Realizing Common Communication Patterns in Partitioned Optical Passive Stars (POPS) Networks

Greg Gravenstreter and Rami G. Melhem, *Senior Member*, *IEEE*

Abstract—We consider the problem of realizing several common communication structures in the all-optical Partitioned Optical Passive Stars (POPS) topology. We show that, often, the obvious or "natural" method of implementing a communication pattern in the POPS does not efficiently utilize its communication capabilities. We present techniques which distribute the communication load uniformly in the POPS for four of the most common communication patterns (all-to-all personalized, global reduction operations, ring, and torus). We prove that these techniques provide optimal performance in the sense that they minimize the time required to deliver the messages from each node to its neighbors.

Index Terms—Optical interconnections, passive stars, embedding, all-to-all communications, reduction operations, multiplexing.

1 Introduction

The ever-increasing needs of new multiprocessor interconnection networks have created an interest in optical fiber technology. The characteristics of optical fiber links are well suited to resolve several of the most serious problems in electronic networking. These advantages include optical fiber's power efficiency, lack of reactive loading factors, and a relatively high noise immunity. The striking rise in demand for network throughput, which is the most challenging network requirement, is driven by both large increases in system size and higher node interactions. Initially, optical fiber links were incorporated into existing network designs. Throughput bottlenecks and high latencies resulting from electronic/optical conversions and processing at intermediate hops still limited the capacities of these networks. "All-optical" networks were specifically developed to address these issues. Those networks can be implemented with passive optical technology and can provide enormous potential throughput with very low latencies.

New all-optical networking designs are needed that present multiple data channels which are physically concurrent [6]. Wavelength division multiplexing (WDM) is one method for achieving these concurrent multiple channels. A large amount of research has been accomplished on WDM-based systems [4], [13]. The majority of this has been to develop multiple-access multiple-channel protocols on star architectures using tunable transmitters and receivers [7]. Theoretically, WDM has remarkable potential, yet substantial progress is needed to achieve acceptable tuning speeds (nanoseconds) over necessary ranges (hundreds of nanometers). More progress is also needed on solutions to power budget problems in large WDM systems.

The Partitioned Optical Passive Stars (POPS) topology [1], [5] is a topological approach to providing multiple physical data channels. It is an interconnection architecture that uses multiple nonhierarchical stars to implement single-hop networks [2], [3]. It is an all-optical topology constructed exclusively with passive optical technology and benefits from all the corresponding characteristics, such as no intermediate electronic/optical conversions, no reactive factors, and high noise immunity. All data channels in a POPS network may use the same fixed wavelength. Fixed transmitters and fixed receivers using current technologies may be chosen based on cost, mechanical, or other factors. Further, the design flexibility of a POPS topology can be used to avoid power budget problems. Established multiplexing methods and control protocols, such as those used in time-division multiplexing (TDM) or wavelength-division multiplexing (WDM) designs, may be applied to any or all of the data channels on a POPS network.

A POPS topology is configured at design time to provide a fixed number of physically concurrent data channels, each of which is capable of high capacity in a circuit-switched system. The number of such channels is not absolutely limited and is a key engineering trade-off. This design flexibility provides for a customized optimization between lower total system complexity versus the combination of higher system throughput, lower power budgets, and lower network control overheads. An analysis in [5] confirms that, independent of the number of channels configured, POPS network data channels can be efficiently utilized for random permutation-based communication patterns.

Beyond random communication patterns, multiprocessor applications frequently exhibit regular communication patterns. Topologies that are appropriate for interconnection networks must efficiently accommodate these communication structures. All-to-all personalized (complete)

- *G. Gravenstreter is with the Software Engineering Institute, Carnegie-Mellon University, Pittsburgh, PA .*
- *R.G. Melhem is with the Department of Computer Science, University of Pittsburgh, Pittsburgh, PA 15260. E-mail: melhem@cs.pitt.edu.*

Manuscript received 15 Nov. 1995; revised 13 June 1996.
For information on obtaining reprints of this article, please send e-mail to: tc@computer.org, and reference IEEECS Log Number 107130.

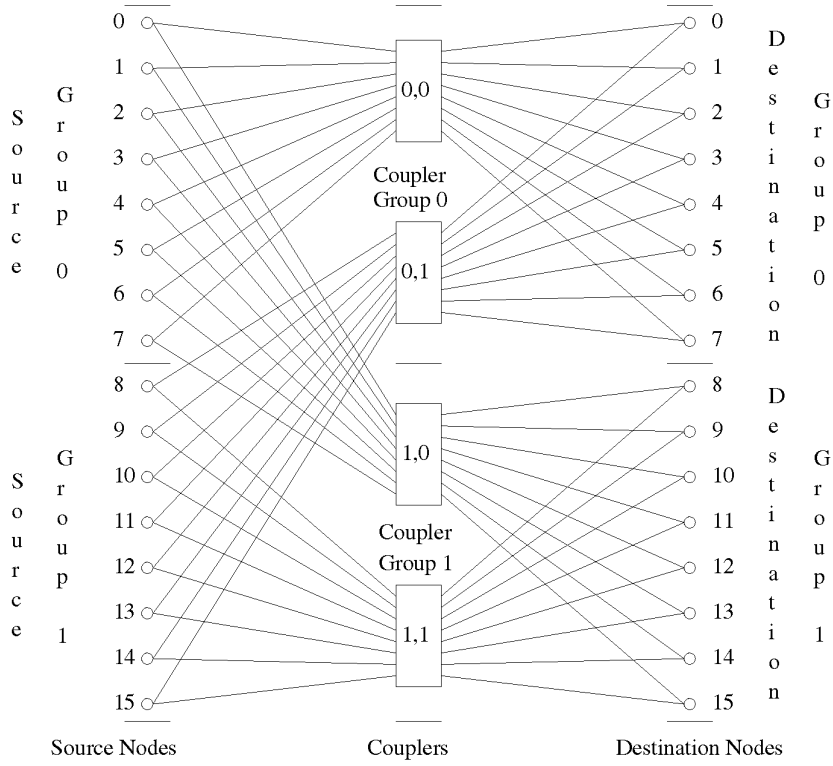


Fig. 1. An $n = 16$, $d = 8$ POPS network ($g = 2$).

communication and global reduction operations are examples of structures that are commonly used in practical applications. Their implementations on different network topologies have been widely studied [9], [11], [15]. Array-type communications are other examples of regular communication patterns that are frequently generated in multiprocessing applications. These applications include image processing, the numeric solution of partial differential equations, and a wide range of scientific modeling problems. Numerous embeddings of rings and meshes into different topologies have been reported in the literature [8], [12], [10], [14], [16]. In this context, an embedding is a mapping of the nodes and edges in a graph representing the communication requirements of an application onto the nodes and links of the physical communication network.

This paper discusses the capability of POPS networks to support application communication structures. After the introduction, the second section describes the components, parameters, and notation for a POPS topology. The third section presents preliminary results about the general capabilities of POPS networks. The fourth section examines all-to-all personalized (complete) communications in a POPS topology. The fifth section considers global reduction operations. The sixth section examines the embeddings of a ring communication structure in a POPS topology. The seventh section considers the embedding of a two-dimensional torus. A conclusion summarizes key points about all these structures.

2 Description of the POPS Topology

A small POPS network is shown in Fig. 1. Source nodes relay messages through optical links to passive optical couplers. These couplers send the messages through other optical links to the destination nodes. A minimal set of two independent parameters completely determines a POPS network implementation. The first parameter, a measure of the system size, is the number of nodes and is denoted by n. The second parameter, a measure of the coupler complexity, is the degree of each coupler and is denoted d . It will be assumed that both n and d are powers of two. Each coupler is a $d \times d$ passive optical star which equally distributes the optical power on any of its d inputs to all of its d outputs.

The set of n source nodes are partitioned into $g = n/d$ equal-sized source groups. Similarly, the set of n destination nodes are partitioned into g destination groups. Each source group and each destination group consists of $d = n/g$ nodes. The g groups of source nodes are called source node-groups and are denoted by SNG_0, ..., SNG_{g-1}, and the g groups of destination nodes are called destination node-groups and are denoted by DNG_0,..., DNG_{g-1}.

