Time-Space-Wavelength Networks for Low-Complexity Processor Interconnection

Khaled A. Aly
Dept. Electrical & Computer Engineering
University of Central Florida
Orlando, FL 32816-2450
kaa@engr.ucf.edu

Patrick W. Dowd
Dept. Electrical & Computer Engineering
State University of New York at Buffalo
Buffalo, NY 14260-2050
dowd@eng.buffalo.edu

Abstract – *This paper studies a flexible hierarchic design approach of large processor networks with distributed media access. The cluster-based interconnection combines passive metal buses and passive optical star couplers at two hierarchic levels, independently employing interleaved TDMA for conflict-free interprocessor communication. The system delay analysis highlights the tradeoffs of arbitrarily combining space-division at the local level, wavelength-division at the global level, with time-division as a conflict-free access scheme at both levels and in the form of a speedup factor associated with optical transmission. The frame synchronization time that dominates the access delay in TDMA-based protocols is broken down into two additive rather than multiplicative factors. The paper proposes a simple distributed slot synchronization scheme that does not require a centralized system clock. It is shown that this hierarchic approach has the advantages of modularity, expansion flexibility, complexity and performance-wise scalability, and spatial bandwidth re-use.*

1 Introduction

Both multi-bus and wavelength-division multiplexed (WDM) interconnection schemes have been considered for high performance multiprocessor interconnection. Multi-bus networks have been considered in the past as a performance-complexity tradeoff between the connectivity of a single bus and a full crossbar switch [1]. Several bus arbitration schemes have been studied. Performance limitations exist due to bus contention, limited bandwidth, and fanout prohibiting the construction of very large processor networks using this approach. On the other hand, the use of large crossbar or multi-stage interconnection networks is limited by high asymptotic switching element complexity, switching overhead, as well as fault tolerance constraints. Optical interconnects have been considered based on optical fiber [2, 3] or wavelength multiplexed star to overcome the limitations of the metal bus [4–7].

WDM star-coupled processor networks offer the potential to realize large-scale high-performance parallel computer systems at low complexity . The low-complexity is due to the passive interconnection fabric and the distributed access scheme. These features contribute to improving the system fault tolerance as well. The excellent latency-throughput characteristics are due to the concurrent transmission over multiple distinct wavelength channels. However, scalability to the massive parallelism region is constrained by the power budget and the number of WDM channels that can be formed according to the tuning range of light sources and filters as well as the crosstalk and power budget considerations [8].

Optical devices such as star couplers are now commercially available at moderate cost, while tunable lasers and filters (with mS-μS switching time) are still expensive. Two alternative approaches to a single tunable laser diode transmitter are LED spectral slicing [9] and laser diode arrays (non-tunable but each operating at a distinct wavelength) [10]. Spectral slicing uses the broad optical spectrum of LEDs and the bandpass filtering characteristics of a passive multiplexer gratings device to establish unique output wavelengths. This results in a low cost solution that is limited in the number of channels and the channel bit rate [11]. However, an advantage of spectral slicing is that switching between channels is achieved electronically and therefore nS switching time is possible through high speed logic. The number of channels that can be formed via an integrated surface-emitting laser array is limited mainly by the fabrication complexity. Lower cost wavelength-selective receivers can be also achieved using either the spectral-sliced LED approach or an integrated photodiode array. Integrated multichannel transmitter and receiver devices based on source/detector arrays and multiplexer gratings have been demonstrated in [12].

It is recognized that a hierarchical network is capable of reducing the complexity and improving the performance

of a large processor network by taking advantage of spatial reference locality. Several hierarchical schemes have been considered as extensions to conventional direct interconnection networks [13, 14]. In [6], an optical hierarchical scheme, denoted as space-wavelength hierarchical architecture (SWHA), that draws on the analogy with the fat-tree network of [13] was studied. The SWHA achieves the Fat-Tree objective of increased bandwidth per link at higher levels of the hierarchy while also obtaining a significant improvement in flexibility, performance and fault tolerance. The SWHA achieves *adaptable bandwidth allocation*: the bandwidth at each level does not need to remain fixed but can be *dynamically reallocated to adapt to changing communication requirements*.

The cost of wavelength-selective light sources and detectors poses a constraint on the feasibility of a massively parallel system with one optical interface per processor. Moreover, the data rate at which a light source can be modulated may well exceed the capability and/or requirement of the communication interface of a single processor. This paper proposes and analyzes a simple scalable configuration that breaks the tie between the maximum system size and the above mentioned constraints. The proposed configuration has significant performance, cost, and packaging advantages. TSW, for *time-space-wavelength interconnection*, is used in this paper to denote the architecture.

The network is divided into a *local electronic* and a *global optical* hierarchic levels. Nodes are grouped into clusters with a multiple bus network acting as the medium for local communication and for global reference transport to the optical interface. Each cluster contains an asynchronous time-division multiplexer (ATDM) which collects global packets from the local network and converts the global packet stream into a higher speed optical stream for transmission through the star coupler. This approach results in complete functional independence between both local and global networks. The network can be expanded by as little as one processor without intervening with the operation or the performance of the other clusters. The slot-channel assignment allows source routing to take place with tree self routing employed at the ATDM stage for the global packets to avoid head-of-line blocking when accessing distinct wavelength channels. Slot synchronization is accomplished in a distributed fashion that minimizes the slot synchronization overhead and significantly enhances the system fault tolerance. Packaging constraints are relaxed due to the passive broadcast-select interconnect and the independence between distinct clusters as well as hierarchic levels. Reconfigurable bandwidth partitioning can be done in a similar way to that employed in the SWHA of [6] at the local and/or global levels, allowing several thousands of processors to be interconnected with excellent sustained performance.

The rest of this paper is organized as follows. Section 2 defines the two-level hierarchical interconnection, the media access protocol, and a distributed slot synchronization scheme. The system delay-throughput performance is analyzed in Section 3 through a detailed queueing model, and compared to the performance of both single and multi-level I-TDMA access protocols. Conclusions are presented in Section 4.

2 Hierarchic Time-space-wavelength Interconnection

A distributed shared memory multiprocessor system environment is considered. Each node consists of a processor, its associated cache, its portion of the distributed shared memory, denoted as the *local global memory* (LGM), and a serial I/O communication port. Nodes are grouped into clusters. Local communication between nodes within the same cluster takes place over a passive multi-bus network, referred to as the *local interconnection network* (LIN). Optical communication is used only between clusters through a passive star coupler. An asynchronous time-division multiplexer (ATDM) collects non-local references from each cluster and interfaces the cluster to the optical *global interconnection network* (GIN). The network is intended to provide low latency access by taking advantage of spatial reference locality via hierarchical time-slotted packet communication. The physical design objectives are to achieve source routing with the passive fabric, maintain independence of the LIN and the GIN, avoid a complex switching function within the ATDM, avoid central slot synchronization, and allow modular scalability. Section 2.1 provides the formal network description, Section 2.2 describes the media access protocol, Section 2.3 discusses distributed slot synchronization, and Section 2.4 evaluates the system characteristics: modularity, scalability, and bandwidth partitioning.

