gain vs Length for Different Pump Powers

From the results of the first experiment in Figure 17, it is seen that as the length of the amplifier is increased, initially the gain also increases. At a certain point the gain saturates. This is the *optimum length*. Beyond the *optimum length*, it is seen that the gain is reduced. From Figure 17, it is also seen that the if the pump power is increased, the gain increases. These results agree with

similar curves published before[Gie:91]. In the published results also, the presence of an *optimum length*, is evident. The difference between the published and the curves obtained here, is that the gain is different in the two cases. But the general shape of the curves is the same.

In the next experiment the variation of the signal out vs the signal in for the ampifier was plotted. The aim in doing this experiment was to find the variation in the gain of the amplifier, as the input signal was varied. This would give an idea about the amplification for the logical one and the zero, on the optical bus.

The input signal was varied from $10\mu W$ to $100\mu W$, in steps of $10\mu W$. This ensured that the range in which the logical one and zero power levels were likely to be, is sweeped. The length of the amplifier was held constant at $1m$ and the Er-ion density at $2.5 * 10^{24}/m^3$. This was done for different values of the pump powers. The pump was varied from $10mW$ to $50mW$, in steps of $10mW$. The results are shown in Figure 18.

From Figure 18, it is seen that, as the pump power is increased the signal out also increases i.e. the amplification increases. Also, for a pump power of $10mW$, it is seen that a signal of $10\mu W$ (logical zero) is attenuated. But, a signal of $100\mu W$ (logical one) gets amplified. This suggests the use of an amplifier with these characteristics, in an optical bus. If such an amplifier could be introduced at specific points in an optical bus, the zero could be made to die out.

5.2 Experiment on the Optical bus

For this experiment on the optical bus, first the one and the zero power levels were decided. The values were based on the general charateristics of laser diodes. These diodes would be used to generate the signal and pump powers. To ensure fast switching, the signal diode has to be at the edge of cut-off and the forward biased region, when its output is logical zero. And for a logical one, it is necessary that the diode is not too much forward biased. Otherwise it would take a long time to switch it off. Taking these facts into account, the logical one and zero levels were taken to be $100\mu W$ and $10\mu W$, respectively.

In an ideal optical bus, the power level of the logical one would be maintained, while that of the logical zero would die out. This implies that the amplification due to the erbium amplifiers, should be controlled. If the amplification at a particular point is high, then both the one and the zero levels are going to increase drastically from their earlier levels. If the erbium amplifiers are placed too frequently, the amplification may exceed the loss due to the couplers. And the signal levels would go on increasing. At the same time, if the erbium amplifiers are placed too sparsely, the signal level would go on falling. These factors decided the placement strategy of the amplifiers on the bus.

For the experiment, an amplifier was needed which would show a high amplification at a higher signal power, and a low amplification or loss at a lower signal power. This criterion is satisfied by an amplifier with the following characteristics: length = $1m$, density of Er-ions = $2.5 * 10^{24}/m^3$, pump = $10mW$. This is evident from Figure 18. A coupling ratio of 0.964 was taken for the signal, and 0.01 for the pump. The low coupling ratio for the pump meant that, most of the pump power from

the diagonal input of the coupler was coupled into the fiber amplifier. The coupling ratio of 0.964 for the signal demanded an erbium amplifier after every four couplers. This would approximately compensate for the loss in the signal power, due to the couplers. The length of the bus was taken to be of 50 couplers. The results of the experiment are shown in Figure 19.

From Figure 19, it can be seen that the logical one power level is approximately maintained through the bus. But the logical zero dies out. Beyond about 40 couplers, the logical one starts falling. This is because as the signal power falls, the gain also falls. And when the signal level goes below a certain value, the loss due to the couplers becomes more than the gain due to the amplifiers.

6. Conclusions

The characteristics of the erbium amplifier in Figure 17 agree with the papers published earlier. From the figure the presence of an *optimum length*, is evident.

From the results of the experiment on the optical bus with Er amplifiers in Figure 19, it is seen that the logical one level is almost maintained upto a certain number of couplers. The lower power level of $10\mu W$, looses power at the couplers as well as the amplifiers, and hence dies out. Thus we have shown that it is possible to build an optical bus with erbium amplifiers. This bus has the advantage of being able to support a higher number of couplers. With a proper choice of the amplifiers charateristics, and the placement strategy, it is possible to build a bus in which the logical one level is approximately maintained, while the zero dies out.

7. Future Work

In this paper the bus could support 50 couplers. Beyond this the logical one level falls rapidly. If the one level could be accurately maintained, the bus could be made to support a larger number of couplers.

Many *talkers* (devices introducing signals into the bus) and *listeners* (devices reading the signal from the bus), could be then connected to the bus. It would be a challenging problem, to design the bus in such a fashion that, no matter who *talks* or *listens*, the power levels on the bus are maintained.

Another interesting aspect would be to integrate wavelength division multiplexing, into the model. This would require two or more signal powers and maybe different pumps to amplify them. The interdependences would also have to be taken into consideration.

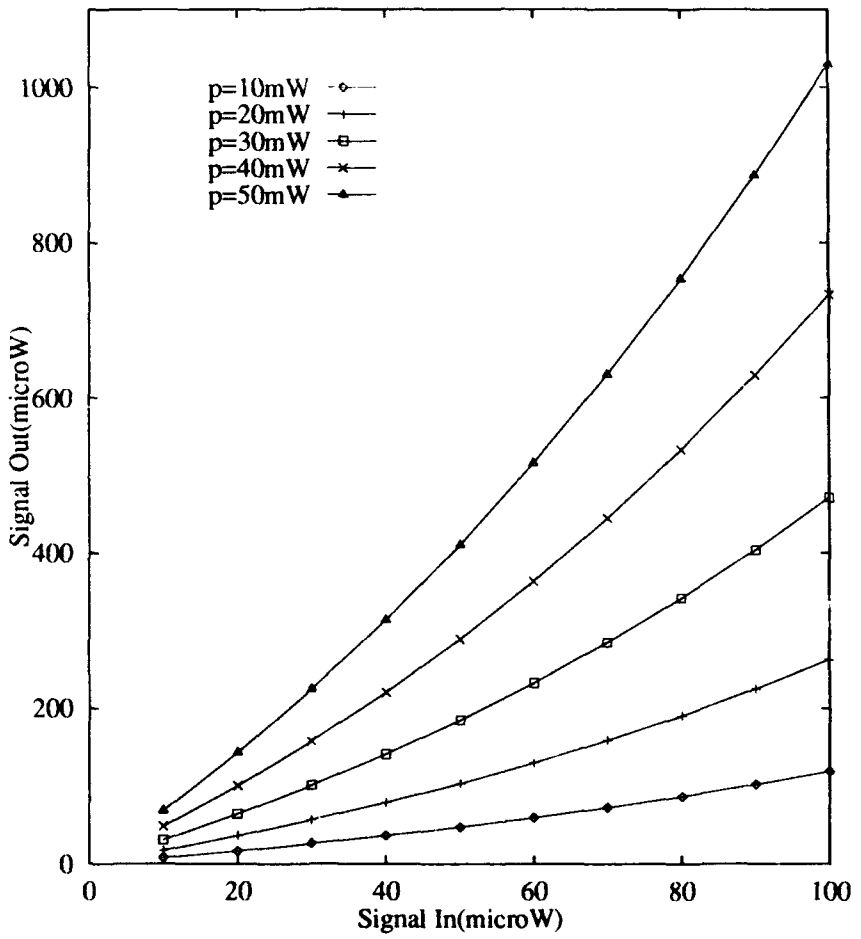


Figure 18: Signal Out vs Signal In for the Amplifier

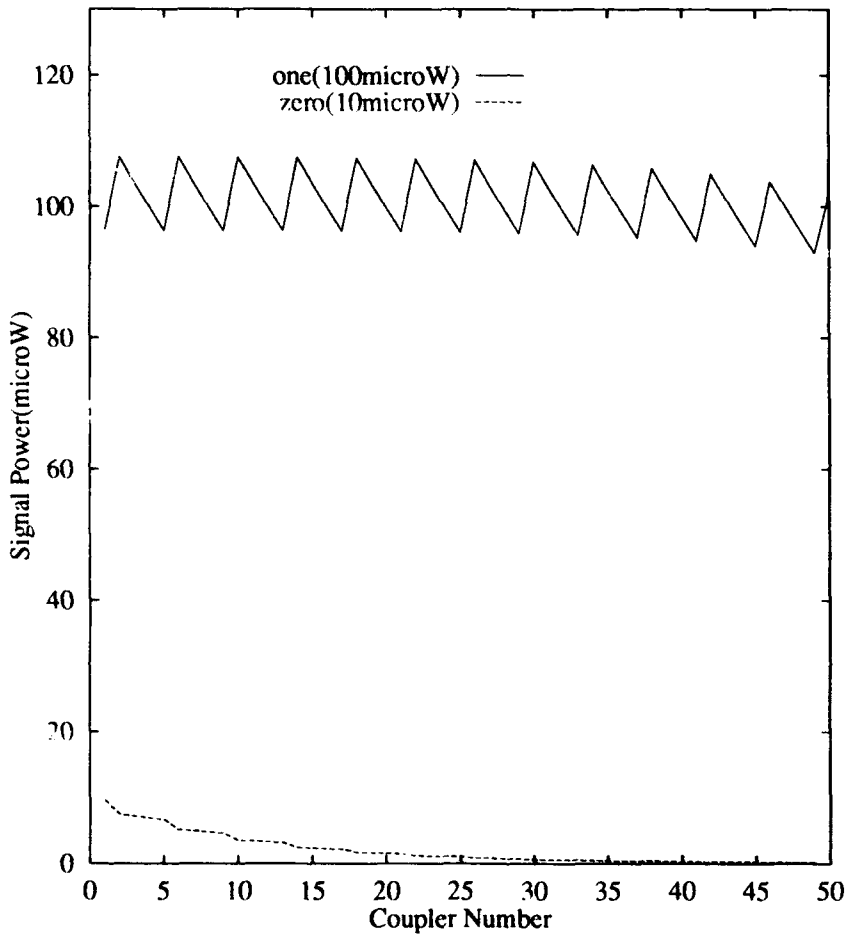


Figure 19: One and Zero Level Variation in an Optical Bus

3.7 Bandwidth as a Virtual Resource in Multiprocessor Interconnections

A significant problem in implementing optoelectronic multiprocessor interconnection networks is the difference in bandwidth between the optical channels which carry the messages and the electronic circuitry which controls the network. To address this problem, we propose a novel control paradigm that is based on dynamic network reconfiguration rather than explicit routing of messages using encoded addresses. This reconfiguration is accomplished by sequencing the interconnection network through a series of states which establish all of the required communication paths. In this paradigm, the routing problem becomes one of transforming a sequence of configurations in response to new connection requests. Thus, it is sufficient that the control system operates at a bandwidth proportional to the *changes* in the communication pattern.

Introduction

Hybrid optoelectronic computing structures are required to provide the information processing capabilities for the next generation of computing and communications systems. This work is focused on reconfigurable optoelectronic interconnection networks: networks constructed of optical waveguides in which messages are switched or routed by means of optoelectronic devices [Goo89]. The dichotomy between the bandwidth of the optical channels which carry information through these networks and the performance of the electronic controllers and decoders which determine the routing and destination of those messages is a significant bottleneck. We introduce a new class of routing algorithms for reconfigurable networks to bridge this gap in optical versus electronic performance. The algorithms are based on a new control paradigm which exploits the locality in multiprocessor communication streams to reduce the control latency inherent in reconfigurable interconnection structures.

Research on electronic reconfigurable interconnections spans almost two decades of computer engineering literature. A number of books on the subject exsist [Sie84] which provide an excellent background. In the optical domain, reconfigurable systems first evolved in the context of free space designs using spatial light modulators as the switching fabric. Guided wave systems have been studied more recently. Specificially, previous work be the authors has focused on the fundamental problems in the design of optical bus based interconnections for multiprocessors: addressing [CML87, MCL89, CLM90b, LCM90, CDLM91], data transfer [GMH+90, GM90, GMH+91] and control [CLM90a]. One of the primary concerns in all our work has been the problem of latency in optical interconnections. Therefore, our addressing results emphasize the ability to select several locations in parallel. The arbitration mechanism is fully distributed in order to avoid the latency of communications with a central controller. The message pipelining results emphasize the amortization of latency over a large group of messages.

Busses, however, are only the simplest of shared resource interconnection networks. For more

complex shared resource networks, addressing and control issues cannot be solved independently. In these networks, addressing and control are implemented jointly as message routing. Unlike addressing, routing considers not just the destination of a message, but also its path, and the network resources needed for that path. Any routing strategy represents a tradeoff between explicit addressing and global control. In most systems, explicit addressing dominates this tradeoff. In other words, source nodes drive the control hardware which arbitrates resources to create the message path.

In this section we present an alternative in which the control system dynamically allocates resources based on global knowledge of the message traffic. Thus, rather than transmitters presenting addresses tc the network, the network establishes a set of paths and presents these to the transmitters and receivers. Since most interconnection networks cannot provide all possible paths simultaneously, the issue in these routing strategies is to allocate the network resources such that they consistently meet the needs of the current message traffic. By using the locality which is inherent in the message traffic, a single control operation may allocate a network resource for use by a sequence of messages between a common source and destination. Thus the latency of making the control decision is amortized over the entire sequence.

Routing by Global Resource Allocation

Routing as selection in a virtual connection space

The *principle of locality* is a well established paradigm governing the way that processes communicate with memory resources. Temporal locality in a program refers to the tendency of programs to work in localized sections of code. Spatial locality refers to the same pattern of localized access to data memory. This principle is the basis for virtual memory systems in which a large memory space is provided using a much smaller amount of physical memory. In multiprocessor interconnection structures, the same principles can be applied in order to allow an interconnection network of lower bandwidth to provide the bandwidth requirements of a fully interconnected system.

We draw a direct analogy between the routing problem for reconfigurable multiprocessor interconnection networks and paging in virtual memory systems. The correspondence between the two is summarized in table 3. As shown in table 3, connections in a fully connected network can be viewed as the analog of memory locations in a complete virtual address space. Just as a physical memory supports a subset of the virtual address space, so a switched interconnection network implements a subset of connections from the fully connected network. Physical memory is shared and reused to create the illusion of large virtual memory space. Similarly, a switched network can be reconfigured to emulate the functionality of full interconnection.

The analogy also extends to addressing. The unit resource in a memory system is a single memory location. In a communication network the unit resource is a single connection path. In memories, an $n - bit$ virtual address defines an address space of $2^n = N$ addressable locations. Paging divides this address space into k pages, each of size m locations such that $N = m \times k$. In communication networks, a full interconnection network for n sources and n destinations provides $n \times n = N$ interconnection paths, i.e., N unique source-destination pairs. While an

arbitrary switched interconnection network may establish any of these N paths, it is only capable of connecting a subset of m paths simultaneously, giving a particular configuration. In order to enumerate all possible paths, a *sequence* of k different configurations is required, such that $N = m \times k$. Just as at any given time, a subset of the k pages resides in physical memory to satisfy the current set of memory requests, a communication network needs only to sequence through a subset of configurations to satisfy the current requests for paths.

Entity	in Virtual Memory	in Communications Network
Addressing Space	Virtual Address	All $n \times n$ connections in a fully connected network
Degree of Sharing	Physical Memory Size	The current configuration sequence length
Unit of Sharing	Page	A single network configuration
Addressable Unit	Memory Location	A particular path from a source to a destination

Table 3: Summary of Virtual Memory vs Switched Interconnection Analogy

Virtual memory systems work because the principle of locality states that if a working set of pages is made available to a group of programs, that set of pages will change slowly over time. By extension, if a sequence of configurations is sufficient to support all of the current traffic in an interconnection network, so will that sequence change slowly over time. Thus, we can use the principle of locality to decouple the performance of control algorithms for interconnection networks from the latency of individual messages. The control algorithm needs only to perform in the time frame of locality changes, not in the time frame of individual message traffic. This is a key concept for high bandwidth optical interconnection networks. Since the routing decisions in these networks are most often made by electronic controllers, the performance of these controllers represents a significant bottleneck. By exploiting locality, message routing can be reduced to a problem of providing a repeated sequence of configurations to the network. Control becomes a problem of transforming that sequence to track the changes in the locality of message traffic.

We call this technique *state sequence routing*. In the next section we describe the technique formally and present three examples. These examples, a time division switched linear bus, a space division switched multistage network, and a wavelength division switched star network, demonstrate that the technique can be used independently of the switching domain. The key research issues involve the control algorithms for these networks. In the remaining sections, we present our preliminary work in which we have identified three classes of control algorithms. The first is a static allocation technique, which is based on a fixed computational structure derived from compilation analysis. The second is a dynamic technique which assumes an explicit request and release protocol. The third is also a dynamic technique, which uses information from concurrent local and global control algorithms.

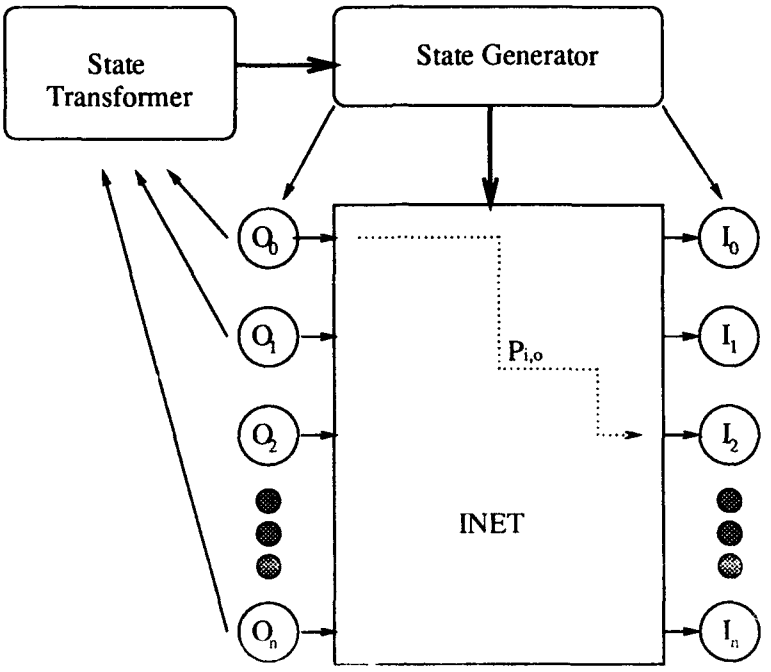


Figure 20: **General Interconnection Block**

State sequence routing

The general structure of a communication network based on state sequence routing is shown in figure 20. Let INET in this figure be an $n \times n$ interconnection network connecting a set I of n input ports to a set O of n output ports, and let $p_{io} = (i, o) \in I \times O$ denote the path between a specific input port $i \in I$ and a specific output port $o \in O$. We assume that the INET may establish any of the possible $N = n^2$ paths, but that it cannot establish them simultaneously. Thus, INET may be a bus, a multistage interconnection network (MIN), a wavelength division switched (WDS) star, or any other type of interconnection network. Within an INET, we define a mapping, M, to be a set of paths that can be established at the same time without conflicts in the INET. For each mapping M there is a corresponding state S which represents the configuration of the network (i.e., switch settings, detector tunings, etc.) corresponding to that mapping.

Since the establishment of two paths at the same time in an INET may cause conflicts, not every set of paths is a mapping. We refer to the establishment of all the paths in a mapping as the *realization* of the mapping. Given a set of paths $P \subseteq I \times O$, it may not be possible to realize all paths in P at the same time without conflicts. However, P can be partitioned into several mappings, $P = M_1 \cup M_2 \cup ... \cup M_t$. Each mapping M_i, $i = 1, 2, ..., t$, may be realized in sequence. Given that each mapping has a corresponding state, the set of paths P may be implemented as an ordered sequence of states, $[S_1, S_2, ..., S_t]$ where t is the length of the sequence.

Returning now to figure 20, the state generator block is responsible for generating the current

state sequence. The control algorithm, which determines the sequence, runs in the state transformer. The current state of the network is also communicated to each of the transmitting nodes. Thus, a transmitting node waits for the network state corresponding to a mapping which contains the required path. When such a state is detected, the node transmits its message. If no such mapping exists within the current state sequence, the control algorithm modifies the state sequence to include a mapping which supports the requested path. We will discuss the specifics of several control algorithms in the next section. First, we present three examples of the application of state sequence routing to each of a time division switched linear bus, a space division switched multistage network, and a wavelength division switched star network. The topologies of these networks are shown in figure 21.

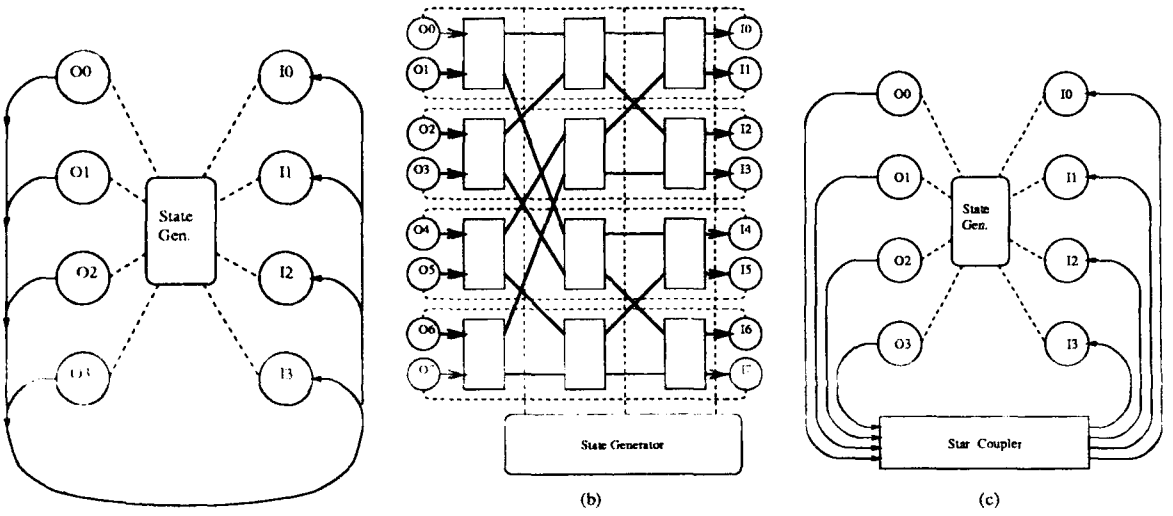


Figure 21: **Three Interconnection Networks**

Figure 21(a) is a linear bus interconnection of four transmitters and four receivers. This bus implements time division switching as follows. Synchronized by a global clock, a time division multiplexed pipeline of four messages is originated simultaneously, one message from each transmitter, O_0 ... O_3. In order for the pipeline to be formed without collisions, the length of fiber between each transmitter is large enough such that an entire message can be stored in its length. In this version of the bus we have opted for a completely passive interconnect. Therefore, we must assume that each receiver is capable of distinguishing between individual messages within the pipeline. In other words, each receiver can be programmed to accept the *ith* message, including null messages, in the pipeline and to ignore all others. If this capability is not available in the receivers we can easily support the same functionality by replacing the passive receiver splitters with active switches. Busses of this type have been studied in detail by the authors in [GMH+91].

For this network, a path between sender O_i and receiver I_j is formed when I_j is programmed to select the $i-th$ message from a pipeline. A mapping is formed by the set of paths supported in a pipeline. The current programming of all receivers is the state. The state generator in this example has two responsibilities, first to transmit the programming for the current state to each receiver, and second to inform each transmitter which receiver has been programmed to receive its

message. If more than one mapping is required, the state generator would sequence through a set of mappings, informing the transmitters and receivers of each state in the sequence.

Figure 21(b) is a multistage interconnection network (MIN). In this network, a unique path can be established between any input port and any output port. Along each path, there are $\log n$ switches, one at each stage. In order to establish a path, each switch along the path has to be set properly to either a "straight" or a "cross" state. Other switches in the network can be in either state without affecting the path. Therefore, in any mapping there will be exactly n paths.

Given a mapping, it is straightforward to find the state S that realizes the mapping. Once the set of mappings is determined, the state generator outputs the sequence of states which establishes the mapping for all current paths. As described earlier, each transmitter monitors the output of the state generator and waits for a mapping which contains the requested path. Unlike the linear bus case, where only a limited set of shared resources were available, supporting a new path in the MIN may not require replacing an existing path or extending the sequence. Instead it may be possible to transform one or more mappings in the current sequence to form a mapping in which the new request is realizable.

Finally, figure 21(c) is a wavelength division multiplexed star network. A variety of interconnection protocols exist for these nets [Dou91]. For this example, we assume the most general case in which any of I transmitters may communicate with any of O receivers on any of W wavelengths. In other words, transmitters and receivers are independently tuned. Since in general $| W | \leq min(| I |, | O |)$, we further assume that any transmitter or receiver may be effectively "turned off" such that it will neither transmit or receive. Under these assumptions, a path consists of a triple $(i, o, w) \in I \times O \times W$ and a mapping consists of any realizable set of paths such that each path communicates on an different wavelength. The state S corresponding to each mapping is therefore a collection of transmitter and receiver tunings. In this case the state generator outputs a sequence of mappings by assigning a wavelength for each path in the mapping. Each source node monitors the state for an assignment of a wavelength to its transmitter. When such a wavelength is assigned, the message is sent using the current mapping. Receivers must also monitor the state for wavelength assignments. In protocols which require the identification of the sender, they must also monitor the state to discover the transmitting node to which that wavelength was assigned. This is the general case, a number of simplifications are also possible. For example, either the transmitters (or receivers) could be assigned fixed wavelengths and only the receivers (or transmitters) could be tuned by the state generator.

For each of these examples, we have demonstrated that the basic context of state sequence routing is consistent across the domains of time, space and wavelength switching. Specifically, paths, mappings and states can be easily identified independent of the nature of the INET. Built upon this basic idea is the concept that mappings can be sequenced and the current sequence matched to actual traffic. Locality suggests that, once matched, sequences tend to change slowly. Therefore, control algorithms need not perform at the speed of individual messages, but instead may operate at speeds proportional to changes in locality. We turn now to the discussion of these control algorithms.

Static allocation algorithms

Given a specific application (a parallel algorithm), the communication requirements of that application can often be estimated and modeled by a bipartite graph, which we call a *Connection Request* (or CR) graph. A directed edge from a source processor i to a destination processor o in a CR indicates a request for the establishment of a path p_{io} in the MIN. We will use the same notation for a path to denote an edge and use the terms "edge" and "path" interchangeably. The maximum number of edges going out from a node or coming into a node in a CR graph is called the *degree* of the graph.

Given a CR graph, we call a mapping sequence $[M_1, M_2, \ldots, M_t]$ a *Minimal Connection* (or MC) sequence for the CR graph if every edge $(i, o) \in CR$ is established in some mapping and no two mappings in the sequence can be merged together. We call a sequence for a CR graph *optimal* if it is an MC sequence for the graph and it has the shortest length among all other MC sequences for the same graph. Note that if t is equal to the degrees of the graph, then the MC configuration is optimal.

When the communication structure of an application is regular, finding a sequence for its CR graph is often called *embedding*. Note that the ability to embed regular communication structures efficiently is important since there are many existing applications designed for them. The sequence length t is a measure of the efficiency of the embedding. Table 4 shows results which have been proven [Guo91, Qia91] for several structures embedded in a time division switched (TDS) bus and a space division switched (SDS) MIN.

Structure	TDS bus	SDS MIN
Binary Tree	3	4
Binary Hypercube	$\log n$	$\log n$
Fully Connected	n	n
Bidirectional Ring	2	2
Mesh	4	4
Cube Connected Cycle		3

Table 4: Sequence Lengths for Various Embeddings of n nodes

For non-regular CR graphs, an MC sequence can always be obtained by selecting a subset of mappings from a complete connection (CC) sequence. Assuming that each mapping is a set of n paths, any CC sequence will consist of n mappings. The problem with this *selection algorithm* is that it restricts itself to the particular set of n mappings in a chosen CC sequence. However, there may be as many as $n!$ mappings possible in the network.

Alternatively, given a CR graph with a set of edges E, a MC configuration can be obtained by composing mappings based on the set E, which may be different from any mapping in a chosen CC configuration. More specifically, the *composition algorithm* composes each mapping in a greedy fashion.

Under low or medium load conditions, the composition algorithm improves over the selection algorithm as expected. However, when the load is extremely high, the sequence length of an MC configuration could exceed n. That is, sequences under high load condition using this algorithm may be worse than simply using a CC configuration. In order to improve the selection algorithm under low or medium load conditions, and yet bound the sequence length of any MC configuration by n under high load conditions, we may use the selection algorithm first to determine a set of up to n mappings needed, then examine each mapping to see if it can be deleted from the configuration by *migrating* paths established in it to other mappings in the configuration. Simulation results [Qia91] show that this last algorithm outperforms both the selection and the composition algorithms on an SDS MIN.

Dynamic allocation

Static control algorithms work well if all paths are required to be established from the beginning to the end of the execution of an application. Unfortunately, most applications generate requests which cannot be determined until run time. Also, even if a CR graph that contains all edges needed during the execution can be constructed a priori, it may be inefficient to perform static control based on such a graph. Some paths are used only for a certain duration of time and are wasted for the remaining time during execution. It may be possible to reduce the total execution time of such an application by dynamically modifying the mapping sequence at run time. In this section we present the first of two algorithms which support dynamic allocation. It is based on an explicit request-release protocol and assumes the sequence length as well as its contents may vary dynamically. As with static allocation we assume an INET connecting n transmitters and n receivers, where each mapping may consist of up to n paths. To support the protocol we further assume that each processor can independently communicate request-release messages to the state transformer.

When the state transformer receives a path request there are three ways in which this path may be included in the current sequence:

- It can be added incrementally. In which case an existing mapping which is capable of realizing both its active paths and the new path is transformed such that it includes the new path. This is the most desirable option.
- A current path can be migrated from one mapping to another in order to create a mapping in which the new path is realizable. This is part of a more general problem of sequence compression which is also associated with the handling of release messages.
- A new mapping can be added to the sequence which contains the requested path.

If a run time request is to release a path, the mapping that currently establishes the path may be deleted if:

- the mapping contains no other active paths, or
- all other active paths can be migrated to other mappings.

Obvious bottlenecks exist in two aspects of this system. First, for adding a new path, a search of the existing mapping sequence is required. Second, in order to reduce the sequence length, compaction of the sequence is necessary. However, one can argue that it may be possible to do the latter, compression, off line. For the former operation, searching a linear list, the worst case latency is proportional to the length of the sequence. Yet, if a path were already in the sequence, the worst case latency for the corresponding mapping to be transmitted to the INET is also proportional to the sequence length. Therefore, if we assume an algorithm can be devised such that the test for realizability can work in time equal to a sequence step, we could expect similar performance for both new and existing paths.