A set of $c = g^2 = n^2/d^2$ couplers is partitioned into g groups of g couplers each. The g couplers in the ith coupler group, $0 \le i < g$, are denoted by $C_{i,0}$,..., $C_{i,g-1}$. The d inputs for a given coupler, $C_{i,j}$, are connected to the d nodes in SNG_j, and the d outputs of $C_{i,j}$ are connected to the d nodes in node-group DNG_i .

Each source node, x , $0 \le x < n$, has g transmitters, denoted by $T_{x,0}$, ..., $T_{x,g-1}$, where $T_{x,j}$ is connected to a coupler in the jth coupler group and, thus, is used to communicate with nodes in DNG_j. Similarly, each destination node y has g receivers, denoted by $R_{y,0}$, ..., $R_{y,g-1}$, where $R_{y,j}$ is connected to a coupler which provides communication with

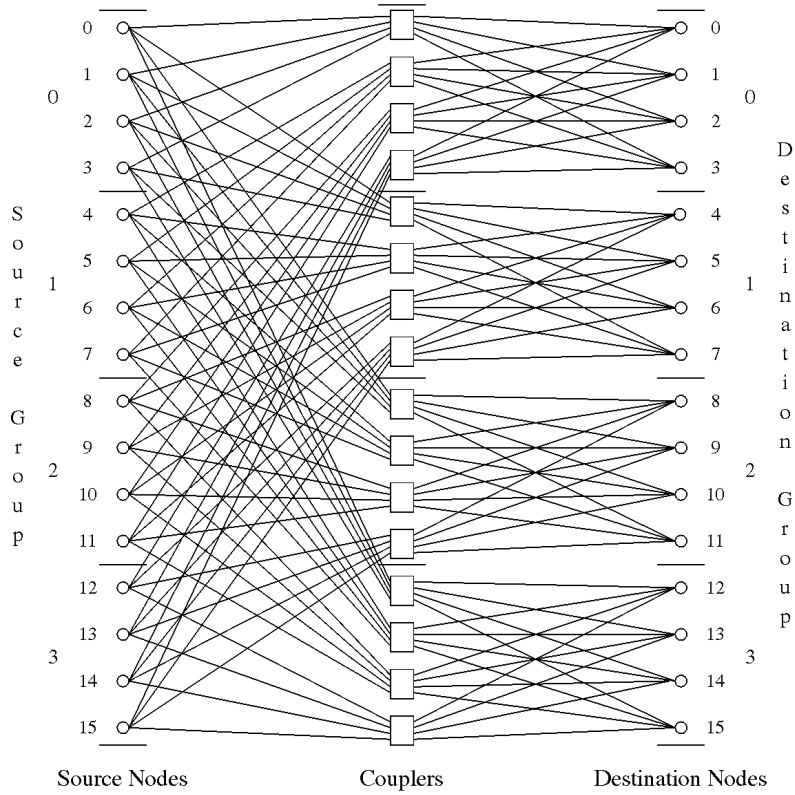


Fig. 2. An $n = 16$, $d = 4$ POPS network ($g = 4$).

nodes in SNG_j. Note that POPS networks are different from electronic switching networks that have similar interconnections. The difference is that POPS use $d \times d$ passive couplers rather than $d \times d$ cross-bar switches. Consequently, at most one input to a given coupler should carry a signal at any one time in order to prevent collisions.

Given a fixed system size, n, the choice of coupler degree, d, allows for a wide range of system characteristics. As the coupler degree approaches n, the system assumes the nature of a single passive star, which includes low system cost and complexity, restricted throughput, and increased power dissipation. As the coupler degree approaches one, the system becomes a completely connected topology, with very high system cost and optimal performance. In this paper, we will assume that $d \geq \sqrt{n}$ since this implies that $c \leq n$. That is, the number of couplers is not larger than the number of nodes. If $d < \sqrt{n}$, then there are more couplers than nodes, and it is not possible to efficiently utilize the communication capabilities of the POPS when each node can send only one message at a time.

The POPS topology shown in Fig. 1 has a fairly large coupler degree ($d = 2\sqrt{n}$). In Fig. 2, another example POPS topology, with the same number of nodes as in Fig. 1 but a smaller coupler degree, is given for comparison.

3 The Group Communication Capabilities of POPS

In a POPS network, the determination of the route for a specified message is simple. Specifically, consider a single message, denoted $M_{x,y}$, which originates at source node x in node-group SNG_ξ and terminates at node y in node-group DNG_η, where $\xi = \lfloor x/d \rfloor$ and $\eta = \lfloor y/d \rfloor$. Source node x uses transmitter $T_{x,\eta}$ to send the message. Coupler $C_{\eta,\xi}$ transports the message, and destination node y uses receiver $R_{y,\xi}$ to get the message. Thus, any specific message has a unique path composed of a transmitter, a coupler, and a receiver, ($T_{x,\eta}$, $C_{\eta,\xi}$, $R_{y,\xi}$). Moreover, for any two given messages, M_{x_i,y_i} and M_{x_j,y_j}, where $x_i \neq x_j$ and $y_i \neq y_j$, the only possible point of conflict in the path of the two messages is the coupler.

Only a single message per coupler can be delivered in a fixed time period, which we call a time slot. Thus, a maximum of c messages can be delivered by a POPS network implementation in a single time slot. Either wave-division multiplexing (WDM) or time-division multiplexing (TDM) may be applied when multiple messages require the same coupler. WDM uses multiple wavelengths, while TDM uses multiple time slots. Thus, a sequence of wavelengths or a sequence of time slots is required. In this paper, we will assume that TDM is used, with the understanding that similar results may be achieved if WDM were to be used. Hence, messages that cannot be delivered in one time slot will be delivered over a sequence of time slots, where, in each time slot, some nonzero number of couplers will deliver one message each.

The merit of an interconnection network can be studied by analyzing its capability for realizing different communication

structures. In general, a communication structure can be defined as a set, $\mathcal{M} = \{M_{x_0,y_0}, \ldots, M_{x_{m-1},y_{m-1}}\}$ of m messages to be delivered. In this paper, we will determine the minimum number of time slots, s, needed to deliver the messages in $\mathcal{M}$. The same results directly apply if M is a set of connections that are to be established in a time-multiplexed manner in the POPS. In this case, s represents the minimum degree of multiplexing which allows all the connections in M to be established in the POPS. Finally, if WDM is to be used, then s would be the minimum number of wavelengths needed to simultaneously establish the connections in M.

In [5], POPS networks delivering random message sets that are permutations or subsets of permutations were shown to be highly efficient. The following sections of this paper examine some regular communication structures. Techniques for realizing all-to-all personalized, global reduction, ring, and torus communication patterns in POPS networks are presented and proven to be optimal. In this context, optimality means 100 percent coupler utilization during every time slot and/or minimum theoretical number of time slots. The following definition and lemmas will simplify subsequent discussions. In the remainder of this paper, we will use $+_u$ and $-_u$ to denote the addition and subtraction *modulo u* operations. That is, $a +_u b = (a + b) \bmod u$ and $a -_u b = (a - b) \bmod u$.

DEFINITION 1. *The* **Group Offset**, *or simply the* **Offset**, *for a message $M_{x,y}$ from source node x in node-group $\xi = \lfloor x/d \rfloor$ to destination node y in node-group $\eta = \lfloor y/d \rfloor$ is $\eta -_g \xi$.*

Any message originating in node-group ξ, with an offset of f, uses $C_{\xi +_g f, \xi}$. That is, the ξth coupler in coupler group $\xi +_g f$. Further, the subset of g couplers that consists of the ξth coupler in each coupler group is required for delivery of g messages originating in node-group ξ when each of the messages have different offsets. Hence, we have the following lemma:

LEMMA 1. *A set of g messages originating at different nodes in the same node-group can be delivered in one time slot if and only if the offsets of the messages are all different.*

Consider g of the sets described in Lemma 1, each originating at a different node-group. Lemma 1 implies that a set of $c = g^2$ messages can be delivered in one time slot if and only if

1) exactly g messages originate in each of the g node-groups, and
2) in every set of g messages originating at the same node-group, each message has a different offset.