2.1 Description

The system consists of m_1 clusters interconnected via a passive star coupler. The number of nodes in each cluster does not need to be identical. For the purposes of formal definition and performance modeling, it is assumed that each cluster contains m_0 nodes, with a total system size of $M = m_1 m_0$ nodes. The LIN consists of B buses, and each node can transmit (receive) to (from) any bus via B passive

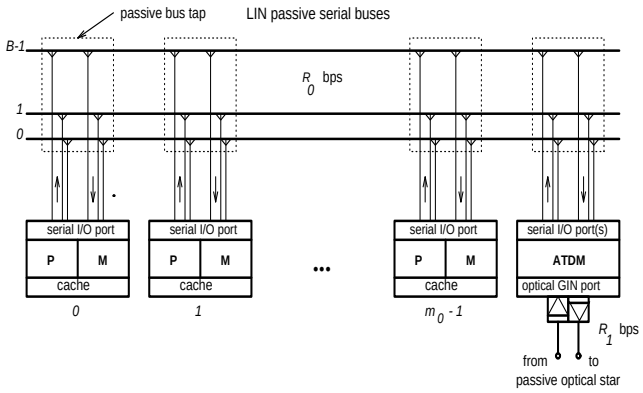

Figure 1: Local interconnection within a cluster through a passive multi-bus network

taps in each direction. The ATDM is assumed to have a_0 *inbound* buffers (transmitting to the LIN) and a_1 outbound buffers (receiving from the LIN). Conversion between the electronic and optical domains at the ATDM is done using wavelength-selective transmitter and receiver, each capable of tuning to one of C wavelength channels. A possible potential approach to achieve feasible tunability to a limited number of channels is by employing an array of light sources (detectors) each operating at a distinct wavelength along with grating multiplexer (demultiplexer) devices. It is assumed that at least the receiver of the ATDM is realized via this approach since all C channels can be monitored and distributed slot synchronization is made possible. The cluster configuration is shown in Fig. 1.

Organization of the node and ATDM I/O ports is shown in Fig. 2. The output port contains B separate buffer spaces to avoid head-of-line blocking when accessing the bus network. The node is allowed to access at most one bus in a given time slot, as described in the following section. The bus selection logic determines the bus a node can access at each time slot. Each node is assigned a *home bus* where it receives packets from other nodes. A source node determines the home bus of the destination node based on its processor id in a decentralized way that does not require the exchange of any state information. The input port has the capability of monitoring all buses so that it can detect the end of transmission of previous slot and keeps the output port synchronized with the rest of the local network. Data is removed only from the node's home bus. The same organization is considered for ATDM optical ports (its LIN inbound and outbound ports are identical to those of a node). Wavelength channels are used in place of buses. At the output port, C buffer spaces are available for queueing global packets, one buffer for each channel. A channel is selected according to a similar slot-channel assignment

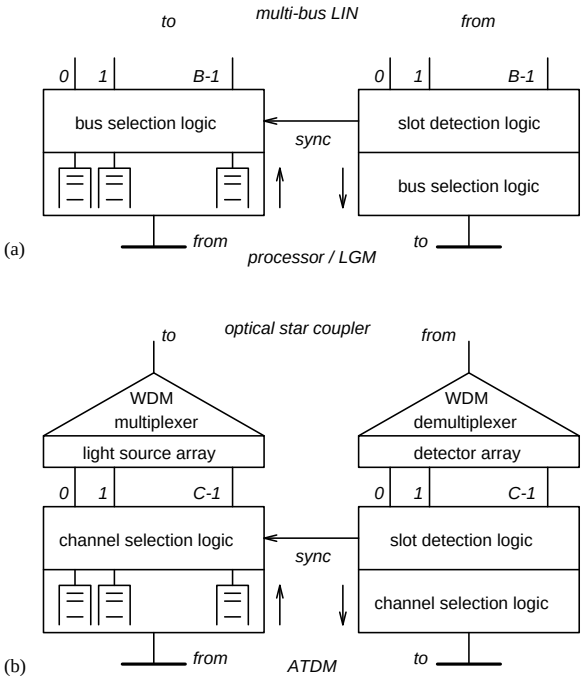

Figure 2: Organization of communication ports (a) local and (b) global ports.

map through the source array and the WDM multiplexer. The input port is assigned a home wavelength channel, and also monitors all channels to achieve slot synchronization at the GIN level.

The number of inbound ATDM ports, a_0, is arbitrary but every port need to maintain B individual buffer spaces. In general the number of outbound ports, a_1, is also assumed to be arbitrary. However, one of the design goals is to avoid creating a bottleneck at the ATDM stage. An incoming packet from the LIN to the ATDM is to be forwarded to one of C buffers for unblocked optical transmission. Each of the outbound ports is treated as a node input port in the sense that it is assigned a home bus for valid packet reception. If $a_1 = 1$, a simple self-routing tree stage can be employed to direct packets to one of the C buffers according to their destination cluster address. If $a_1 = C$, each outbound port will be directly connected to one of the C channels. Each outbound port is assigned its own home bus. The source node determines which outbound port to forward a packet to depending on the destination cluster. If $1 < a_1 < C$, an $a_1 \times C$ switch is needed at the ATDM to direct packets to the channel that is home of their destination cluster. To avoid the complexity of an internal space switch, with the possible blocking, output contention, and fault tolerance implications, only the cases of $a_1 = 1$ and $a_1 = C$ are considered.

2.2 Access Protocol

The communication protocol described below provides conflict-free communication, source-routing and does not require any complex switching function to be incorporated into the ATDM. It is also applied uniformly, but independently, at both the local and global levels. Interleaved time-division multiple access (I-TDMA) has been studied in [15] for a multiple-wavelength star-coupled network, as an extension to the cyclic single-channel TDMA. A multi-level version of the protocol was used in [6]. The single-level version is being applied in this paper to both the multi-bus LIN and the multi-wavelength GIN, with the term *channel* used to indicate either of them.

To allow source routing, destinations are assigned *home* channels and each channel is accessed by one source on a cyclic basis. No source has access to more than one channel and no more than one source can access the same channel in a given time slot. The LIN bus $i \in [0, B-1]$ is the home bus for all node output ports and ATDM outbound ports $k \in [0, m_0 + a_1 - 1]$ where

$$k \bmod B = i$$

Port k can transmit over bus j (targeting any destination whose home bus is j) during the

$$[(k-j) \bmod (m_0 + a_0 - 1)]^{th}$$

LIN time slot. Similarly at the global level, wavelength channel $i \in [0, C-1]$ is a home channel for all clusters $k \in [0, m_1 - 1]$ where

$$k \bmod C = i$$

Cluster k can transmit over channel j (targeting any destination whose home channel is j) during the

$$[(k-j) \bmod m_1]^{th}$$

GIN time slot. Channel-slot assignment maps for both communication levels are illustrated in Figure 3.

I-TDMA as defined in [15] achieves uniform low latency for loads below saturation, and high throughput that scales with the offered traffic load. The maximum throughput equals the number of channels because communication is conflict-free. It has two main drawbacks: the requirement of global slot synchronization among all network nodes, and the latency being dominated by the frame synchronization time which is directly proportional to the network size. Breaking down the network into independent local and global hierarchical levels contributes to overcoming both problems. Slot synchronization is done independently at each level with a smaller number of nodes/clusters (distributed synchronization is considered in Section 2.3). The average frame synchronization time is broken down into two additive factors corresponding to m_1 and m_0 If both m_1 and m_0 are of the same order, it is reduced by a factor of order $O(\sqrt{M})$. Moreover, advantage is taken of both the electronic and optical domains (space-division and wavelength-division) where each is more suitable and without complete reliance on either domain.