Finally, one reasonable tradeoff is to buffer run time requests and periodically execute a static reconfiguration algorithm. At each selected instance, a snapshot of the CR graph is constructed based on all current paths that need to be established. This approach may also be useful as an off line compression algorithm which constructs a complete optimal sequence for all current paths and uses it to periodically replace the current sequence.

Cooperating local/global control

In this section we examine control structures suitable to the most general case, for example, heterogeneous program environments where the CR graphs cannot be predicted in advance, or environments where an explicit request/release protocol cannot be supported. In such systems, the interconnection network is invisible to both systems and application software.

In this discussion we use a system model similar to the previous dynamic allocation algorithm. Once again, the INET connects n transmitters and n receivers with each mapping capable of establishing up to n paths. Unlike the previous model, there is no explicit request/release protocol and the sequence length t is fixed at some value $t \leq n$. Once a path is established in the sequence, it remains in place until forcibly removed by replacement with a new path. Like previous models, each transmitting node holds a message until the state of the INET represents a mapping which contains the required path. In the absence of explicit request/release signals, it is the responsibility of the state transformer to guarantee that the state generator will output such a state in a finite amount of time.

The control algorithm in the state transformer must make two decisions. On one hand, new paths must be implemented in minimum time. However, with a fixed length sequence, optimal performance requires that the state removed by this transformation have the smallest impact on the current traffic. In order to satisfy both of these constraints, the control algorithm is implemented as two cooperating processes. One process uses only local information to make fast decisions regarding path replacement. This process uses a predetermined set of rules which require only information on the new path and the node from which it originates. The second process looks at global information such as path and transformation histories, as well as the current sequence. This process periodically updates the rules used by the local process.

Any number of rule systems could be devised for the local process. For example, the local rules can be based on partitioning the sequence positions into subsets Therefore, the local rule

would state that the transmitting node will contend only with nodes with which it shares a common subset. Partitioning the sequence positions into subsets is not unlike set associative cache mapping. For example, under these rules two nodes may be directed to always contend for position one in the mapping sequence. The global process has the power to rearrange the subsets periodically. Continuing with the example, should both nodes begin to contend heavily for position one, either or both can be regrouped such that they share a new subset with a less active node, or are placed by themselves in separate subsets.

Impact on system performance

The importance of reducing control latency for high speed networks can be seen from the following argument. As the bandwidth of networks increases, latency, the delay for a single message to travel through the network, does not necessarily decrease proportionally. This is because latency is composed of several components: message control time, message transmission time, and the physical delay of the network. Only the message transmission time (i.e., the bit rate of the message) is directly tied to the bandwidth of the network. The physical delay of the network is limited by the wave propagation speed and the length of the path. We cannot reduce this time. The control time is a function of the speed of the control hardware and the complexity of the control task. For high speed optoelectronic networks, the control time dominates the latency of the system. Therefore, we must reduce the control time, and thus the latency, to fully utilize the high bandwidth of these networks.

Ideally an interconnection network would provide all possible N^2 paths between every pair of processors. Since all paths are available at all times, control time would be reduced to zero. While optical and optoelectronic networks provide enormous temporal and spatial bandwidth, full interconnect is still an unrealistic expectation for large networks. Therefore, we have been investigating shared resource networks, which can provide only a subset of all possible paths at any given time.

There are two models for sharing the interconnection resource: explicit addressing and reconfiguration. Addressing allows messages to use a common path (e.g., a bus) but additionally requires arbitration to determine which source gets access to that path. Reconfiguration of the network, while seemingly more complex, has the advantage of solving both arbitration and addressing problems at the same time.

Reconfiguration of the network can be done either by local or global control. If local decisions are made at switches throughout the network, the message must be received, decoded, and used to control the switch. These operations can be done in parallel for different nodes in the network; however, each control operation adds to the total delay time of the individual message.

In the domain of optoelectronic reconfigurable networks, global control has two advantages. First, we can make the routing decisions once for each message, and second we can use global information to optimize the allocation of the resources of the network. Additionally, we can use the *locality* of the message traffic to reduce the number of routing decisions which have to be made. By reducing the number of decisions, we can reduce the control time, and thus reduce the average

latency for messages in the network. The details of this technique is the focus of the next section.

While it is clear that locality can increase the effective throughput of a reconfigurable interconnection network, it is not clear what impact this has on the scalability of these systems. We know that locality has an effect on relative sizes of virtual and physical memory resources. In this study we will determine if the same principle holds in communication resources. In other words, we want to answer the question: *Is it possible to scale communications systems in the virtual domain, while using a limited* (or even fixed) *set of physical resources?*

Reconfiguration Latency

For any path through a reconfigurable communications network the end to end latency, t_{path}, can be characterized as follows:

$$t_{path} = t_{control} + t_{launch} + t_{fly}. \tag{16}$$

In other words, latency is the sum of the time necessary to establish a path, launch a message into the channel, and propagate that message for the length of the channel. For circuit switched systems each of these terms are single values. For packet switched, or multi-hop, the end-to-end latency is simply the sum over each hop. Thus, if all three of the control, launch, and propagation phases of message transmission operate sequentially, the maximum message throughput, T_{path}, is given by:

$$T_{path} = 1/t_{path} = 1/(t_{control} + t_{launch} + t_{fly}). \tag{17}$$

In high bandwidth networks these operations are commonly overlapped in a pipeline fashion and, thus, the maximum throughput is:

$$T_{path} = 1/max(t_{control}, t_{launch}, t_{fly}), \tag{18}$$

in other words, the inverse of the longest of the pipeline stages. If the transmission media is capable of holding multiple messages, then we need only consider the maximum of launch time and control time:

$$T_{path} = 1/max(t_{control}, t_{launch}). \tag{19}$$

For both electronic and optical networks, control time is by far the dominating term. Electronic network designers have attempted to reduce this term by distributing control through the network, thereby performing multiple control operations in parallel[Sie84]. Each of these operations requires that an intermediate control node receives, decodes, and makes a routing decision based on a portion of the incoming message in real time. Thus, distributed control ultimately places an upper limit on network bandwidth given by the ability of the the local controllers to act on the address of the message without introducing delay.

To avoid this limit, high bandwidth optoelectronic networks must implement distributed control by a store and forward of each message at the control points. This correspondingly increases the fly time and therefore can only be justified in wide area networks. In such networks, fly time is already large and the latency of collecting global control information is prohibitive. For small networks such as a multiprocessor interconnect, fly time is typically a few multiples of the launch time. Collecting global information is less difficult and the latency cost of store and forward routing is high.

Since a central control unit establishes the entire path in a single operation, control time no longer places an upper limit on bandwidth. As bandwidth increases, launch times become smaller and the result is that much of the available bandwidth in the network is wasted. One solution is to simply increase the message size, which means transmitting more bits per control operation. However, this solution is limited in multiprocessor applications where message traffic is characterized by bursts of small messages, typically a few words in length.

Specifically, consider an arbitrary reconfigurable network using a centralized controller to arbitrate access to m channels. From equation 19 we know that the throughput of any end-to-end connection is inversely proportional to the maximum of the control and launch times. Since the overall network is capable of launching m packets simultaneously, the network throughput is:

$$T_{network} = m/max(t_{control}, t_{launch}). \tag{20}$$

A network is considered to be saturated when the rate of packets entering the network is equal to $T_{network}$.

For purposes of this analysis, we consider that all resources in the network are shared equally by all packet sources. Therefore, we ignore topologically induced wait times. Allows a simple definition for the packet arrival rate, T_{packet}, from n processors as a linear function of the average packet generation rate per processor, $G_{processor}$:

$$T_{packet} = n \times G_{processor}. \tag{21}$$

By combining these expressions, the saturation point of the network can be expressed in terms of the number of processors, n, as:

$$nG_{processor} = m/max(t_{control}, t_{launch}). \tag{22}$$

Therefore, for any network with a fixed set of resources, system scalability in the number of processors is an inverse function of the control time. This function is bounded at zero for infinite control time and at m/t_{launch} for $t_{control} \leq t_{launch}$.

Clearly, since m/t_{launch} is an expression of the bandwidth capacity of a network, equation 22 tells us that increasing network bandwidth will not lead to a proportionate increase in the number of processors that can be supported. Only by reducing the control time can we effectively increase the number of processors supportable by the network. Further, since the function is bounded by a maximum which is a function of bandwidth, an incremental reduction in control time will have a larger effect for higher bandwidth systems.

In this proposal we have introduced a routing paradigm based on locality of message traffic. We divide the message traffic into two classes of messages, those messages that follow the same path as their immediate predecessors and those that require the establishment of new paths. The former are local messages and the latter are nonlocal messages. The message paths are established by a central controller, and as argued above, the ability of such a controller to process connection requests determines the throughput of the entire interconnection system.

Assuming that P_c is the probability that a connection request requires the establishment of a new path (non-local messages), then $1 - P_c$ is the probability that a connection request does not

require the establishment of a new path (local messages). Hence, the effective rate by which the central controller can service connection requests is:

$$t_{control} = t_{non-local} \times P_c + t_{local} \times (1 - P_c). \tag{23}$$

where t_{local} and $t_{non-local}$ represent the control time for local and and non-local messages respectively. For the state sequence algorithm, non-local control involves a sequence replacement while local control is essentially a no-op. Hence the maximum bandwidth of the central controller and thus the maximum throughput of the network is approximately:

$$T_{network} = 1/t_{local} \times (1 - P_c). \tag{24}$$

In other words, higher throughputs are obtained when P_c is smaller, that is when the probability that a message will require a sequence transformation is low. This probability, P_c, depends on three factors: the actual message locality, the relative time between successive local messages, and the probability that a sequence slot will be preempted. The first factor is a property of the application, as is the second which is additionally dependent on process granularity and the relative bandwidth of the processors and the network. The third factor is a function of the network traffic, the network topology, the sequence length and the global sequence transformation algorithm.

Although an analytical model that considers all these factors is extremely complex, our preliminary simulation results [QMCL] show that high effective network bandwidth may be obtained by applying our control algorithms because of the inherent locality exhibited in the communication patterns of most practical applications. These simulations were conducted with synthetic loads, using assumptions of normally distributed traffic patterns. The benefit of this research will be to more accurately characterize P_c using experimental data from a prototype system.

To summarize, recall that in section 3.7 we defined the set of interconnections between N communicating processes as a virtual connection space. Within this space a reconfigurable network dynamically provides the necessary paths to support the actual traffic. To the extent that we can use locality to increase the ability of the network to support a larger number of processors without an increase in the physical resources, we can achieve scaling in the virtual domain.

Conclusions and Future Research

This paper has presented a mechanism by which we can exploit locality to reduce the complexity of control in multiprocessor interconnection structures. We have shown three approaches to network reconfiguration techniques based on the paradigm of state sequence routing. The first technique can be used when a compile time analysis of the application is appropriate. The second and third techniques apply to the more general case of runtime message requests. The performance impact of these techniques for optoelectronic interconnection structures have also been analyzed to show the importance of these techniques for high bandwidth interconnections.

3.8 References

[All90] F.C. Allard. *Fiber Optics Handbook For Engineers and Scientists.* McGraw–Hill, 1990.

[CCCS90] Y.C. Chen, W.T. Chen, G.H. Chen, and J.P. Sheu. Designing efficient algorithms on mesh-connected computers with multiple broadcasting. *IEEE Transactions on Parallel and Distributed Systems*, 1(2):241–245, April 1990.

[CDLM91] 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.

[CLM90a] 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.

[CLM90b] D.M. Chiarulli, S.P. Levitan, and R.G. Melhem. Self routing interconnection structures using coincident pulse techniques. In *SPIE OE/Boston'90*, Boston, MA, November 4-9 1990.

[CML87] D.M. Chiarulli, R.G. Melhem, and S.P. Levitan. Parallel memory using coincident optical pulses. *IEEE Computer*, 20(12):48–57, December 1987.

[Dou91] P. Doud. High performance interprocessor communication through optical wavelength division multiple access channels. In *Proceedings of the 18th Symposium on Computer Architecture*, volume ACM 19(3), May 1991.

[DP78] A.M. Despain and Patterson. X-tree: A tree structured multi-processor computer architecture. In *5th International Symposium on Computer Architecture*, pages 144–151, 1978.

[GDT+89] C.R. Giles, E. Desurvire, J.R. Talman, J.R. Simpson, and P.C. Becker. 2-gbit/s signal amplification at $\lambda = 1.53\mu m$ in a erbium-doped single-mode fiber amplifier. *Journal of Lightwave Technology*, 7(4):651–656, April 1989.

[GM90] Zicheng Guo and Rami Melhem. Embedding pyramids in array processors with piplined busses. In *Intl. Conf. on Application Specific Array Processors*, pages 665–676, Princeton, N.J., 1990.

[GMH+90] Z. Guo, R. Melhem, R. Hall, D. Chiarulli, and S. Levitan. Array processors with pipelined optical busses. In *Frontiers'90: 3rd Symposium on the Frontiers of Massively Parallel Computation*, University of Maryland College Park, MD, October 8-10 1990.

[GMH+91] 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.

[Goo89] J. Goodman. Switching in an optical interconnect environment. In *OSA Proceedings on Photonic Switching*, volume OSA 3, March 1989.

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

[Guo91] Zicheng Guo. *Array Processors with Piplined Busses and Their Implication in Optically and Electronically Interconnected Multiprocessors.* PhD thesis, Department of Electrical Engineering, University of Pittsburgh, 1991.

[LCM90] S.P. Levitan, D.M. Chiarulli, and R.G. Melhem. Coincident pulse techniques for multiprocessor interconnection structures. *Applied Optics*, 29(14):2024-2033, May 10 1990.

[Lev87] Steven P. Levitan. Measuring communication structures in parallel architectures and algorithms. In *The Characteristics of Parallel Algorithms*, pages 101–137. MIT Press, Cambridge, MA, 1987.

[LFR+89] R.I. Laming, M.C. Farries, L. Reekie, D.N. Payne, P.L. Scrivener, F. Fontana, and A. Righetti. Efficient pump wavelengths of erbium-doped fiber optical amplifiers. *Electronic Letters*, 25(1):12-14, January, 5 1989.

[MCL89] R.G. Melhem, D.M. Chiarulli, and S.P. Levitan. Space multiplexing of optical waveguides in a distributed multiprocessor. *The Computer Journal, British Computer Society*, 32(4):362-369, 1989.

[NTM85] M. Nassehi, F. Tobagi, and M. Marhic. Fiber optic configurations for local area networks. *IEEE Journal on Selected Areas in Communications*, SAC-3(6):941-949, November 1985.

[Qia91] Chunming Qiao. *Time-Division Multiplexed Communication in Multiprocessors and its Applications to Optical Interconnection Networks.* PhD thesis, Department of Computer Science, University of Pittsburgh, 1991. Dissertation Proposal.

[QMCL] Chunming Qiao, R.G. Melhem, D.M. Chiarulli, and S.P. Levitan. Reconfiguring optically interconnected networks with time division multiplexing. *Journal of Parallel and Distributed Computing.* (submitted).

[Sie84] Howard Jay Siegel. *Interconnection Networks for Large Scale Parallel Processing: Theory and Case Studies.* Lexington Books, 1984.

[Gie:91] Giles, Randy C. and Desurvire, Emmnuel. Modelling Erbium Doped Fiber Amplifiers. *Journal of Lightwave Technology*, Vol. 9, No.2, (February, 1991), pp. 271-283.

[Raj:90] Liu, Karen and Ramaswami, Rajiv, *Analysis of Optical Bus Networks Using Doped-Fiber Amplifiers*, Yorktown Heights, New York: IBM Research Division, June, 1990.

[Sunak] Bastien, Steven P., Sunak, Harish R. D., Balakrishnan, Sridhar and Kalomiris, Vas, E., *Analysis of Erbium Doped Fiber Amplifiers* Kingston, Rhode Island: University of Rhode Island.

[Lee:86] Lee, Donald L., *Electromagnetic Principles of Integrated Optics* New York: John Wiley & Sons, 1986, pp. 283-288.

[Min:91] Miniscalco, William J., Erbium-Doped Glasses for Fiber Amplifiers at 1550 nm, *Journal of Lightwave Technology*, Vol. 9, No. 2, (February,1991), pp. 234-250.

[Bar:91] Barnes, William J., Laming, Richard I., Tarbox, Eleanor J. and Morkel, P. R., Absorption and Emission Cross Section of Er^{3+} Doped Silica Fibers, *IEEE Journal of Quantum Electronics*, Vol. 27, No. 4, (April, 1991), pp. 1004-1010.

[Agg:91] Poulain, M., Fluoride Glass Composition and Processing, in *Fluoride Glass Fiber Optics*, Aggarwal, Ishwar D., Lu, Grant, ed., San Diego, Academic Press, 1991.

4 Project Publications: Previous, Current and in Preparation

Publications in Preparation

- "Bandwidth as a virtual resource"; D. M. Chiarulli, S. P. Levitan, R. G. Melhem.
- "A Time Domain Approach for Avoiding Crosstalk in Multistage Interconnection Networks"; C. Qiao, R. Melhem, D. M. Chiarulli, S. P. Levitan.
- "A Lossless Electro-Optical Bus using Erbium Amplifiers", M. Bidnurkar, S. Levitan, D. M. Chiarulli, R. Melhem.

Publications (total funding period)

- "Reconfiguring Optically Interconnected Networks with Time Division Multiplexing"; Chunming Qiao, R.G. Melhem, D.M. Chiarulli, S.P. Levitan; (submitted) *Journal of Parallel and Distributed Computing.*
- "Optical Multicasting in Linear Arrays"; Chunming Qiao, R.G. Melhem, D.M. Chiarulli and S.P. Levitan, *International Journal on Optical Computing, Vol. 2*, No. 1, pp. 31-48, April, 1991.
- "Efficient channel allocation for routing in optically interconnected multiprocessor systems"; C.Qiao, R.Melhem, D. M. Chiarulli, S. P. Levitan; SPIE Symposium on OE/Aerospace Sensing'92, Conf. on Advances in Optical Information Processing V; Orlando, Fl.; April 20-24, 1992.
- *Array processors with pipelined busses and their implication in optically and electronically interconnected multiprocessors*; Zicheng Guo; Ph.D. thesis; Department of Electrical Engineering, University of Pittsburgh, 1991.
- "Demonstration of an all optical addressing circuit"; D. Chiarulli, S. Levitan, R. Melhem; Optical Society of America Topical Meeting on Optical Computing; Technical Digest Vol. 6, TuC3-1, pp.235-238; Salt Lake City, UT, March 4-6, 1991.
- "An all optical addressing circuit: experimental results and scalability analysis"; Donald M. Chiarulli, Robert M. Ditmore, Steven P. Levitan, and Rami G. Melhem;*IEEE Journal of Lightwave Technology, Vol. 9*, No. 12, pp. 1717-1725, 1991.
- "Pipelined communications in optically interconnected arrays"; Z. Guo, R.G. Melhem, R.W. Hall, D.M. Chiarulli, and S.P. Levitan; *Journal of Parallel and Distributed Computing, Vol. 12*, No. 3, pp. 269-282, 1991.
- "Multicasting in optical bus connected processors using coincident pulse techniques"; Chunming Qiao, R. Melhem, D. Chiarulli and S. Levitan; International Conference on Parallel Processing; (poster); St. Charles, IL, August 20-23, 1991.

- "Time-division optical communications in multiprocessor arrays"; C. Qiao, R. Melhem; Proc. of the Supercomputing 91 Conference, Albuquerque, NM, November, 1991; IEEE Press (1991).

- "Self routing interconnection structures using coincident pulse techniques;" D.M. Chiarulli, S.P. Levitan, and R.G. Melhem; In *SPIE OE/Boston'90*, 1390-25, S4, pp. 403-414; Boston, MA, November 4-9 1990.

- "Pipelined communications on optical busses"; Z. Guo, R. Melhem, R. Hall, D. Chiarulli, and S. Levitan; In *SPIE OE/Boston'90*, 1390-26, S4, pp 415-426; Boston, MA, November 4-9 1990.

- "Embedding pyramids in array processors with Pipelined busses"; Zicheng Guo and Rami Melhem; In *Intl. Conf. on Application Specific Array Processors*, pages 665–676, Princeton, N.J., 1990.

- "Array processors with pipelined optical busses"; In *Frontiers'90: 3rd Symposium on the Frontiers of Massively Parallel Computation*, Z. Guo, R. Melhem, R. Hall, D. Chiarulli, and S. Levitan; pp.333-324; University of Maryland College Park, MD, October 8-10 1990.

Previous (Related) Publications

- "Coincident pulse techniques for multiprocessor interconnection structures"; S.P. Levitan, D.M. Chiarulli, and R.G. Melhem; *Applied Optics*, 29(14):2024–2033, May 10 1990.

- "Optical bus control for distributed multiprocessors"; D.M. Chiarulli, S.P. Levitan, and R.G. Melhem; *Journal of Parallel and Distributed Computing*, 10:45–54, 1990.

- R.G. Melhem, D.M. Chiarulli, and S.P. Levitan; "Space multiplexing of waveguides in optically interconnected multiprocessor systems", *The Computer Journal, British Computer Society*, 32(4):362–369, 1989.

- "Asynchronous Control of Optical Busses in Closely Coupled Distributed Systems"; Donald M. Chiarulli, Rami Melhem, Steven P. Levitan; Technical Report 88-2, Department of Computer Science, University of Pittsburgh, 1988.

- D.M. Chiarulli, R.G. Melhem, and S.P. Levitan. "Using coincident optical pulses for parallel memory addressing", *IEEE Computer*, 20(12):48–57, December 1987.

5 Project Personnel

5.1 Current Vita of Principal Investigators

- Donald M. Chiarulli, Co-PI
- Rami G. Melhem, Co-PI
- Steven P. Levitan, Co-PI

Curriculum Vitae

Donald M. Chiarulli

Department of Computer Science
University of Pittsburgh
Pittsburgh, PA 15260
412-624-8839
INTERNET: don@cs.pitt.edu

Research Interests

Architectures and algorithms for the design and implementation of highly parallel computing systems are the focus of my research. The most recent work concentrates on the development of optical technologies for interconnection networks and processing. Additional interests include algorithms for the simulation and verification of systems implemented in VLSI and methodologies for testing these systems.

Education

Ph. D. in Computer Science, 1986, Louisiana State University, Baton Rouge.
Dissertation: "A Horizontally Reconfigurable Architecture for Extended Precision Integer Arithmetic."

M. S. in Computer Science, 1979, Virginia Polytechnic Institute, Blacksburg.
Thesis: "An Automated Personnel Identification System for the VPI Computing Center."

B. S. in Physics, 1976, Louisiana State University, Baton Rouge. Minor: Chemistry

Academic Experience

Associate Professor, 1992-Present, Department of Computer Science, University of Pittsburgh, Pittsburgh, Pennsylvania.

Assistant Professor, 1986-1992, Department of Computer Science, University of Pittsburgh, Pittsburgh, Pennsylvania.

Instructor/Research Associate, 1979-1986, Department of Computer Science, Louisiana State University, Baton Rouge, Louisiana.

Research Assistant, 1977-1979, Hardware and Network Services, Virginia Polytechnic Institute Computing Center, Virginia Polytechnic Institute, Blacksburg, Virginia.

Current and Completed Research Grants:

"Reconfigurable Optoelectronic Multiprocessor Interconnection Structures" Air Force Office of Scientific Research, October 1992-January 1996, $575,842, (Co-PI with Rami Melhem and Steven Levitan).

"Coincident Pulse Techniques for Hybrid Optic.. electronic Computer Systems" Air Force Office of Scientific Research, July 1989-July 1992, $479,511, (Co-PI with Rami Melhem and Steven Levitan).

"Parallel Memory Addressing Using Optical Pulse Delay Modulation" Air Force Office of Scientific Research, July 1988-July 1989, $50,132. (Co-PI with Rami Melhem and Steven Levitan).

"Optical Technology for Network Based Multiprocessors" NSF/MIPS, October 1988-October 1989, $49,983, (Co-PI with Rami Melhem and Steven Levitan).

"A VLSI Design and Test Facility for the University of Pittsburgh, NSF, CISE Research Instrumentation, January 1988, $65,597. (Co-PI with Steven P. Levitan).

"A Hig. Performance Computer for Factoring Large Numbers," National Security Agency MDA904-85-H-0006, February 1985 February 1987, $198,296. (Research Associate with Walter G. Rudd).

"A High Performance Computer for Factoring Large Numbers" (equipment grant), National Science Foundation DCR 83-115-80, April 1985, $40,261 (Research Associate with Walter G. Rudd).

"Provision of Aquifer and Subsurface Database System", Louisiana Department of Natural Resources, Interagency Contract #21541-80-01, April 1980 - September 1980, $67,000 (Investigator with Walter G. Rudd).

T. avel Grants

SIGDA, 28th Design Automation Conference, San Francisco, CA, June 1991.

Industrial and Internal Grants

"A Microcomputer Lab for the Department of Computer Science", (equipment grant) NCR Corporation, $100,000, December 1988.

"A Medium Speed Print Server for the Computer Science Department and the NCR Microcomputer Lab", University of Pittsburgh Provost's Departmental Infrastructure Grant, December 1990, $18,091.

"Terminal Server Replacement on the Computer Science Department LAN" University of Pittsburgh Provost's Departmental Infrastructure Grant, December 1990, $30,798.

Pending Grants

"Novel Statistical and Computational Algorithms for Genetic Mapping", project within "A Human Genome Research Center for the Mapping of Chromosome 13", National Institutes of Health, five year project budget 1,136,485. (Co-PI with D. Weeks and H. Nicholaus).

"Optical Bandwidth as a Virtual Resource for Multiprocessor Interconnections", DARPA/MTO, three year project budget 1,452,927, (Co-PI with Steven Levitan and Rami Melhem).

Patents

Processor Utilizing Reconfigurable Process Segments, Co-inventors: W. G. Rudd, and D. A. Buell, Reg #4748585, May 1988.

An Optical Selector Switch, Co-inventors: R. Melhem, and S. Levitan, Reg #4883344, September 1988.

Refereed Publications

Chunming Qiao, R.G. Melhem, D.M. Chiarulli, S.P. Levitan, "Reconfiguring Optically Interconnected Networks with Time Division Multiplexing", (submitted) *Journal of Parallel and Distributed Computing*

C. Qiao, R.G. Melhem, D.M. Chiarulli, and S.P. Levitan, "Efficient Channel Allocation for Routing in Optically Interconnected Multiprocessor Systems", SPIE Symposium on OE/Aerospace Sensing'92, Conference on Advances in Optical Information Processing V, April 1992.

D. M. Chiarulli, R. Ditmore, S. Levitan, and R.G. Melhem, "An All Optical Addressing Circuit: Experimental Results and Scalability Analysis", *IEEE Journal of Lightwave Technology*. Vol. 9, No. 12, 1991.

C. Qiao, R.G. Melhem, D.M. Chiarulli, and S.P. Levitan, "Optical Multicasting in Linear Arrays", *International Journal of Optical Computing*. Vol. 2, No. 1, 1991

Z. Guo, R. Melhem, R. Hall, D. Chiarulli, S. P. Levitan. "Pipelined Communications in Optically Interconnected Arrays", *Journal of Parallel and Distributed Computing*, Vol. 12, No. 3, 1991.

D. M. Chiarulli, S. Levitan and R. Melhem, "Demonstration of an All Optical Addressing Ciruit" Technical Digest: OSA Topical Meeting on Optical Computing, 1991.

C. Qiao, R.G. Melhem, D.M. Chiarulli, and S.P. Levitan, "Multicasting in Optical Bus Connected Processors Using Coincident Pulse Techniques", *(Poster) International Conference on Parallel Processing*, August 1991.

D. M. Chiarulli, R. Melhem, S. P. Levitan, "Optical Bus Control for Distributed Multiprocessors" *Journal of Parallel and Distributed Computing* Vol. 10, No. 1, September 1990.

Z. Guo, R. Melhem, R. Hall, S. Levitan and D. Chiarulli, "Array Processors with Pipelined Optical Buses," *Proceedings* Frontiers 90 Conference on Massively Parallel Computation, October 1990.

D. M. Chiarulli, S. Levitan and R. Melhem, "Self Routing Interconnection Structures Using Coincident Pulse Techniques," *Proceedings* SPIE International Symposium on Advances in Interconnections and Packaging, November 1990.

Z. Guo, R. Melhem, R. Hall, S. Levitan and D. Chiarulli, "Pipelined Communications on Optical Busses", *Proceedings* SPIE International Symposium on Advances in Interconnections and Packaging, November 1990.

A. Martello, S.P. Levitan, D. Chiarulli, "Timing Verification Using HDTV", *Proceedings* Design Automation Conference, 1990.

S.P. Levitan, D. Chiarulli, R. Melhem, "Coincident Pulse Techniques for Multiprocessor Interconnection Structures", *Applied Optics* Vol. 29, May 1990.

R. Melhem, D. Chiarulli, S.P. Levitan, "Space Multiplexing of Pptical Waveguides in a Distributed Multiprocessor" *The Computer Journal, British Computer Society*, Vol. 3, No. 4, 1989.

D. M. Chiarulli, R. Melhem, S. P. Levitan, "Parallel Memory Addressing Using Coincident Optical Pulses", *IEEE Computer*, Vol. 30, No. 12, December 1987.

M. Martin, D. Chiarulli, and S. Iyengar, "Parallel Processing of Quadtrees on a Horizontally Reconfigurable Architecture Computing System, *Proceedings*, 15th International Conference on Parallel Processing, Chicago, Illinois, August 1986, 895-902.

D. M. Chiarulli, W. G. Rudd, and D. A. Buell, "A Hierarchical Condition Code Structure for Parallel Architectures", SIAM Conference on Parallel Processing for Scientific Computing, Norfolk, Virginia, November 1985.

D. M. Chiarulli and D. A. Buell, "Parallel Microprogramming Tools for a Horizontally Reconfigurable Architecture," *International Journal of Parallel Programming*, Vol 15, No. 2, May 1987.

D. M. Chiarulli, W. G. Rudd, and D. A. Buell, "DRAFT--A Dynamically Reconfigurable Processor for Integer Arithmetic," *Proceedings*, 7th International Symposium on Computer Arithmetic, Urbana, Illinois, June 1985, 309-317.

W. G. Rudd, D. A. Buell, and D. M. Chiarulli, "A High Performance Factoring Machine," *Proceedings*, 11th Annual International Symposium on Computer Architecture, Ann Arbor, Michigan, June 1984, 297-300.

Curriculum Vitae

Steven Peter Levitan

Department of Electrical Engineering
Benedum Engineering Hall
University of Pittsburgh
Pittsburgh, PA 15261

PROFESSIONAL INTERESTS

The design, modeling, and simulation of highly parallel systems, including parallel computer architectures, parallel algorithms, and VLSI. Additional interests include design tools and methodology for software, hardware, and VLSI.

CURRENT POSITION

Wellington C. Carl Associate Professor in the Department of Electrical Engineering at the University of Pittsburgh.