The following lemma states these two conditions in a slightly different form.

LEMMA 2. *A set, $\mathcal{M}$,* of $c = g^2$ *messages that originate at different nodes and terminate a different nodes can be delivered in one time slot if and only if:*

1) *for each f, $0 \le f < g$, exactly g messages in $\mathcal{M}$ have offset f,*
2) *the g messages that have a given offset f originate at different node-groups.*

In other words, in a single time slot, each source node-group can communicate with each destination node-group exactly once. This all-to-all personalized (complete) group communication is possible for all POPS topologies. Lemma 2 directly leads to the following more general result:

LEMMA 3. *A set, $\mathcal{M}$, of kg^2 messages can be delivered in k time slots if:*

1) *For each f, $0 \le f < g$, exactly kg messages in $\mathcal{M}$ have offset f,*
2) *Exactly k messages with a given offset f originate at each node-group,*
3) *At most k messages originate at each node and at most k messages terminate at each node.*

Lemma 3 is easily proven by decomposing the set $\mathcal{M}$ into k sets of g^2 messages each, such that each set satisfies the conditions of Lemma 2.

Given a specific set of messages, these messages are delivered in a short sequence of time slots if communication in each time slot emphasizes both intergroup communications and no more than one message between any source/destination node-group pair. High coupler utilizations are achieved when communication in each time slot approaches a complete group communications pattern.

The following property is also a direct result of the ability of POPS to achieve complete group communication in one time slot.

LEMMA 4. *If a set of messages, $\{M_{x,y}\}$ can be delivered in one time slot, then the set of messages, $\{M_{y,x}\}$, which is obtained by interchanging sources and destinations, can also be delivered in one time slot.*

An n-node communication structure can be formally specified by a graph $G = (S, D, E)$, where S and D are sets of n nodes and $E = \{\langle u, v\rangle : u \in S, v \in D\}$ is a set of edges connecting nodes from S to nodes from D. A POPS can realize the structure G in a sequence of s time slots if the set of messages $\mathcal{M} = \{M_{u,v} : \langle u, v\rangle \in E\}$ can be decomposed into s subsets, $\mathcal{M} = \mathcal{M}_1 \cup \ldots \cup \mathcal{M}_s$, where each subset $\mathcal{M}_i$, $i = 1, \ldots, s$, satisfies the conditions of Lemma 2. If the POPS network is used to interconnect the inputs and outputs of n processing elements, then $S = D$, and the structure is specified by $G = (V, E)$, where $V = S = D$. In this case, the source nodes and destination nodes in the POPS are the same and, thus, both a source node-group, SNG_i, and the corresponding destination node-group, DNG_i, will be denoted by NG_i.

Given a POPS network, some communication structures cannot be efficiently implemented without appropriately remapping the nodes in V to the nodes in the networks, that is, without embedding the structure in the network. Formally, an embedding of a structure $G = (V, E)$ onto an n-processor POPS topology is a one-to-one mapping function $\mu : V \to \{0, \ldots, n-1\}$ which maps each node in V onto a node in the POPS network. An optimal embedding is thus defined as an embedding which minimizes the length of the sequence needed to deliver the message set $\mathcal{M} = \{M_{\mu(u),\mu(v)} : \langle u, v\rangle \in E\}$, thus satisfying the requirement of the communication structure in minimal time.

Given that nodes in POPS networks are divided into equal size node-groups, a mapping function μ can be

expressed in terms of two functions, μ_g and μ_n. For any node v in V, $\mu_g(v)$ specifies the node-group to which v is mapped, and $\mu_n(v)$ specifies a particular node within the node-group. That is, $\mu(v) = d * \mu_g(v) + \mu_n(v)$. From Lemma 2, it is clear that the conditions necessary to deliver a set of messages in one time slot do not depend on the actual source nodes and destination nodes of the messages. Rather, they depend on the source node-groups and the destination node-groups, as long as the messages originate and terminate at different nodes. Thus, the sequence length (number of time slots) needed to deliver the messages in an embedded communication structure does not depend on μ_n, as long as μ is a one-to-one function.

The above discussion argues that the sequence length resulting from embedding communication structures onto POPS networks can be derived even if only the group-mapping function, μ_g, is specified. Hence, in the following sections, we will only specify group-mapping functions. However, only functions, μ_g, that map exactly d nodes of V to each node-group will be considered, since such group-mapping functions will lead to one-to-one mapping functions, μ.

4 All-to-All Personalized Communications

A major communications pattern is all-to-all personalized (complete) communications represented by a completely connected graph G. Here, each node sends a unique message to every node in the system. That is, $\mathcal{M} = \{M_{u,v} : u, v = 0, ..., n-1\}$. This complete communication requires a total of n^2 messages. If all couplers are utilized during every time slot required to deliver these messages, the sequence length is the minimum possible and the all-to-all personalized communications is considered optimal.

For a POPS topology, an optimal complete communications requires all c couplers to be used during each time slot needed to deliver the n^2 messages. If $d < \sqrt{n}$, then $n < c$. Thus, since only n messages are possible per time slot (one per node), not all couplers can be utilized simultaneously, and all-to-all personalized communications can not be optimal. However, as we stated earlier, we will assume that $d \geq \sqrt{n}$. This implies that $n \geq c$ and that an optimal complete communications is always possible, as stated in the following theorem.

THEOREM 1. *In any POPS topology where n and d are powers of two and $d \geq \sqrt{n}$, all-to-all personalized communications is optimal, and requires a sequence of length $s = d^2$.*

PROOF. Partition the n nodes into $p = n/c$ subsets, denoted $\mathcal{N}_i$, $0 \leq i < p$, where each subset contains $n/p = c$ nodes evenly distributed among the node groups. Specifically, each $\mathcal{N}_i$ contains $c/g = g$ nodes from each of NG_0, ..., NG_{g-1}. For example, we may define $\mathcal{N}_i = \bigcup_{j=0}^{g-1} \mathcal{N}_i^j$, where $\mathcal{N}_i^j = \{ig + jd + k : 0 \leq k < g\}$ contains g nodes from node-group NG_j. Since $d^2/n \geq 1$ from $d \geq \sqrt{n}$, and both d^2 and n are powers of two, then p is a positive integer.

The all-to-all personalized communication pattern will be accomplished in p^2 phases denoted $\mathcal{P}_r$, $r = 0, \ldots, p^2 - 1$. One source subset $\mathcal{N}_{i_s}$, $0 \leq i_s < p$, completely communicates with one destination subset $\mathcal{N}_{i_d}$, $0 \leq i_d < p$, in a single phase. That is, consecutive time slots in a phase are exclusively used to deliver all messages in a set $\mathcal{M}_{i_s,i_d} = \left\{ M_{u,v} : u \in \bigcup_{j=0}^{g-1} \mathcal{N}_{i_s}^j, v \in \bigcup_{q=0}^{g-1} \mathcal{N}_{i_d}^q \right\}$, which is the set of messages originating from any node in $\mathcal{N}_{i_s}$, and terminating at any node in $\mathcal{N}_{i_d}$. One possible choice for i_s and i_d for phase $\mathcal{P}_r$ is $i_s = \lfloor r/p \rfloor$ and $i_d = r \bmod p$.

Now, $\mathcal{M}_{i_s,i_d}$ contains g^4 messages. Specifically, for each value of j and q, where $0 \leq j < g$ and $0 \leq q < g$, there are g^2 messages in $\mathcal{M}_{i_s,i_d}$ originating at nodes in $\mathcal{N}_{i_s}^j$ and terminating at nodes in $\mathcal{N}_{i_d}^q$. These g^2 messages have offset $q -_g j$. Moreover, for a given f between 0 and $g-1$, there are exactly g possible pairs of values for q and j such that $q -_g j = f$. Hence, the first two conditions of Lemma 3 are satisfied for $\mathcal{M}_{i_s,i_d}$ with $k = g^2$. The third condition of Lemma 3 is also satisfied since $\mathcal{M}_{i_s,i_d}$ contains exactly g^2 messages originating at each node in $\bigcup_{j=0}^{g-1} \mathcal{N}_{i_s}^j$. Hence, the set $\mathcal{M}_{i_s,i_d}$ can be delivered in $k = g^2$ time slots.