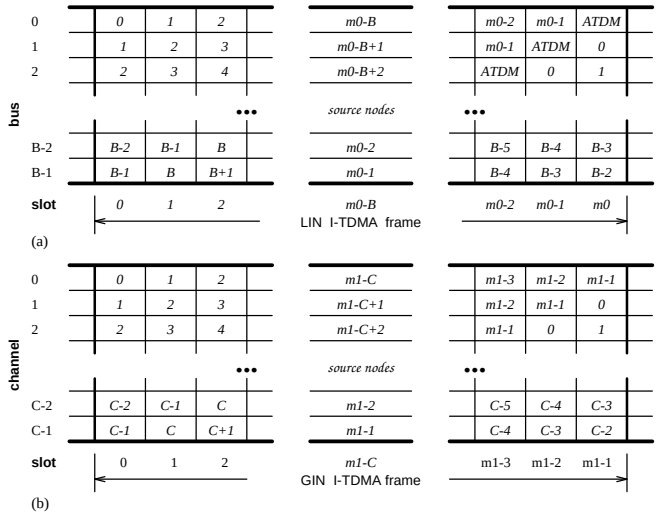

Figure 3: I-TDMA slot assignment maps: (a) bus-slot assignment map to source nodes in LIN with B buses and one ATDM inbound port, (b) channel-slot assignment map to source clusters (ATDM's) with C wavelengths

Time-division may be used not only in conjunction with each level as the means of access scheduling, but also at the ATDM stage in the form of a possible speedup factor when converting from electronic to optical transmission. Assuming that the bus serial bit rate is R_0 bps, and the serial bit rate of each wavelength channel is R_1 bps, the *speedup factor* is defined as

$$S = R_1/R_0 \geq 1$$

and is not necessarily an integer. $S > 1$ implies a smaller GIN slot and frame periods. Packets are transmitted from the C ATDM buffers over the optical network independent of their arrivals to the outbound port(s).

2.3 Distributed Slot Synchronization

Distributed slot synchronization is considered due to two important goals: improving *fault tolerance* and reducing

synchronization overhead. It is well known in switching and multiaccess protocols that the performance improves significantly if time is slotted and fixed length packets are transmitted within valid slot boundaries. This is the case with I-TDMA described in Section 2.2. Generally, there are several ways to achieve slot synchronization:

Centralized Clock: fan out a centralized clock to all nodes. This is difficult to accomplish with a large number of nodes, and there is a serious fault tolerance degradation since the failure of the central clock will disable the entire system. Slot overlap may occur if propagation delay is not uniform between the clock and all nodes.

Frame Synch: a second alternative is to have each node adjust its clock once per frame according to a signal from one reference designated node and have the slot period larger than the packet size by an amount that corresponds to the maximum possible skew of the local clock.

Slot Synch: another possible approach is to require all nodes to transmit a packet in each slot, that is enveloped by a *start-of-transmission* and *end-of-transmission* characters. If a node is not backlogged, the packet data field is blank or a specified bit pattern. In a single-bus system, all stations monitor the bus, and therefore node $i+1$ starts transmission only after detecting the end-of-transmission character from node i. The synchronization overhead consists of the propagation delay between two successive nodes in the TDMA frame plus the header decoding time. This time may still be less than a maximum security guardband to allow for clock skew during a frame. The most important feature is that there is no reliance on a single node to resynchronize every frame. For example, if node i fails node $i+1$ can detect the absence of the synchronizing character and assume the role of node i in the frame.

The third approach can be generalized for use with the multi-channel I-TDMA as follows. The node and ATDM organization of Fig. 2 shows that every node's input port and ATDM's outbound port taps all B buses. The desired bus fanout is $m_0 + a_1$. All buses are monitored for synchronization character detection and the transmit port is supplied with the *sync* strobe to inform it of the beginning of its assigned slot. Due to the fact that bus allocation to source nodes is interleaved in time, the following is a simple analysis to determine the maximum synchronization overhead and show that the buses remain synchronous with a bounded drift.

Let the slot period according to the real packet transmission time be denoted as T. The actual slot period at node i, which may vary due to a positive or negative clock skew of δ_i, is $T+\delta$. Let $t^-_{b,i}$ denote the time epoch that corresponds to the end of transmission over bus b at slot $i-1$ and $t^+_{b,i}$ denote the time epoch corresponding to the beginning of transmission over bus b at slot i. These epochs can be defined recursively as follows (for a frame length of m):

$$t^-_{b,i} = t^+_{b,i} + T + \delta_{i \bmod m} \quad (1)$$

$$t^+_{b,i} = \begin{cases} \max\{t^-_{b,i}, t^-_{b+1,i}\} & \text{for} \quad 0 \le b < B-1 \\ t^-_{b,i} & \text{for} \quad b = B-1 \end{cases} \quad (2)$$

The expressions result directly from the slot assignment maps of Fig. 3. A node that is assigned slot i for transmission over bus $b < B-1$ can go ahead only after it has detected the end-of-transmission character of the successor node in the frame of bus b *and* after it has completed transmission over bus $b+1$ during slot $i-1$. For example, assuming that the network is initialized at the beginning of slot 0 so that transmission over all buses b start simultaneously at time $t^+_{b,0} = 0$, then the end of transmission epochs are:

$$t^-_{b,1} = T + \delta_b$$

The beginning of slot 1 transmission takes place at times:

$$t^+_{b,1} = \begin{cases} \max\{t^-_{b,1}, t^-_{b+1,1}\} = \\ T + \max\{\delta_b, \delta_{b+1}\} & \text{for} \quad 0 \le b < B-1 \\ T + \delta_{B-1} & \text{for} \quad b = B-1 \end{cases}$$

Assuming that the maximum positive clock skew over all nodes is δ, and since every node contributed equal skew to all buses, Equations 1 and 2 imply that all buses will maintain synchronous frame within a variation bounded by δ and that the maximum slot enlargement over the packet transmission time T is also δ. Therefore distributed slot synchronization is achieved with a maximum overhead of δ/T. The difference between this scheme and the scheme that assumes resynchronization every frame is that the overhead is not embedded into the slot time, but is rather variable dictated by the individual node skews with the mentioned bound. Also, the frame is recursively initialized for all buses if only one node started transmission over bus $B-1$ at an arbitrary time (initialization takes $B-1$ slots, after which all buses become active).

2.4 Network Characteristics

This section evaluates the system's modularity, scalability and bandwidth partitioning. *Modularity* refers to the capa-

bility of the system to be expanded with arbitrarily small increments without disturbing the access protocol functionality and delay performance. The system *scalability* refers to the corresponding increase of interconnection complexity and/or degradation in access delay performance when the system is expanded by a certain increment. *Bandwidth partitioning* in a multi-channel network implies partitioning the available channels to several hierarchical levels, with the result of concurrent spatial re-use of channels by groups of nodes within a certain hierarchic level.