EDUCATION

Ph. D. May 1984 University of Massachusetts Department of Computer and Information Science (COINS). Dissertation title: "Parallel Architectures and Algorithms: A Programmer's Perspective". Advisor: Caxton C. Foster.
M. S. September 1979 University of Massachusetts, (COINS), Specialization: Computer Systems.
B. S. June 1972 Case Western Reserve University, School of Engineering. Major: Computer Science, Minor: Electrical Engineering

PROFESSIONAL POSITIONS HELD

Wellington C. Carl Assistant Professor, 1987-1992: Department of Electrical Engineering, University of Pittsburgh.
Assistant Professor, 1984-1986: Department of Electrical and Computer Engineering (ECE), University of Massachusetts, Amherst.
Research/Teaching Assistant/Lecturer, 1977-1984: Department of Computer and Information Science (COINS), University of Massachusetts, Amherst.
Senior Systems Engineer, 1972-1977: Xylogic Systems Inc.

ACADEMIC AWARDS

Wellington C. Carl Faculty Fellowship: 1987-1993
IEEE Computer Society Distinguished Visitor:1989-1994
Regents' Doctoral Fellowship: 1984

PATENTS

An Optical Selector Switch, (with R. Melhem, and D. Chiarulli), Approved September 1988, Number 4,883,334.

CURRENT AND PENDING GRANTS

National Science Foundation, 5/92-5/95, "A Research Experiences for Undergraduates Site: Training Students to Model Polymer Behavior Through Computer Simulations"; $150,000 DMR-9200174 (CI) with A.C. Balazs (PI), and R.L. Pinkus.

National Science Foundation, 1/92-12/94, "Temporal Specification Verification"; $218,292 (PI) MIP-9102721.

Association for Computing Machinery - SIGDA, 12/92-6/93, ACM/SIGDA "Creation of a SIGDA Internet Server"; $23,834 (PI).

National Science Foundation, 7/91-7/93, "Distribution of VLSI Design Software for Education and Research"; $97,618 (PI) MIP-9101656.

National Institute of Mental Health (ADAMHA), 4/90-3/93 "Contribution of PCP and NMDA Receptors to Network Properties of the Hippocampal Formation"; $600,000 (CI) with T. W. Berger(PI), R. J. Sclabassi(CI), G. Barrionuevo, D. N. Krieger.

Air Force Office of Scientific Research, 7/89-7/92, "Coincident Pulse Techniques for Hybrid Electronic/Optical Computer Systems"; $479,511 (CO-PI) with D.M. Chiarulli, R. Melhem; AFOSR-89-0469.

Office of Naval Research, 6/87-5/90 "Changes in Neuronal Network Properties Induced by Learning and Synaptic Plasticity: A Nonlinear Systems Approach"; $531,943 ((CI) with T.W. Berger(PI), R.J. Sclabassi, G. Barrionuevo, D.N. Krieger). N00014-87-K-0472 / N00014-90-J-4000.

Air Force Office of Scientific Research, 7/92-7/95, "Reconfigurable Opto/Electronic Multiprocessor Interconnection Structures"; $655,690 (CO-PI) with D. M. Chiarulli, R. G. Melhem (Pending).

National Science foundation "A Systems Theoretic Approach to Neural Network Function", $356,257, 1/1/93-12/31/95 Co-PI with Robert J. Sclabassi, Donald J. Weisz (Pending).

National Science foundation "An Integrated Engineering Design Sequence", $352,601, 9/1/92-8/31/95, Co-PI with Ronald Hoelzeman, James Cain, Dorothy Setliff (Pending).

PUBLICATIONS (Journal Articles and Book Chapters)

1. "SPAR: A Schematic Place and Route System"; Stephen T. Frezza and Steven P. Levitan; (in press) *IEEE Transactions on Computer Aided Design of Integrated Circuits.*
2. "Optical Multicasting in Linear Arrays"; Chunming Qiao, R.G. Melhem, D.M. Chiarulli and S.P. Levitan, *International Journal on Optical Computing, Vol. 2,* No. 1, pp. 31-48, April, 1991.
3. "An All Optical Addressing Circuit: Experimental Results and Scalability Analysis"; Donald M. Chiarulli, Robert M. Ditmore, Steven P. Levitan, and Rami G. Melhem;*IEEE Journal of Lightwave Technology, Vol. 9,* No. 12, pp. 1717-1725, 1991.
4. "Pipelined Communications In Optically Interconnected Arrays"; Z. Guo, R.G. Melhem, R.W. Hall, D.M. Chiarulli, and S.P. Levitan; *Journal of Parallel and Distributed Computing, Vol. 12,* No. 3, pp. 269-282, 1991.

5. "An Interactive Toolset for Characterizing Complex Neural Systems"; D.N. Krieger, T.W. Berger, S.P. Levitan, and R.J. Sclabassi; *Computers and Mathematics, Vol. 20,* Mathematic·: Models in Medicine, No.4-6, pp. 231-?46, 1990.
6. "Coincident Pulse Techniques for Multiprocessor Interconnection Structures"; S.P. Levitan, D.M. Chiarulli, R.G. Melhem; *Applied Optics; Vol. 29,* No. 14, pp. 2024-2033, May, 1990.
7. "Optical Bus Control for Distributed Multiprocessors"; D.M. Chiarulli, S.P. Levitan, R.G. Melhem; *Journal of Parallel and Distributed Computing; Vol. 10,* No. 1, pp. 45-54, 1990.
8. "Nonlinear Systems Analysis of Network Properties of the Hippocampal Formation"; T.W. Berger, G. Barrionuevo, S.P. Levitan, D.N. Krieger, and R.J. Sclabassi; pp. 283-352; *Learning and Computational Neuroscience: Foundations of Adaptive Networks,* M. Gabriel and J. W. Moore (Eds.), M.I.T. Press, 1990.
9. "Space Multiplexing of Optical Waveguides in a Distributed Multiprocessor"; R.G. Melhem, D.M. Chiarulli and S.P. Levitan; *The Computer Journal, British Computer Society, Vol. 32,* No. 4, pp. 362-369, 1989.
10. "The Image Understanding Architecture"; C. C. Weems, S. P. Levitan, A. R. Hanson, E. M. Riseman, J. G. Nash, D. B. Shu; *International Journal of Computer Vision Vol. 2,* pp. 251-282 (1989).
11. "Theoretical Decomposition of Neuronal Networks"; R.J. Sclabassi, D.N. Krieger, J. Solomon, J. Samosky, S.P. Levitan, and T.W. Berger; (in) *Advanced Methods of Physiological System Modeling, Vol. 2,* V.Z. Marmarelis (Ed.), pp. 129-146, Plenum Press, New York, 1989.
12. "Using VHDL as a Language for Synthesis of CMOS VLSI Circuits"; S.P. Levitan, A.R. Martello, R.M. Owens, M.J. Irwin; (in) *Computer Hardware Description Languages and their Applications* J.A. Darringer and F. J. Ramming, Eds.; Elsevier, Amsterdam, 1989; pp. 331-346; IFIP WG 10.2, 9th Intl. Symp. on Computer Hardware Description Languages; Washington D.C., June, 1989.
13. "Using Coincident Optical Pulses for Parallel Memory Addressing"; D. Chiarulli, R. Melhem, S. Levitan; *Computer* Vol. 20, No. 12, pp. 48-57. December, 1987.
14. "Measuring Communication Structures in Parallel Architectures and Algorithms"; Steven P. Levitan (in) *The Characteristics of Parallel Algorithms;* L. Jamieson, D. Gannon, and R. Douglass (Eds.), Cambridge, MA; MIT Press, 1987; pp. 101-137.
15. "The UMass Image Understanding Architecture"; Steven P. Levitan, Charles C. Weems, Allen R. Hanson and Edward M. Riseman; (in) *Parallel Computer Vision;* Leonard Uhr (Ed.). Academic Press, New York, 1987; pp. 215-248.
16. "Signal to Symbols: Unblocking the Vision Communications/Control Bottleneck"; Steven P. Levitan, Charles C. Weems, Edward M. Riseman; (in) *VLSI Signal Processing* (proceedings of the 1984 IEEE Workshop on VLSI Signal Processing at University of Southern California, Los Angeles, CA; November 27-29, 1984); IEEE Press; New York, NY; 1984; pp. 411-420.
17. "A Content Addressable Array Parallel Processor and Some Applications"; Charles C. Weems, Steven P. Levitan, Daryl T. Lawton, and Caxton C. Foster; (in) *Image Understanding,* Proceedings of the DARPA Workshop, Arlington, Virginia; June 23, 1983; Science Applications, Inc. Report Number SAI-84-176-WA.

Vita of Rami G. Melhem

EDUCATION:

1983	Ph.D.	Computer Science, University of Pittsburgh
1981	M.S.	Computer Science, University of Pittsburgh
1981	M.A.	Mathematics, University of Pittsburgh
1978	B.S.	Mathematics, Ein-Shams University, Cairo, Egypt
1976	B.S.	Electrical Engineering, Cairo University, Egypt

PROFESSIONAL ACTIVITIES:

Editor	IEEE Transactions on Computers,
Guest editor	J. of Parallel and Dist. Comp.- Special issue on Optical Comp. - 1993.
Chairman	Prog. committee - ISMM Conf. on Parallel and Dist. Comp. & Sys. - 1992,
Member	Prog. committee - Int. Conf. on Application Specific Array Processors - 1991.
	Prog. committee - ISMM Conf. on Parallel and Dist. Comp. & Sys. - 1991.
	Prog. committee - Int. Workshop on Defect/Fault Tolerance in VLSI - 1992.
Organizer	Symposium on PCCG methods and Supercomputing - 1989.

PROFESSIONAL EXPERIENCE:

1984	Research Associate, University of Pittsburgh (January-September)
1984-1987	Assistant Professor of Computer Science, Purdue University (on leave from Sept. 1985 to Sept. 1987)
1986-1989	Assistant Professor of Computer Science, University of Pittsburgh
1989-	Associate Professor of Computer Science, University of Pittsburgh

GRANT AWARDS:

ONR:	"Application of Comp. Networks and Systolic Arrays to Scientific Computation". With W. C. Rheinboldt. $233,402.00. June 1985 to September 1988.
AFOSR:	Investigator, (C. Hall and T. Porsching, principal Investigators); "Computational Fluid Dynamics at the Institute for Comp. Math. & Applications" $587,858.00, June 1984 to June 1987.
AFOSR:	"Parallel Memory Addressing Using Coincident Optical Pulses". With D. Chiarulli and S. Levitan. $50,132.00. July 1988 to July 1989.
NSF:	"Optical Technology in Network Based Multiprocessors". With D. Chiarulli and S. Levitan. $49,983.00. July 1989 to June 1990.
AFOSR:	"Coincident Pulse Techniques for Hybrid Optical-Electronic Computer Systems". With D. Chiarulli and S. Levitan. $479,511.00. August 1989 to July 1992.
NSF:	"Bi-level Reconfigurations of Fault Tolerant Arrays in Bi-modal Environments".

$61,547.00. September 1989 to August 1991.

NSF: "CISE Research Instrumentation grant for the acquisition of an Intel Hypercube". With M.L. Soffa and T. Znati. $124,300.00. March 1990 to March 1991.

Hewlett-Packard Labs: "Real-time Protocols for Multimedia Applications". With T. Znati and R. Sclabassi. $28,218.00. May 1992 - Sept. 1993.

AFOSR: "Reconfigurable Opto/Electronic Multiprocessor Interconnection Structures". With D. Chiarulli and S. Levitan. 575,276.00. Oct. 1992 to Jan. 1996.

PUBLICATIONS IN ARCHIVED JOURNALS:

1) R. G. Melhem and W. C. Rheinboldt, "A Comparison of Methods for Determining Turning Points of Non-linear Equations", *Computing*, vol. 29, pp.201-226, (1982).

2) R. G. Melhem and W. C. Rheinboldt, "A Mathematical Model for the Verification of Systolic Networks", *SIAM Journal on Computing*, vol. 13, no. 3, pp. 541-565, (1984).

3) R. G. Melhem, "On the Design of a Pipelined/Systolic Finite Element System", *Computers and Structures*, vol. 20, pp.67-75, (1985).

4) R. G. Melhem, "Formal Analysis of a Systolic System for Finite Element Stiffness Matrices", *Journal of Computer and System Sciences*, vol. 31, no. 1, pp. 1-27, (1985).

5) R. G. Melhem, "A Study of Data Interlock in Computational Networks for Sparse Matrix Multiplication", *IEEE Transactions on Computers*, vol 36, no 9, pp.1101-1107, (1987).

6) R. G. Melhem, "Parallel Gauss/Jordan Elimination for the Solution of Dense Linear Systems". *Parallel Computing*, vol 4, no 3, pp.339-343, (1987).

7) R. G. Melhem, "Determination of Stripe Structures for Finite Element Matrices", *SIAM Journal on Numerical Analysis*, vol 24, no 6, pp.1419-1433, (1987).

8) R. Melhem, "Toward Efficient Implementations of PCCG Methods on Supercomputers", *The Int. Journal on Supercomputer Applications*, vol 1, no 1, pp.71-98, (1987).

9) D. Chiarulli, R. Melhem and S. Levitan, "Using Coincident Optical Pulses for Parallel Memory Addressing", *IEEE Computer*, vol 20, no 12, pp.48-58, (1987).

10) R. G. Melhem, "Verification of a Class of Self-timed Computational Networks", *BIT*, vol 27, no 4 (1987), pp.480-500.

11) R. G. Melhem, "Parallel Solution of Linear Systems with Striped, Sparse Matrices", *Parallel Computing*, vol 6, no 2, pp. 165-184, (1988).

12) R. G. Melhem, "A Modified Frontal Technique Suitable for Parallel Systems", *SIAM J. on Scientific and Statistical Computing*, vol 9, no 2 (1988), pp. 289-303.

13) K. Ramarao, R. Daley and R. Melhem, "Message Complexity of the Set Intersection Problem", *Information Processing Letters*, vol 27, no 4, pp.169-174 (1988).

14) R. Melhem and K. Ramarao, "Multicolor Ordering of Sparse Matrices Resulting from Irregular Grids", *ACM Tran. on Mathematical Software*, vol 14, no 2, pp. 117-138 (1988).

15) R. Melhem and C. Guerra, "The Application of a Sequence Notation to the Design of Systolic Computations", *BIT*, vol 29, no 3, pp. 409-427 (1989).

16) R. Melhem, "A Systolic Accelerator for the Iterative Solution of Sparse linear systems", *IEEE Trans. on Computers*, vol 38, no 11, pp.1591-1595 (1989).

17) R. Melhem, D. Chiarulli and S. Levitan, "Space Multiplexing of Waveguides in Optically Interconnected Multiprocessors", *The Computer Journal*, vol 32, pp. 362-369 (1989).

18) C. Guerra and R. Melhem, "Synthesis of Systolic Algorithm Designs", *Parallel Computing*, vol 12, no. 2, pp. 195-207 (1989).

19) S. Levitan, D. Chiarulli and R. Melhem, "Coincident Pulse Techniques for Multiprocessor Interconnection Structures", *Applied Optics*, vol 29, pp. 2024-2033, (1990)

20) Y. Pan and R. Melhem, "Short Circuits in Buffered Multi-stage Interconnection Networks". *The Computer Journal*, vol 33, no. 4, pp. 323-329 (1990).

21) R. Melhem and G. Hwang, "Embedding Rectangular Grids into Square Grids with Dilation Two". *IEEE Transactions on Computers*, vol. 39, no. 12, pp. 1446-1455, (1990).

22) D. Chiarulli, S. Levitan and R. Melhem, "Optical Bus Control for Distributed Multiprocessors". *The Journal of Parallel and Distributed Computing*, vol. 10, no. 1, pp. 45-54 (1990).

23) M. Alam and R. Melhem, "An Efficient Spare Allocation Scheme and its Application to Binary Hypercubes". *IEEE Trans. on Parallel and Dist. Sys.* vol 2, no 1, pp. 117-126 (1991).

24) Z. Guo, R. Melhem, R. Hall, S. Levitan and D. Chiarulli, "Pipelined Communication in Optically Interconnected Arrays", *J. of Parallel and Dist. Comp.* vol 12, no 3, (1991).

25) D. Chiarulli, R. Ditmore, S. Levitan and R. Melhem, "An All Optical Addressing Circuit: Experimental Results and Scalability Analysis", *IEEE J. of Lightwave Tech.* vol. 9, no. 12, (1991).

26) C. Qiao, R. Melhem, S. Levitan and D. Chiarulli, "Optical Multicasting in Linear Arrays", *International Journal of Optical Computing*, vol 2, no. 1, pp 31-48 (1991).

27) F. Provost and R. Melhem, "A Dist. Alg. for Embedding Trees in Hypercubes", *Journal of Parallel and Distributed Computing*, vol. 14, no. 1, pp. 85-89, (1992).

28) R. Melhem, "Bilevel Reconfigurations of Fault Tolerant Arrays", *IEEE Trans. on Computers*, vol 41, no 2, pp. 231-239 (1992).

29) C. Qiao and R. Melhem, "Time-Division Optical Communications in Multiprocessor Arrays", Accepted for publication in

5.2 Students Funded During Current Period

- Bolanile Onodipe (Ph.D. April 1990) Thesis Title: Analytical Model for the Effect of Transverse Magnetic Fields on GaAs MESFETS. Supported: September 1989 - April 1990.
- Zicheng Guo (Ph. D. May 91) Thesis Title: Array Processors with Pipelined Busses and Their Implication in Optically and Electronically Interconnected Multiprocessor Architectures. Supported: Summer of 1989, Jan-Dec 1990.
- Robert Ditmore (M.S. May 1991) Thesis Title: Analysis of Power Distribution in Optical Buses. Supported: September 1989 - May 1991.
- Chunming Qiao (Ph.D.) Expected completion date: April 1993. Topic: Optically Interconnected Networks with Time and Space Multiplexed Communication. Supported: September 1990 - August 1992.
- David George (M.S. April 1991) Topic: Synthesis of Asynchronous Finite State Machines. Supported: Summer of 1990.
- Tom George (M.S. April 1991) Topic: Extended Simulation Models for VHDL. Supported: Summer of 1990.
- Manoj Bidnurkar (M.S. September 1992) Topic: Erbium Doped Fiber Amplifiers and Optical Buses. Supported: January 1992 - August 1992.
- James Tezza (B.S.) (M.S) Expected completion date: December 1992. Topic: Power Distribution in Lossless Tapped Erbium Doped Fiber Busses for Multiprocessors. Supported: Spring 1991 - August 1992.
- Michael Bigrigg (M.S.) Expected completion date: August 1993. Topic: Analysis of Message Locality in Multiprocessor Interconnect. Supported: Spring 1992 - August 1992.

6 Project Interactions

6.1 Conferences and Workshops

- Levitan attended a DARPA Computing Systems Technology P.I. meeting. Daytona Beach, FL, September, 1992.
- Qiao attended the Frontiers'92 Conference, October 1992, Washington, D.C.
- Qiao attended the International Conference on Parallel Processing, August, 1992
- Melhem attended and presented at SPIE Symposium on OE/Aerospace Sensing'92, Conf. on Advances in Optical Information Processing V; Orlando, Fl.; April 1992.
- Chiarulli and Levitan attended and presented at the AFOSR sponsored workshop on Reconfigurable Optical Interconnects, Boulder, CO, March, 1992.
- Levitan and Chiarulli organized and chaired a panel discussion on "The Future of Optics in Computing" at the November 1991 Supercomputing Conference. Qiao also presented a paper at that conference.
- Melhem and Qiao attended and presented at the International Conference on Parallel Processing, August, 1991.
- Chiarulli and Levitan attended and presented at the Optical Society of America Topical Meeting on Optical Computing Salt Lake City, March, 1991.
- Chiarulli, Guo, and Levitan attended and presented at SPIE Boston/OE'90, November 1990, Boston, MA.
- Guo and Melhem attended and presented at Frontiers'90 Conference, October 1990, College Park, MD.
- Guo and Melhem attended and presented at Intl. Conf. on Application Specific Array Processors, September 1990, Princeton, N.J.
- Chiarulli and Levitan attended and presented at Workshop on Optical Neural Networks, February 7-10, 1990, Jackson Hole, WY.
- Chiarulli, Ditmore, Guo, Levitan, Melhem, and Onodipe attended SPIE OE/Laser Conference January 1990, Los Angles, CA.

6.2 Invited Presentations

- Levitan has been invited to present the group's work at the University of California, San Diego, November, 1992.
- Levitan presented the group's work at the IBM T.J. Watson Research Center, October 1991.
- Levitan presented the group's work to the Department of Electrical Engineering at the University of Pittsburgh, January 1991.
- Chiarulli and Levitan gave invited talks at University of Colorado at Boulder, Optoelectronic Computing Systems Center, Boulder, CO, February 6, 1990.

6.3 Other Interactions

- Levitan met with David Misunas and Richard Otte of the Microelectronics and Computer Technology Corporation (MCC), to discuss possible collaborative efforts. October 2, 1992.

- Prof. Vincent Heuring, from University of Colorado at Boulder, Optoelectronic Computing Systems Center visited the research group at Pittsburgh. March 4, 1992
- Professor H.K. Kim of the Department of Electrical Engineering served on the M.S. Thesis committee for Manoj Bidnurkar.
- Chiarulli and Melhem are Guest Editors for a special issue of the *Journal of Parallel and Distributed Computing* on Optical Computing.
- Levitan has served on the Ph.D. committee's of Brian Telfer, Sanjay Natarajan and John-Scott Smokelin, students of Prof. David Casasent from Carnegie Mellon University.
- Melhem is the program chair, and Levitan is on the program committee of the Fifth ISSM International Conference on Parallel and Distributed Computing and Systems, Pittsburgh, October, 1992.
- Chiarulli ran an Optical Computing graduate seminar in the Department of Computer Science.
- Richard Thompson has agreed to be on the Ph.D. committee of Chunming Qiao. Professor Thompson also gave a talk at the Department of Electrical Engineering.
- We have been interacting with Dr. William Miniscalco from GTE Labs on our work with erbium doped glass fiber. GTE has given us samples of doped fiber for our experiments.
- Professor Harry Jordan from University of Colorado at Boulder. Optoelectronic Computing Systems Center visited the research group at Pittsburgh.
- David George visited the University of Colorado at Boulder. Optoelectronic Computing Systems Center in October 1990.

7 Project New Discoveries

- We have demonstrated an all optical addressing circuit and determined the temporal and physical scalibility limits of the system.
- We have quantified the advantages and general applicability signal pipelining for interconnection networks.
- We have resolved (from both a theoretical and a practical point of view) the issue of shadows in multi-dimensional structures.
- We have established the feasibility of lossless optical buses for distributed multiprocessors.
- We have realized the generalization of signal pipelines to TDM structures and further to SDM and WDM based networks as well.
- We have developed a theoretical framework for the use reconfigurable networks to provide high bandwidth, low latency, cost effective interconnection networks for multiprocessors.

8 Project Evaluation

We have made significant progress with regards to our fabrication of prototype structures to verify the applicability of our ideas to multiprocessor interconnection networks. We believe that our generalization of coincident structures to pipelined structures, and pipelined structures to more general reconfigurable structures will be a significant contribution to the theory and practice of high speed multiprocessor interconnection networks.

A Copies of Several Recent Papers from the Research Group

1. An all Optical Addressing Circuit: Experimental Results and Scalability Analysis *IEEE Journal of Lightwave Technology 9:12, 1991*

2. Pipelined Communications in Optically Interconnected Arrays *Journal of Parallel and Distributed Computing, 12:3, 1991*

3. Optical Multicasting in Linear Arrays *International Journal on Optical Computing*, 2, 31-48 (1991).

An All Optical Addressing Circuit: Experimental Results and Scalability Analysis

Donald M. Chiarulli, *Member, IEEE*, Robert M. Ditmore, Steven P. Levitan, *Member, IEEE*, and Rami G. Melhem, *Member, IEEE*

***Abstract*—In this paper, we present results from a demonstration of both single and parallel selection in a one of four optical addressing circuit operating at 250 MHz using coincident pulse addressing. We then present an analysis of power distribution in two different tapped fiber structures. Based on our results, we discuss issues of scalability with respect to synchronization and power distribution in larger systems.**

I. Introduction

TWO properties of optical signals, unidirectional propagation and predictable path delay, make it possible to devise logic systems in which information is encoded as the relative timing of two optical signals. Coincident pulse addressing is an example of such a system. In this technique, 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.

Stated more formally, consider an optical fiber of length L with two optical pulse sources, P_1 and P_2 coupled to each end. Each source generates pulses of width τ and height h. Define $l = \tau c_g$ 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 2×2 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 $(L - (n + 1)l)/2 + il$, from the left end of the bus. The optical bus in the center of Fig. 1 shows such an arrangement for $n = 3$. 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_r 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_r, then, according to the addressing equation above, to select a specific detector i each t_i will be in the range $-(n - 1)\tau \le t_r - t_i \le (n - 1)\tau$. 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 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. Therefore, the system throughput is $\nu = 1/2\tau$.

In previous papers, we have discussed the general application of coincident pulse techniques to both memory addressing and multiprocessor network applications [2], [3], [7], [8]. In this paper, we emphasize the practical limits on the applicability of this technique for large systems. In order to design large scale computer systems, we need to know the realistic limits on the speed, size, and cost of such systems. Our long term goal is to build high-speed multiprocessor interconnection networks using off the shelf optical components and tapped fiber busses.

Tapped fiber busses, those with one or more transmitter and multiple receivers, have been less widely adopted than simple point-to-point fibers, primarily because of scalability limits based on power distribution [9]. However, the recent development of low ratio passive couplers [5] and the prospect of fiber based optical amplifiers [4], [6] suggest a closer examination of the power distribution problem. Therefore, we have constructed a prototype system for conducting experiments from which we can extrapolate reasonable limits on the speed and size of practical multicomputer systems.

In this paper, we first present results from two laboratory experiments on a prototype coincident pulse addressing system. The two questions to be answered by the experiments are: how do synchronization error and power loss effect the scalability of such systems. Therefore, the first experimental is an examination of the coincident pulse

Manuscript received March 25, 1991, revised July 23, 1991. This work was supported, in part, by the Air Force Office of Scientific Research under Grant number: AFOSR-89-0469.

D. M. Chiarulli and R. G. Melhem are with the Department of Computer Science, University of Pittsburgh, Pittsburgh, PA 15260.

R. M. Ditmore is with Intergraph Corporation, Huntsville, AL.

S. P. Levitan is with the Department of Electrical Engineering, University of Pittsburgh, Pittsburgh, PA 15261.

IEEE Log Number 9103633.

power, as a function of the synchronization of the arriving pulses. The second experiment demonstrates our ability to select arbitrary subsets of the detectors with a select pulse train operating at 250 MHz. The ability to perform selections of multiple detectors is key to various computer system applications we have investigated. However, the second experiment highlights a more fundamental problem: the power loss due to the tapping couplers on the bus diminishes the ratio of coincident to noncoincident pulse heights for long bus structures. The two experiments contrast the temporal and physical scalability of coincident pulse systems and show that the dominant effects are, in fact, power distribution limits on the physical scale of these systems.

Section III expands on the power distribution issue with an analysis of power distribution in two tapped fiber network structures. The first is the same linear structure that we use in our experiments. The second is a dual-level structure that consists of a main fiber and a series of secondary distribution fibers from which power is tapped. We conclude with a discussion of the implications of these findings to the construction of large systems.

II. Experimental Results

Fig. 1 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 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 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 (and in later experiments four) detectors. The use of two shift registers allows us flexibility in the positioning of the reference pulse relative to the select pulse train.

A. Pulse Synchronization

In our first experiment, measurements were made to characterize the effect of synchronization error between the reference and select pulses on the power of the coincident pulse. Since 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.

A coincident pulse structure with three detectors was used, as shown in Fig. 1. This allowed detector D_2 to be located in the center of the bus resulting in exactly equal noncoincident pulse heights, as shown in the oscilloscope trace of Fig. 2.

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.

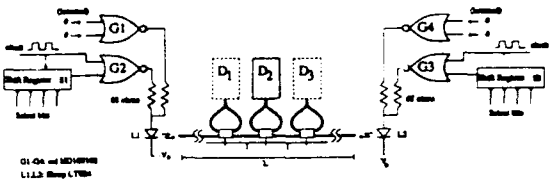


Fig. 1. Synchronization experiment.

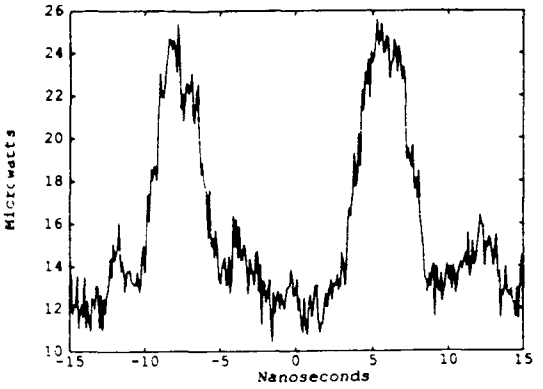


Fig. 2. Synchronization experiment waveform.

Fig. 3 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 units, 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 Fig. 3 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 manufacturing 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 as well as any jitter introduced by the electrical clock generators. For the former, recent results [10] have shown that the variability of the index of refraction of the fiber versus temperature is on the order of 40 ps/km-degree C, and that this is the dominant tempera-

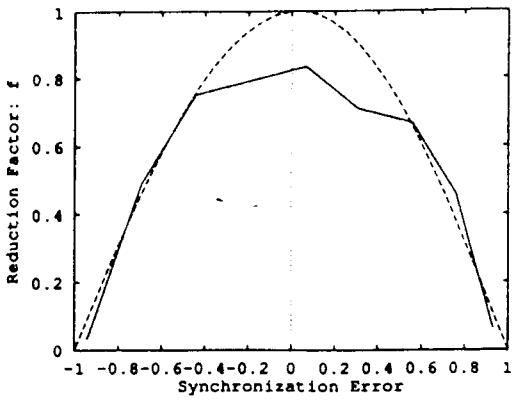


Fig. 3. Synchronization error reduction factor.

ture effect. This represents a very minor variation in effective optical path length. Obviously, jitter and thermal effects in the electronics will be the predominant sources of synchronization error.

However, from Fig. 3 we can see that a timing synchronization 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 electronic pulse generation can be tolerated without significant degradation of the coincident signal. This result characterizes a temporal limit on scalability, based on a limit of achievable pulse widths. Timing errors of several hundred picoseconds are tolerable in gigahertz systems. Therefore, using off the shelf components operating in the one gigahertz range, $\tau = 1$ ns and a system throughput of $\nu = 1/2\tau = 500 \times 10^6$ addressing operations per second is feasible.

The other primary limit, which we need to address, is optical power distribution. Since we are using a passive bus structure, the optical signals are not amplified at any point on the bus. Therefore, sufficient optical power must be available at each detector to individually discriminate coincidence from noncoincidence in the presence of selection pulses for other detectors and noise. This is the subject of the second experiment.