Thus, all-to-all communications can be accomplished in $(n/c)^2$ phases, each requiring $g^2 = c$ time slots, for a total of n^2/c time slots. This is optimal for the delivery of n^2 messages using c couplers. □

Fig. 3 illustrates the all-to–all communications in the POPS given in Fig. 1 for which $g = 2$ and $c = 4$. Messages are drawn as arrows from sources to destinations. Out of the $(n/c)^2 = 16$ phases needed for the all-to-all communications, only phases $\mathcal{P}_0$, $\mathcal{P}_2$, and $\mathcal{P}_{15}$ are shown. Each phase requires four time slots, and the messages in each phase satisfy the conditions of Lemma 3, while the four messages within each time slot satisfy the conditions of Lemma 2. The sets $\mathcal{N}_i^j$, $i = 0, ..., 3$, $j = 0, 1$, are indicated in the figure where

$$\mathcal{N}_0 = \mathcal{N}_0^0 \cup \mathcal{N}_0^1 = \{0, 1\} \cup \{8, 9\},$$
$$\mathcal{N}_1 = \mathcal{N}_1^0 \cup \mathcal{N}_1^1 = \{2, 3\} \cup \{10, 11\},$$
$$\mathcal{N}_2 = \mathcal{N}_2^0 \cup \mathcal{N}_2^1 = \{4, 5\} \cup \{12, 13\},$$
$$\mathcal{N}_3 = \mathcal{N}_3^0 \cup \mathcal{N}_3^1 = \{6, 7\} \cup \{14, 15\}.$$

5 Global Reduction (Tree-Oriented) Communications

Many applications require global data reductions. These operations generate a single data value from data spread across all system nodes. In systems where each node can transmit and receive only one message at a time, a global reduction algorithm requires a minimum of log n phases,

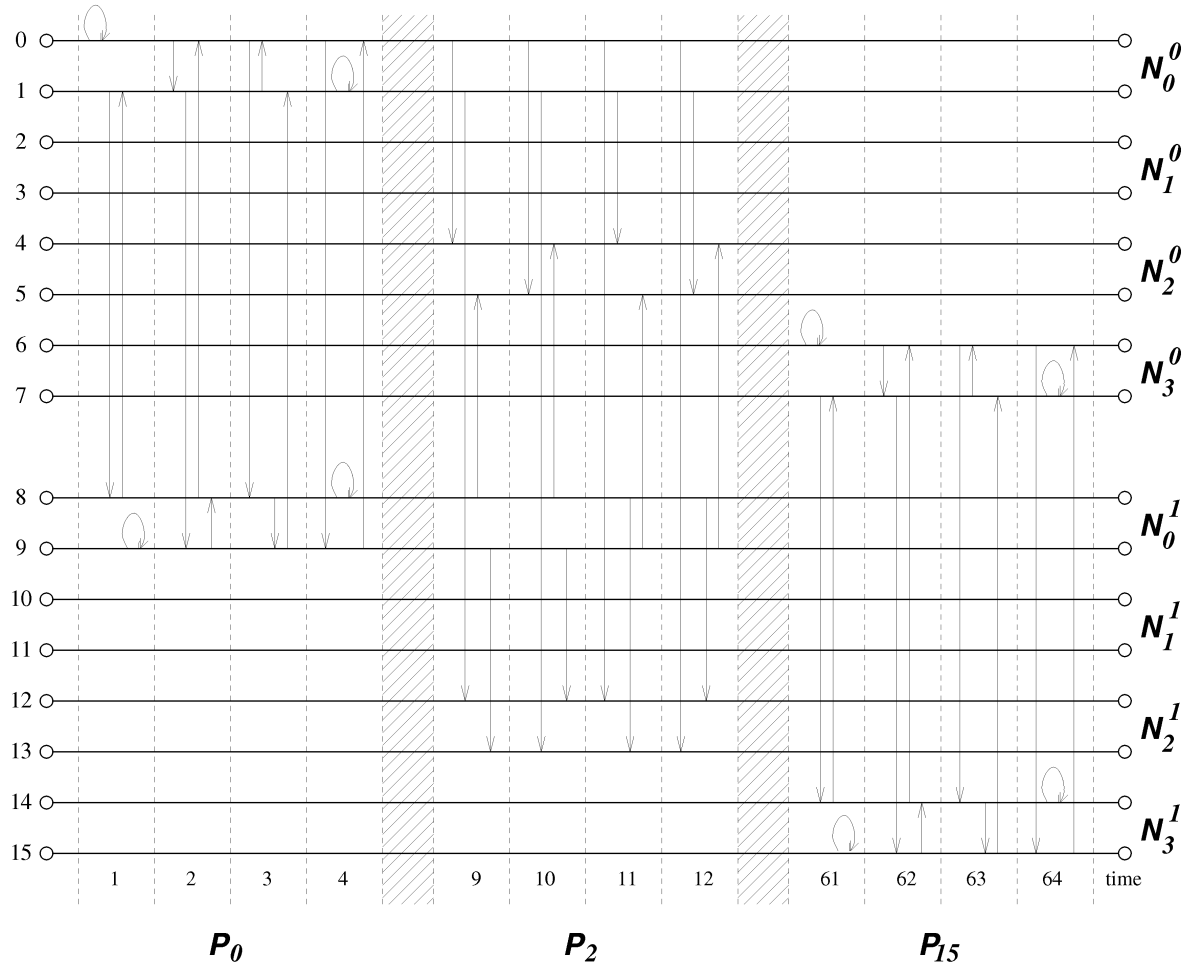


Fig. 3. Complete communications in a POPS with $n = 16$ and $d = 8$ ($g = 2$).

where the logarithm is base 2. When all phases are completed, a single node (typically, node 0) in the system has the result of the reduction.

The simplest implementation in a POPS topology of a global reduction operation, called a natural global reduction, is as follows: During phase $\mathcal{P}_i$ of the reduction algorithm, $1 \le i \le \log n$, the set of messages is given by

$$\left\{ M_{k+2^{i-1},k} : k = j2^i,\ 0 \le j < 2^{(\log n)-i} \right\}.$$

Fig. 4a shows the messages required in the five phases of the global reduction operation for a network with $n = 32$. For example, in the first phase, $\mathcal{P}_1$, the messages in set $\{M_{1,0}, M_{3,2}, M_{5,4}, ..., M_{29,28}, M_{31,30}\}$ are delivered. $\mathcal{P}_2$ has the message set $\{M_{2,0}, M_{6,4}, M_{10,8}, ..., M_{26,24}, M_{30,28}\}$, and the last phase $\mathcal{P}_5$ has the message set $\{M_{16,0}\}$.

THEOREM 2. *In any POPS topology, a natural global reduction requires a sequence of length* $s = (d - 1) + (\log g) \ge \log n$.