Modularity and Scalability

Modularity of the TSW interconnection is achieved at both the local and global levels. The number of clusters and the number of nodes per cluster are independent. Global inter-cluster communication may take place with any number of wavelength channels ($C \geq 1$) and local cluster communication may also take place with any number of buses ($B \geq 1$). The multiprocessor network upgrades can be achieved with maximum flexibility. Increasing the number of nodes by a small fraction is done by adding one (or more) nodes to some (or all) clusters and/or adding more clusters. Increasing the network capacity can be achieved by increasing the number of local and/or global channels. Adding clusters or wavelength channels requires only modifying the channel-slot assignment map, which is computed only once at each ATDM. The number of logical buffers of the outbound ATDM port would be correspondingly increased to match the new number of channels. Increasing the number of nodes per cluster is simple because the bus network is non-switched and only involves extending the passive taps to cover the added bus(es). All nodes must be aware of local or global frame changes since this may involve changes to the home channels of nodes (or clusters). Packaging constraints are relaxed because metal buses exist only within a cluster, and inter-cluster communication is via fiber links to and from the star coupler. The metal buses could be implemented on cluster "backplanes", directly tapped by the nodes and ATDM PCBs. Due to the distributed slot synchronization scheme, the operation of the global network is insensitive to propagation delay and clusters do not need to be packed within a tight enclosure.

The TSW interconnection has a lower than linear complexity. The number of fiber links between clusters is of order $O(m_1)$ and the number of crosspoints in each LIN is of order $O(m_0)$, since B does not increase linearly as m_0 increases. The overall interconnection complexity for a network size of $M = m_1 m_0$ is of order $O(m_1 + m_0)$. The dominating factor of the access latency below saturation, which is the frame synchronization time, is also of the same order. Therefore, the interconnection is efficiently scalable from both cost and performance standpoints. Detailed performance analysis is conducted in Section 3.

Reconfigurable Bandwidth Partitioning

Bandwidth partitioning is an important feature that enables more effective utilization of the available bandwidth, space or wavelength channels, especially in the existence of spatial reference locality. Programmability of the interconnection network allows modifying the bandwidth partition based on variations on cross-reference traffic. For example, consider a flat network with $\mathcal{M}$ nodes and $\mathcal{C}$ channels. This can be hierarchically partitioned into two levels, consisting of $\mathcal{M}_1$ clusters of $\mathcal{M}_0$ nodes each, with $\mathcal{C}_0$ channels allocated to the local cluster references, and $\mathcal{C}_1 = \mathcal{C} - \mathcal{C}_0$ channels allocated to inter-cluster references. The effective number of channels being concurrently utilized is increased to $\mathcal{C}_1 + \mathcal{M}_1\mathcal{C}_0$, with an improvement factor of $1 + \mathcal{C}_0(\mathcal{M}_1 - 1)/\mathcal{C}$. This hierarchical bandwidth partitioning was introduced and analyzed in [6]. In general, for an r-level hierarchy, the number of effective concurrent channels is:

$$\mathcal{C} = \sum_{i=0}^{r-1} \mathcal{C}_i \prod_{j=i+1}^{r-1} \mathcal{M}_j \qquad (3)$$

Both the LIN and GIN bandwidths can be partitioned according to this principle. Partitioning the optical bandwidth of the GIN may be done using a tree network of spatial wavelength routers and r-channel receivers, as described in [6]. Hierarchical partitioning of the LIN may be done similarly if space switches replaced the passive bus taps. Introducing this type of hierarchy to the LIN and/or GIN is arbitrary and depends on the network configuration. The two-level (LIN/GIN) hierarchy studied in this paper is a generalization that allows independent operation of both levels, where each may be hierarchical.

3 Access Delay Analysis

This section presents the performance analysis of the hierarchical interconnection scheme under consideration. Section 3.1 describes the detailed model based on tandem $Geom/D/1$ queues representing local, outbound, and inbound references. Section 3.2 reduces the performance results to closed form under certain conditions regarding the network configuration. Section 3.3 analyses the delay-throughput characteristics and compares them with other related architectures.

3.1 The Model

The delay model is based on $Geom/D/1$ queues which are commonly used with TDMA-based protocols [16]. It is assumed that nodes generate packets independently in successive slots, with a Bernoulli parameter g. The packet may be destined to either a node within the cluster or to a node in another cluster with probabilities p_0 and $p_1 = 1 - p_0$, respectively. When the generated traffic at a node references all nodes uniformly, $p_{0,uniform} = (m_0 - 1)/(M - 1)$ and $p_{1,uniform} = m_0(m_1 - 1)/(M - 1)$. It is assumed in general that $p_0 \geq p_{0,uniform}$ and $p_1 \leq p_{1,uniform}$, to account for reference locality within a cluster. It is also assumed that local references target nodes within the clusters with uniform probability and global references target nodes outside the cluster with uniform probability. The normalized durations of the local and global I-TDMA frames are:

$$L_0 = m_0 + a_0 \quad (4)$$

$$L_1 = m_1/S \quad (5)$$

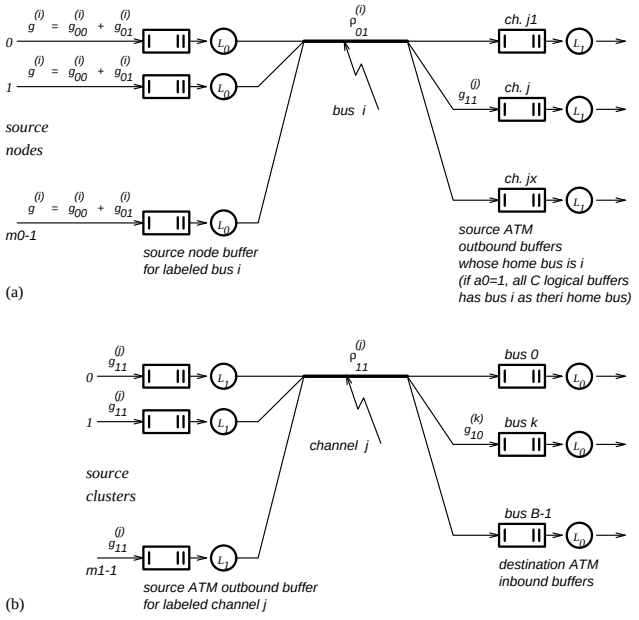

Figure 4: Queueing model: (a) local and outbound traffic (b) inbound traffic

Every node (or inbound ATDM port) has B buffer spaces, where buffer $i \in [0, B-1]$ is used for FIFO queueing of packets addressed to a node (or an outbound ATDM port) whose home bus is i. The arrival rate to each buffer depends on the relative number of destination nodes assigned to that buffer. Since there are two types of destinations (local and global) and because the number of each type is not necessarily an integer-multiple of B, the assignment of home buses to destination nodes and home wavelength channels to destination clusters need to be individually considered.