B. Coincident Pulse Power

Our second set of experiments were used to characterize the effect of detector position on the available coincident pulse power. A similar experimental setup was used, this time with four detectors, as shown in Fig. 4.

Figs. 5–8 show the output waveforms for detectors D1 and D3 for various selection patterns. Note that for each selection pattern (pair of waveforms) the experimental equipment was adjusted so that the absolute values of pulse heights for different selection patterns varied. Figs. 5 and 6 show coincident and noncoincident waveforms at detectors D1 and D3, respectively. Note that in both cases, the noncoincident waveforms (shown in (b)) are of unequal power. This is due to the fact that each pulse has passed through a different number of couplers and, hence,

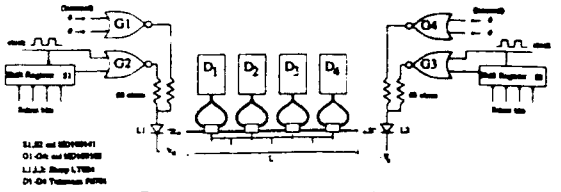


Fig. 4. Detector power experiment.

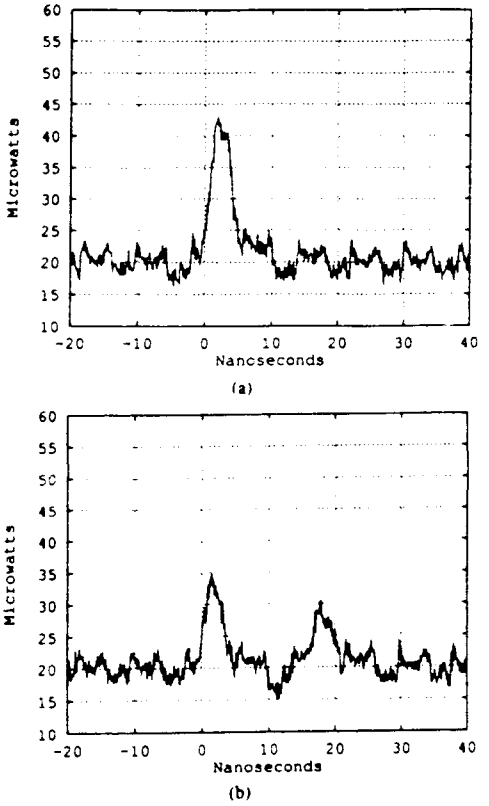


Fig. 5. (a) Selection of D1 measured at D1. (b) selection of D3 measured at D1.

has become attenuated to different levels. This clearly shows that the relative power between coincident and noncoincident pulses is a function of the detector location.

Figs. 7 and 8 are examples of parallel selections. The waveform in Fig. 7(a) shows a parallel selection waveform at detector site D_3 for the selection of three detectors, including D_3. This incident waveform peak is comparable to the noncoincident waveform, in Fig. 7(b), in which D_3 has been removed from the set of selected locations. Similarly Fig. 8 shows parallel selection of all four detectors at sites D_1 and D_3.

To quantify the power degradation that we observed in these experiments, we define the amount of additional power in a coincident pulse relative to the largest noncoincident pulse seen by a detector as the power margin, *Pm*. This is given as a fraction of the maximum noncoincident pulse power:

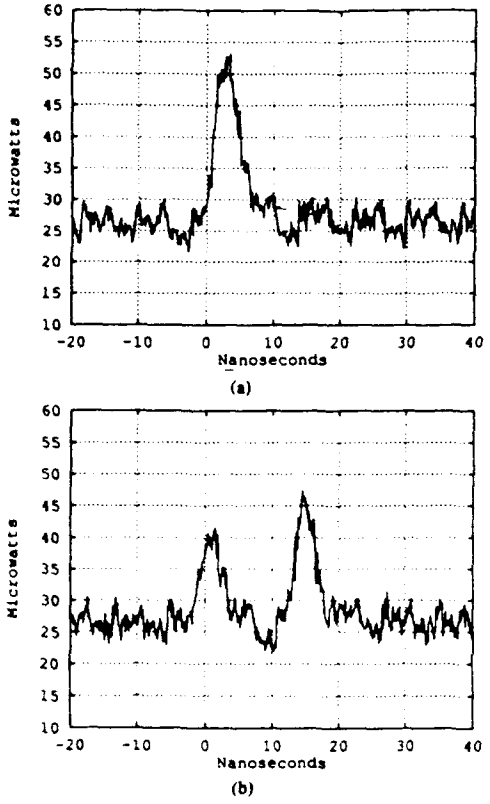


Fig. 6. (a) Selection of D3 measured at D3, (b) selection of D1 measured at D3.

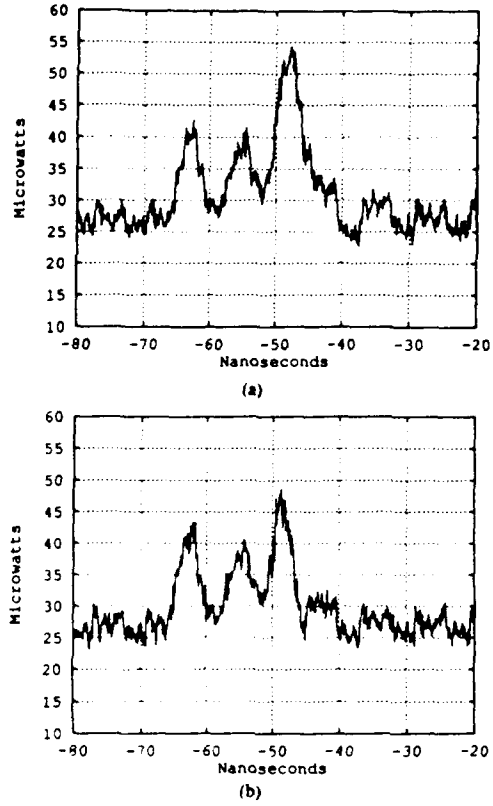


Fig. 7. (a) Selection of D1, D2, D3 measured at D3, (b) selection of D1, D2 measured at D3.

$$Pm = (p_1 + p_2 - \max(p_1, p_2))/\max(p_1, p_2)$$
$$= \min(p_1, p_2)/\max(p_1, p_2). \quad (1)$$

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

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

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. For all the selection experiments shown in Figs. 5 through 8, the power margin is in excess of 30%. That is, coincident power is greater than 130% of peak single pulse power. This is measured at D_1, which is the leftmost detector on the bus.

In the next section we discuss the implications of power margin on scalability issues.

III. Analytical Study of Power Distribution

In this section, we present an analysis of power distribution in each of two tapped fiber network structures. The first is a simple linear structure with a single backbone and a series of passive coupler taps, as used in the experiment above. The second is a dual level structure, which consists of a backbone fiber and a series of secondary distribution fibers from which power is tapped.

In this analysis, we use passive, bidirectional, 2 × 2, symmetric fiber couplers as shown in Fig. 9 [1], [5]. These are identical to the couplers we used in our previously discussed experiments, except that in our analysis we assume no excess loss in the couplers. Since the couplers are bidirectional, we arbitrarily let A, B be the input ports and A', B' be the output ports. Equation (2) shows power distribution from the input to the output:

$$\begin{pmatrix} A' \\ B' \end{pmatrix} = \begin{pmatrix} r & (1-r) \\ (1-r) & r \end{pmatrix} \begin{pmatrix} AB \end{pmatrix} \quad (2)$$

where r is the coupling ratio. Using these couplers, we now discuss the linear and dual level structures.

A. The Linear Structure

As is shown in Fig. 10, a linear bus 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:

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

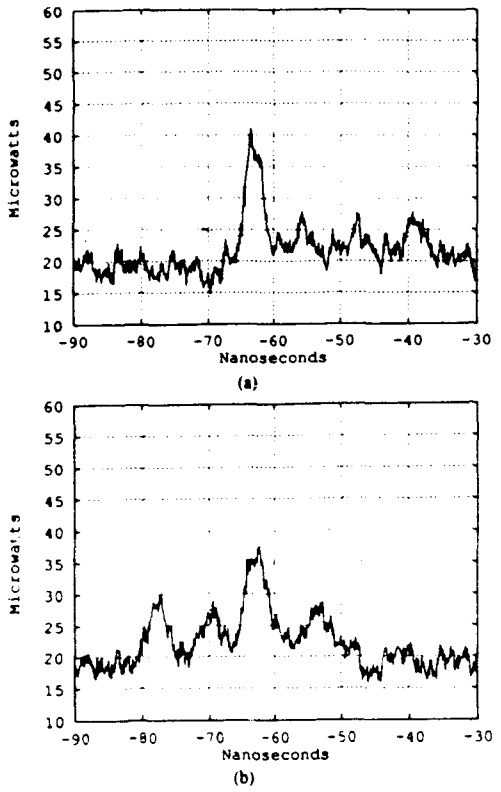


Fig. 8. (a) Selection of D1, D2, D3, D4 measured at detector D1. (b) selection of D1, D2, D3, D4 measured at detector D3.

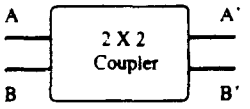


Fig. 9. Symmetric fiber coupler.

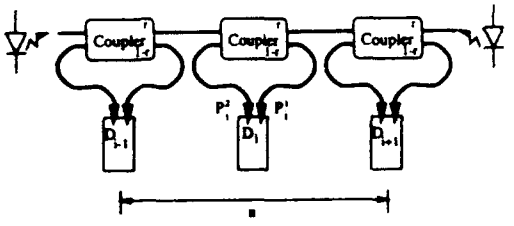


Fig. 10. Linear optical bus.

Since the bus is symmetrical, we can analyze one signal that originates on the left from a single transmitter and propagates to the right as shown in Fig. 10.

Fig. 11 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

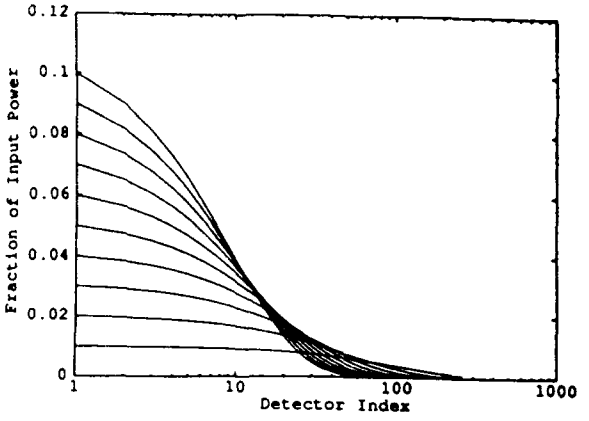


Fig. 11. Power p_i^1 at detector D_i for $90\% \le r \le 99\%$.

the curves, the absolute power falls off geometrically with increasing i, $1 \le i \le n$.

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 $P\text{min}$, then the maximum number of detectors supportable is

$$n = \frac{\log\left(\frac{P\text{min}}{1 - r}\right)}{\log (r)} + 1. \tag{4}$$

Equation (4) is shown graphically in Fig. 12 for a set of coupling ratios $r = 90\%, 95\%, 97\%, 98\%, 99\%$, and $0.01\% \le P\text{min} \le 1\%$ 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 $P\text{min}$, we will be able to support more detectors on the bus. We also note the sharp drop in n for high values of $P\text{min}$ and r, which reflects the situation where much of the available power flows off the end of the bus and is wasted.

However, for our experimental setup, it is clear that it is not the absolute power but rather the power margin that imposes a bound on the size of the system. In addition, since the bus configuration chosen for this structure requires bidirectional propagation, we are constrained to use a single tapping ratio, r, for all couplers. Based on these two constants, the graph shown in Fig. 13, 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 Fig. 12 that using commercially available 95% couplers, and assuming we can tolerate a $P\text{min}$ of 0.0001 of input power we could achieve bus lengths of abut 120 detectors. This would be the case for an input of 100 mW of power injected into the bus, and a detector sensitivity of 10 μW, operating at 250 MHz. However, Fig. 13 shows that for a power margin of $Pm = 20\%$ we could only reach lengths of 32 detec-

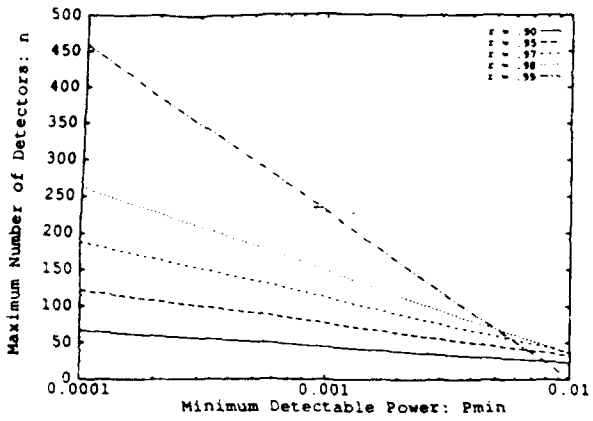


Fig. 12. Number of detectors versus *Pmin* for various values of *r*.

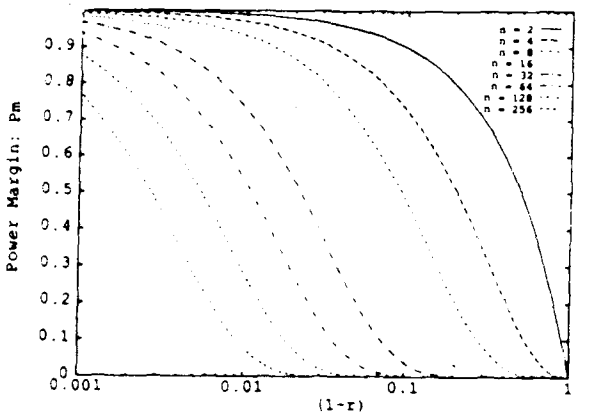


Fig. 13. Power margin *Pm* versus $0.001 \geq (1 - r) \geq 1$ for various bus sizes.

tors. Therefore, due to both minimum power constraints and power margin issues the system scale is highly sensitive to the fixed value of *r*. Further, we note that power margin imposes a tighter constraint than absolute power.

To help alleviate this problem, we propose a two level bus structure. By using two levels, we can essentially increase the tapping ratios on our buses and more effectively control the amount of power at each detector.

B. The Dual-Level Structure

The basis for the power distribution problem in the linear system is the fact that detectors at the start of the bus use more power than needed and, therefore, detectors at the end of the bus are starved. If we were to relax the requirement of fixed ratio taps in favor of varying the coupling ratios, we would need a number of distinct, precisely tuned couplers approaching the number of detector sites [9]. Yet, no couplers exist that would allow tuning to a precision of more than one or two percent. Of course, the use of tuned couplers forces the network to be unidirectional since coupling ratios must decrease in the direction of propagation.

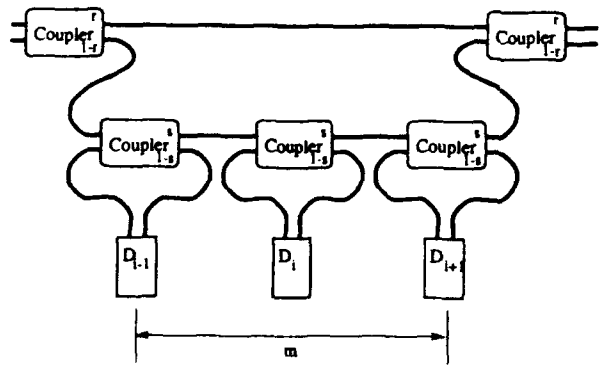


Fig. 14. Dual-level optical bus.

An alternative method that does not require multiple coupling ratios is to adopt a dual-level bus structure. As shown in Fig. 14, we split the bus into a main fiber and a sublevel to create a section of the bus, labeled *m*. The sublevel contains *m* detectors in a linear arrangement except for the last detector, which feeds back the remaining power into the main fiber and the next section. In the main fiber, care must be taken to ensure that the optical path length is the same as the subsection so that the two parts of the signal arrive synchronized at the next section. The dual-level bus consists of a series of these sections.

Once again, we start with an analysis of absolute power for this structure, and then proceed to power margin issues. Thus, we assume the input is from the left (into the upper leg to the first coupler) and propagates to the right. The detectors are numbered linearly in the direction of propagation.

We further assume two types of couplers with splitting ratios of *r* and *s* for the main level and sublevel, respectively. The power at any given detector site in Fig. 14 is given by

$$p_i = (1 - r \quad r)A^k \begin{pmatrix} 1 \\ 0 \end{pmatrix} s^l(1 - s) \tag{5}$$

where p_i is the power at site *i*, *r*, and *s* are coupling ratios, *k* is *i* div *m*, *l* is *i* mod *m*, *m* is the number of detectors in a sublevel, and

$$A = \begin{pmatrix} r & 1 - r \\ (1 - r)s^m & rs^m \end{pmatrix}. \tag{6}$$

From linear algebra [11], we know that a vector of the form $u_k = A^k u_o$ can be rewritten as $u_k = \Sigma_{i=1}^{n} c_i \lambda_i^k x_i$, where λ_i are the eigenvalues of matrix *A*, the x_i's are the associated eigenvectors and the coefficients c_i are determined from the initial condition u_o.

For our analysis, we rewrite the matrix of (5) in the form

$$A^k \begin{pmatrix} 1 \\ 0 \end{pmatrix} = c_1 \lambda_1^k x_1 + c_2 \lambda_2^k x_2 \tag{7}$$

and the coefficients are determined by $c_1 x_1 + c_2 x_2 = u_o$.

The x_i are vectors and they are given by $x_i = \binom{\xi_i}{1}$. Rewriting the coefficient equation gives

$$\begin{pmatrix} c_1\xi_1 \\ c1 \end{pmatrix} + \begin{pmatrix} c_2\xi_2 \\ c2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

which has the solution

$$c_i = \frac{1}{\xi_1 - \xi_2}.$$

Assuming, without loss of generality, that $\lambda_1 > \lambda_2$ as k increases, then the $c_1\lambda_1^k x_1$ term in (7) quickly dominates. Therefore, a good approximation is given by

$$p_i = [(1 - r) \quad r]c_1\lambda_1^k x_1 s^i(1 - s). \tag{8}$$

Fig. 15 shows a comparison of a linear bus and a dual-level structure for the particular case of $r = s = 90\%$, $n = 256$, and $m = sqrt(n)$. Clearly, the power at the detectors for the linear bus falls off much more rapidly than for the dual-level bus. The dual-level bus shows a characteristic "saw-tooth" pattern of power distribution. At the beginning of each section, power is restored by injection of power from the main backbone. This more evenly distributes all of the available power down the length of the bus.

In the linear structure, we examined the bounds for the minimum power needed at the last detector. For the dual-level structure, we will examine the minimum power seen at the last detector of the last section. This minimum power is given by the equation

$$Pmin = \lambda_1^k(\xi_1(1 - r) + r)c_1 s^{m-1}(1 - s). \tag{9}$$

As with the linear case, the ability to support large systems is dependent upon maximizing the values of r and s. However, in the dual-level case, we additionally may vary m, the number of detectors per section. The relationship between r, s, and m is captured in λ_1, which is a monotonically increasing function of r, and s but is not monotonic in m. Therefore, it is desirable to fix r and s to be as large as possible and adjust m to maximize the total number of detectors in the system.

This relationship is shown in Fig. 16. The two families of curves represent coupling ratios of $r = s = 90\%$ and $r = s = 95\%$. The curves are the number of detectors (length of the bus) supportable at different *P*min values. For the 90% curves, *P*min = 0.0001, 0.0002, 0.0004, 0.0008, 0.0016, 0.0032, and 0.0064. For the 95% curves, *P*min = 0.0001, 0.0002, 0.0004, 0.0008, and 0.0016. Note that the dual-level structure with 95% couplers cannot support high minimum power detectors since the power into the first detector $p_1 = 0.05 \times 0.05 = 0.0025$. The long tails on the curves reflect the condition where $m \geq n$.

Having chosen values for r, s, and m, we can rewrite equation (9) to compute the number of detectors supportable as a function of *P*min:

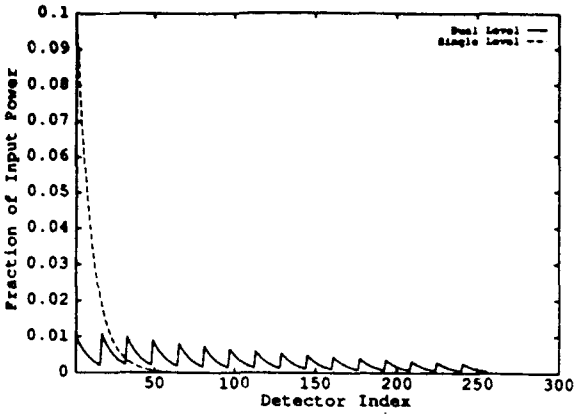


Fig. 15. Power at detector sites for single- and dual-level 256 node buses.

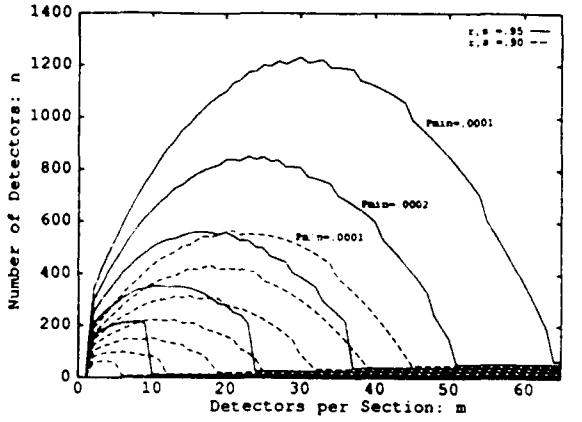


Fig. 16. n versus m for r, $s = 0.90$, 0.95, and various *P*min.

$$n = m\frac{\log\left(\dfrac{Pmin}{((1 - r)x_1 + r)c_1 s^{k-1}(1 - s)}\right)}{\log(\lambda_1)}. \tag{10}$$

A plot of numerical solutions for (10) is shown in Fig. 17.

Equation (10) and Fig. 17 allow a direct comparison of the dual-level bus performance shown in Fig. 17 with linear bus performance derived in (4) and plotted in Fig. 12. From this comparison, we can see that, in terms of *P*min, the optimized dual level bus gives approximate factors of between 4 and 10 improvement (depending on the coupler ratios) over the simple linear configuration.

To perform the analysis of power margin for the two level structure, we compare the maximum power at any detector to the minimum power at any detector on the bus. This simplifies the calculation and gives a bound on the "envelope" of the saw-tooth power curve (as shown in Fig. 15). For these curves (shown in Fig. 18) we are again using $m = sqrt(n)$ and $r = s$. Unlike the curves for the linear bus, these curves have a peak and approach an asymptotic value for very large values of r and s. This is

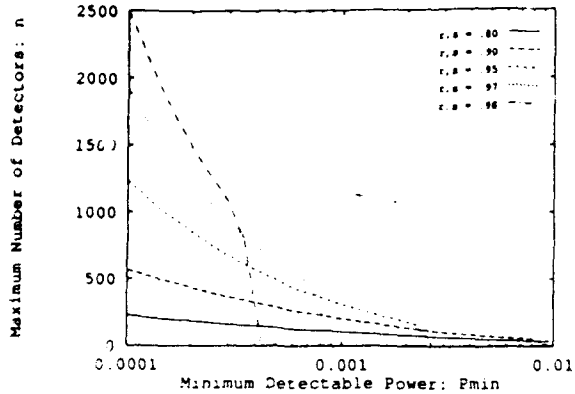


Fig. 17. Pmin versus n for different values of r, s.

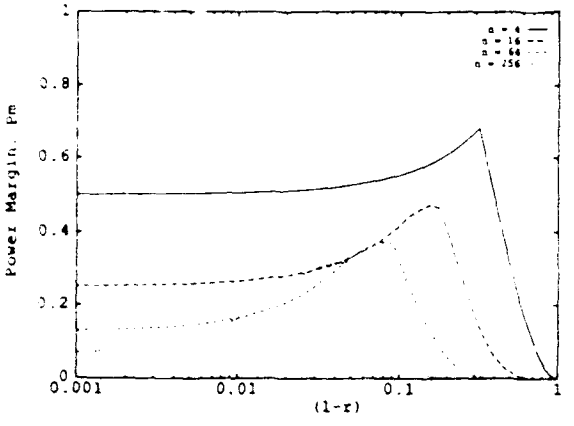


Fig. 18. Pm versus $0.001 \geq (1 - r) = (1 - s) \geq 0$ for dual-level buses.

because, similar to the linear case, we reach a point at which a significant percentage of the power must be thrown away at the end of the bus, in order to account for the large coupling ratio of the final tap. However, it is still the case that Pm, the power maximum margin, limits the scalability of the system more tightly than absolute power. As a practical example similar to the linear case, using available 95% percent couplers, the power margin limits bus size to about 300 detectors, rather than the 1250 detectors we could expect based on minimum power requirements of $Pmin = 0.0001$.

IV. Summary

Clearly, three factors, threshold power margin, synchronization error, and coupling ratio determine system scale. Our experiments have shown 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. However, we believe near term technologies (e.g., fiber amplifiers) and alternate bus structures will alleviate this problem. The fact that the temporal scalability bounds show significantly shorter pulses can be supported, is very encouraging for the long-term application of this technique.

References

[1] F. C. Allard, *Fiber Optics Handbook For Engineers and Scientists*. New York: McGraw-Hill, 1990.

[2] D. M. Chiarulli, S. P. Levitan, and R. G. Melhem, "Self routing interconnection structures using coincident pulse techniques," in *SPIE OE/Boston '90* (Boston, MA), Nov. 4-9, 1990.

[3] D. M. Chiarulli, R. G. Melhem, and S. P. Levitan, "Parallel memory using coincident optical pulses," *IEEE Comput.*, vol. 20, no. 12, pp. 48-57, Dec. 1987.

[4] C. R. Giles, E. Desurvire, J. R. Talman, J.R. Simpson, and P. C. Becker, "2 Gb/s signal amplification at $\lambda = 1.53$ μm in a erbium-doped single-mode fiber amplifier," *J. Lightwave Technol.*, vol. 7, no. 4, pp. 651-656, Apr. 1989.

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

[6] R. I. Laming *et al.*, "Efficient pump wavelengths of erbium-doped fiber optical amplifiers," *Electron. Lett.*, vol. 25, no. 1, pp. 12-14, Jan. 1989.

[7] S. P. Levitan, D. M. Chiarulli, and R. C. Melhem, "Coincident pulse techniques for multiprocessor interconnection structures," *Appl. Opt.*, vol. 29, no. 14, pp. 2024-2033, May 1990.

[8] R. G. Melhem, D. M. Chiarulli, and S. P. Levitan, "Space multiplexing of optical waveguides in a distributed multiprocessor," *Comput. J. British Comput. Society*, vol. 32, no. 4, pp. 362-369, 1989.

[9] M. Nassehi, F. Tobagi, and M. Marhic, "Fiber optic configurations for local area networks," *IEEE J. Select. Areas Commun.*, vol. SAC-3, no. 6, pp. 941-949, Nov. 1985.

[10] D. Sarrazin, H. Jordan, and V. Heuring, "Digital fiber optic delay line memory," in *Digital Optical Computing II*, vol. 1215 (Los Angeles, CA), Jan. 1990. SPIE.

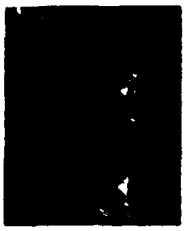
[11] Gilbert Strang, *Linear Algebra and Its Applications, 2nd ed.* New York: Academic, 1980.

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

He is an Assistant Professor of Computer Science at the University of Pittsburgh. His research interests include architectures and algorithms for the design and implementation of highly parallel computing systems, and optical technologies for interconnection networks and processing.

Dr. Chiarulli is a member of the IEEE Computer Society, SPIE, and the Optical Society of America.

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



Steven P. Levitan (S'83-M'83) received the B.S. degree from Case Western Reserve University in 1972 and the M.S. and Ph.D. degrees both in computer science, from the University of Massachusetts, Amherst in 1979 and 1984, respectively.

He is the Wellington C. Carl Assistant Professor of Electrical Engineering at the University of Pittsburgh, Pittsburgh, PA. 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. His research interests include optical computing, parallel computer architecture, parallel algorithm design, and computer aided design for VLSI.

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

Rami G. Melhem received the B.E. degree in electrical engineering from Cairo University, Egypt, in 1976, the M.S. degree in mathematics/computer science from the University of Pittsburgh in 1981, and the Ph.D. in computer science from the University of Pittsburgh in December 1983.

He is an Associate Professor of Computer Science at the University of Pittsburgh, Pittsburgh, PA. He has been an Assistant Professor of Computer Science at Purdue University from 1984 to 1986 and at the University of Pittsburgh from 1986 to 1989. His research interests include optical computing parallel systems, fault tolerant systems and the application of large computational arrays to scientific problems.

JOURNAL OF PARALLEL AND DISTRIBUTED COMPUTING 12, 269–282 (1991)

Pipelined Communications in Optically Interconnected Arrays*

ZICHENG GUO, RAMI G. MELHEM, RICHARD W. HALL, DONALD M. CHIARULLI, AND STEVEN P. LEVITAN

Departments of Electrical Engineering and Computer Science, University of Pittsburgh, Pittsburgh, Pennsylvania 15261

Two synchronous multiprocessor architectures based on pipelined optical bus interconnections are presented. The first is a linear pipeline with enhanced control strategies which make optimal use of the available communication bandwidth of the optical bus. The second is a two-dimensional architecture in which processors are placed in a square grid and interconnected to one another through horizontal and vertical pipelined optical buses. These architectures allow any two processors to communicate with each other using one (for the linear case) or two (for the two-dimensional case) pipelined bus cycles. Further, they permit all processors to have simultaneous access to the buses using slots within a pipelined cycle. We show that the architectures have simple control structures and that well-known processor interconnections, e.g., the complete binary trees and the hypercube networks, can be efficiently embedded in them. These architectures have an effectively higher bandwidth than conventional bus configurations and appear to be good candidates for a new generation of hybrid optical–electronic parallel computers.

1. INTRODUCTION

Two-dimensional meshes of processors have been extensively studied in various forms and augmentations [23, 26, 37]. Large-scale implementations of two-dimensional meshes have been built [2, 10, 17]. However, since the communication diameter of an $n \times n$ mesh is $O(n)$, different approaches have been considered to augment the communication capabilities of the mesh to reduce this diameter. Meshes have been augmented with global buses [3, 10, 11, 35], reducing the communication diameter but giving only very small bandwidth improvements. Row and column bus augmentations [29, 30] have yielded both a low communication diameter and adequate bandwidth for certain classes of algorithms. Interconnection networks have been considered for augmenting rows and columns in a mesh including trees [27, 28, 39] and compounded graphs [18, 19]. The binary hypercube can also be viewed in this context as a two-dimensional mesh with horizontal and vertical hypercube interconnections [18, 19].