PROOF. For the natural global reduction described above, each of the first $\log d$ phases requires the delivery of messages for which the destination node-group is the same as its source node-group. Specifically, in each group, there are d data values to be reduced to one, with only a single message possible per time slot. Thus, the $\log d$ phases require

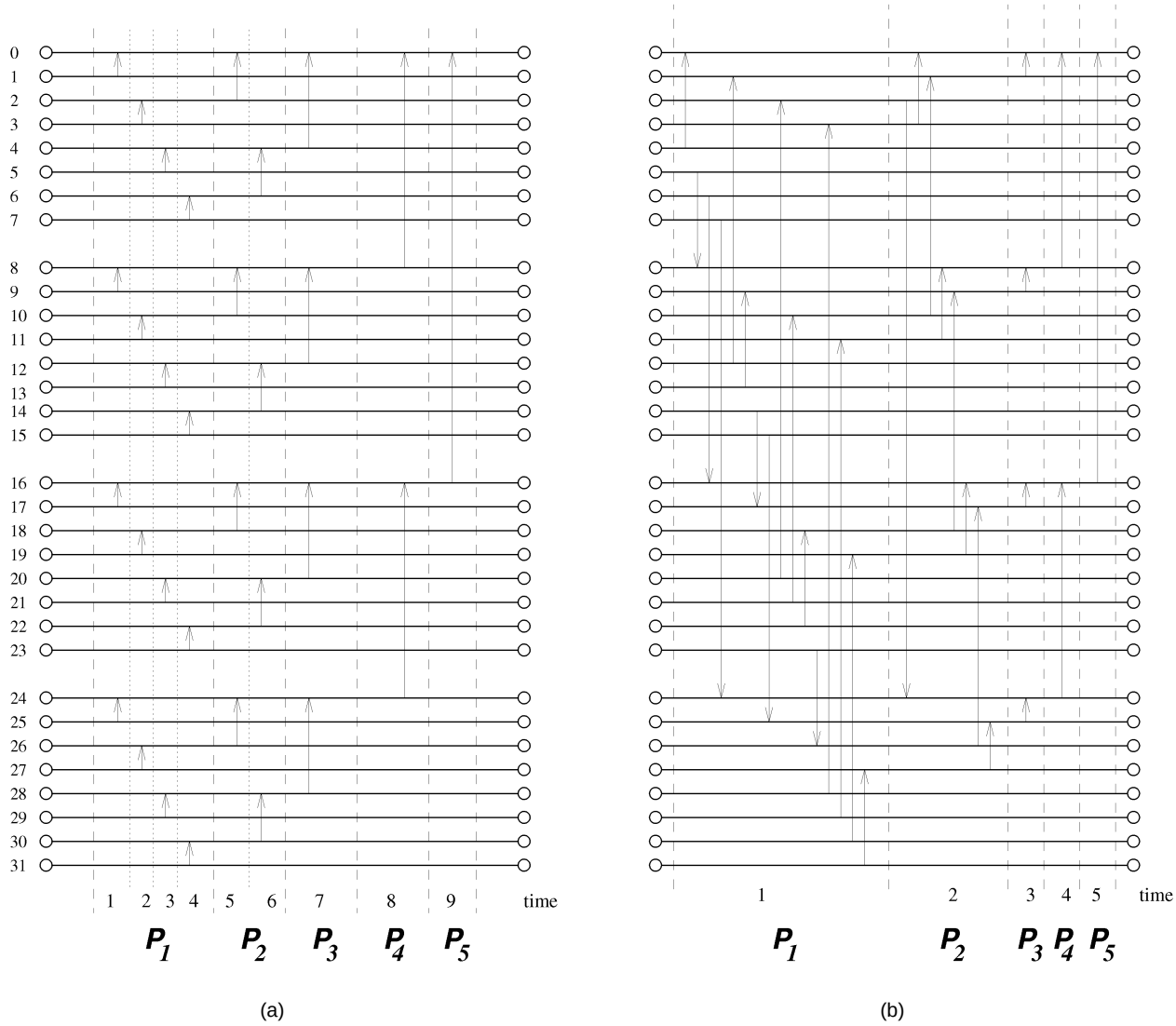
$$d/2 + d/4 + \ldots + d/d = \sum_{i=1}^{\log d} d/2^i = d - 1$$

time slots.

During the phases $\log d + 1$, ..., $\log n$, each phase requires communication where each message's destination node-group is different than its source node group. Hence, each of these phases can be completed in one time slot, and the entire natural global reduction takes $(d - 1) + (\log n - \log d) = (d - 1) + (\log g)$ time slots. □

As is clear from Theorem 2, the natural global reduction operation is not optimal because each of the first $\log d - 1$ phases require more than one time slot each, while some of the couplers are not utilized. If the POPS has at least $n/2$ couplers, then there are enough couplers to deliver the $n/2$ messages of the first phase in one time slot. Given that the number of messages delivered in subsequent phases is less than $n/2$, then, for a POPS with at least $n/2$ couplers ($d \le \sqrt{2n}$), an optimal global reduction should be completed in $\log n$ time slots. This is achievable, as described next.

THEOREM 3. *In any POPS topology where* $d \le \sqrt{2n}$, *global reduction is possible in a sequence of length* $s = \log n$.



(a) (b)

Fig. 4. Global reduction communications in a POPS with $n = 32$ and $d = 8$. (a) Natural reduction. (b) Optimal reduction.

PROOF. Let $d = \sqrt{2n/b}$, for some integer $b > 0$. Thus, each of the g groups in the POPS contains $d = 2g/b$ nodes. That is, they contain at most $2g$ nodes. Partition the nodes in each node-group NG_j, $0 \le j < g$, into two partitions. The first partition,

$$\mathcal{N}_j^{1,R} = \{dj + t : t = 0, \ldots, (d/2) - 1\}$$

contains the first $d/2$ nodes, and the second partition,

$$\mathcal{N}_j^{1,T} = \{dj + (d/2) + t : t = 0, \ldots, (d/2) - 1\}$$

contains the last $d/2$ nodes. During the first phase, $\mathcal{P}_1$, nodes in $\mathcal{N}_j^{1,T}$ transmit messages and nodes in $\mathcal{N}_j^{1,R}$ receive messages. However, in order to distribute the communication load on all the couplers, each of the $d/2$ messages originating from $\mathcal{N}_j^{1,T}$ is sent to a different $\mathcal{N}_j^{1,R}$, $i = 0, ..., g - 1$. This can be easily accomplished if node $dj + (d/2) + t$ sends a message to node $d(j +_g t) + t$, for $j = 0, ..., g - 1$ and $t = 0, ..., (d/2) - 1$. This type of communication pattern has $gd/2 = n/2 \le g^2$ messages which all involve a different source/destination node-group pair. These $n/2$ messages form a subset (proper subset only if $b > 1$) of the set of g^2 messages described in Lemma 2, thus ensuring that $\mathcal{P}_1$ requires one time slot.

After $\mathcal{P}_1$ is complete, all nodes in $\mathcal{N}_j^{1,T}$, $0 \le j < g$, become inactive during succeeding phases, while all nodes in $\mathcal{N}_j^{1,R}$, remain active. Each phase, $\mathcal{P}_i$, $1 < i \le \log d$, follows a process similar to $\mathcal{P}_1$. At the beginning of the phase, $\mathcal{N}_j^{i-1,R}$ is partitioned into two equal-sized partitions,

$$\mathcal{N}_j^{1,R} = \{dj + t : t = 0, \ldots, (d/2i) - 1\}$$

of receiving nodes, and

$$\mathcal{N}_j^{1,T} = \{dj + (d/2i) + t : t = 0, \ldots, (d/2i) - 1\}$$

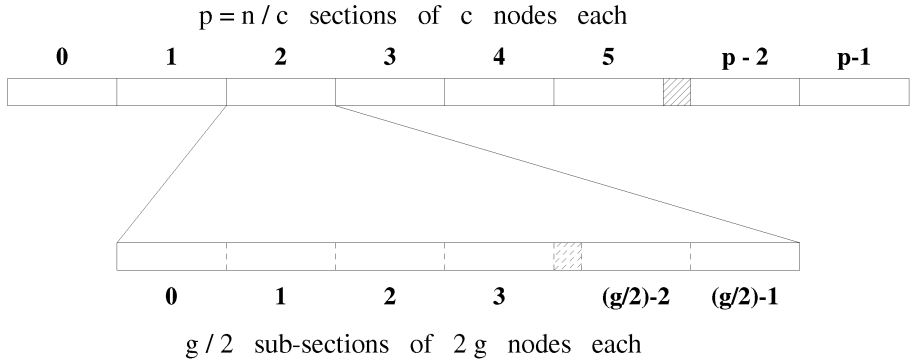


Fig. 5. Overview of the alternating-pair embedding for a ring.

of sending nodes. Each node $dj + (d/2i) + t$ sends a message to node $d(j +_g t) + t$, for $j = 0, ..., g - 1$, $t = 0, ..., d/2i$. The resulting message set is always a subset of the one specified in Lemma 2 and, thus, communication in $\mathcal{P}_i$ can be accomplished in one time slot.