Let $u_{00}^{(i)}$ denote the number of local destinations whose home bus is i, and $u_{01}^{(i)}$ indicate the number of outbound ATDM ports whose home bus is i. The destination-to-buffer assignment is

$$u_{00}^{(i)} = \begin{cases} \lceil m_0/B \rceil & \text{for} \quad 0 \leq i < m_0 \bmod B \\ \lfloor m_0/B \rfloor & \text{for} \quad m_0 \bmod B \leq i < B \end{cases} \quad (6)$$

for local destinations. For global destinations, the outbound ATDM ports need to be considered in two cases. If $\lceil m_0 \bmod B \rceil \leq \lceil (m_0 + a_1) \bmod B \rceil$ then

$$u_{01}^{(i)} = \begin{cases} \lceil a_1/B \rceil & \text{for} \\ & m_0 \bmod B \leq i \leq m_0 + a_1 \bmod B \\ \lfloor a_1/B \rfloor & \text{otherwise} \end{cases} \quad (7)$$

Otherwise, when $\lceil m_0 \bmod B \rceil > \lceil (m_0 + a_1) \bmod B \rceil$, we have

$$u_{01}^{(i)} = \begin{cases} \lceil a_1/B \rceil & \text{for} \quad 0 \leq i \leq (m_0 + a_1) \bmod B \\ & \text{and} \quad m_0 \bmod B \leq i \leq B - 1 \\ \lfloor a_1/B \rfloor & \text{otherwise} \end{cases} \quad (8)$$

The geometric packet arrival process to buffer $i \in [0, B-1]$ of any of the m_0 nodes in a cluster can be expressed as a superposition of local and global arrivals to that buffer, with parameter $g^{(i)} = g_{00}^{(i)} + g_{01}^{(i)}$, where

$$g_{00}^{(i)} = gp_0 u_{00}^{(i)}/(m_0 - 1) \quad (9)$$

$$g_{01}^{(i)} = gp_1 u_{01}^{(i)}/a_1 \quad (10)$$

The utilization factor of buffer i in a given node is $\rho_0^{(i)} = g^{(i)} L_0$. This factor can be expressed as a summation $\rho_0^{(i)} = \rho_{00}^{(i)} + \rho_{01}^{(i)}$, where

$$\rho_{00}^{(i)} = g_{00}^{(i)} L_0 \quad (11)$$

$$\rho_{01}^{(i)} = g_{01}^{(i)} L_0 \quad (12)$$

indicate the probabilities of bus i carrying a local and global packet during a given slot, respectively. The average waiting times in buffer i for local and global traffic are given by:

$$\overline{W}_{00}^{(i)} = \frac{gp_0 u_{00}^{(i)}(m_0+a_0)(m_0+a_0-1)}{2[(m_0-1)-gp_0 u_{00}^{(i)}(m_0+a_0)]} \quad (13)$$

$$\overline{W}_{01}^{(i)} = \frac{gp_1 u_{01}^{(i)}(m_0+a_0)(m_0+a_0-1)}{2[(m_0-1)-gp_1 u_{01}^{(i)}(m_0+a_0)]} \quad (14)$$

The behavior of an ATDM outbound buffer need to be analyzed next. We consider only the cases where $a_1 = 1$ and $a_1 = C$, to avoid the need for an internal $(a_1 \times C)$ switch fabric within the ATDM. Therefore, a global packet either arrives to a single ATDM outbound port and then gets routed to one of C buffer spaces, or directly arrives to one of C outbound ATDM ports, each having an LIN home bus assigned to it. In both cases, the number of clusters assigned to buffer $j \in [0, C-1]$ (those whose home channel is j) is:

$$u_{11}^{(j)} = \begin{cases} \lceil m_1/C \rceil & \text{for} \quad 0 \le j < m_1 \bmod C \\ \lfloor m_1/C \rfloor & \text{for} \quad m_1 \bmod C \le j < C \end{cases} \quad (15)$$

Each of the C outbound buffer spaces is modeled by a $Geom/D/1$ queue with a geometric arrival parameter that corresponds to the fraction of clusters whose home channel is j. If $a_1 = C$:

$$g_{11}^{(j)} = \rho_{01}^{(i)} u_{11}^{(j)}/(m_1-1) \quad (16)$$

where $i = (m_0 + j) \bmod B$ and if $a_1 = C$:

$$g_{11}^{(j)} = \frac{\rho_{01}^{(i)} u_{11}^{(j)}}{\sum_{x=1}^{u_{01}^{(i)}} u_{11}^{(j_x)}} \quad (17)$$

where $i = (m_0 + j_x) \bmod B \ \forall\, j_x \in \{j_1, \ldots, j, \ldots, j_{u_{01}^{(i)}}\}$. Note that i is the home bus of the single outbound port (all C buffer spaces) in the first case and is the home bus of all outbound ports, each having a single buffer space, j_x in the second case (including the labeled channel j).

The utilization factor of ATDM outbound buffer j is given by

$$\rho_{11}^{(j)} = g_{11}^{(j)} L_1 \quad (18)$$

and the waiting time is:

$$\overline{W}_{11}^{(j)} = \frac{g_{11}^{(j)} m_1 (m_1 - S)}{2S[1 - g_{11}^{(j)} m_1 S]} \quad (19)$$

At the destination ATDM, the optical packet stream is converted to electronic and is cyclically demultiplexed to the a_0 inbound ports, each of which has B separate buffer spaces. The arrival parameter to each of these buffers can be expressed in terms of the utilization of the wavelength channel that is a home channel for the considered cluster. For buffer k in a given inbound port;

$$g_{10}^{(k)} = \frac{1}{a_0} \frac{\rho_{11}^{(j)}}{u_{11}^{(j)}} \frac{u_{00}^{(k)}}{m_0} \quad (20)$$

where j is the home channel of the considered cluster. Since all clusters are assumed identical, j denotes an arbitrary labeled channel (there is no relation between j and k). The utilization factor of an inbound buffer k is:

$$\rho_{10}^{(k)} = g_{10}^{(k)} L_0 \quad (21)$$

and the waiting time is:

$$\overline{W}_{10}^{(k)} = \frac{g_{10}^{(k)}(m_0+a_0)(m_0+a_0-1)}{2[1 - g_{10}^{(k)}(m_0+a_0)]} \quad (22)$$

The average local and global reference latencies are found by adding the packet transmission and mean frame synchronization time and averaging the waiting time over all home buses (and channels for global reference latency);

$$\overline{D}_0 = \frac{m_0+a_0+1}{2} + \frac{1}{B}\sum_{i=0}^{B-1} \overline{W}_{00}^{(i)} \quad (23)$$

$$\overline{D}_1 = \frac{2S(m_0+a_0+1)+(m_1+1)}{2S} + \frac{1}{B}\sum_{i=1}^{B-1} \overline{W}_{01}^{(i)} + \frac{1}{C}\sum_{j=0}^{C-1} \overline{W}_{11}^{(j)} + \frac{1}{BC}\sum_{k=0}^{B-1}\sum_{j=0}^{C-1} \overline{W}_{10}^{(k)} \quad (24)$$

In Eqn. 24, the first summation represents the average waiting time for a global reference in the initiating node buffer, the second summation represents its average waiting time in the ATDM outbound buffer (each labeled channel j uniquely determines a corresponding bus i, Eqn. 17), and the third double summation represents the average waiting time in the destination cluster ATDM inbound buffer (the channel of arrival and the bus leading to final destination

node are independent). The average access latency is given in terms of the probability of reference locality within the cluster;

$$\overline{D} = p_0(\overline{D}_0 - \overline{D}_1) + \overline{D}_1 \quad (25)$$

The total system throughput is:

$$\Gamma = \left[m_1 \sum_{i=0}^{B-1} \rho_0^{(i)} + \sum_{j=0}^{C-1} u_{11}^{(j)} \sum_{k=0}^{B-1} \rho_{10}^{(k)} \right] + \sum_{j=0}^{C-1} \rho_{11}^{(j)} \quad (26)$$

where the first term represents the combined LIN throughput of all clusters due to both local and global references and the second term represents the star coupler throughput. The maximum attainable system throughput is $M = m_1 m_0$.