* This work was, in part, supported by Air Force Grant AFOSR-89-0469 and by NSF Grant MIP-8901053.

One of the simplest mesh augmentation schemes is the row and column bus augmentation. However, exclusive write access to buses is a major contributor to the low bandwidth of bus interconnections. A unique property of optics provides an alternative to this exclusive access, namely, the ability in optics to pipeline the transmission of signals through a channel. In electronic buses, signals propagate in both directions from the source, while optical channels are inherently directional and have precise predictable path delays per unit distance. Hence, a pipeline of optical signals may be created by the synchronized directional coupling of each signal at specified locations along the channel. This property has been used to parallelize access to shared memory [5], to enhance the bandwidth in bus-connected multiprocessor systems [22], and to minimize the control overhead in networking environments [38].

In this paper, we present two multiprocessor architectures, called *Array Processors* with *Pipelined Buses* (APPB), which employ optical bus interconnections in processor arrays. In Section 2 we review the basic principle of pipelining messages on optical buses. In Section 3 we introduce our linear APPB, where processors are connected with a single optical bus. We present efficient approaches to message routing and network embedding for the linear APPB as well as techniques for enhancing the bus utilization through enhanced control functions. In Section 4 we introduce our two-dimensional APPB, where processors are interconnected with horizontal and vertical optical buses. We discuss routing and embedding issues for this new architecture. We show how binary tree and hypercube interconnections can be effectively embedded and identify key design issues for effective embeddings of arbitrary interconnections. In Section 5 we compare the efficiency of the pipelined bus communication model with that of nonpipelined buses and of store and forward communications in nearest-neighbor structures. Finally, Section 6 contains concluding remarks.

2. MESSAGE PIPELINING ON OPTICAL BUSES

Consider the system of Fig. 1a, where n processors, each having a constant number of registers, are connected through a single optical waveguide (bus). Each processor is coupled to the optical waveguide with two passive couplers, one for injecting (writing) signals on the waveguide and the other

0743-7315/91 $3.00

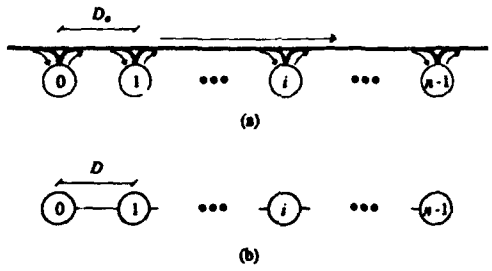


FIG. 1. (a) A system of n processors connected with a single optical waveguide (bus). (b) A linear array of n processors with nearest-neighbor connections.

for receiving (reading) signals from the waveguide [20, 40]. Each receiving coupler passively taps a percentage (typically 5–10%, depending on the coupling ratio) of the optical signal power available on the bus. Thus the couplers do not introduce any delay to the propagation of optical signals along the bus. However, the degradation of signal power does place an upper limit on the number of processors that can be connected on the bus [8]. As in the case of electronic buses, each processor j communicates with any other processor i by sending a message to i through the common bus. However, because optical signals propagate in one direction, a processor j may send signals to another processor i only if $i > j$.

Assume that a message on an optical bus consists of a sequence of optical pulses, each having a width w in seconds. The existence of an optical signal of width w represents a binary bit 1, and the absence of such a signal represents a 0. Note that w includes a time for electro-optical conversions, rise and fall times, and propagation delay in the latch of the receiver circuits [6]. For analytical convenience, we let D_o be the optical distance between each pair of adjacent nodes (it will become clear that the distance between two adjacent nodes need not be equal) and τ be the time taken for an optical pulse to traverse the optical distance D_o. To transfer a message from a node j to node i, $i > j$, the sender j writes its message on the bus. After a time $(i - j)\tau$ the message will arrive at the receiver i, which then reads the message from the bus.

The properties of unidirectional propagation and predictable path delays of optical signals may be used advantageously. Specifically, unlike the electronic case, where the writing access to the bus by each node must be mutually *exclusive*, all nodes in the system of Fig. 1a can write on the bus *simultaneously*, provided that the following collision-free condition [22] is satisfied,

$$D_0 > bwc_g, \tag{1}$$

where b is the number of binary bits in each message, and c_g is the velocity of light in the waveguide. Clearly if this condition is satisfied and the system is synchronized such that every node starts writing a message on the bus at the same instant, then no two messages injected on the bus by any two distinct nodes will collide. Here by colliding we mean that two optical signals injected on the bus by any two distinct nodes arrive at some point on the bus simultaneously. This kind of synchronized pulse generation is restrictive but it can be met in several ways [21]. An optically distributed clock can be broadcast without skew to each node, or electro-optical switches can be used in place of sources to "switch in" pulses generated from a single source. With this condition satisfied, every node can, in parallel, send a message to some other node, and the messages will all travel from left to right on the bus in a pipelined fashion, as shown in Fig. 2. Thus we use the term *pipelined bus*. In the rest of this paper we always assume that the collision-free condition (1) is satisfied.

To facilitate our discussion in subsequent sections we define some terms. Let τ be defined as before and n be the number of nodes on the pipelined optical bus. We define $n\tau$ as a *bus cycle* and correspondingly τ as a *petit cycle*. Note that a bus cycle is the time taken for an optical signal to traverse the entire length of the optical bus. For the discussion in this section, we do not include in a bus cycle the time taken to prepare and process a message before it can be injected on the bus. This time is explicitly introduced in our performance analysis in Section 5. If every node is writing a message simultaneously on the bus, then each node has to wait for at least a bus cycle to inject its next message. Note that each cycle on the pipelined bus may be emulated by n cycles in a linear array with nearest-neighbor communications shown Fig. 1b. Comparison of the two interconnection schemes is made in Section 5.

Let us look at a simple routing task where each node transmits a message and each node is programmed to receive a message from the kth node (if it exists) to its left. All nodes start injecting messages at the beginning of a bus cycle, and all the messages travel on the optical bus in pipelined fashion without collision. By waiting for a specific interval of time, a node can selectively read the message intended for it as that message passes by the node. In our example, each node i is to receive a message from node $i - k$ and thus must read its message from the bus after $k\tau$ time from the beginning of the bus cycle. In this way, a message routing pattern in which each node sends a message to the kth node to its right has been realized. In fact, as will be seen, we can realize various message routing patterns in a simple, straightforward way.

3. LINEAR ARRAY PROCESSORS WITH PIPELINED BUSES

In the system of Fig. 1a, messages can be transmitted only from left to right. To allow message passing from right to

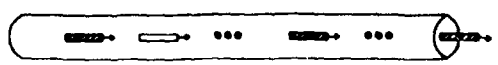


FIG. 2. Message pipelining on the optical bus. A blank rectangle indicates "no signal," implying that some processor is not sending a message.

left, another optical bus is used, as shown in Fig. 3a. In this figure, we have two optical buses; the upper one is used for sending messages from left to right, and the lower one is used for sending messages from right to left. Each node can write and read messages on either bus as desired. Obviously signals in different buses do not disturb one another; that is, the two buses can support two separate pipelines. The system in Fig. 3a is our architecture of linear APPB. For convenience the linear APPB in Fig. 3a is schematically drawn as in Fig. 3b.

To specify the time at which a node should receive a message, we introduce a control function $twait(i)$, which is defined as the time that node i should wait, relative to the beginning of the bus cycle, before reading the message sent to it from some other node j. Thus

$$twait(i) = (i - j)\tau.$$

If τ is considered as a time unit, then *twait* can be interpreted in terms of the number of such time units and thus be written $twait(i) = i - j$. Clearly if $twait(i) > 0$, then the message is to be received from the left; if $twait(i) < 0$, then the message is to be received from the right. If $twait(i) = 0$, then no message should be received by node i. The value of $twait(i)$ can be stored in a *wait register*, and more than one such register may be used if a node is to receive more than one message in one bus cycle.

This *twait* control function, however, has the disadvantages that it depends crucially on timing accuracy and is sensitive to the optical distance D_0 between two adjacent nodes. An equivalent control function, *mwait*, that does not have these disadvantages may be defined if we require that each node inject a message, real or dummy, every bus cycle. In this case we define $mwait(i)$ as the number of messages that node i should skip before reading its message. For example, if $mwait(i) = \gamma$, then node i should receive the $|\gamma|$th message that passes i on the bus. That is, it has to wait until $|\gamma| - 1$ messages have passed and then it reads its own message. The sign of γ determines on which bus the message should be received. Clearly *mwait* is equivalent to *twait* and either control function may be used. For convenience we simply write the control function as *wait*, and we assume that the optical distance between each pair of adjacent nodes i and $i + 1$ is constant.

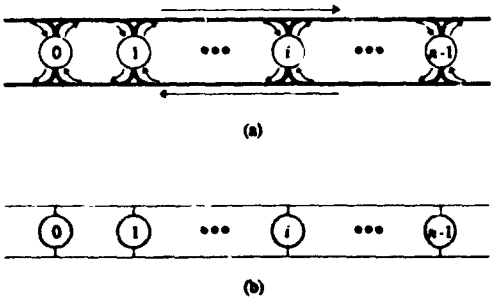


FIG. 3. (a) Linear array processors with pipelined buses (APPB). (b) A schematic drawing of (a).

The control function *wait* can only be used when the communication pattern is known to the receiver in the sense that the receiver knows from which node the message is to be received. In cases where the communication pattern is unknown to the receiver, the coincident pulse techniques [5, 21] may be used such that an addressing pulse and a reference pulse coincide at the detector of the receiver, thereby addressing it. In this paper we use *wait* for addressing since the communication patterns which we discuss are known to the receiver.

In the following we present techniques for message routing and network embedding in the linear APPB. For the purpose of evaluating the communication efficiency, we note that a lower bound on the number of bus cycles needed to transfer H messages in the linear APPB is $\lceil H/n \rceil$, where n is the number of nodes on the optical bus. This lower bound is obtained by assuming a perfectly even distribution of messages along the bus at each bus cycle, that is, *every* node has one message to send at *each* bus cycle.

3.1. *Message Routing in Linear APPB*

Various message routing patterns can be realized in a simple, straightforward way. Since a routing pattern is determined by the *wait* functions, we need only determine these *wait* functions for each routing pattern. The most common patterns are:

One-to-One. The system executes a $SEND(j, i)$ instruction, which means that a message is to be transferred from node j to node i. Thus, $wait(i) = i - j$, where i is a *single* specific node.

Broadcast. The system executes $BROADCAST(j)$, which means that node j broadcasts a message, and all other nodes i will receive that message. In this case, $wait(i) = i - j$ for *all* $i \neq j$.

Semigroup Communication [4]. The system executes a $SEMIGROUP(i)$ instruction, which says that some global information, e.g., extrema and sum, is to be computed and stored at node i. This task can be accomplished by having the linear APPB logically function as a tree with the root being node i. Later in this section we present embeddings of binary trees which facilitate such a tree emulation task.

Permutations. For each node j to send a message to a node $i = PERM(j)$, where $PERM(\ )$ is an arbitrary permutation, we set $wait(i) = i - j$ for *all* i.

We see that the computation of $wait(i)$ is very simple and uniform. The only difference among the *wait* functions for different message routing patterns is that the nodes involved

are different. It is clear that all these communication tasks can be performed using a single bus cycle, except the semigroup communication, which takes $\log(n)$ bus cycles. Note that, in the linear APPB, message passing between two nonneighboring nodes is nearly as efficient as that between two neighbors. Specifically, a message takes τ more time to pass one more node on the optical bus. This is not the case in the linear array with nearest-neighbor connections shown in Fig. 1b, where to pass a node, en route to another node, a message has to go through a router. In this sense we may say that the APPB is communication efficient, and in particular global-communication efficient.

3.2. *Embedding Binary Tree and Hypercube Networks into Linear APPB*

In this subsection we show how to embed other interconnection networks into the linear APPB. Our first example is the embedding of complete binary tree networks. To show that a binary tree network can be embedded in the linear APPB it is sufficient to find the *wait* function for each processor in the linear APPB such that the desired routing pattern is accomplished.

Let L be the number of levels of a complete binary tree and let the root of the tree be node 1. Each node i, $i \geq 1$, which is not a leaf node has two children, $2i + \delta$, where $\delta = 0, 1$, corresponding to i's left and right child, respectively (see Fig. 4a for an example). Consider an embedding in which node i in the tree is mapped to node $i - 1$ in the linear APPB. For convenience, we call this embedding E_{t1} (see Fig. 4b). In E_{t1}, the wait functions for node i to receive a message from its children are:

$$wait_{c,\delta}(i) = \begin{cases} i - (2i + \delta) = -(i + \delta), & i < 2^{L-1}, \\ 0, & \text{otherwise.} \end{cases}$$

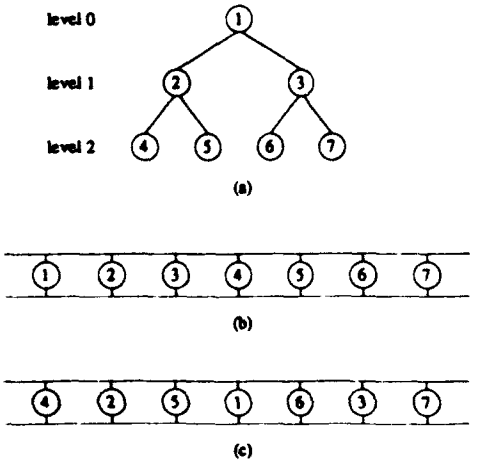


FIG. 4. Embeddings of complete binary trees in the linear APPB. (a) A binary tree. (b) The first embedding, E_{t1}. (c) The second embedding, E_{t2}.

Thus, to realize children-to-parent message routing each parent should wait for $wait_{c,0}(i)$ and $wait_{c,1}(i)$ time to read the messages from its left and right child, respectively. Clearly this routing task can be performed using one bus cycle.

For parent-to-children message transfer in E_{t1}, each parent has two messages to send to its two children, respectively. In this case, two bus cycles are needed to carry out such a routing task, one to send messages to left children and one to send messages to right children. Let $wait_{p,0}(j)$ and $wait_{p,1}(j)$ be the wait functions for a left child and right child, respectively, to receive a message from its parent. Then, during the first cycle we have

$$wait_{p,0}(j) = \begin{cases} j - \dfrac{j}{2} = \dfrac{j}{2}, & j = \text{even}, \\ 0, & \text{otherwise}, \end{cases}$$

and during the second cycle we have

$$wait_{p,1}(j) = \begin{cases} j - \dfrac{j-1}{2} = \dfrac{j+1}{2}, & j = \text{odd, and } j \neq 1, \\ 0, & \text{otherwise.} \end{cases}$$

Mapping each node i in the binary tree network onto node i (or $i - 1$ as was just done above) in the linear APPB is a straightforward approach. Using this straightforward approach we can embed any type of network in the linear APPB. This approach, however, may not give a good embedding in the sense that it may take more time than needed, in number of bus cycles, to accomplish a given communication task. As is seen next, another tree embedding, E_{t2}, has a better communication efficiency than E_{t1}.

Embedding E_{t2} may be viewed as pressing the binary tree from the root down until all the nodes fall in the level of the leaf nodes (see Fig. 4c). In this embedding the two children of a node i are on opposing sides of i. Thus the parent-to-children routing pattern, as well as the children-to-parent routing pattern, may be accomplished in one bus cycle. Specifically, if i is a node at level l, where l is the integer satisfying $2^l - 1 < i < 2^{l+1}$, then the wait functions for i to receive the messages from its two children are

$$wait_{c,\delta}(i) = \begin{cases} (-1)^{\delta} 2^{L-l-2}, & i < 2^{L-1}, \\ 0, & \text{otherwise.} \end{cases}$$

The parent-to-children message routing pattern in E_{t2} is different from that in E_{t1} in that the two messages from a parent will travel on two different buses. Then the two messages from each parent node can be simultaneously injected on the two buses, respectively, in the same bus cycle. Hence, the parent-to-children routing pattern can be accomplished in one bus cycle. $wait_{p,\delta}$ can be determined by noting that

$wait_{p,s}(j) = -wait_{c,s}(i)$, where i is the parent of j. That is, the wait functions for parent-to-children message transfer are

$$wait_{p,s}(j) = \begin{cases} (-1)^{s+1}2^{s-l-1}, & j > 1, \\ 0, & j = 1. \end{cases}$$

Next, we consider a k-dimensional binary hypercube in which the nodes are numbered such that if nodes i and j are neighbors across dimension h, $1 \leq h \leq k$, then $|i-j| = 2^{h-1}$ (see Fig. 5a). Let E_{cl} be the embedding of this k-cube into a linear APPB such that each node i in the hypercube is mapped into node i in the linear APPB. With this embedding, a node in the hypercube may send a distinct message to each of its k neighbors if each node sends one message to one neighbor in each bus cycle. For example, at the hth bus cycle a message is sent from each node to its neighbor at distance 2^{h-1}. To accomplish this, the time that a node i has to wait during the hth bus cycle before receiving a message from its neighbor along the hth dimension is

$$wait_h(i) = \pm 2^{h-1}.$$

In our discussions so far, we have allowed each node to send only one message on each bus during each bus cycle. In other words after placing a message on the bus in the current cycle, all nodes must wait until the next cycle to initiate the next message. In the following subsection, we show that such a wait is not always necessary.

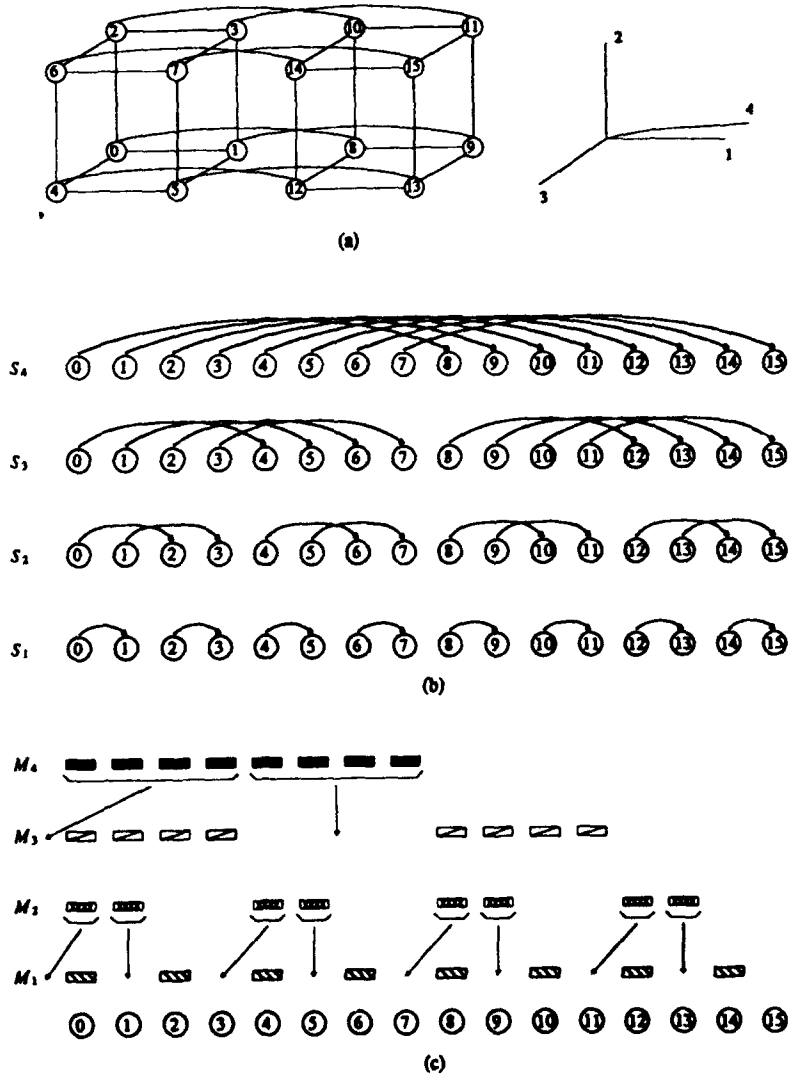


FIG. 5. (a) A binary hypercube and its dimension assignment. (b) Message routing patterns in the hypercube. (c) Message distribution in the hypercube.

3.3. *Interleaved and Overlapped Pipelining*

Up until now, we have required that each node send only one message on each bus in one bus cycle and that the transmission of messages be initiated at the beginning of a bus cycle. Given these two restrictions, no specific control function was needed for the initiation of messages. However, if some node does not have a message to send during a bus cycle, a slot of one petit cycle in duration will be created. Interleaved pipelining is a technique which tries to fully utilize the communication capacity of the pipelined bus by inserting a message into any available slot. This may be accomplished if a node is allowed to place more than one message on the same bus within a bus cycle, but at different petit cycles. To allow for this flexibility, a control function $send_g(j)$ must be used to specify the time, relative to the beginning of a bus cycle, at which node j should write its gth message on the bus.

To show how interleaved pipelining works, let us now examine the routing patterns in E_{c1}. Since message transfers in opposite directions on the two buses of the linear APPB form two separate and symmetric pipelines, we need to look at only one direction. Consider the left-to-right message transfer in E_{c1}, and define k sets, $S_h = \{j \mid 0 \leq j < n, 0 \leq (j \bmod 2^h) < 2^{h-1}\}$, $1 \leq h \leq k$, of nodes for the k-cube. That is, S_h is obtained by partitioning the n nodes of the hypercube into 2^h-node groups and including in S_h the first 2^{h-1} nodes in each group. For example, for the 4-cube in Fig. 5a, we have $S_1 = \{0, 2, 4, 6, 8, 10, 12, 14\}$, $S_2 = \{0, 1, 4, 5, 8, 9, 12, 13\}$, $S_3 = \{0, 1, 2, 3, 8, 9, 10, 11\}$, and $S_4 = \{0, 1, 2, 3, 4, 5, 6, 7\}$. Note that all the k sets, S_h, have the same cardinality 2^{k-1}, and each contains node 0. Hence, in the realization of the binary k-cube using a linear APPB, there are k routing patterns. In the hth pattern, $1 \leq h \leq k$, the nodes in set S_h send messages to their neighbors along the hth dimension in the hypercube, as indicated with the arrowed curves in Fig. 5b. Correspondingly, the messages can be divided into k sets, M_h, $1 \leq h \leq k$, which are sent by the k sets of nodes S_h, respectively. For the routing patterns in Fig. 5b, these message sets are shown in Fig. 5c.

Using interleaved pipelining, the messages in the two sets M_{2s-1} and M_{2s}, $1 \leq s \leq k/2$, are interleaved and sent in the same bus cycle. Let $send_1(j)$ and $send_2(j)$ be the times at which node j writes its messages in M_{2s-1} and M_{2s}, respectively, on the bus during bus cycle s. Correspondingly, let $wait_1(i)$ and $wait_2(i)$ be the wait functions for a node i to receive the messages in M_{2s-1} and M_{2s}, respectively, during bus cycle s. Then, for interleaved pipelining we have the following *send* functions for a node j at bus cycle s, $1 \leq s \leq k/2$:

$$send_1(j) = 0, \quad j \in S_{2s-1},$$

$$send_2(j) = \begin{cases} 0, & j \in S_{2s} \text{ and } 2^{2s-2} \leq (j \bmod 2^{2s}) < 2^{2s-1}, \\ 2^{2s-2}, & j \in S_{2s} \text{ and } 0 \leq (j \bmod 2^{2s}) < 2^{2s-2}. \end{cases}$$

The corresponding *wait* functions are

$$wait_1(i) = i - j, \quad j \in S_{2s-1},$$

$$wait_2(i) = \begin{cases} i - j, & j \in S_{2s} \text{ and } 2^{2s-2} \\ & \quad \leq (j \bmod 2^{2s}) < 2^{2s-1}, \\ i - j + 2^{2s-2}, & j \in S_{2s} \text{ and} \\ & \quad 0 \leq (j \bmod 2^{2s}) < 2^{2s-2}. \end{cases}$$

A node for which the *send* or *wait* function is not defined above should not send or receive any message. Note that the times determined by these *send* and *wait* functions are with respect to the beginning of each bus cycle s. Also note that since the receiving node i knows the id of the sending node j (since they are neighbors in the k-cube), it knows which of the two values of $wait_2(i)$ should be used. As an example, the interleaved pipelining for the messages in Fig. 5c is achieved by interleaving message sets M_1 and M_2 in the first bus cycle and M_3 and M_4 in the second bus cycle. The arrowed lines in Fig. 5c show how the messages are being interleaved, and the resulting message pipelines are shown in Fig. 6a.

It can be seen that using interleaved message pipelining, the total communication time taken for each node to send a message to each of its neighbors is $k/2 + 1$ bus cycles, where the last bus cycle is due to the time needed to clear out the first $n/4$ messages (sent by nodes 0 through $n/4 - 1$) in M_k that were inserted in front of M_{k-1}. Comparing with k bus cycles, the time needed if each node sends one message per bus cycle, our savings in the communication time is $(k - 2)/2$ bus cycles. Although this savings is significant there are still unused slots from the rightmost nodes on the bus, as can be seen from the message pipeline at time $t = 16$ in Fig. 6a. We next show how to utilize these empty slots using overlapped pipelining.

In overlapped pipelining, we pipeline the message pipelines obtained from interleaved pipelining by allowing the messages for bus cycle s to be initiated before bus cycle $s - 1$ terminates, as long as message collision does not occur. For this purpose we define a new control function, $step_s$, which specifies the time, with respect to the beginning of the first bus cycle, at which the messages for bus cycle s are initiated. Clearly, savings in communication time is possible if $step_s - step_{s-1} < n\tau$. In this case, we avoid confusion by calling the bus cycles *message transfer steps*.

In E_{c1}, the control function $step_s$, $1 \leq s \leq k/2$, specifies when S_{2s-1} and S_{2s} should start sending their messages. Specifically let $step_1 = 0$ and let $step_s$, $1 < s \leq k/2$, be the time interval in number of petit cycles between the initiations of steps 1 and s. Then, messages from step s and step $s - 1$ will not collide if

$$step_s = step_{s-1} + n - \frac{3}{4}2^{2s-2}, \quad 1 < s \leq \frac{k}{2}.$$

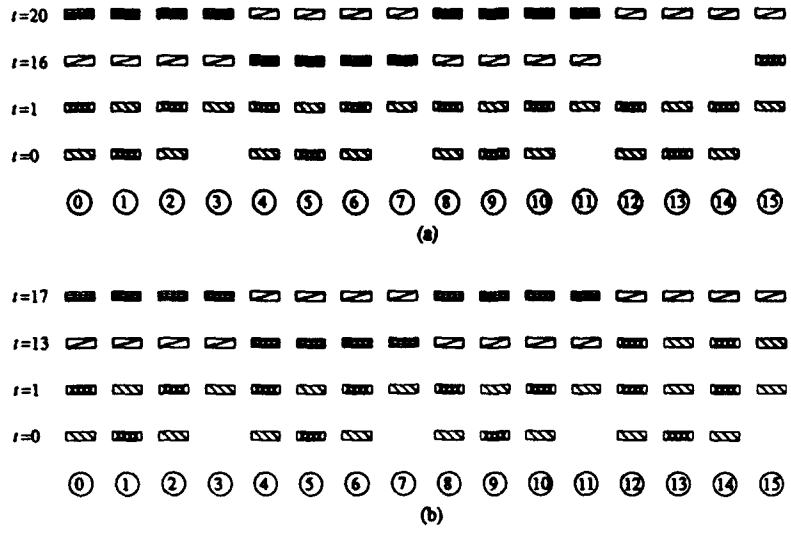


FIG. 6. (a) Interleaved pipelining and (b) overlapped pipelining of messages in the 4-cube. (t is measured in petit cycles.)

The *send* and *wait* functions defined in the previous subsection are still applicable here, but they are now defined with respect to the time determined by $step_s$, the beginning of transfer step s, rather than the beginning of each bus cycle s. Figure 6b shows the result of overlapped pipelining of the message pipelines in Fig. 6a. Note that in interleaved pipelining there was also some overlapping between the two message pipelines generated in two consecutive bus cycles, as can be seen from the message pipeline at time $t = 16$ in Fig. 6a. But, as has been mentioned previously, interleaved pipelining does not fully utilize the pipelined bus.

These control functions *step*, *send*, and *wait* together result in a minimized total communication time. To show this we first note that since the cardinality of M_h, $1 \leq h \leq k$, is $n/2$, the total number of messages is $kn/2$. Thus, if we assume that the message distribution over processors is perfectly even in each bus cycle (every processor has a message to send in each bus cycle), then the time needed for transferring these messages is at least $\lceil kn/2n \rceil = k/2$ bus cycles, or equivalently $kn/2$ petit cycles. In our case, however, such an assumption of even message distribution does not hold. For example, no message can be inserted on the bus at processor $n-1$ in the first bus cycle, as can be seen from the message pipeline at time $t = 0$ in Fig. 6b. Now we compute the total time, in number of petit cycles, using the control functions determined above. It can be shown that

$$step_{k/2} = \sum_{s=2}^{k/2} \left(n - \frac{3}{4} 2^{2s-2} \right) = \left(\frac{k}{2} - \frac{5}{4} \right) n + 1.$$

The time due to $send_2$ at step $k/2$ is $2^{k-2} = n/4$. Finally it takes n petit cycles for the bus to clear out. Therefore the total time in number of petit cycles is

$$\left(\frac{k}{2} - \frac{5}{4} \right) n + 1 + \frac{n}{4} + n = \frac{k}{2} n + 1.$$

Finally, we note that interleaved message pipelining may also be applied to binary tree routing patterns. From our previous discussion we know that the parent-to-children message routing in E_{t1} has to be done in two bus cycles and that the same message routing task can be performed using a single bus cycle in E_{t2}. Communication efficiency in E_{t2} can be further improved by using interleaved message pipelining because during parent-to-children message transfer only every other node is sending a message. Thus each parent can send two messages to each child in one bus cycle.

4. TWO-DIMENSIONAL ARRAY PROCESSORS WITH PIPELINED BUSES

Linear optical buses have the disadvantage that message transfer may incur $O(N)$ time delay in an N-processor system. To reduce this delay to $O(\sqrt{N})$, we consider two-dimensional APPBs. In a two-dimensional APPB, each node is coupled to four buses as shown in Fig. 7a, where the two horizontal buses are used for passing messages horizontally in the same way as before, and the two vertical buses are used for passing messages vertically in a similar way. For convenience we diagram our two-dimensional APPB as in Fig. 7b. Each node in a two-dimensional APPB of size $N = m \times n$ will be given two identifications, one being a pair of numbers (x, y), $0 \leq x < m$, $0 \leq y < n$, indicating the row–column position of the node in the two-dimensional APPB, and the other being the row-major index, $i = xn + y$, $0 \leq i < N$, of the node. Corresponding to the bus cycle defined for the linear case, in the two-dimensional APPB we define $n\tau$ and $m\tau$ as a *row bus cycle* and a *column bus cycle*, respectively, where τ is a petit cycle as defined previously. When there is no confusion, e.g., while talking about message transmissions in a row, we simply say a *bus cycle* instead of a *row bus cycle*.