After $\log d$ phases, there is a single active node left in each node-group. Similar to the natural embedding above, these g nodes may be reduced to one in $\log g$ time slots. Thus, the entire sequence requires $\log d + \log g = \log n$ time slots. □

Fig. 4b illustrates a minimal global reduction, as defined above, in a POPS topology with $n = 32$, $d = 8$ ($g = 4$). The reduction from the nine time slots needed for the natural reduction of Fig. 4a to the five time slots for the optimal embedding of Fig. 4b is obvious.

If $d > \sqrt{2n}$, then $c = g^2 < n/2$, and the number of couplers is smaller than the number of messages that are to be exchanged in phase 1. Specifically, if $d = \sqrt{2bn}$, where $b > 1$, then $d = 2bg$ and $gd/2 = bg^2$ messages are to be exchanged in Phase 1 of the global reduction. By using the same partitioning described in the proof of Theorem 3, we obtain a message set which satisfies the conditions of Lemma 3 with $k = b$. That is, Phase 1 of the global reduction can be accomplished in b time slots. In general, phase i for $i < \log b$ will require b/i time slots, which means that the first $b - 1$ phases require a total of $2(b - 1)$ time slots. Each of the remaining phases will require one time slot and, thus, the total global reduction can be accomplished in $\log n + 2(b - 1) - \log b$ time slots.

6 Ring Embeddings in a POPS Topology

In a ring communication structure, each node in the ring sends a unique message to the succeeding node. That is, in an n-node ring, node k sends a message to node $k +_n 1$, where $+_n$ is the addition operation modulo n. Formally, a ring communication structure is given by (V, E), where $V = \{0, \ldots, n - 1\}$, and $E = \{<u, u +_n 1> : u \in V\}$. We will consider only unidirectional rings in this section, since Lemma 4 provides a means for extending the results to bidirectional rings in which a node, k, sends messages to both node $k +_n 1$ and node $k -_n 1$.

An obvious embedding of a ring, which we shall call the natural embedding, is given by $\mu(k) = k$. With that embedding, the set of messages for the ring communication is given by $\mathcal{M} = \{M_{k,k+_n 1} : k = 0, \ldots, n - 1\}$. The following lemma shows that the natural embedding of the ring onto a POPS network is not optimal.

Lemma 6. *Using the natural embedding of ring nodes onto a POPS network, the ring communication requires a sequence of length* $s = d - 1$.

Proof. For the natural embedding described above, each POPS topology node-group contains d consecutive ring elements. During each iteration of the ring communication, d messages originating from nodes in a given node-group are sent to the next node in the ring. Of these d messages, all but one are sent to nodes within the given node-group. The $d - 1$ messages from the given node-group to the same node-group all use the same coupler. At most one message, however, is sent from any specific node-group to a different node-group. Thus, the natural embedding of the ring communication structure requires a sequence length of $d - 1$. □

Ideally, a sequence length equal to $max\{1, n/c\}$ should be enough to deliver the n messages in the ring communication structure, where $c = g^2$ is the number of couplers in the POPS. Given that $d = n/g$, it follows that $d - 1 > max\{1, n/c\}$ for $d > 2$. Thus, the natural embedding does not lead to a minimal sequence length. Another embedding, which we shall call the alternating-pair embedding, will be shown to be optimal.

The idea of the alternating-pair embedding is as follows: First, partition the n ring elements into $p = n/c$ subsets, called sections, and denoted $\mathcal{N}_i$, $0 \le i < p$. Note that p is a positive integer if $d \ge \sqrt{n}$ and both d^2 and n are powers of two. Each section, $\mathcal{N}_i$, contains $n/p = c$ consecutive ring elements. The assignment algorithm given below, assigns the c ring elements in each section $\mathcal{N}_i$ to node-groups in the POPS network. This algorithm yields an identical sequence of group assignments for each section in the ring.

Second, within a given single ring section $\mathcal{N}_i$, partition the c ring elements into $g/2$ subsets, called subsections, and denoted $\mathcal{N}_i^J$, $0 \le J < g/2$. Since g is a power of two, $g/2$ is a

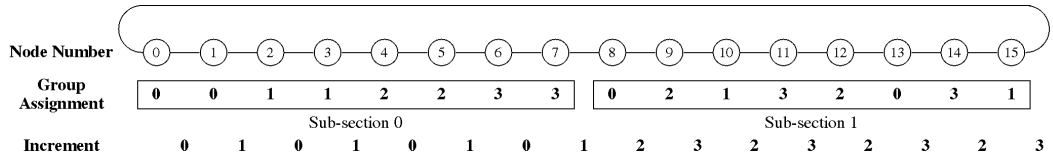


Fig. 6. Node group assignments for a ring embedding ($n = 16$, $d = 4$).

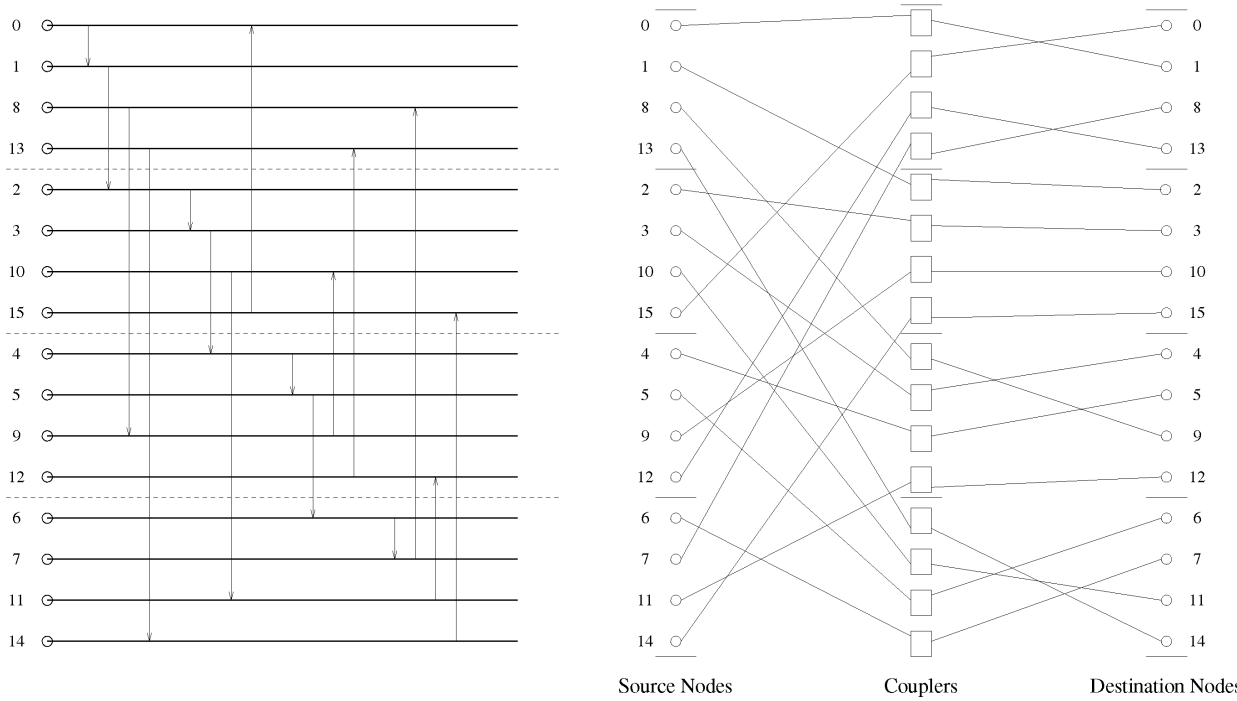


Fig. 7. A ring embedding for an $n = 16$, $d = 4$ POPS network.

positive integer. Each subsection, $\mathcal{N}_i^J$, contains $c/(g/2) = 2g$ consecutive ring elements. Fig. 5 shows the partitioning of the ring elements into sections and subsections.