3.2 Simplification

The above detailed model may be relaxed by uniformly treating all buffers regardless of assignment of home buses (channels) to nodes (clusters). It can be seen that this assumption is reasonable if $m_0 + a_0 \gg B$, $a_1 \gg B$, and $m_1 \gg C$. It is actually valid if B divides both $m_0 + a_0$ and a_1 and C divides m_1. By simply using $u_{00} = m_0/B$, $u_{01} = a_1/B$, and $u_{11} = m_1/C$; the waiting time results become:

$$\overline{W}_{00} = \frac{gp_0 m_0(m_0 + a_0)(m_0 + a_0 - 1)}{2[(m_0 - 1)B - gp_0 m_0(m_0 + a_0)]} \quad (27)$$

$$\overline{W}_{01} = \frac{gp_1 a_1(m_0 + a_0)(m_0 + a_0 - 1)}{2[(m_0 - 1)B - gp_1 m_0(m_0 + a_0)]} \quad (28)$$

$$\overline{W}_{11} = \frac{gp_1 m_1(m_0 + a_0)(m_1 - S)}{2S[BC - gp_1(m_0 + a_0)m_1 S]} \quad (29)$$

$$\overline{W}_{10} = \frac{gp_1(m_0 + a_0)^2(m_0 + a_0 - 1)}{2[a_0 B^2 S - gp_1(m_0 + a_0)^2]} \quad (30)$$

The local and global delay expressions become

$$\overline{D}_0 = \frac{m_0 + a_0 + 1}{2} + \overline{W}_{00} \quad (31)$$

$$\overline{D}_1 = \frac{2S(m_0 + a_0 + 1) + (m_1 + 1)}{2S} + \overline{W}_{01} + \overline{W}_{11} + \overline{W}_{10} \quad (32)$$

The mean access latency is given by Eqn. 25, and the total system throughput can be expressed as

$$\Gamma = m_1 B(\rho_0 + \rho_{10}) + C\rho_{11}$$

$$= g\frac{m_1(m_0 + a_0)[a_0 BS + (1 - p_0)(m_0 + 2a_0)]}{a_0 BS} \quad (33)$$

where both $\rho_0 + \rho_{10}$ and ρ_{11} are smaller than one. The maximum system throughput is, as intuitively expected,

$$\Gamma_{max} = m_1 B + SC \quad (34)$$

Note that both delay and throughput measures are normalized to the LIN slot time. The delay-throughput characteristics are analyzed in the following section for various system configurations.

3.3 Analysis

The system performance is analyzed via three sets of graphs shown in Figures 5, 6, and 7. The graphs examine the delay-throughput performance and system scalability characteristics and compares them to both flat and hierarchical I-TDMA-based systems with only one node per cluster. The variables considered in this evaluation are the hierarchic configuration, the number of buses and channels, the number of ATDM ports, the speedup due to optical transmission, and the reference locality.

Delay-throughput: The set of graphs of Fig. 5 illustrate the delay-throughput performance characteristics for a fixed system size of 1 K-nodes, 32 clusters with 32 nodes each. The horizontal and vertical axes represent the system throughput and mean access delay normalized to the system size and the LIN time slot, respectively. The plots are shown with $B = C = \{2, 4, 8\}$ buses (wavelength channels). In each graph, the number of outbound ATDM ports is varied between 1 and C and that of the inbound ports between 1 and 2 ports. The reference probability is varied within each graph from uniform reference (with $p_0 = (m_0 - 1)/(M - 1)$ to higher reference localities with $p_0 = 0.5$ and 0.9. The dominating delay component is the summation of the LIN and GIN frame synchronization times ($m_0 + m_1/S$). If $m_1 = m_0 = \sqrt{M}$, the mean frame synchronization time is reduced by a factor of $\sqrt{M}$ from that of a regular TDMA-based scheme.

For a given number of channels $B = C$, the impact of reference locality is noted as relaxing the delay versus throughput profile (lower average below saturation delay and higher maximum throughput). As the reference locality increases, the queueing time at both the ATDM outbound and inbound ports becomes less significant. The GIN utilization decreases, which is desirable since every cluster contributes global references from m_1 nodes. On the average, the LIN utilization is not dramatically affected. The reason is that global references eventually generate local traffic at their destination cluster. The load presented to

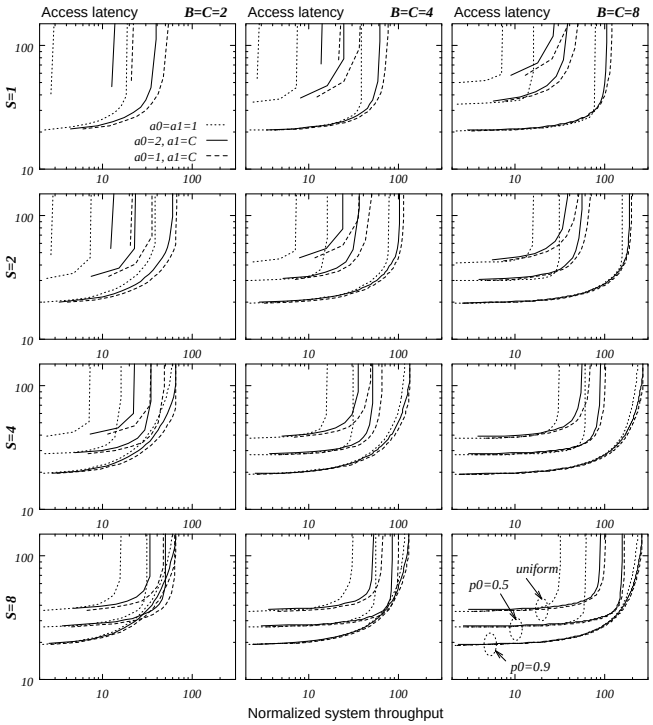

Figure 5: Delay (normalized to LIN slot) verses normalized system throughput characteristics of a 1024 node TSW-based system with $m_1 = m_0 = 32$, $B = C = \{2, 4, 8\}$, $a_0 = \{1, 2\}$, $a_1 = \{1, C\}$, $S = \{1, 2, 4\}$, and $p_0 = \{\text{uniform}, 0.5, 0.9\}$

node buffers should not dominate that presented to ATDM inbound buffers since local traffic is generated by m_0 nodes while global traffic is generated by $m_0(m_1 - 1)$ nodes. Using the simplification of Section 3.2, these loads are $g_{00} = gp_0/B$ and $g_{10} = gp_1(m_0 + a_0)/a_0B^2S$, respectively. It can be seen that a probability of a local reference $p_0 = p_1(m_0+a_0)/a_0B^2S$ maintains a balanced load at both points, and therefore a balanced saturation throughput.