4.1. *Message Routing in Two-Dimensional APPB*

A unique issue that arises in the two-dimensional APPB is the relay of messages. Specifically, suppose a message is to be transferred from node (x_1, y_1) to node (x_2, y_2), with $x_1 \neq x_2$ and $y_1 \neq y_2$. Then the message may first be sent from (x_1, y_1) to (x_1, y_2), which is the node at the intersection of row x_1 and column y_2, in the first bus (a row bus cycle) and then from (x_1, y_2) to (x_2, y_2) in the second bus cycle (a column bus cycle). That is, the message has to be buffered at node (x_1, y_2) at the end of the first bus cycle and then relayed to its destination in the second bus cycle. For the purpose of relaying the message, we define a control function *relay* for node (x_1, y_2) as

$$relay[(x_1, y_2)] = y_2 - y_1,$$

which indicates that node (x_1, y_2) will read a message from a row bus at time $|y_2 - y_1|$ (relative to the start of the row bus cycle) and then write that message on the proper column bus at the beginning of the following column bus cycle. If $relay[(x_1, y_2)] = 0$, then no message is to be relayed by node (x_1, y_2). Clearly, in the worst case up to n messages have to be relayed and, therefore, n relay buffers are needed at the relaying node. Now we are ready to show how the four most commonly used message routing patterns discussed in the previous section can be realized in the two-dimensional APPB.

One-to-One. The system executes a $SEND[(x_1, y_1), (x_2, y_2)]$ instruction, which requires that node (x_1, y_1) send a message to node (x_2, y_2). We have $relay[(x_1, y_2)] = y_2 - y_1$ (in row bus cycle), and $wait[(x_2, y_2)] = x_2 - x_1$ (in column bus cycle). This communication takes two bus cycles.

Broadcast. The system executes a $BROADCAST[(x, y)]$ instruction, which states that node (x, y) broadcasts the same message to all other nodes (x_j, y_j). In a row bus cycle, (x, y) broadcasts the message to nodes (x, y_j), $y_j \neq y$. Then in the following column bus cycle all (x, y_j), including (x, y), broadcast the message in their corresponding columns. Thus $relay[(x, y_j)] = y_j - y$, and $wait[(x_j, y_j)] = x_j - x$. This communication also takes two bus cycles.

Semigroup Communication. This corresponds to the execution of $SEMIGROUP[(x, y)]$, which says that some global information is to be computed and stored at node (x, y). This task can be accomplished using two linear semigroup operations, one in rows and the other in a column. That is, first we view each row as a linear APPB and do $SEMIGROUP(y)$ in all rows. Then in column y, we perform $SEMIGROUP(x)$. Thus $2 \log(n)$ bus cycles are needed for this task.

Permutations. Let $PERM[(x, y)]$ be an arbitrary permutation. To avoid using n relays at each node, we can use a three-phase routing approach [24, 32] or equivalently a three-bus-cycle approach in the two-dimensional APPB. In this approach the first bus cycle is a "preprocessing" step which distributes messages in each row such that the messages going to the same row will occupy different columns. Then the second and third bus cycles will route the messages to their destination row and destination node, respectively. We note that for arbitrary permutations this approach implies the use of a centralized controller which would compute the message destinations for the preprocessing step. This calculation requires the construction of a bipartite graph and its partitioning into complete matchings, which would dominate the time complexity for the total task of computing and implementing an arbitrary permutation. In applications where a permutation can be precomputed, this time cost can be amortized over many subsequent applications of the permutation.

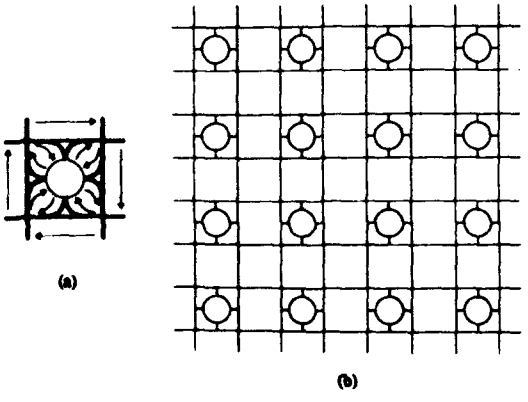


FIG. 7. Two-dimensional APPB. (a) A processor coupled to four waveguides in the two-dimensional APPB. (b) A schematic drawing of the two-dimensional APPB.

4.2. *Embedding Binary Trees in Two-Dimensional APPB*

As mentioned previously, arbitrary message routing and permutations in two-dimensional APPB may require n relaying buffers in each node in the worst case. In this subsection we present an embedding for a binary tree network in which only one relay buffer is needed to route messages. An embedding of an L-level complete binary tree into a two-dimensional APPB with $n = 2^k$ columns may be obtained by (i) mapping levels $0, \ldots, k-1$ of the tree to row 0 of the two-dimensional APPB and (ii) mapping level l, $k \leq l < L$, of the tree to the 2^{l-k} rows, $2^{l-k}, 2^{l-k}+1, \ldots, 2^{l-k+1}-1$, of the APPB such that the two children of the same parent are mapped into two adjacent rows in the same column as the parent. Specifically we define our embedding of a binary tree network into the two-dimensional APPB by a mapping $F(i) = (F_x(i), F_y(i))$, which maps each node i, $1 \leq i < 2^L$, in the tree to a node $(F_x(i), F_y(i))$ in the two-dimensional APPB. Let i be a node at level l, $0 \leq l < L$, in the binary tree. The mapping is defined by

$$F_x(i) = \begin{cases} 0, & 1 \leq i < 2^k, \\ 2^{l-k} + i \bmod 2^{l-k}, & 2^k \leq i < 2^L, \end{cases}$$

and

$$F_y(i) = \begin{cases} i, & 1 \leq i < 2^k, \\ \left\lfloor \dfrac{i \bmod 2^l}{2^{l-k}} \right\rfloor, & 2^k \leq i < 2^L. \end{cases}$$

As an example the embedding for the 4-level binary tree in Fig. 8a is shown in Fig. 8b. Let us call this embedding E_{t3}. E_{t3} has the following properties: (i) Parent nodes i, $1 \leq i < 2^{k-1}$, and their children are in row 0; (ii) parent nodes i, $2^{k-1} \leq i < 2^k$, which are in row 0, have their children in row 1; and (iii) parent nodes i, $2^k \leq i < 2^{L-1}$, and their children are in the same column. Properties (i) and (ii) are obvious. Here we prove only (iii). Since in the binary tree each parent node i has two children $2i + \delta$, $\delta = 0, 1$, to prove (iii) we need only show that $F_y(i) = F_y(2i + \delta)$ for $2^k \leq i < 2^{L-1}$. For that, let i be a parent node at level l, where $k \leq l < L - 1$ and $i = p2^l + q$ for some integers p and q such that $0 \leq q < 2^l$. Then

$$F_y(2i + \delta) = \left\lfloor \frac{(2(p2^l + q) + \delta) \bmod 2^{l+1}}{2^{l+1-k}} \right\rfloor$$

$$= \left\lfloor \frac{(p2^{l+1} + 2q + \delta) \bmod 2^{l+1}}{2^{l+1-k}} \right\rfloor$$

$$= \left\lfloor \frac{2q + \delta}{2^{l+1-k}} \right\rfloor = \left\lfloor \frac{q}{2^{l-k}} \right\rfloor = F_y(i).$$

It is now clear that the relay function is not needed for message transfer between parent nodes i and their children if $1 \leq i < 2^{k-1}$ or $2^k \leq i < 2^{L-1}$. However, such a relay is needed if $2^{k-1} \leq i < 2^k$. The wait and relay functions for E_{t3} are obtained in the following.

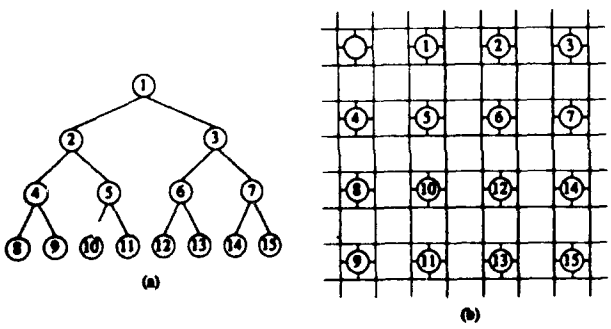


FIG. 8. (a) A 4-level binary tree. (b) Its embedding, E_{t3}, in the two-dimensional APPB.

Let $wait_{c,\delta}[(x, y)]$, where $(x, y) = F(i)$, be the *wait* functions for a parent node i to receive a message from its left and right child for $\delta = 0$ and 1, respectively. For the case $1 \leq i < 2^{k-1}$, the results for the linear APPB directly give $wait_{c,\delta}[(x, y)] = -(y + \delta)$. For the case $2^k \leq i < 2^{L-1}$, let i be at level l, $k \leq l < L - 1$, and $i = p2^{l-k} + q$. Then

$$F_x(i) = 2^{l-k} + i \bmod 2^{l-k} = 2^{l-k} + q;$$

$$F_x(2i + \delta) = 2^{l+1-k} + (2i + \delta) \bmod 2^{l+1-k}$$

$$= 2^{l+1-k} + (p2^{l+1-k} + 2q + \delta) \bmod 2^{l+1-k}$$

$$= 2^{l+1-k} + 2q + \delta; \; wait_{c,\delta}[(x, y)] = F_x(i) - F_x(2i + \delta)$$

$$= (2^{l-k} + q) - (2^{l+1-k} + 2q + \delta) = -(x + \delta).$$

$wait_{p,\delta}(j)$ can be obtained by recalling that $wait_{p,\delta}(j) = -wait_{c,\delta}(i)$, where i is the parent of j.

For the case where $2^{k-1} \leq i < 2^k$, a *wait* and a *relay* function are needed. Let $relay_{c,\delta}[(0, y)]$, $0 \leq y < 2^k$, be the *relay* function of node $(0, y)$ for relaying the message from a child node (again, $\delta = 0$ for the left child and $\delta = 1$ for the right child) to its parent. Then we can show that

$$relay_{c,\delta}[(0, y)] = -1, \quad 0 \leq y < 2^k,$$

$$wait_{c,\delta}[(0, y)] = 2^k - y - \delta, \quad 2^{k-1} \leq y < 2^k.$$

Note that each node $(0, y)$ needs to relay only *one* child-to-parent message with the message from left (right) child being relayed by $(0, y)$ with y even (odd), and that even though node 0 is not a node in the tree, it helps relay messages. Also note that $relay_{c,\delta}$ is applicable to column bus cycles. Now let $relay_{p,\delta}[(0, y)]$, y even (odd), be the relay function for node $(0, y)$ to relay the message from a parent $(0, Y)$ to its left (right) child for $\delta = 0$ (1). Then $relay_{p,\delta}$ is easily obtained from $relay_{p,\delta}[(0, y)] = -wait_{c,\delta}[(0, Y)]$. And $wait_{p,\delta}$ is determined as in the linear case.

4.3. *Network Embeddings Requiring No Relays*

Embedding E_{t3} still requires one message relay for communication between two neighboring nodes in binary trees. To further improve the communication efficiency, in this subsection we show how to obtain embeddings of binary trees as well as hypercubes such that no such message relay is needed. Two approaches may be used to eliminate message relays by intermediate nodes: a hardware approach and a "software" approach. In the hardware approach, optical switches are used at the intersections of row and column buses to switch an optical signal, say, from a row bus to a column bus, without requiring relay by an intermediate processor [15]. In this paper we consider the "software" ap-

proach, which relies on designing embeddings such that all neighboring processors in a network are mapped into the same row or column in the two-dimensional APPB. Thus, no message relay is needed and no *relay* function is required. This improves the communication efficiency significantly. However, it has the disadvantage that nodes in the APPB may not be fully utilized.

A basic measure that is usually used to evaluate the quality of an embedding of a source graph $G_1 = \{V_1, U_1\}$ with a set of nodes V_1 and a set of edges U_1 into a mesh architecture with a set of nodes V_2 is the *expansion cost*, which is defined as the ratio of the number of nodes in the target mesh to the number of nodes in the embedded graph. Another measure useful for such evaluation is the *dilation cost*. Specifically, the dilation of an edge $u \in U_1$, which is mapped to a path Q in the target mesh, is $|Q| - 1$, where $|Q|$ is the number of nodes on Q. However, the mesh model corresponding to that of APPBs is different from those studied previously [1, 13, 34] because the efficiency of the communication between any two nodes in the same row or column in an APPB does not depend on the distance between these two nodes. Therefore the criterion that is to be satisfied by an embedding is different from previously studied criteria. Specifically, it is desirable to obtain an embedding in which any two neighboring nodes in the source graph are mapped into either the same row or the same column in the two-dimensional APPB, thus allowing them to communicate with each other using a single bus cycle. An embedding which satisfies this requirement will be said to satisfy the *alignment condition*. Note that E_{t3} obtained in the previous subsection does not satisfy the alignment condition and thus requires message relays. That embedding, however, does have an optimal expansion cost of $2^L/(2^L - 1)$. In contrast, the binary tree embedding presented in the following satisfies the alignment condition, but its expansion cost is not optimal. This demonstrates a trade-off between the expansion cost and the dilation cost for network embeddings in the two-dimensional APPB.

Consider Fig. 9a and assume that we already have an embedding of an s-level binary tree with $N_s = 2^s - 1$ nodes into a two-dimensional APPB of size $a_s \times b_s$. The embedding is assumed to satisfy the alignment condition. That is, all the neighboring nodes in the s-level tree are mapped into the same row or column in the two-dimensional APPB. Using this level s embedding (starting level) as building blocks, the embedding for an $(s + 2)$-level tree is obtained as shown in Fig. 9b. Clearly in this embedding the neighboring nodes are again on the same row or column. A still larger tree is obtained by repeating this modular building procedure until the desired size is achieved. Let us call this embedding E_{t4}. Assuming that in E_{t4} the embedding of an L-level tree, $L = s + 2\sigma$, $\sigma = 0, 1, \ldots$, occupies an area, in number of nodes, equal to A_L in the two-dimensional APPB, we may inductively prove that

$$A_L = 2^{L-s}[A_s + (1 - 2^{-(L-s)/2})b_s].$$

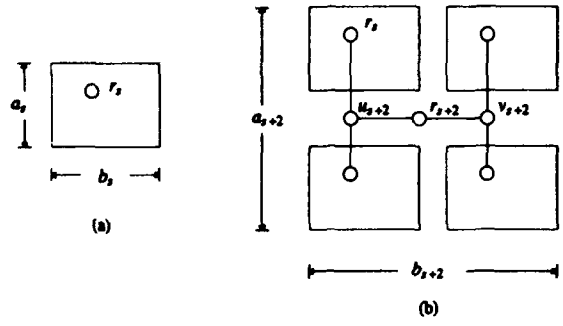


FIG. 9. Modular embedding of binary trees, E_{t4}, in the two-dimensional APPB. (a) A building block in which an s-level binary tree is embedded. (b) Embedding of an $(s + 2)$-level binary tree.

With this result, the expansion cost for the embedding of an L-level tree is

$$C_L = \frac{A_L}{N_L} = \frac{2^{L-s}[A_s + (1 - 2^{-(L-s)/2})b_s]}{2^L - 1} = \frac{2^{L-s}[A_s + (1 - 2^{-(L-s)/2})b_s]}{2^{L-s}(N_s + 1) - 1}.$$

It can be checked that C_L is monotonically increasing with L. However, for large L, the value of C_L asymptotically equals

$$C_{L,\max} = \frac{A_s + b_s}{N_s + 1}.$$

Note that, if $N_s \gg 1$ and $A_s \gg a_s$, the value of C_L simplifies to $C_L \approx A_s/N_s = C_s \geq 1$. That is, the expansion cost for the entire embedding is determined by the expansion cost of the building block. Thus low expansion costs may be obtained if the starting building block satisfies $N_s \gg 1$, $A_s \gg b_s$, and $C_s \to 1$. Some examples of building blocks are shown in Fig. 10 with their corresponding expansion cost $C_{L,\max}$. Note that in this modular embedding scheme, as the embedding goes one level higher, the number of levels of the tree increases by 2. Thus if s is even (odd) then L is even (odd). Therefore according to whether the desired level L of the tree is even or odd, the starting level s must be chosen properly.

To determine the control functions for E_{t4}, let r_l be the root at an embedding level l, $l = s + 2, s + 4, \ldots, L$, and (x_l, y_l) be the coordinate, i.e., the row–column position, of r_l in the two-dimensional APPB. Then from Fig. 9b, the coordinate of r_l is

$$(x_l, y_l) = (a_{l-2}, b_{l-2} - 1),$$

where a_l and b_l can be found to be equal to $2^{(l-s)/2}(a_s + 1) - 1$ and $2^{(l-s)/2}b_s$, respectively. Thus,

$$(x_l, y_l) = (2^{(l-s-2)/2}(a_s + 1) - 1, 2^{(l-s-2)/2}b_s - 1).$$

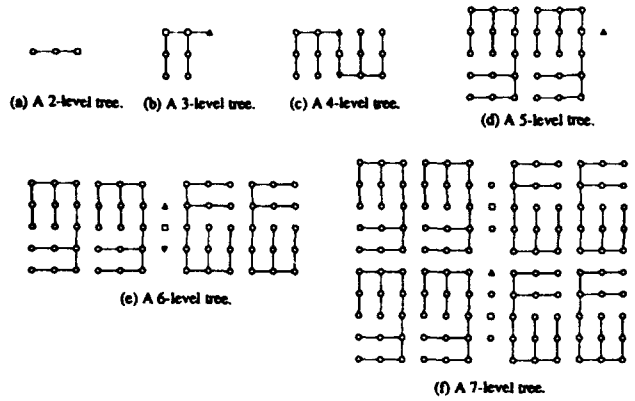


FIG. 10. Example building blocks for the modular embedding of binary trees, E_{t4}, and their corresponding expansion costs $C_{L,\max}$: (a) 1.5, (b) 1.5, (c) 1.25, (d) 1.31, (e) 1.22, (f) 1.12.

Now the control functions can be determined as follows. First, within the building block determine the *wait* functions according to the specific building block in use. Let (x_s, y_s) be the coordinate of r_s, the root in the building block. We then need only determine the *wait* functions for the new nodes which appear as we go to a higher-level embedding. For example, in Fig. 9b when we go from level s to $s + 2$, the new nodes are r_{s+2}, u_{s+2}, and v_{s+2}. By letting $wait_{c,u_l}(r_l)$ be the *wait* function for node r_l to receive a message from child node u_l, we have

$$wait_{c,u_l}(r_l) = y_l - y_{l-2},$$

$$wait_{c,v_l}(r_l) = -(y_l - y_{l-2} + 1),$$

$$wait_{c,r_{l-2}}(u_l) = wait_{c,r_{l-2}}(v_l) = \pm(x_l - x_{l-2}),$$

where the coordinate (x_l, y_l) is as determined previously. These are the *wait* functions for the new parents to receive messages from their children. The *wait* functions for the children to receive messages from these new parents are obtained by recalling that $wait_p = -wait_c$.

Next we show that the binary hypercube of 2^{2k} nodes can also be embedded in a two-dimensional APPB of size $2^k \times 2^k$ such that the alignment condition is satisfied. As in the case of binary trees, the embedding is again modular with the basic module being the binary 2-cube shown in Fig. 11a. A 3-cube embedding is obtained by putting together two such 2-cubes side-by-side as shown in Fig. 11b, and a 4-cube embedding is obtained by putting together two 3-cubes one above the other as shown in Fig. 11c and so on. Note that the nodes in Fig. 11c correspond to the cube nodes of Fig. 5a. In this way the embedding, denoted E_{c2}, of the binary hypercube of the desired size is obtained modularly.

It is observed that in embedding E_{c2}, each row and column is itself a binary k-cube. For example, if we take the column number y as the node id for the nodes in any row x, then row x is a binary k-cube consisting of nodes y, $0 \leq y < 2^k$. Let us call each row or column a *subcube*. Then we have 2^{k+1} such subcubes. For each subcube, if we use the column id y (or the row id x) to identify its nodes, all the control functions *step*, *send*, and *wait* are exactly the same as those derived for E_{c1} in the linear APPB. Thus the total communication time for emulating the hypercube can be minimized through overlapped pipelining as presented in the previous section. It can be seen that all the neighboring nodes in the hypercube are mapped to either the same row or the same column in the two-dimensional APPB. Therefore E_{c2} satisfies the alignment condition and thus requires no message relay for communications between neighboring nodes in the hypercube. Finally, since the number of nodes used in the two-dimensional APPB is equal to that of the hypercube, we achieve a minimal expansion cost of unity.

5. BANDWIDTH ANALYSIS

In this section, we evaluate the merit of the pipelined communication structure by comparing it with linear arrays which utilize nearest-neighbor and exclusive access bus interconnections. We evaluate the different models irrespective of the technology used to implement them. In other words, we assume that the transmission rate and the propagation delay are the same for both optical and electronic communication links.

Consider the linear array of n processors with nearest-neighbor connections as shown in Fig. 1b and assume that the physical separation between each pair of neighboring processors is D. Such an array may emulate one cycle of a pipelined bus in a time $n(T_p + T_D)$, where T_D is the propagation time required for a signal to travel the distance D and T_p is the time required to process a message at the sending and the receiving ends of a communication link. T_p includes synchronization, message generation, buffering, and routing. We note that for the cases of interleaved and overlapped pipelining discussed in Section 3.3, at most two messages might be processed in this time. The bandwidth of the nearest-neighbor connected array, B_a, defined as the maximum number of messages that may be transmitted per second, is thus given by

$$B_a = \frac{n}{n(T_p + T_D)} = \frac{1}{T_D}\frac{1}{\rho + 1},$$

where $\rho = T_p/T_D$.

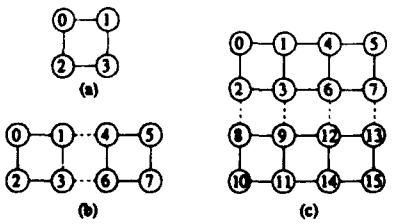


FIG. 11. Modular embedding of binary hypercubes, E_{c2}, in the two-dimensional APPB. (a) 2-cube. (b) 3-cube. (c) 4-cube with the node labeling corresponding to that in Fig. 5a.

For the pipelined linear APPB, the optical distance, D_0, between two consecutive processors should be larger than the message length bwc_g (see Eq. (1) in Section 2). In other words, if $D \geq bwc_g$, we set $D_0 = D$; otherwise D_0 should be made equal to bwc_g (for example, by coiling an optical fiber) so that each processor can inject a message into the bus without collision. Thus, the signal propagation time, T_{D_0}, between two consecutive processors is $\max\{T_D, \alpha T_D\}$, where $\alpha = (bwc_g)/D$. The pipelined bus cycle time is then $T_p + nT_D\max\{1, \alpha\}$. Given that n messages may be transmitted during a pipelined bus cycle, the bandwidth of the pipelined bus is

$$B_p = \frac{n}{T_p + nT_D\max\{1, \alpha\}}, \tag{2}$$

and thus,

$$\frac{B_p}{B_a} = \frac{n(\rho + 1)}{\rho + n\max\{1, \alpha\}}. \tag{3}$$

In Fig. 12a a parametric plot showing the relation between B_p/B_a and ρ is given in terms of n for $\alpha \leq 1$ and $\alpha > 1$. The curve for $\alpha \leq 1$ corresponds to the case where the message length is less than or equal to the physical separation between processors, while the curve for $\alpha > 1$ reflects the case where message length is longer than the physical separation between processors, and thus the optical path has been extended to accommodate the entire message. By taking the limit of Eq. (3) as $\rho \to \infty$, it is clear that, for fixed α and large ρ, the ratio B_p/B_a approaches n. Also, when $\rho = 1$ and $\alpha \leq 1$, we obtain $B_p/B_a \approx 2$. In Fig. 12b we plot B_p/B_a versus ρ for a fixed-size array with $n = 64$ and for several values of α. These plots show that the pipelined bus is more effective for larger values of ρ and smaller values of α.

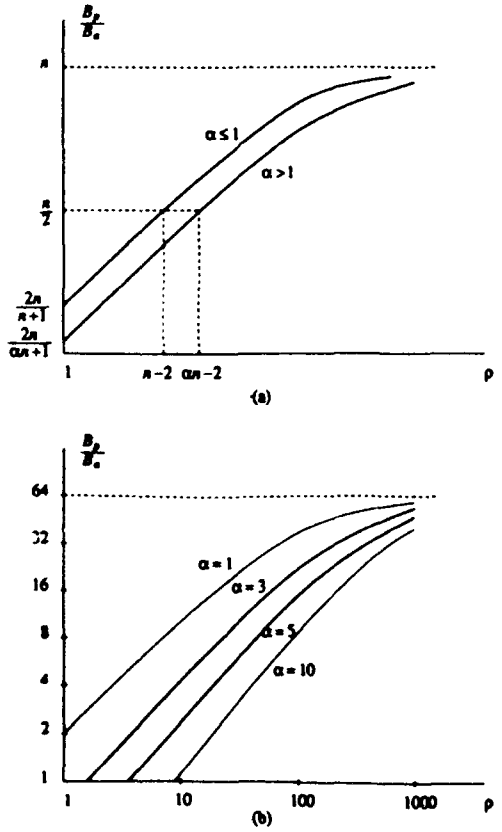


FIG. 12. The ratio, B_p/B_a, of the bandwidth of a pipelined bus to that of a linear array with nearest-neighbor connections as a function of ρ, α, and n. (a) A parametric curve. (b) For a fixed-size system with $n = 64$.

For multiprocessor interconnections, D is determined by placement and routing within VLSI chips, by PC board connections, or by back-plane interconnections. In all cases, D, and therefore T_D, is relatively small. Given that T_p is, at least, on the order of microseconds, the ratio, ρ, of processing to communication times should be much larger than 1 (on the order of 10–1000). Also, with current technology it is reasonable to assume that α is relatively small (between 1 and 10). For example, for board-to-board communications ($D \approx 10$ cm), it is possible to drive an optical communication line at the speed of 10 GHz. Assuming that the speed of light in optical fibers is $c_g = 2 \times 10^8$ m/s, and that each message contains $b = 16$ bits, we obtain $\alpha \approx 3$. The same value of α is obtained if optical communications are implemented on GaAs wafers at 100 GHz and a physical processor separation of 1 cm. Note that the value of α may be reduced if parallel buses are used to reduce b.

Next we compare the bandwidth of a pipelined bus with that of an exclusive access bus. Given that the bandwidth of an exclusive access bus is $B_e = 1/(T_p + nT_D)$, we have

$$\frac{B_p}{B_e} = \frac{n(\rho + n)}{\rho + n\max\{1, \alpha\}}.$$

This shows that as α approaches 1, the pipelined bus can accommodate n messages in the same cycle time as the exclusive access bus. For larger α, the pipelined bus cycle will be stretched to accommodate the length of the messages, and thus, the performance gain due to pipelining will be less than n.

The above analysis is independent of the media used for communication. If optical pipelined buses are to be compared with electronic buses, then the physical constraints on the electronic propagation speed should be taken into account. Specifically, the effect of capacitive loading and mutual inductance on the signal propagation speed (the transmission line effect) should be considered. Thus, message pipelining using electro-optical technology offers a potential for substantially enhancing bandwidth utilization. Further, pipelining techniques will be of increasing effectiveness because this technology offers the capability of generating very short pulses [12, 33], thus reducing w and decreasing α.

6. CONCLUDING REMARKS

We have presented efficient communication architectures which exploit the optical signal's properties of unidirectional propagation and predictable path delays in order to pipeline messages on optical buses. As shown in Section 5, the pipelined model has its merits irrespective of the technology in which it is implemented. Although the presentation in this paper is based on an optical model in which delays inherent in optical fibers serve as slots for space multiplexing, it is possible to use shift registers as buffer memories for these slots [36]. Thus pipelined buses may be implemented in either optics or electronics. However, for the electronic implementation, the signal propagation delay, T_D, will depend on the speed of the shift registers, resulting in a relatively small value for the ratio of processing to communication times, ρ.

We proposed efficient approaches to fundamental message routings including one-to-one, broadcast, semigroup communications, and permutations for the APPB architectures. Such efficient accomplishment of these commonly used message routing patterns can significantly improve the efficiency of many parallel algorithms. We presented here efficient embeddings of the binary trees and hypercube networks. Embeddings for other well-known interconnection networks, including pyramids, shuffle-exchange networks, X-binary-trees [9], and X-quad-trees, have also been obtained [14, 16]. Such efficient embeddings of these well-known communication structures allow all algorithms designed for these structures to be efficiently executed on the APPB architectures. They also allow an APPB to be logically reconfigured as an architecture which is more suitable for a given computation task.

We have not considered in this paper several issues that are relevant to the implementation of the proposed architectures. Such issues include the synchronization of the processors to the accuracy implied by the speed of optics, temporal pulse positioning, optical fanout, and the distribution of optical power in a way that allows the detector at each processor to detect the optical signals correctly. These issues must be addressed with regard to the reliability, scale, and device technology which is appropriate for computing applications. Some of these issues have been presented in [7, 25, 31].

In our experimental work [6, 8, 21] we are investigating the practical limits to these technological concerns. We have shown that three factors, threshold power margin, synchronization error, and coupling ratio, determine the system scale. On the basis of current and near-term technology, our experiments show that synchronization error does not contribute significantly to the bounds of system size. Rather, power distribution effects dominate. Preliminary investigations show that by using off-the-shelf optical components we can currently build linear buses operating at 300 MHz and containing about 100 processors. Using more sophisticated electro-optics (gallium arsenide, custom couplers, and dual level bus structures) we believe that 10-GHz buses of over 400 processors are feasible. Further, we believe that near-term technologies such as fiber amplifiers as well as alternate bus structures will alleviate the power distribution problem.