Third, within a given single subsection of a given single ring section, the algorithm assigns the first element in the subsection to node-group 0 , then increments of $2J$ and $2J + 1$ are alternately (and cumulatively) added to generate successive node-group assignments. Following is the assignment algorithm.

node = 0 ;
  for section = 0,. .., $n/c - 1$ do
    for $J = 0, ..., g/2 - 1$ do /* map nodes in subsection J*/
      μ_g (*node*) = 0; /* assign the first node in subsection J to node-group 0 */
      node = *node* + 1;
      for $i = 1, ..., 2g - 1$ do
        if i is odd
          then $\mu_g(node) = \mu_g(node - 1) +_g 2J$
          else $\mu_g(node) = \mu_g(node - 1) +_g 2J +_g 1$;
        node = *node* + 1;

A different way of presenting the above algorithm is to assign the first node in the ring to node-group zero ($\mu_g(0) = 0$) and, then, specify the offsets $\phi(k) = \mu_g(k +_n 1) -_g \mu_g(k)$ for the nodes $k = 0, \ldots, n - 1$. The following algorithm determines the offsets.

node = 0;
for section = 0, ..., $n/c - 1$ do
  for $J = 0, ..., g/2 - 1$ do /* consider nodes in subsection J*/
    for $i = 0, ..., 2g - 1$ do
      if i is odd
        then $\phi(node) = 2J +_g 1$
        else $\phi(node) = 2J$;
      node = *node* + 1;

Note that if u and v are the first nodes in two consecutive subsections, then $v - u = 2g$ and, thus, $\mu_g(v)$ is equal to $\mu_g(u)$ plus the sum of all the $2g$ offsets $\phi(i)$, $i = u, \ldots, v - 1$. Out of these offsets, g are equal to J and g are equal to $2J + 1$ and, thus, their sum (modulo g) is zero. Hence, $\mu_g(v) = \mu_g(u) = 0$, since $\mu_g(0) = 0$. This proves that the offsets calculated by the above algorithm leads to the node-group assignment given in the previous algorithm.

In Fig. 6, we show the application of the algorithm to the embedding of a 16-node ring into the POPS shown in Fig. 2. In this case, $g = 4$ and $n = c = 16$. Thus, there is only one section, which is composed of two subsections. The first node in each subsection is assigned to node-group 0, and the group assignments of subsequent nodes in subsection 0 are obtained by incrementing the group assignments by the offsets 0 and 1 alternatively. The group assignments of the nodes in subsection 1 are obtained by incrementing the group assignments by the offsets 2 and 3. The assignment of nodes within groups is done arbitrarily. From Fig. 7, it is

clear that all 16 messages from node $\mu(u)$ to node $\mu(u +_{16} 1)$, $u = 0, \ldots, 15$, can be delivered in one time slot without conflict. The following proves the optimality of the alternating-pair embedding of rings onto POPS networks.

THEOREM 4. *In any POPS topology where n and d are powers of two, if $d \geq \sqrt{n}$, then the alternating-pair embedding of a ring is a one-to-one embedding which requires a sequence of length $s = n/c = d^2/n$ to accomplish ring communication. This sequence length is optimal.*

PROOF. We will prove that the set of messages originating from the c nodes within any single ring section satisfies the conditions of Lemma 2.

First, the offsets used in the alternating-pair embedding assignment algorithm are the offsets for the c messages in the section. Each of the g possible offsets is used in only one of the $g/2$ subsections in the section. Within the subsection where a given offset is used, it alternates with one other offset, thus occurring g times and satisfying the first hypothesis of Lemma 2.

Second, within the subsection J where a particular offset is used, this offset occurs in every other message. Thus, the group assignment for the sources of the messages with the given offset increases by units of $(2J) + (2J + 1) = 4J + 1$. Since $4J + 1$ is odd, and the number of possible group assignments g is a power of two, they are relatively prime. Consequently, exactly one of the g messages that has the given offset originates from each node-group. This satisfies the second hypothesis of Lemma 2, thus showing that the alternating-pair embedding leads to a one-to-one mapping of nodes $\{0, \ldots, n-1\}$ onto a POPS.

Lemma 2 now implies that the communication generated by a single section can be achieved in a single time slot. Thus, the p sections in the entire ring require $p = n/c$ time slots. This is the minimum theoretical sequence length to deliver n messages with c couplers, and implies a 100 percent coupler utilization. □

A modified version of the proof of Theorem 4 may be applied to prove that the alternating-pair embedding is optimal even if $d \leq \sqrt{n}$. In this case, $n < c$ and there is only one section. If we partition this section into $n/2g$ subsections of $2g$ nodes each, we may apply the same embedding algorithm and show that the n messages in the ring structure can be delivered in one time slot.

7 TORUS EMBEDDINGS IN A POPS TOPOLOGY

Another particularly important regular communication structure is the two-dimensional torus. Here, each torus element sends four unique messages, one to each of its immediate neighbors (up, down, left, and right). We will consider only unidirectional tori since Lemma 4 can extend our result to bidirectional tori. Specifically, we will consider the communication structure $(V, E_h \cup E_v)$, where $V = \{0, \ldots, n-1\}$ and

$$E_h = \left\{ < u, u+1 > : u \in V \ \& \ u +_{\sqrt{n}} 1 \neq 0 \right\} \cup \left\{ < u, u - \sqrt{n} + 1 > : u \in V \ \& \ u +_{\sqrt{n}} 1 = 0 \right\}$$

represents wrapped around horizontal communications, and $E_v = \left\{ < u, v +_n \sqrt{n} > : u \in V \right\}$ represents wrapped around vertical communications. Given an embedding, μ_g, of the nodes in V onto a POPS, let $\mathcal{M}_h$ and $\mathcal{M}_v$ be the set of messages corresponding to E_h and E_v, respectively.

As for the ring communication structure, obvious embeddings exist for the two-dimensional torus in a POPS topology. Also, as for the ring, these simple embeddings for the torus can be shown to have relatively high sequence lengths and poor coupler utilization. For example, in Fig. 8, we consider the 4×4 torus communication in the 16-node POPS shown in Fig. 1, in which $d = 8$. In Fig. 8a, where the natural embedding is used, it is clear that row communication requires eight time slots while column communication can be done efficiently in four time slots. In Fig. 8b, where the alternating-pair embedding is used, it is clear that row communication can be done efficiently in four time slots, while the column communication requires eight time slots. Neither embedding is optimal since, in an optimal embedding, both row and column communication should be accomplished in four time slots each. Such an optimal embedding is shown in Fig. 8c.

The alternating-pair embedding, introduced for the ring communication structure and shown to be optimal there, can be modified for embedding the two-dimensional torus in a POPS topology. An additional restriction, namely $d \geq 2\sqrt{n}$, is required in the proof that the modified alternating-pair torus embedding is optimal.

The modified alternating-pair embedding is as follows: First, partition the n torus elements into $p = n/c$ subsets, called sections, and denoted $\mathcal{N}_i$, $0 \leq i < p$. Each section, $\mathcal{N}_i$, contains c consecutive torus elements, or, stated differently, $c/\sqrt{n}$ consecutive torus rows. The assignment algorithm given below, assigns the c torus elements in each section $\mathcal{N}_i$, $0 \leq i < p$, to node-groups in the POPS network. This algorithm yields an identical sequence of group assignments for each section in the torus.

Second, within a given single torus section $\mathcal{N}_i$, partition the c torus elements into $g/2$ subsets, called subsections, and denoted $\mathcal{N}_i^J$, $0 \leq J < g/2$. Since g is a power of two, $g/2$ is a positive integer. Each subsection, $\mathcal{N}_i^J$, contains $c/(g/2) = 2g$ consecutive torus elements. It is a requirement that at least one complete subsection fits within each row of the torus, thus, $2g \leq \sqrt{n}$ and $d \geq 2\sqrt{n}$. Fig. 9 shows the partitioning of the torus elements into sections and subsections.