The impact of varying the number of ATDM ports is examined in all cases. Outbound buffers are varied between 1 and C only since other arbitrary numbers would require an internal switch within the ATDM. When the reference locality is not too high (p_0 uniform and 0.5), it is noted that using C outbound ports consistently results in a significant improvement of system capacity, in terms of saturation throughput. There is no notable impact on the average delay below saturation and also the capacity improvement is small when the locality is too high ($p_0 = 0.9$) since the outbound ports are then lightly loaded. Increasing the number of inbound ports a_0 from 1 to 2 generally has a negative impact on the performance. The reason is that the local frame time would increase corresponding to the increase in a_0. There is no benefit in terms of alleviating congestion because a single port has B logical buffers, which corresponds to the maximum LIN capacity. Therefore, it is concluded that one inbound port is sufficient and this parameter is not varied anymore.

The speedup factor contributes to reducing the normalized global frame time and to enable rapid clearing of outbound traffic. Improvement can be most observed for all plotted cases in each graph when S increases from 1 to 2 and 4. The increase from 4 to 8 improves the delay near the saturation point and also contributes to the reduction of the normalized global frame synchronization time. However, it appears that lower speedups for the considered system size are sufficient to handle outbound traffic without significant queueing delay.

Scalability: Fig. 6 examines the performance scalability as the system size grows. The system can grow either via unbalanced or balanced increments. An unbalanced growth implies starting with a limited number of clusters, each with a fixed number of nodes, and expand by increasing the number of clusters. Balanced growth would maintain consistent and lower frame synchronization time because $m_1 + m_0 = M$ is minimum when $m_1 = m_0 = \sqrt{M}$. Figures 6(a), (b), and (c) show the delay versus throughput plots for an unbalanced growth of from 4 to 64 32-node clusters, doubling the system size at each step. It can be seen that $M = 32 \times 32$ nodes has the optimal profile and that the saturation point is decreased when the system size is increased resulting in its becoming unbalanced again. Figures 6(d), (e), and (f) show the performance scalability characteristics corresponding to balanced system growth from 8×8 to 32×32 nodes. The interconnection scheme considered in this paper does not pose any constraints on the method of system expansion and increments as small as 1 node to arbitrary clusters are allowable. Scalability is examined also as the reference locality changes from uniform to $p_0 = 0.5$ and 0.9 for both methods of growth, in Figures 6 (a,d), (b,e), and (c,f), respectively.

Comparison to TDMA-based architectures: Fig. 7 compares the performance of the considered approach to that of the hierarchical SWHA architecture of [6] and to a non-hierarchical (one-level) TDMA-based systems. The SWHA has a similar hierarchical spatial re-use approach which results in a larger maximum throughput than the limited number of channels, as discussed in Section 2.4 (Eqn. 3). Figures 7(a) and (b) provides the performance comparison with varying system size and number of channels, respectively. Figures 7(a) considers TSW with $B = 4$, $a_1 = C$, and $C = 4$ and m_1 as compared to a two-

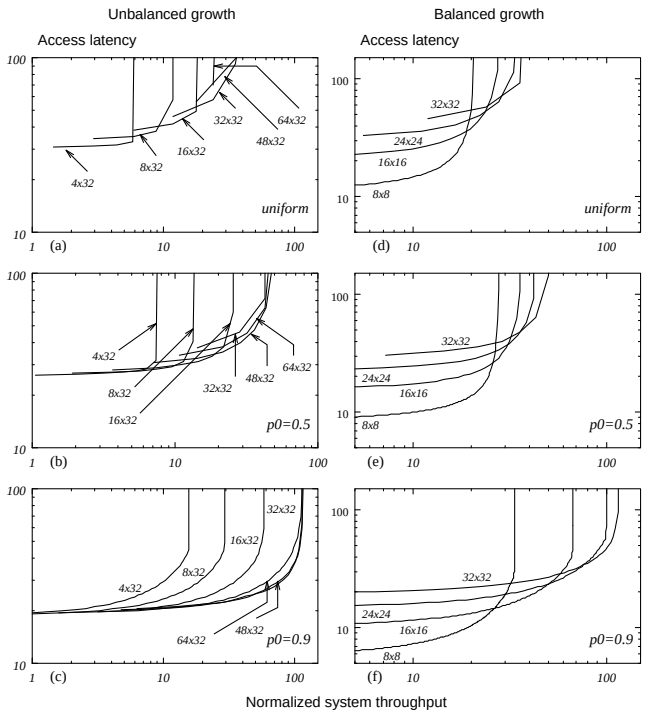

Figure 6: Scalability of TSW interconnection showing both unbalanced and balanced growth methods. Delay (normalized to LIN slot) verses normalized system throughput with $B = C = 4$, $S = 2$, $a_0 = 1$, $a_1 = 4$, and $p_0 = \{\text{uniform}, 0.5, 0.9\}$

level SWHA with equal size using 32 channels and optimal channel partitioning for a reference locality of $p_0 = 0.8$. In the TSW the number of wavelength channels is usable only for global communication, and therefore $m_1 \leq m_1$. Due to reducing the frame synchronization time domination, The TSW has lower delay in both cases, which is not very sensitive to size increase. Throughput is generally lower than that of SWHA because the considered number of channels is smaller. When $C = 32$, the TSW maximum throughput approaches that of the SWHA. Figures 7(b) considers a fair relative number of channels in the comparison of TSW SWHA , and a flat I-TDMA system. C is varied as 8, 16, and 32 channels for the I-TDMA and SWHA, while the same variation is applied to $B + C$ for the TSW.

The main feature of SWHA is extending the system capacity beyond the available number of channels through wavelength spatial re-use, which is clear in the improvement over the I-TDMA. The other important feature is that the bandwidth partition is reconfigurable and can be varied to suit different system configurations and reference localities. The main feature of the TSW is the significant delay reduction due to breaking down the frame time. The maximum throughput is close to that of the SWHA when the number of channels is comparable. The ratio of local-to-global bandwidth partition is not reconfigurable in the TSW, and the shown configuration ($B = 4$) is not necessarily optimal with regards to the considered reference locality. The number of buses can be increased independently of the number of channels with some hardware modification to the cluster. In terms of complexity and fault tolerance, both architectures employ passive and distributed interconnection schemes. The SWHA employs spatial wavelength switches for reconfigurable bandwidth partitioning, while the TSW employs ATDM multiplexers at the local-global boundary.

When the SWHA was introduced in [6], it was the intention that the term *node* may represent a cluster of nodes sharing the optical interface. The architecture presented in this paper is actually an extension of the previous work. The GIN in this paper may be an SWHA network itself, enabling very low-latency and high-capacity system scalability to the massive parallelism region of several thousand processors. The concept of reconfigurable bandwidth partitioning can be applied to the LIN bus network as well. The TSW architecture of this paper is intended to treat the nodes of the SWHA as clusters and examine the performance of the resulting system.