REFERENCES

1. Bailey, D., and Cuny, J. An efficient embedding of large trees in processor grids. *Proc. 1986 International Conference on Parallel Processing.* IEEE Computer Society, Silver Spring, MD, 1986, pp. 819–822.
2. Batcher, K. E. Design of a massively parallel processor. *IEEE Trans. Comput.* C-29, 9 (1980), 836–840.
3. Bokhari, S. H. Finding maximum on an array processor with a global bus. *IEEE Trans. Comput.* C-32, 2 (1984), 133–139.
4. Chen, Y. C., Chen, W. T., Chen, G. H., and Sheu, J. P. Designing efficient parallel algorithms on mesh-connected computers with multiple broadcasting. *IEEE Trans. Parallel Distrib. Systems* 1, 2 (1990), 241–245.
5. Chiarulli, D. M., Melhem, R. G., and Levitan, S. P. Using coincident optical pulses for parallel memory addressing. *IEEE Comput.*, (Dec. 1987), 48–57.
6. Chiarulli, D. M., Levitan, S. P., and Melhem, R. G. Self routing interconnection structures using coincident pulse techniques. *SPIE Proc. International Symposium on Advances in Interconnects and Packaging,* Boston, MA, 1990, Vol. 1390.
7. Chiarulli, D. M., Levitan, S. P., and Melhem, R. G. Optical bus control for distributed multiprocessors. *J. Parallel Distrib. Comput.* 10 (1990), 45–54.
8. Chiarulli, D. M., Levitan, S. P., and Melhem, R. G. Demonstration of an all optical addressing circuit. *Proc. OSA Topical Meeting on Optical Comput.*, Salt Lake City, UT, 1991, pp. 235–238.
9. Despain, A. M., and Patterson, D. A. X-tree: A tree structured multiprocessor computer architecture. *Proc. 5th International Symposium on Computer Architecture,* 1978, pp. 144–151.
10. Duff, M. J. B., Watson, D. M., Fountain, T. J., and Shaw, G. K. A cellular logic array for image processing. *Pattern Recognition* 5 (1973), 229–237.
11. Duff, M. J. B., and Fountain, T. J. *Cellular Logic Image Processing.* Academic Press, New York, 1986.
12. Fujimoto, J., Weiner, A., and Ippen, E. Generation and measurement of optical pulses as short as 16 fs. *Appl. Phys. Lett.* 44 (1984), 832–834.
13. Gordon, D., Koren, I., and Silberman, G. Embedding tree structures in VLSI hexagonal arrays. *IEEE Trans. Comput.* C-33, 1 (1984), 104–107.
14. Guo, Z. Array processors with pipelined busses and their implication in optically and electronically interconnected multiprocessor architectures. Ph.D. thesis, Department of Electrical Engineering, University of Pittsburgh, 1991.
15. Guo, Z., Melhem, R. G., Hall, R. W., Chiarulli, D. M., and Levitan, S. P. Array processors with pipelined optical busses. *Proc. 3rd Symposium on Frontiers of Massively Parallel Computation,* 1990, pp. 333–342.
16. Guo, Z., and Melhem, R. G. Embedding pyramids in array processors with pipelined busses. *Proc. International Conference on Application Specific Array Processors,* 1990, pp. 665–676.
17. Hunt, D. J. The ICL DAP and its application to image processing. In Duff, M. J. B., and Levialdi, S. (Eds.). *Languages and Architectures for Image Processing.* Academic Press, San Diego, CA, 1981.
18. Jrad, A. M., and Hall, R. W. The OFC enhanced mesh architecture: A performance study. *Proc. 1987 Workshop on Computer Architecture for Pattern Analysis and Machine Intelligence,* 1987, pp. 184–191.

19. Jrad, A. M., and Hall, R. W. Orthogonal fast channels: An enhanced mesh architecture. *Proc. 1987 International Conference on Parallel Processing*. IEEE Computer Society, Silver Spring, MD, 1987, pp. 828–831.
20. Kawasaki, B. S., Hill, K. O., and Lamont, R. G. Biconical-taper single-mode fiber coupler. *Opt. Lett.* **6**, 7 (1981), 327–328.
21. Levitan, S. P., Chiarulli, D. M., and Melhem, R. G. Coincident pulse techniques for multiprocessor interconnection structures. *Appl. Opt.* **29**, 14 (1990), 2024–2033.
22. Melhem, R. G., Chiarulli, D. M., and Levitan, S. P. Space multiplexing of waveguides in optically interconnected multiprocessor systems. *Comput. J.* **32**, 4 (1989), 362–369.
23. Miller, R., and Stout, Q. F. Mesh computer algorithms for computational geometry. *IEEE Trans. Comput.* **C-38**, 3 (1989), 321–340.
24. Misra, M., and Prasanna-Kumar, V. K. Efficient VLSI implementation of iterative solutions to sparse linear systems. Tech. Rep. IRIS 246, University of Southern California, 1988.
25. Nassehi, M., Tobagi, F., and Marhic, M. Fiber optic configurations for local area networks. *IEEE J. Selected Areas Commun.* **SAC-3**, 6 (1985), 941–949.
26. Nassimi, D., and Sahni, S. Data broadcasting in SIMD computers. *IEEE Trans. Comput.* **C-30**, 5 (1981), 101–107.
27. Nath, D., Maheshwari, S. N., and Bhatt, P. C. P. Efficient VLSI networks for parallel processing on orthogonal trees. *IEEE Trans. Comput.* **C-32**, 6 (1983), 569–581.
28. Prasanna-Kumar, V. K., and Eshaghian, M. M. Parallel geometric algorithms for digitized pictures on mesh of trees. *Proc. 1986 International Conference on Parallel Processing* IEEE Computer Society, Silver Spring, MD, 1986, pp. 270–273.
29. Prasanna-Kumar, V. K., and Raghavendra, C. S. Array processor with multiple broadcasting. *J. Parallel Distrib. Comput.* **4** (1987), 173–190.
30. Prasanna-Kumar, V. K., and Reisis, D. Image computations on meshes with multiple broadcast. *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-11**, 11 (1989), 1194–1202.
31. Prucnal, P., Blumenthal, D., and Perrier, P. Self routing photonic switching demonstration with optical control. *Opt. Engrg.* **26**, 5 (1987), 473–477.
32. Raghavendra, C. S., and Prasanna-Kumar, V. K. Permutations on Illiac-IV type networks. *IEEE Trans. Comput.* **C-37**, 7 (1986), 662–669.
33. Shank, C. The role of ultrafast optical pulses in high speed electronics. In Morou, G., Bloom, D., and Lee, C. (Eds.). *Picosecond Electronics and Opto-Electronics*. Springer-Verlag, New York, 1985.
34. Singh, A. Near optimal embedding of binary tree architecture in VLSI. *Proc. 8th Symposium on Distributed Computing Systems*, 1988, pp. 86–93.
35. Stout, Q. F. Mesh connected computers with broadcasting. *IEEE Trans. Comput.* **C-32**, 9 (1983), 826–830.
36. Tanenbaum, A. S. *Computer Networks*. Prentice-Hall, Englewood Cliffs, NJ, 1981.
37. Thompson, C. D., and Kung, H. T. Sorting on a mesh-connected parallel computer. *Commun. ACM* **20**, 4 (1977), 263–271.
38. Tobagi, F., Borgonovo, F., and Fratta, L. Expressnet: A high-performance integrated-services local area network. *IEEE J. Selected Areas Commun.* **SAC-1**, 5 (1983), 898–912.
39. Ullman, J. D. *Computational Aspects of VLSI*. Computer Science Press, Rockville, MD, 1984.
40. Whalen, M. S., and Wood, T. H. Effectively nonreciprocal evanescent-wave optical-fibre directional coupler. *Electron. Lett.* **21**, 5 (1985), 175–176.

Received April 6, 1990; accepted January 11, 1991

ZICHENG GUO is finishing his Ph.D. in the Department of Electrical Engineering at the University of Pittsburgh. His current research interests include parallel computer architectures and algorithms, optical communications in multiprocessor networks, and image computation and pattern recognition.

RAMI G. MELHEM is an associate professor of computer science at the University of Pittsburgh. He received a B.E. in electrical engineering from Cairo University, Egypt, in 1976, an M.S. in mathematics/computer science from the University of Pittsburgh in 1981, and a Ph.D. in computer science from the University of Pittsburgh in December 1983. He has been an assistant professor of computer science at Purdue University from 1984 to 1986 and at the University of Pittsburgh from 1986 to 1989. His research interests include optical computing, parallel systems, fault-tolerant systems, and the application of large computational arrays to scientific problems.

RICHARD W. HALL received the B.S.E. degree in electrical engineering from The Evening College of the Johns Hopkins University in 1969 as part of the Westinghouse–Johns Hopkins Awards Program and the M.S. and Ph.D. degrees in electrical engineering from Northwestern University in 1971 and 1975, respectively. He joined the Department of Electrical Engineering at the University of Pittsburgh in 1975 and is currently an associate professor in that department. His current research interests are in the study of parallel algorithms and architectures for visual information processing.

DONALD M. CHIARULLI is an assistant professor of computer science at the University of Pittsburgh. He received a B.S. degree in physics from Louisiana State University in 1976, an M.S. degree in computer science from Virginia Polytechnic Institute in 1979, and a Ph.D. in computer science from Louisiana State University in 1986. From 1979 to 1983, he was President of Datanet Services Inc., a consulting and software development firm. While at Louisiana State he was responsible for the design and construction of The Factoring Machine, a reconfigurable VLIW machine for factoring large numbers. Dr. Chiarulli's current research interests include hybrid optical/electronic computer architecture, optical interconnects, VLSI design, and parallel computation. He is a member of the IEEE Computer Society, ACM, SPIE, and the Optical Society of America.

STEVEN P. LEVITAN is the Wellington C. Carl Assistant Professor of Electrical Engineering at the University of Pittsburgh. He received the B.S. degree from Case Western Reserve University (1972) and his M.S. (1979) and Ph.D. (1984) degrees, both in computer science, from the University of Massachusetts, Amherst. He worked for Xylogic Systems, designing hardware for computerized text processing systems, and for Digital Equipment Corp. on the Silicon Synthesis project. 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. Dr. Levitan's research interests include computer-aided design for VLSI, parallel computer architecture, parallel algorithm design, and VLSI design. He is a member of the IEEE Computer Society, ACM, SPIE, and OSA.

INTERNATIONAL JOURNAL OF OPTICAL COMPUTING, Vol. 2, 31–48 (1991)

OPTICAL MULTICASTING IN LINEAR ARRAYS

CHUNMING QIAO, RAMI G. MELHEM, DONALD M. CHIARULLI AND STEVEN P. LEVITAN

Department of Computer Science and Department of Electrical Engineering, University of Pittsburgh, Pittsburgh, PA 15260, U.S.A.

SUMMARY

In this paper, we use coincident pulse techniques to implement multicasting among processors connected by optical buses. First, we discuss two basic models of a unary addressing implementation. To reduce addressing latency and overcome system size limits, we propose a two-level addressing implementation in which multicasting introduces the problem of possibly addressing unintended processors (called *shadows*). We show how additional addressing pulses can be used to reduce these *shadows*. For regular multicasting patterns such as those often found in image processing and scientific applications, a shadow-free partition of the group to be multicasted can be systematically constructed. For arbitrary multicasting patterns, a simple, incremental partitioning algorithm is introduced. In summary, the two-level addressing implementation results in higher efficiency, lower minimum optical path requirements and potentially large speed-ups over the unary addressing.

KEYWORDS Multicasting Coincident pulse addressing Optical waveguides Shadow-free partition

1. INTRODUCTION

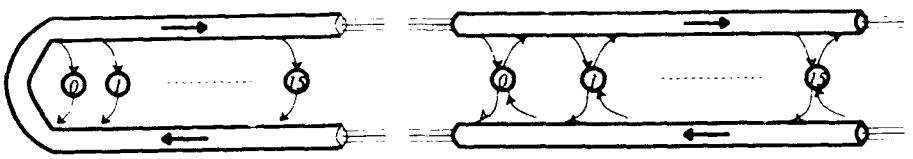
Coincident pulse techniques are based on two properties of optical pulse transmission, namely unidirectional propagation and predictable propagation delay per unit length. The technique was first introduced in the context of parallel memory addressing but was also applied to multiprocessor interconnection structures.[3,9,14] In this paper, coincident pulse techniques will be applied as an addressing mechanism for multicasting among optical bus connected processors.

In Section 2, we first review coincident pulse techniques as addressing mechanisms in two models of optical bus connected multiprocessor systems. We then show how multicasting can be implemented using unary addressing. In Section 3, two-level addressing is proposed to reduce the addressing latency and overcome the system size limit imposed by the unary addressing. However, multicasting with two-level addressing introduces the problem of possibly addressing unintended processors (called *shadows*). Simulation results of shadow reduction using additional addressing pulses are given. In Section 4, we show how shadows can be avoided by partitioning the group to be multicasted into shadow-free (SF) subgroups. We also show that speed-ups over unary addressing can be achieved when multicasting using two-level addressing. Finally, we conclude the paper in Section 5.

‡ A preliminary short version of this paper appears in the proceedings of 1991 International Conference on Parallel Processing

1047-8507/91/010031-18$09.00

Received 25 January 1991
Revised 29 May 1991



(a) Folded Bus (b) Dual Bus

Figure 1. Two basic models

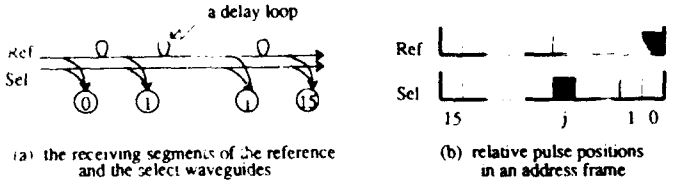
2. COINCIDENT PULSE ADDRESSING

As an introduction to using coincident pulse techniques as an addressing mechanism, we discuss two models of multiprocessor systems in which processors are connected by optical buses.[11] In the first model, called the *folded bus* model (see Figure 1(a)), each processor transmits on the lower half segment of a bus, while receiving from the upper half segment. In the second model, called the *dual bus* model (see Figure 1(b)), each processor is connected to two buses, one for downstream transmitting and upstream receiving and the other for upstream transmitting and downstream receiving.

An optical bus consists of three waveguides, one for carrying messages, one for carrying *reference pulses* and one for carrying *select pulses*, which we call the *message waveguide*, the *reference waveguide* and the *select waveguide* respectively. Messages are organized as *message frames*, which have a certain fixed length. The propagation delay on the reference waveguide is the same as that on the message waveguide but not the same as that on the select waveguide. A fixed amount of additional delay, which we show as loops (see Figure 2), are inserted onto the reference waveguide and the message waveguide.

The basic idea of using coincident pulse techniques as an addressing mechanism is as follows. Addressing of a destination processor is done by the source processor which sends a reference pulse and a select pulse with appropriate delays, so that after these two pulses propagate through their corresponding waveguides, a coincidence of the two occurs at the desired destination. The source processor also sends a message frame which propagates synchronously with the reference pulse. Whenever a processor detects a coincidence of a reference pulse and a select pulse, it reads the message frame. In essence, the address of a destination processor is unary encoded by the source processor using the relative transmission time of a reference pulse and a select pulse.

More specifically, let w be the pulse duration in seconds, and let c_b be the velocity of light in these waveguides. Define a unit delay to be the spatial length of a single optical pulse, that is $w \times c_b$. Starting with the fact that all three waveguides have equal intrinsic



(a) the receiving segments of the reference and the select waveguides (b) relative pulse positions in an address frame

Figure 2. A unary addressing implementation

propagation delays, we add one unit delay (shown as one loop) on the reference waveguide and the message waveguide between any two adjacent receivers. In the folded bus model, this means that one unit delay is added between any two processors on the upper half (receiving) segment of the reference waveguide and the message waveguide as shown in Figure 2(a). Since there are no changes on the lower half (transmitting) segments of any waveguide and the message waveguide has exactly the same length as the reference waveguide, Figure 2(a) shows only the upper receiving segments of the select waveguide and the reference waveguide for a 16-processor system with a unary addressing implementation. Let T_{ref} be the time when processor i transmits its reference pulse and $T_{sel}(j)$ be the time when it transmits a select pulse. With delays added on the reference waveguide as in Figure 2(a), these two pulses will coincide at processor j if and only if

$$T_{sel}(j) = T_{ref} + j \tag{1}$$

where $0 \leq i, j < N$ and N is the total number of processors in the system.

This means that for a given reference pulse transmitted at time t, the presence of a select pulse at time $t + j$ will address processor j while the absence of a select pulse at that time will not. Since we have $0 \leq (T_{sel}(j) - T_{ref}) < N$, it is clear that N time units are needed to encode the complete address information of N processors with a unary addressing implementation. We define an *address frame* to be the address information in the form of a sequence of either the presence or the absence of select pulses relative to a given reference pulse. With a unary addressing implementation, an address frame has a length of N units long.

Figure 2(b) shows the position of the reference pulse and the select pulse addressing processor j at the transmission time of an address frame in the folded bus model. The relative position of the select pulse to the reference pulse will remain the same from the time the address frame is transmitted to the time it finishes propagation through the transmitting segment of the bus. However, the relative position will be changed as the address frame propagates through the receiving segment of the bus.

From the value of the term $(T_{sel}(j) - T_{ref})$, it is also clear that the relative positions of the reference pulse and select pulses are independent of the sending processor at the transmission time in this model. However, in the dual bus model, where a sending processor could transmit downstream and upstream using two buses, the necessary and sufficient condition for a select pulse to coincide with the reference pulse is

$$T_{sel}(j) = T_{ref} + j - i, \quad \text{if} \quad j \geq i \tag{2a}$$

or

$$T_{sel}(j) = T_{ref} + i - j, \quad \text{if} \quad j < i \tag{2b}$$

Noting that these two models have equivalent functionalities and similar operations, we will concentrate our discussions on the first model, namely the folded bus model throughout the rest of this paper.

One advantage of using coincident pulse techniques as an addressing mechanism is its applicability to multicasting. Traditional addressing mechanisms for multicasting, such as separate-addressing, multi-destination addressing and source routing[1,6,15] have been mainly developed for point-to-point networks and are inefficient, especially in bus connected

systems. Most recent research work[10] exploits *tree forwarding*[7] on broadcast networks which constructs multicast trees and uses group identifiers when multicasting. It requires explicit group formations of communicating processors.

Using coincident pulse addressing however, the sender can multicast to an arbitrary group of processors by sending an address frame containing one or more select pulses which are properly positioned so that each of them coincides with the reference pulse at one of the multicasting destinations. Once a processor detects a coincidence, it picks up a copy of the multicasted message frame which is synchronous with the reference pulse.

3. TWO-LEVEL ADDRESSING

Using unary addressing, an address frame is N units long. There are two reasons why we want to reduce the length of address frames by using two-level addressing. One has to do with efficiency. Unary addressing could be very inefficient in a large multiprocessing system where the address frame is longer than the message frame. The other reason has to do with the physical limitation of optical path length between two adjacent processors. One way to ensure that the frames sent by one processor do not collide with other frames sent by other processors is to arbitrate the bus to allow exclusive access by one processor at a time, as in References 4 and 11. Another way is to pipeline the bus. That is, to synchronize all processors such that they will send messages at the beginning of each cycle. The propagation delays between two adjacent processors should be large enough to prevent frames from overlapping as in References 8 and 11. If unary addressing is used, it is necessary for the optical path between any two adjacent processors to have a length of at least $N \times w \times c_b$ to prevent overlapping of the address frames. Although the required minimum optical path length can be reduced by shortening the pulse width w, the address frame length, which is linear in the system size, becomes a limiting factor.

A two-level addressing implementation divides the whole system into logical clusters. Addressing of a single destination is accomplished by using one level of unary addressing to select a particular cluster and another level of unary addressing to select an individual processor within the selected cluster. Two trains of select pulses are used, one for each level of addressing and their pulse trains are sent in parallel. Therefore, the length of address frames can be reduced as neither the number of clusters nor the size of any cluster is larger than the system size.

Assume that $N = n^2$ processors are linearly connected. If every n consecutive processors constitute one logical cluster, two-level addressing in this linear system is logically equivalent to addressing a two-dimensional array. More specifically, we can view the linear system as the result of embedding an $n \times n$ array in row major fashion. Each row of processors of the array is embedded into n consecutive processors in the linear system. Hence, selecting a logical cluster is equivalent to selecting a row while selecting an individual processor within a cluster is equivalent to selecting a column processor within a row.

3.1. *Two-level addressing in a linear system*

As mentioned above, we will view a linear system with $N = n \times n$ processors as a result of embedding an $n \times n$ array in row major fashion and use terms such as 'row', 'column' and 'diagonal' logically. As a logical equivalent to two-level addressing, a two-dimensional

addressing implementation uses two select waveguides: one to select a row, and another to select a column. A pulse sent on the row-select waveguide will coincide with the reference pulse at all the processors in a particular row, while a pulse sent on the column-select waveguide will coincide with the reference pulse at all processors in a column. In other words, a row-select pulse causes a row select trace and a column-select pulse causes a column select trace. Pulses sent on these two select waveguides are denoted by $W1$ and $W2$ respectively.

A coincidence is said to occur at a given processor only if all *three* pulses, namely a reference pulse, a $W1$ pulse and a $W2$ pulse, coincide with each other at that processor. Since unary addressing is used when selecting a row, each pulse in the pulse train of $W1$ corresponds to one row. Similarly, each pulse in the pulse train of $W2$ corresponds to one column. Therefore, sending a reference pulse and a pair of one $W1$ pulse and one $W2$ pulse causes a coincidence at a processor that is located at the intersection of the corresponding row and column. More specifically, we denote L_1^i to be the pulse of $W1$ selecting row i and denote L_2^j to be the pulse of $W2$ selecting column j. By sending these two pulses and the reference pulse, the processor at row i and column j, which is processor $i \times n + j$ in an $N = n^2$ linear structure, is addressed. Figure 3 shows a logical two-dimensional view of addressing processor 10 with these two select pulses when $i = j = 2$ and $N = 16$.

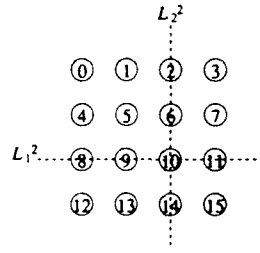
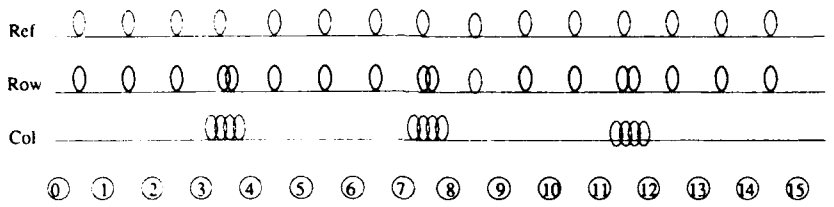


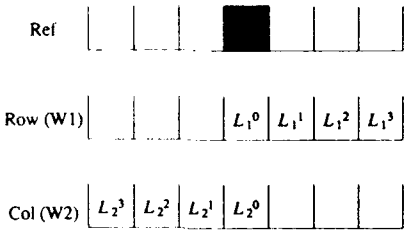
Figure 3. A logical view of addressing with two select pulses

In order to achieve the above coincidence pattern, we add, as in unary addressing, one unit delay between two processors on the receiving segments of the reference waveguide and the message waveguide. In addition, on the receiving segment of the row select waveguide ($W1$ waveguide), one unit delay is added between successive processors and an extra unit delay is added between two processors of successive rows. On the receiving segment of the column select waveguide ($W2$ waveguide), n unit delays are added between two receivers of successive rows. The amount of delay added on these two select waveguides can be obtained by solving a set of underconstrained equations (see Appendix). Again, because the message waveguide has exactly the same length as the reference waveguide and the lower half-(transmitting)-segments of all waveguides do not have any delays, only the receiving segments of the reference waveguide and the two select waveguides are shown in Figure 4(a). Taps from the waveguides to the processors are also omitted from the figure.

Let r_j and c_j be the row number and column number of processor j respectively. That is, $j = r_j \times n + c_j$ where $0 \le j < N$ and $0 \le r_j, c_j < n$. Let T_{ref} be the time when a processor transmits its reference pulse. And further assume a processor transmits a $W1$ pulse selecting row r_j at time $T_{sel}(L_1^{r_j})$ and a $W2$ pulse selecting column c_j at time $T_{sel}(L_2^{c_j})$. Given the



(a) the receiving segments of the reference, the row select and the column select waveguides



(b) Relative pulse positions of $W1$ and $W2$ in an address frame

Figure 4. A two-level addressing implementation

added delays on three waveguides as shown in Figure 4(a), a coincidence of these three pulses will occur at processor j if and only if

$$T_{sel}(L_1^{r_j}) + (j + r_j) = T_{ref} + j \tag{3a}$$

and

$$T_{sel}(L_2^{c_j}) + n \times r_j = T_{ref} + j \tag{3b}$$

That is,

$$T_{sel}(L_1^{r_j}) = T_{ref} - r_j \tag{4a}$$

and

$$T_{sel}(L_2^{c_j}) = T_{ref} + c_j \tag{4b}$$

Since $0 \leq r_j < n$, a $W1$ pulse is ahead of a reference pulse by 0 up to $n - 1$ units. The presence of a W1 pulse r_j units ahead of a reference pulse selects processors at row r_j while the absence does not. Similarly, a $W2$ pulse is 0 up to $n - 1$ units beyond a reference pulse and the presence of a $W2$ pulse c_j units beyond reference pulse selects processors at column c_j at each row while the absence does not. An address frame in the two-level addressing

implementation contains a train of $W1$ pulses and a train of $W2$ pulses and has a length of $2 \times n - 1$ units long. Figure 4(b) shows relative positions of the reference pulse and two trains of select pulses in an address frame at the time of transmission. Again, owing to the fact that an address frame will remain the same as it propagates through the transmitting segment of the bus in the folded bus model, the relative positionings of the reference pulse and both $W1$ and $W2$ pulses in an address frame at the time of transmission are independent of the sending processor.

Multicasting to a group of processors can be accomplished by sending a reference pulse, one train of $W1$ pulses with one or more pulses present and one train of $W2$ pulses with one or more pulses present along with a message frame. However, by doing so, coincidences may also occur at unintended processors, which we call *shadows*.[9] For example, when both processor i and j are addressed, the $W1$ train consists of two pulses, one for row r_i and another for row r_j. Similarly, the $W2$ train also consists of two pulses, one for column c_i and another for column c_j. In addition to processor i and j, the processor at row r_i and column c_j also detects a coincidence and picks up a copy of the multicasted message. So does the processor at row r_j and column c_i. Figure 5 shows a logical two-dimensional view of shadows at processor 1 and 10 as a result of multicasting to processors 2 and 9 in a 16 processor system.

3.2. *Shadow reduction*

As can be seen from the above example, shadows are created because of the unintended couplings of a $W1$ pulse with a $W2$ pulse. One way to reduce shadows is to further identify the intended pairs by using additional select waveguides, called *check* waveguides, for carrying select pulses called *check* pulses. Check pulses are arranged such that they do not coincide with the reference pulse at places where shadows were created. Only processors at which coincidences of a reference pulse and *all* select pulses occur are addressed. This technique for shadow reduction was introduced in Reference 9 for addressing a two-dimensional memory structure. In the remainder of this section, we will show how to apply this technique to two-level addressing in a linear system. Note that having an additional check waveguide in two-level addressing is different from having three-level addressing. The latter would be logically equivalent to addressing a three-dimensional array. That is, addressing a single destination would require three select pulses. The address frame length would be further reduced while more shadows would be likely when multicasting with three-level addressing.

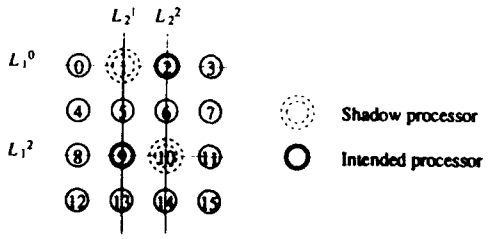
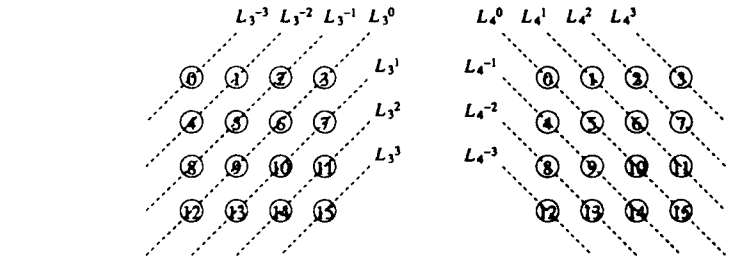
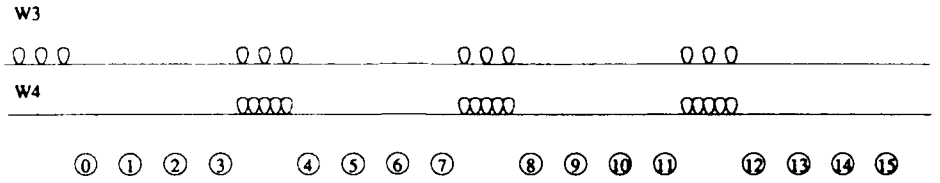


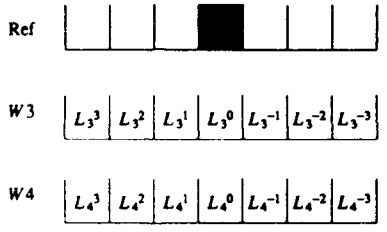
Figure 5. A logical view of shadows created at processor 1 and 10 as a result of multicasting to processor 2 and 9



(a) a logical view of traces of $W3$ pulses (b) a logical view of traces of $W4$ pulses



(c) the receiving segments of the $W3$ and the $W4$ waveguides



(d) relative pulse positions of $W3$ and $W4$ in an address frame

Figure 6. Adding two check pulses $W3$ and $W4$

One such set of check pulses are 45° diagonal select pulses, which will be called $W3$ hereafter. Another are $-45°$ diagonal select pulses, which will be called $W4$. Each pulse in a $W3$ or $W4$ train coincides with the reference pulse at all the processors that are on a 45° or $-45°$ diagonal line respectively, and therefore selects all processors on that diagonal line. Let $L_3^{\pm k}$ be the $W3$ pulse selecting a 45° diagonal line which is k lines below or above the main 45° diagonal line respectively. Figure 6(a) shows a logical two-dimensional view of traces of $W3$ pulses. Similarly, let $L_4^{\pm k}$ be the $W4$ pulse selecting a $-45°$ diagonal line which is k lines above or below the main $-45°$ diagonal line respectively. Figure 6(b) shows a logical two-dimensional view of traces of $W4$ pulses.