4 Conclusions

This paper develops a general hierarchical approach for high-performance massively parallel computer systems that possess the following characteristics: distributed conflict-free access with source routing, physical modularity, low-complexity passive interconnect, scalable performance, small constant expansion increment, and spatial bandwidth re-use that accommodates spatial reference locality. A cluster-based approach is considered in a flexible two-level hierarchy that does not impose configuration constraints. Processors within the same cluster are interconnected via a passive multi-bus network, and clusters are linked through a passive optical star. Interleaved TDMA is employed as the media access protocol at both local and global network levels. Each cluster accesses the global optical network through an asynchronous time-division multiplexer that taps into the local multi-bus network as an ordinary node. The independence between both levels allows more flexible growth and provides better fault tolerance. Most important, from the delay performance standpoint, the TDMA frame synchronization time is broken down into two additive factors. This time, which dominates the average access latency at loads below saturation, is dramatically reduced by an order of $O(\sqrt{M})$, for a network size of M, when the cluster size is of the same order as the number of clusters. Spa-

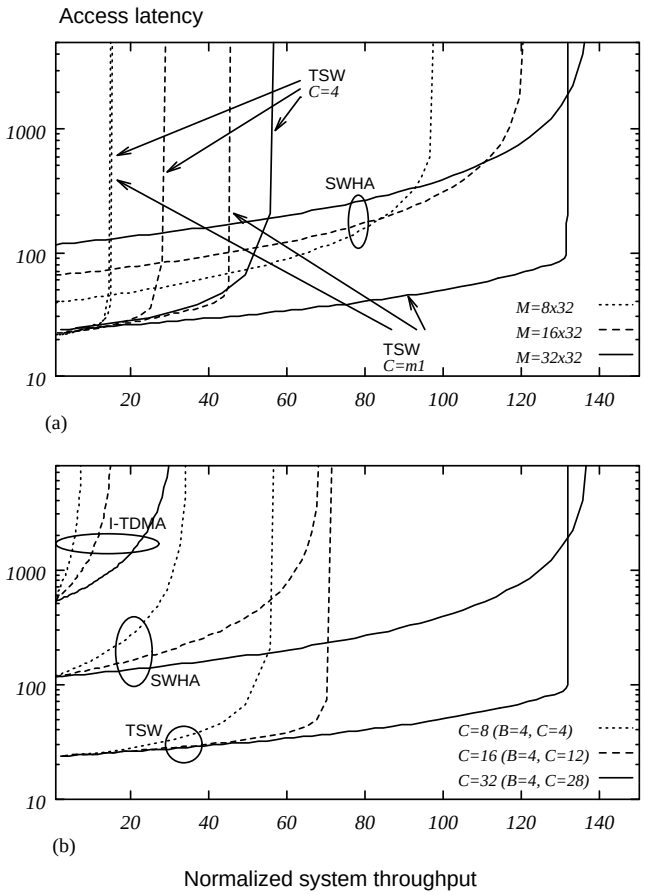

Figure 7: Performance comparison with hierarchical architectures. Delay (normalized to LIN Slot) vs. normalized system throughput with $p_0 = 0.8$. (a) TSW with $C = \{4, m_1\}$, $a_0 = 1$, and $a_1 = C$ compared to SWHA with optimal channel partition ($m_1 = \{8, 16, 32\}$ and $m_0 = 32$ nodes); (b) TSW with $B = 4$ and $B+C = \{8, 16, 32\}$ compared to flat I-TDMA and SWHA with optimal channel partition using $C = \{8, 16, 32\}$ channels ($M = 32 \times 32 = 1$ K-nodes)

tial bandwidth re-use is achieved since each cluster owns its own local interconnection network, taking advantage of reference locality when it is exhibited. The proposed general scheme can be used to build expandable distributed processor networks ranging from a few to several thousand processors, at lower than linear corresponding complexity, and without compromising the system delay-throughput performance.

References

[1] M. Marson, G. Balbo, and G. Conte, *"Performance Models of Multiprocessor Systems"*. The MIT Press, 1986.

[2] D. Chiarulli, S. Levitan, and R. Melhem, "Optical bus control for distributed multiprocessors," *Journal of Parallel and Distributed Computing*, vol. 10, pp. 45–54, Oct. 1990.

[3] T. Szymanski, "A fiber-optic hypermesh for SIMD/MIMD machines," in *Proc. IEEE Supercomputing'90*, pp. 710–719, Nov. 1990.

[4] P. W. Dowd, "Random access protocols for high speed interprocessor communication based on a passive star topology," *IEEE Journal on Lightwave Technology*, vol. 9, pp. 799–808, June 1991.

[5] P. W. Dowd, "Wavelength division multiple access channel hypercube processor interconnection," *IEEE Transactions on Computers*, vol. 41, pp. 1223–1241, Oct. 1992.

[6] P. W. Dowd, K. Bogineni, K. A. Aly, and J. Perreault, "Hierarchical scalable photonic architectures for high-performance processor interconnection," *IEEE Transactions on Computers*, vol. 42, pp. 1105–1120, Sept. 1993.

[7] K. A. Aly and P. W. Dowd, "Scalability of discrete broadcast-select multi-domain optical networks for partitionable multiprocessor networks," *IEEE Transactions on Parallel and Distributed Systems*, (Under Review), 1993.

[8] C. A. Brackett, "Dense wavelength division multiplexing networks: Principles and applications," *IEEE Journal on Selected Areas of Communications*, vol. 8, pp. 948–964, Aug. 1990.

[9] R. Rund and L. Bersiner, "Experimental demonstration of bidirectional WDM transmission with LED spectral slicing," in *8th Annual European Fibre Optic Communications and Local Area Network Conference (E-FOC90)*, (Munich, Germany), June 1990.

[10] M. Maeda *et al.*, "Multigigabit/s operation of 16-wavelength vertical-cavity surface-emitting laser array," *IEEE Photonic Technology Letters*, vol. 3, pp. 863–865, Oct. 1991.

[11] M. Girard, C. Husbands, and P. Dowd, "Media access protocols for spectral sliced WDMA testbed," *SPIE Proceedings (High-Speed Fiber Networks and Channels)*, vol. 1784, pp. 169–180, Sept. 1992.

[12] P. Kirkby, "Multichannel wavelength-switched transmitters and receivers- new component concepts for broadband networks and distributed switching systems," *IEEE Journal on Lightwave Technology*, vol. 8, pp. 202–211, Feb. 1990.

[13] C. E. Leiserson, "Fat-trees: Universal networks for hardware-efficient supercomputing," *IEEE Transactions on Computers*, vol. c-34, pp. 892–901, Oct. 1985.

[14] K. Hwang and J. Ghosh, "Hypernet: A Communication-Efficient Architecture for Constructing Massively Parallel Computers," *IEEE Transactions on Computers*, vol. C-36, pp. 1450–1466, Dec. 1987.

[15] K. Bogineni, K. M. Sivalingam, and P. W. Dowd, "Low complexity multiple access protocols for wavelength-division multiplexed photonic networks," *IEEE Journal on Selected Areas of Communications*, vol. 11, pp. 590–604, May 1993.

[16] R. Rom and M. Sidi, *Multiple Access Protocols – Performance and Analysis*. Springer Verlag, 1990.