Again, the amount of delay that should be added on the $W3$ and $W4$ waveguides can be obtained by solving a set of underconstrained equations. Figure 6(c) shows only the receiving segments of both $W3$ and $W4$ waveguides with added delays. As a result, a $W3$ pulse and

a $W4$ pulse will coincide with the reference pulse at the desired corresponding locations if and only if:

$$T_{sel}(L_3^{+k}) = T_{ref} \pm k \tag{5a}$$

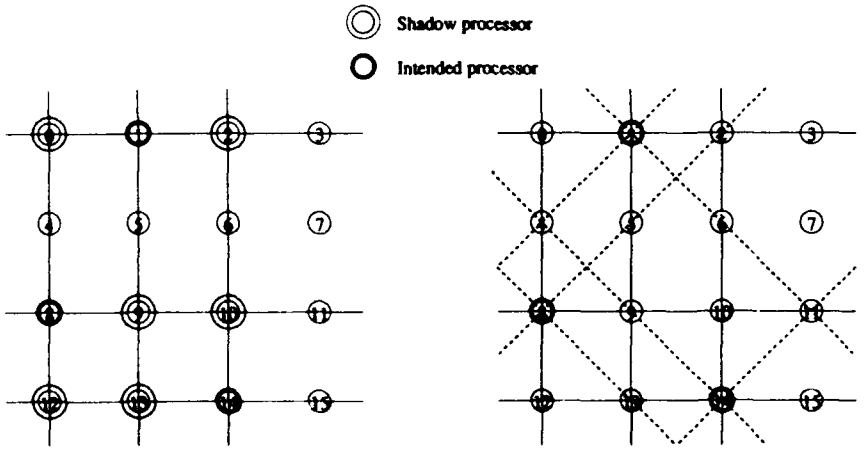
and

$$T_{sel}(L_4^{+k}) = T_{ref} \pm k \tag{5b}$$

These two equations can be derived as follows. First, any processor at a 45° diagonal line which is k lines below the main diagonal line has the index number of $k \times n + i \times (n - 1)$, for $1 \leq i < (n - k)$. This means that the reference pulse will go through $k \times n + i \times (n - 1)$ unit delays to arrive at the processor. Given that there are $n - 1$ added unit delays at the beginning of each row on the receiving segment of the $W3$ waveguide as in Figure 6(c) (with $n = 4$), the $W3$ pulse L_3^k will coincide with the reference pulse at the processor if and only if $T_{sel}(L_3^k) + k + i \times (n - 1) = T_{ref} + n \times k + i \times (n - 1)$, that is, $T_{sel}(L_3^k) = T_{ref} + k$. Similarly, we can completely derive the above equations.

In addition to a $W1$ train and a $W2$ train, an address frame now also contains a $W3$ train as well as a $W4$ train. Adding a $W3$ train and a $W4$ train does not change the length of an address frame, nor it does change the content of the $W1$ train or the $W2$ train. Figure 6(d) shows the relative positions of two trains of $W3$ and $W4$ pulses at the time of transmission of an address frame.

As an example, the two shadows in Figure 5 can be eliminated by using two $W3$ pulses, namely pulse L_3^{-1} and pulse L_3^0. A more complicated example is shown in Figure 7. Figure 7(a) shows a logical view of six shadows created at processors 0, 2, 9, 10, 12 and 13 when multicasting to three processors 1, 8 and 14 without using any check pulses. Figure 7(b)



(a) a logical view of six shadows created without using any check pulses

(b) a logical view of eliminating shadows in (a) with three $W3$ pulses and three $W4$ pulses

Figure 7. Shadow reduction with check pulses

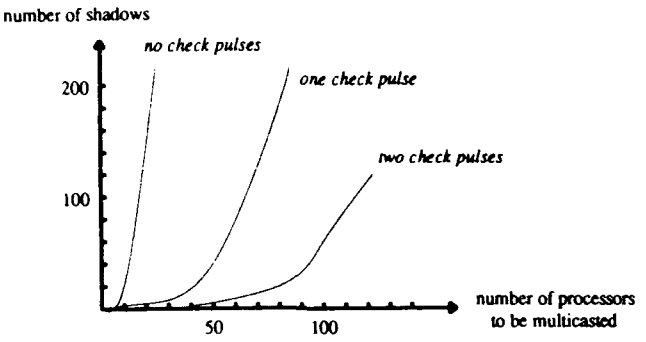


Figure 8. Shadows in a 512 × 512 array

shows a logical view of eliminating these shadows by adding three $W3$ pulses and three $W4$ pulses.

Figure 8 shows simulation results on the numbers of shadows created with different number of check pulses used. It is clear that the addition of check pulses cannot introduce new shadows. It can only reduce the number of existing shadows. However, adding a fixed number of check pulses cannot always completely eliminate shadows, since the theorem given in Reference 9 for a two-dimensional parallel memory structure also holds here.

4. SHADOW AVOIDANCE

Having established the relationship between a particular two-level addressing structure and its logically equivalent two-dimensional addressing representation, we will adopt the usual notion of two-dimensional addressing in an $n \times n$ array in the following discussions for the purpose of simplicity. However, it is worth noting that techniques developed will be applied in physically linear systems with two-level addressing.

One way to avoid shadows when multicasting to a group of processors is to partition the whole group into several subgroups such that each subgroup is multicasted within one cycle without creating any shadows. More formally, assume a group of m processors is a set of linearly ordered processors denoted by $G = \{P_1, P_2, \ldots, P_m\}$. That is, for all $1 \le i \le m$, $0 \le P_i < N$ and $P_{i-1} < P_i$. Define a *shadow free* (SF) partition of the set G to be a number of subgroups S_1 through S_g such that for all $1 \le i, j \le g$ the following conditions are satisfied. (a) $S_i \cap S_j = \phi$ if $i \ne j$, (b) $\bigcup_{i=1}^{g} S_i = G$ and (c) each S_i can be multicasted in one cycle without any shadows.

A number of subgroups is called a *maximal* SF partition if it is a SF partition and if multicasting to more than one of the subgroups within one cycle will create a shadow. Therefore the number of subgroups of a maximal partition is the number of cycles needed to complete the multicasting to the whole group G.

Let processor X, where $0 \le X < N$, be a shadow created when multicasting to a group G in one cycle, clearly $X \notin G$. Define four shadow conditions (SC_i, $i = 1, 2, 3$ and 4) as follows.

SC_1: there is at least one processor in G that is in the same row as X
SC_2: there is at least one processor in G that is in the same column as X
SC_3: there is at least one processor in G that is in the same 45° diagonal line as X
SC_4: there is at least one processor in G that is in the same −45° diagonal line as X.

It can be verified that each of the conditions SC_i is necessary for X to be a shadow if the corresponding select pulse Wi is used to multicast the group G. The logical AND of all the necessary conditions becomes the sufficient condition. For example, if two pulses, $W1$ and $W2$, are used, then both condition SC_1 and SC_2 are the necessary conditions for a shadow to occur. The logical AND of the two is the sufficient condition.

4.1. *Regular multicasting patterns*

In some applications, such as finite element analyses and image processing, multicasting patterns can be quite regular. For example, a convolution of an $n \times n$ array involves multicastings of an element to its $w \times w$ neighbours, where w is the current window size. A group to be multicasted could also be all processors of a row, or of a column or of a diagonal line. By embedding a physical 2-D structure into our linear structure in row-major fashion, these regular 2-D patterns can be characterized by a group of four parameters. More formally, in an embedded $n \times n$ system, we consider a group G of m processors starting with the processor numbered as k (called offset) with increment of d (called stride). Using the general notation in the beginning of the section, we have $G = \{k, k + d, \ldots, k + (m - 1) \times d\}$, where $0 \le k \le k + (m - 1) \times d < N = n^2$. We can use $G(k, d, m, n)$ to uniquely represent such a regular group. We call a group a *dense* group if d is less than n, a *sparse* group otherwise.

While we can make tradeoffs between the number of select waveguides used and the number of cycles needed to multicast to a group of processors, we will first analyse simple cases in which only two select waveguides are used. The results will be extended to cases in which four select waveguides are used.

Definition 1. A row of processors in a logical two-dimensional array is *incomplete* with regard to a group $G(k, d, m, n)$ if and only if the row contains two processors i and j such that $i \in G$, $j \notin G$ and $|j - i| = b \times d$ for some integer $b > 0$. A row is *complete* if and only if the row contains at least one processor of the group G and is not an *incomplete* one.

Definition 2. Define $I(k, d, m, n)$ to be the number of *incomplete* rows with regard to the group G.

Let the first processor of the group be $k = r_k \times n + c_k$ and the last processor of the group $l = k + (m - 1) \times d = r_l \times n + c_l$ for some integers $0 \le r_k, c_k, r_l, c_l < n$. And let condition 1 be that $c_k > d$, condition 2 be that $n - c_l > d$ and condition 3 be that $r_k \ne r_l$. There will be two *incomplete* rows, namely row r_k and row r_l if and only if all three conditions are true. There will be no *incomplete* rows if and only if neither condition 1 nor condition 2 is true. Otherwise, there will be only one *incomplete* row. Therefore, I has an upper bound of 2. Noting that for a *sparse* group $G(k, d, m, n)$ where $d \ge n$, neither condition 1 nor condition 2 is true, therefore $I = 0$.

Lemma 1. Two processors of a group $G(k, d, m, n)$ numbered as i and j, $i < j$, will be in the same column if and only if $j - i = b \times \text{LCM}(n, d)$ for some integer $b > 0$.†

Proof. Let $i = r_i \times n + c_i$, and $j = r_j \times n + c_j$ as before. On the one hand, if $c_j = c_i$, then $j - i = (r_j - r_i) \times n$ and $r_j > r_i$. Since both processors i and j are in the same group, $j - i$ must be a multiplier of d, therefore $j - i$ should be a common multiplier of both n and d.

On the other hand, if $j - i = b \times \text{LCM}(d, n)$ for some integer $b > 0$, then clearly, $j - i$ is a multiple of n, which means processor i and j are at the same column. ■

Lemma 2. If there is a processor i in a group $G(k, d, m, n)$, $i = r_i \times n + c_i$, then processor j at the same column, $j = r_j \times n + c_i$, is also in the group G if

$$|r_j - r_i| = \frac{b \times d}{\text{GCD}(n, d)}$$

and row r_j is a *complete* one.

Proof. (By contradiction.) According to the definition of *complete* row, there must be a processor $\tilde{j}$ at row r_j and $\tilde{j} \in G$. Clearly, $|\tilde{j} - i|$ should be a multiple of d. In addition, since processor j and i are at the same column and are $(b \times d)/[\text{GCD}(n, d)]$ rows apart, $|j - i| = |r_j - r_i| \times n$. That is,

$$|j - i| = b \times \frac{d \times n}{\text{GCD}(n, d)} = b \times \text{LCM}(n, d)$$

which is also a multiple of d. Therefore, $|j - j|$ must be a multiple of d also. If $j \notin G$, then row r_j is not a *complete* one, which contradicts the condition stated above. Therefore, $j \in G$. ■

According to the above lemmas, for a given group G, if we draw one vertical line at each processor of the group G, then all processors at the intersections of these vertical lines with *complete* rows which are $(b \times d)/[\text{GCD}(n, d)]$ apart will belong to the group G, and therefore can be multicasted without any shadows using row select pulses $W1$ and column select pulses $W2$.

Theorem 1. For a *dense* group $G(k, d, m, n)$ where $d < n$, the number of subgroups of a maximal partition with select pulses $W1$ and $W2$ has an upper bound of $d/[\text{GCD}(n, d)] + 1$.

Proof. We will prove the theorem by constructing a SF partition of the group.

First, we use one subgroup for processors of the group G at each *incomplete* row. According to the definition, there are 1 such subgroups and no shadows will occur in any of these subgroups because of the shadow condition SC_2. We partition the rest of group at remaining *complete* rows as follows.

Let $R = d/[\text{GCD}(n, d)]$. Starting at the first *complete* row, we put processors of the group at every R rows apart into a subgroup, creating exactly R subgroups. It can be

† LCM stands for least common multiplier and GCD stands for greatest common divider.

similarly proved as in the Lemma 2 that no shadows will occur in any of these R subgroups. Therefore, a maximal SF partition will have at most $d/[\text{GCD}\ (n, d)] + 1$ subgroups. ■

It can be shown that the SF partition constructed in the above theorem is indeed a *maximal* SF partition. We can similarly prove the following theorem for a *sparse* group. First of all, I is equal to zero when $d \geq n$. Secondly, between any two rows which are R rows apart, where $R = d/[\text{GCD}\ (n, d)]$, there could be at most $(R \times n)/d$ processors belonging to the group and hence at most $[\text{LCM}\ (n, d)]/d$ *complete* rows. Therefore, when putting processors of the group every R rows apart into one subgroup, there are at most $[\text{LCM}\ (n, d)]/d$ subgroups in such a SF partition of the group G. This is stated in Theorem 2.

Theorem 2. For a *sparse* group $G\ (k, d, m, n)$ where $d \geq n$, the number of subgroups of a maximal partition with select pulses $W\ 1$ and $W\ 2$, has an upper bound of $[\text{LCM}\ (n, d)]/d$. ■

We can also prove the following theorems when using either one of the two diagonal select pulses, namely $W\ 3$ or $W\ 4$, with $W\ 1$ instead of using $W\ 2$ with $W\ 1$ as in the above theorems.

Theorem 3. For a *dense* group $G\ (k, d, m, n)$ where $d < n$, the number of subgroups of a maximal partition with select pulses $W\ 1$ and either $W\ 3$ or $W\ 4$, has an upper bound of $d/[\text{GCD}\ (n - 1, d)] + 1$ or $d/[\text{GCD}\ (n + 1, d)] + 1$ respectively. ■

Theorem 4. For a *sparse* group $G\ (k, d, m, n)$ where $d \geq n$, the number of subgroups of a maximal partition with select pulses $W\ 1$ and either $W\ 3$ or $W\ 4$ has an upper bound of $[\text{LCM}\ (n - 1, d)]/d + 1$ or $[\text{LCM}\ (n + 1, d)]/d$ respectively. ■

The idea used to construct SF partitions in Theorem 3 and 4 is similar to the one used in Theorem 1 and 2. Processors of a group at certain number of rows apart will be put into one subgroup. These processors are at the intersections of these rows with either 45° diagonal lines or −45° diagonal lines depending on which one of the select pulses, $W\ 3$ or $W\ 4$, is used.

Since the additions of select pulses will not create any new shadows, we can choose a SF partition which has the *least* number of subgroups when all four select pulses discussed above are used.

Theorem 5. For a *dense* group $G\ (k, d, m, n)$ where $d < n$, the number of subgroups of a maximal partition with four select pulses $W\ 1$, $W\ 2$, $W\ 3$ and $W\ 4$ has an upper bound of

$$\frac{d}{\max\left(\text{GCD}\left((n-1), d\right), \text{GCD}\left(n, d\right), \text{GCD}\left((n+1), d\right)\right)} + 1 \quad ■$$

Theorem 6. For a *sparse* group $G\ (k, d, m, n)$ where $d \geq n$, the number of subgroups of a maximal partition with four select pulses $W\ 1$, $W\ 2$, $W\ 3$ and $W\ 4$ has an upper bound of

$$\min\left(\frac{\text{LCM}\left((n-1), d\right)}{d} + 1, \frac{\text{LCM}\left(n, d\right)}{d}, \frac{\text{LCM}\left((n+1), d\right)}{d}\right) \quad ■$$

By applying Theorem 5 or Theorem 6 to some special instances of a group $G\ (k, d, m, n)$, such as a group of processors at one row with $d = 1$, a group of processors at one column with $d = n$, a group of processors at either diagonal lines with $d = n - 1$ or

$d = n + 1$, we know that each multicasting operation to such groups can be done in only one cycle without any shadows. These multicasting patterns are often seen in matrix manipulations, among many other applications.

The two theorems above can also be extended to a group of processors located in a small area of an $n \times n$ array. We delimit the area by an $\tilde{n} \times \tilde{n}$ array with $\tilde{n} \leq n$. The processors of the $\tilde{n} \times \tilde{n}$ array can be renumbered in a row major fashion from 0 to $\tilde{n}^2 - 1$. If a group of processors can be represented as $G(k, d, m, \tilde{n})$, we can partition the group similarly to what we did before. Since no shadow is possible outside the $\tilde{n} \times \tilde{n}$ area, by replacing every occurrence of n with $\tilde{n}$, the two theorems Theorem 5 and Theorem 6 can be applied to a group $G(k, d, m, \tilde{n})$ when $d < \tilde{n}$ or $d \geq \tilde{n}$ respectively.

The importance of this extension is that some of the most frequently used multicasting patterns in image processing[13] can now be analysed in term of SF partitions. For example, a four-neighbour group around any processor can be represented by a group with $k = 1$, $d = 2$, $m = 4$ and $\tilde{n} = 3$, or $G(1, 2, 3, 4)$, and each multicasting operation to the group can be done in one cycle. If we allow a processor to send a multicasting message to itself, then multicasting to its eight neighbours and itself, which is a group of $G(0, 1, 9, 3)$, can be done in one cycle. Similarly, multicasting to neighbouring $w \times w$ processors, as mentioned at the beginning of this section, can also be done in one cycle. By mapping hierarchical multigrids[2] or pyramid structures properly onto the logical 2-D structure, each processor can multicast to neighbouring processors or processors at the next level in one cycle.

4.2. *Arbitrary multicasting patterns*

As discussed above, there is a systematic way to construct a SF partition for any regular group. In order to construct a *maximal* SF partition, we need to merge subgroups of a SF partition together as long as the newly merged subgroups can still be multicasted without shadows. Similar partitioning procedures can also be used for arbitrary multicasting patterns. The proposed partitioning algorithm presented below consists of two parts. The first part is to construct a SF partition and the second part is to merge subgroups to construct a *maximal* SF partition.

For purposes of simplicity, we assume that three select pulses are used, namely row select pulses $W1$ and two diagonal select pulses $W3$ and $W4$. The first part of the algorithm partitions the whole group into at most $\lceil \frac{n}{2} \rceil$ non-empty subgroups. This is done by putting processors of the group at $\lceil \frac{n}{2} \rceil$ rows apart into one subgroup. Such a partition is a SF partition as stated in the following lemma.

Lemma 3. No shadows are possible when multicasting to a group of processors located at two rows which are at least $\lceil \frac{n}{2} \rceil$ apart in an $n \times n$ array with three select pulses $W1$, $W3$ and $W4$.

Proof. (By contradiction.) Let the two rows be r_1 and r_2. According to the condition SC_1 above, a shadow X must be at one of the two rows, assume it is row r_1. According to the shadow conditions SC_3 and SC_4 above, there must be two processors at row r_2 such that their respective 45° and −45° diagonal lines intersect at X. Therefore, the distance between these two processors should be two times the distance between the two rows r_1 and r_2, that is $2 \times \lceil \frac{n}{2} \rceil$, which is no less than n. Since these two processors are at the same row r_2, their distance could never exceed $n - 1$, hence no shadows are possible. ■

The second part of the algorithm first computes a forbidden set for each subgroup of the above SF partition. A row is said to be forbidden by a subgroup S_i if multicasting to S_i and processors of the group G located at that row will create at least one shadow. A forbidden set for a subgroup S_i is a set of rows forbidden by S_i. Two subgroups are merged together if neither contains processors at a row which is forbidden by the other. The new forbidden set for the merged subgroup is computed by adding certain rows into the union of the two original forbidden sets. When no further merges are possible, the algorithm stops and a maximal SF partition is constructed.

The time complexity of the algorithm is $O(n + m^2)$ where m is the number of processors in a group. This is because the first part of the algorithm takes $O(n)$ time. Careful analysis shows that the second part of the algorithm takes $O(m^2)$ time.

It is clear that no shadow is possible in a group of less than three processors when three select pulses are used. Therefore, any *maximal* SF partition will always have less than $\lceil \frac{m}{2} \rceil$ subgroups. Hence the algorithm will generate at most $\min(\lceil \frac{n}{2} \rceil, \lceil \frac{m}{2} \rceil)$ subgroups.

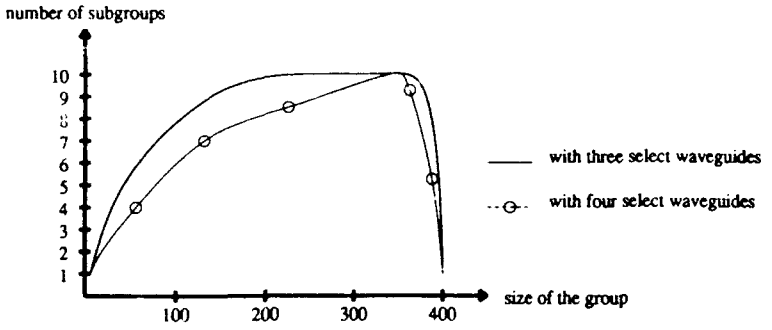
The partitioning algorithm can be executed incrementally when the multicasting pattern changes. Adding or deleting a processor from a group to be multicasted involves possibly spliting an existing subgroup, forming a new subgroup which contains processors at the same row and finally re-merging any subgroups which have since been changed.

If the column select pulses $W2$ are used instead of the row select pulses $W1$, the algorithm above can be adapted accordingly by partitioning columns which are $\lceil \frac{n}{2} \rceil$ columns apart into subgroups and merging them into a *maximal* SF partition. If all four select pulses mentioned above are used, we can start with either SF partitions and augment the condition which determines if a row (or a column) is forbidden by a certain subgroup and merge subgroups together to achieve a *maximal* SF partition.

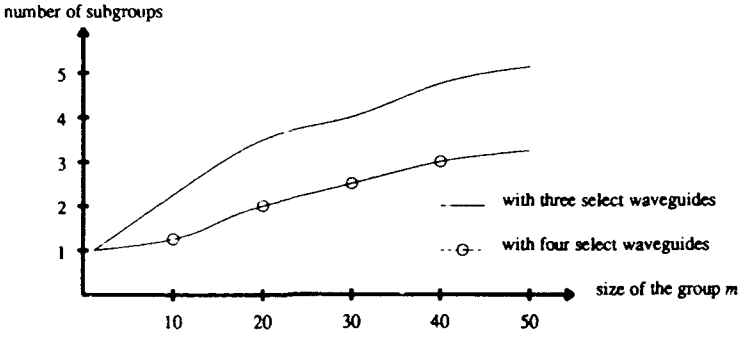
Figure 9 shows the simulation results on the number of subgroups generated by the algorithm. For a 20×20 processor system in Figure 9(a), when all 400 processors are multicasted, there are no unintended processors at all and that is why the number of subgroups is reduced to 1. If the number of subgroups in a maximal SF partition of a group G is g, then the time needed to transmit g address frames, one in each cycle, is $g \times (2 \times n - 1) \times c_h$ in the two-level addressing implementation. However, $N \times c_h$ is needed in a unary addressing implementation. Therefore, the speed-up is at least $n/(2 \times g)$. Noting that g can not exceed $n/2$, the worst-case speed-up is 1. As shown in Figure 9(b), with three select pulses, the average number of subgroups needed to multicast to 50 processors in a 50×50 processor system is only about 5. Thus, a speed-up of 5 is obtained.

5. CONCLUSION

In this paper, coincident pulse techniques have been applied as an efficient addressing mechanism for multicasting among multiprocessors connected by optical buses. Two basic models of a unary addressing implementation have been discussed, and a two-level addressing implementation has been proposed to reduce the address frame length. Two approaches to deal with the shadow problem have been presented. One approach reduces the number of shadows by using check pulses. Another approach avoids possible shadows by constructing SF partitions. It has been shown that for regular multicasting patterns, SF partitions can be constructed systematically and processors can multicast to their communicating processors within one cycle in many applications. A partitioning algorithm has also been presented for arbitrary multicasting patterns. The overall results of the two level addressing implementation are higher efficiency, lower minimum optical path requirements and potential speed-ups.



(a) number of subgroups vs. size of the group in a 20×20 system



(b) number of subgroups for a group of $m = n$ processors in an $n \times n$ system

Figure 9. Simulation results of the partitioning algorithm

This reinforces our belief that coincident pulse techniques are a promising addressing mechanism which can be applied in both parallel memory structures and multiprocessor systems. Finally, we note that in this paper, many technical aspects of pulse generation, coincidence detection, power distribution and other related issues have not been discussed. They can be found in References 5, 9 and 12.

APPENDIX

As mentioned in Section 3.1, the amount of delay that is added on the row select and the column select waveguides can be obtained by solving a set of underconstrained equations. In the following discussion, a folded bus model of $N = n^2$ processor is assumed and some of the previous defined notation is used. Additional notation is defined as follows.

Row (i), *Col* (i):
the number of delay loops added on the receiving segment of the row select and the column select waveguides, respectively, between processor $i - 1$ and processor i, where $0 \leq i < N$.

$D(L_1^i)$, $D(L_2^i)$:
the transmission time of the row select pulse L_1^i and the column select pulse L_2^i, respectively, relative to the transmission time of the reference pulse, where $0 \le i < n$.

Given that there is one delay loop between any two adjacent processors on the receiving segment of the reference waveguide, we have the following set of equations for the row select waveguide:

$$T_{sel}(L_1^i) + \sum_{j=0}^{k} Row(j) \begin{cases} = T_{ref} + k, & \text{if } i \times n \le k < (i+1)n \\ \neq T_{ref} + k, & \text{otherwise} \end{cases} \tag{A.1a}$$

Note that, the number of delay loops added on the reference waveguide has been fixed, therefore a degree of freedom has been removed already. Equation (A.1a) states that the pulse L_1^i should coincide with the reference pulse at all processors at row i but should not coincide with the reference pulse at any other processors. Since $T_{sel}(L_1^i) - T_{ref} = D(L_1^i)$, we can simplify the above equation to

$$D(L_1^i) + \sum_{j=0}^{k} Row(j) \begin{cases} = k, & \text{if } i \times n \le k < (i+1)n \\ \neq k, & \text{otherwise} \end{cases} \tag{A.1b}$$

Clearly, two different pulses cannot be transmitted at the same time, therefore, the following equation has to be satisfied also:

$$D(L_1^i) \neq D(L_1^j), \quad \text{if } i \neq j \tag{A.1c}$$

These two equations, namely (A.1a) and (A.1b), are underconstrained since both $D(L_1^i)$s and $Row(j)$s are variables. Note that the values of $D(L_1^i)$s will determine the address frame length and therefore we choose to fix them first and solve the equation for $Row(j)$s. If $D(L_1^i)$s are fixed such that $D(L_1^i) = i$, we can solve the above equations to get

$$Row(j) = \begin{cases} 0, & \text{if } j = i \times n \\ 1, & \text{otherwise} \end{cases}$$

Note that, if the $D(L_1^i)$s are fixed such that $D(L_1^i) = -i$, we will get the same result as in Section 3.

Similarly, we can have the following set of equations for column select waveguide:

$$D(L_2^i) + \sum_{j=0}^{k} Col(j) \begin{cases} = k, & \text{if } k = i + mn, \text{ where } 0 \le m < n \\ \neq k, & \text{otherwise} \end{cases} \tag{A.2a}$$

$$D(L_2^i) \neq D(L_2^j), \quad \text{if } i \neq j \tag{A.2b}$$

It can be shown that by fixing $D(L_2^i) = i$, we can get the result as in Section 3. Equations and their solutions for check waveguides can be similarly constructed.

ACKNOWLEDGEMENTS

This research is supported in part by a grant from the Air Force Office of Scientific Research under contract #AFOSR-89-0469 and in part by a grant from the National Science Foundation under contract MIP 89 01053.

REFERENCES

1. L. Aguilar, 'Datagram routing for Internet multicasting', *ACM Sigcomm 84 Computer Communications Review*, **14**, 58–63 (1984).
2. T. Chan and R. Schreiber, 'Parallel networks for multi-grid algorithms: architecture and complexity', *SIAM Journal on Scientific and Statistical Computing*, **6**, 698–711 (1985).
3. D. Chiarulli, R. Melhem and S. Levitan, 'Using coincident optical pulses for parallel memory addressing', *IEEE Computer*, **20**, 48–58 (1987).
4. D. Chiarulli, S. Levitan and R. Melhem, 'Optical bus control for distributed multiprocessors', *Journal of Parallel and Distributed Computing*, **10**, 45–54 (1990).
5. D. Chiarulli, R. Ditmore, R. Melhem and S. Levitan, 'An all optical addressing circuit: experimental results and scalability analysis', *IEEE Journal of Lightwave Technology*, **9**(12), (1991).
6. Y. Dalal and R. Metcalfe, 'Reverse path forwarding of broadcast packets', *Communications of ACM*, **21**, 1040–1048 (1978).
7. A. Frank, L. Wittie and A. Bernstein, 'Multicast communication on network computers', *IEEE Software*, **2**, 49–61 (1985).
8. Z. Guo, R. Melhem, R. Hall, D. Chiarulli and S. Levitan, 'Array processors with pipelined optical busses', *Journal of Parallel and Distributed Computing*, **12**(3), 269–282 (1991).
9. S. Levitan, D. Chiarulli and R. Melhem, 'Coincident pulse techniques for multiprocessor interconnection structures', *Applied Optics*, **29**, 2024–2039 (1990).
10. P. McKinley and J. Liu, 'Multicast tree construction in bus based networks', *Communications of ACM*, **33**, 29–42 (1990).
11. R. Melhem, D. Chiarulli and S. Levitan, 'Space multiplexing of waveguides in optically interconnected multiprocessor systems', *The Computer Journal*, **32**, 362–369 (1989).
12. M. Nassehi, F. Tobagi and M. Marhic, 'Fiber optic configurations for local area networks', *IEEE Journal on Selected Areas in Communication*, **SAC-3**, 941–949 (1985).
13. A. Netravali and J. Limb, 'Picture coding: a review', *Proc IEEE*, **68**, 336–406 (1980).
14. C. Qiao and R. Melhem, 'Time-division optical communications in multiprocessor array', *Proceedings of the Supercomputing 91 Conference*, IEEE Computer Society Press, 1991, pp. 644–653.
15. D. Wall, 'Selective broadcast in packet-switched networks', *Proc. Sixth Berkeley Workshop Distributed Data Management and Computer Networks*, 1982, pp. 239–258.

Authors' biographies:

Chunming Qiao is currently pursuing his Ph.D. degree at the Department of Computer Science, University of Pittsburgh. He received a B.S. in Computer Science and Engineering from University of Science and Technology of China in 1985 and an M.S. in Computer Science from University of Pittsburgh in 1990. His current research interests include parallel computer architecture, parallel processing and optical interconnection.

Rami G. Melhem is an Associate Professor of Computer Science at the University of Pittsburgh. He received a B.E. in Electrical Engineering from Cairo University, Egypt, in 1976, an M.S. in Mathematics/Computer Science in 1981, and a Ph.D. in Computer Science in 1983, both from the University of Pittsburgh. He has been an Assistant Professor of Computer Science at Purdue University from 1984 to 1986. His research interests include optical computing, parallel systems and fault tolerant systems.

Donald M. Chiarulli is an Assistant Professor of Computer Science at the University of Pittsburgh. He received a B.S. degree in Physics from Louisiana State University in 1976, an M.S. degree in Computer Science from Virginia Polytechnic Institute in 1979, and a Ph.D. in Computer Science from Louisiana State University in 1986. His research interests include hybrid optical/electronic computer architecture, optical interconnects, VLSI design and parallel computation.

Steven P. Levitan is the Wellington C. Carl Assistant Professor of Electrical Engineering at the University of Pittsburgh. He received a B.S. degree from Case Western Reserve University (1972), his M.S. (1979) and Ph.D. (1984), all in Computer Science, from the University of Massachusetts, Amherst. From 1984 to 1986 he was an Assistant Professor in the ECE Department at the University of Massachusetts. In 1987 he joined the University of Pittsburgh. His research interests include parallel computer architecture, parallel algorithm design, VLSI design and computer architectures for image understanding, and optical computing.