Third, the algorithm which assigns torus elements to POPS network node-groups is similar to that for the ring embedding, but with an additional final step. In essence, as is the case for the ring, the first element in any subsection is assigned to node-group 0, then increments of $2J$ and $2J +_g 1$ are alternately (and cumulatively) added to generate successive node-group assignments. Unlike the ring case, however, a final step is added in which assignments in row r are

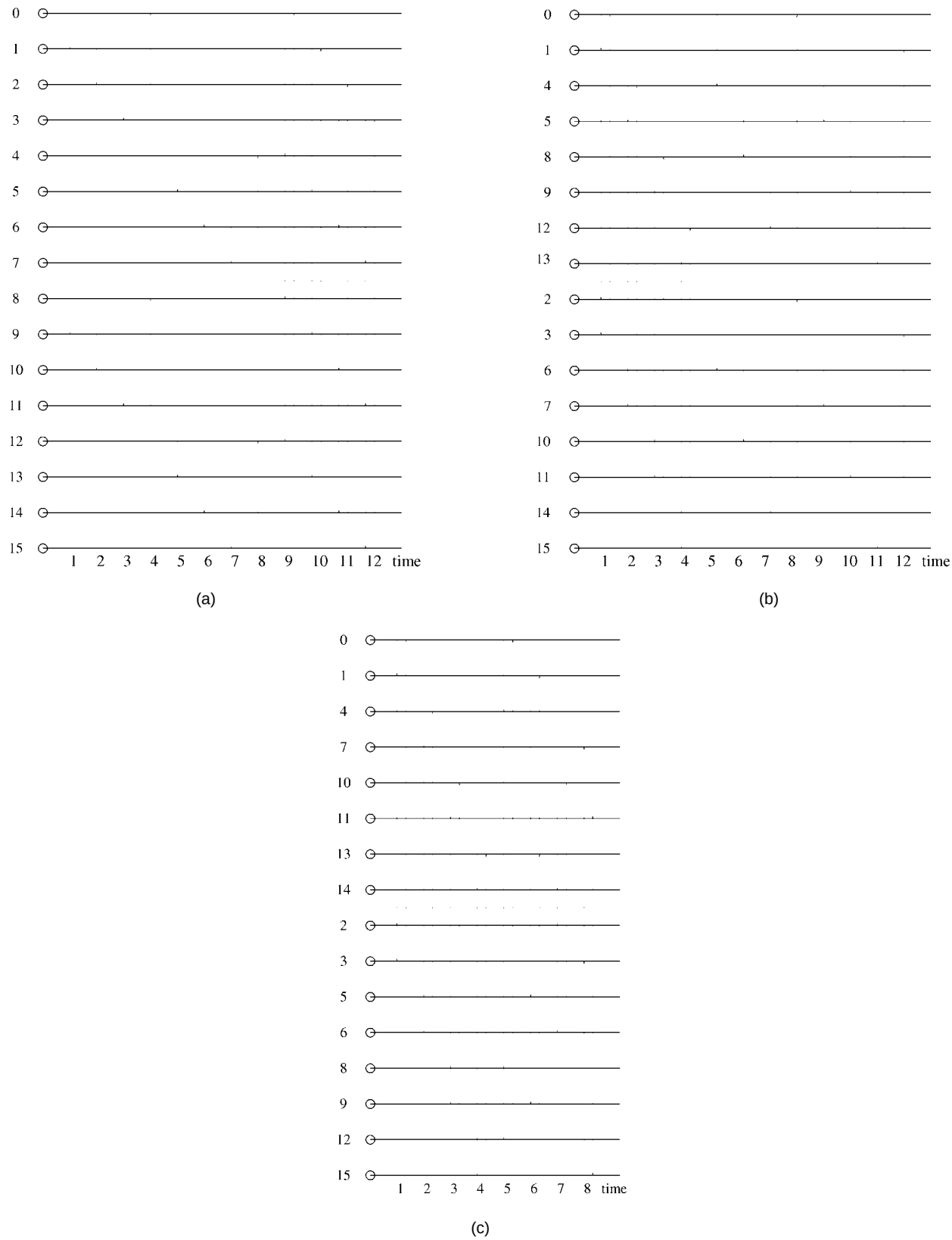


Fig. 8. Torus communication in POPS with $n = 16$ and $d = 8$. (a) Natural embedding. (b) Alternating pair embedding. (c) Optimal embedding.

rotated r positions to the left. Following is the algorithm that assigns node-groups to the elements of a section.

node = 0;
for $J = 0, ..., g/2 - 1$ do /* map nodes in subsection J */
τ(*node*) = 0; /* assign the first node in subsection J to node-group 0 */
node = *node* + 1;
for $i = 1, ..., 2g - 1$ do

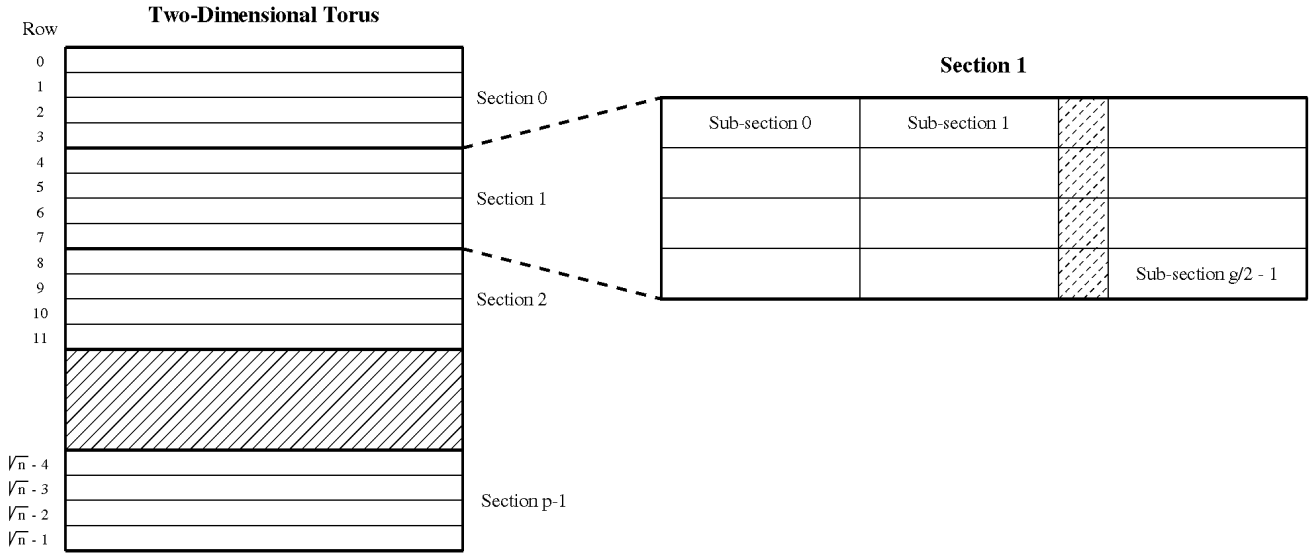


Fig. 9. Overview of the alternating-pair embedding for a torus.

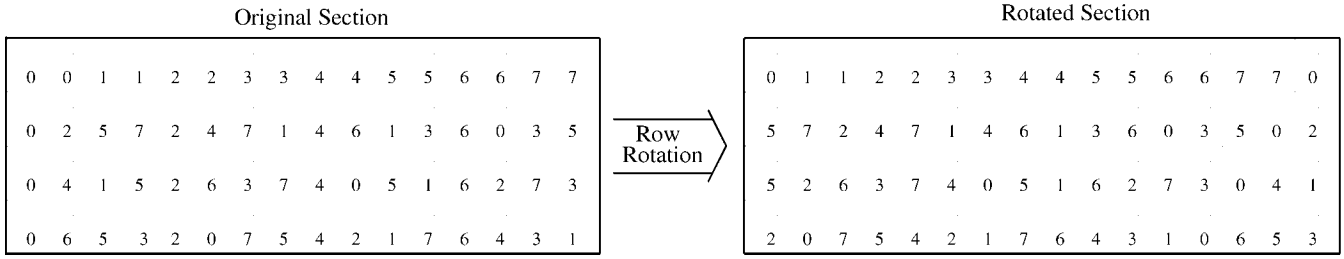


Fig. 10. Node group assignments for a torus embedding ($n = 256$, $d = 32$).

if i is odd
then $\tau(node) = \tau(node - 1) +_g 2J$
else $\tau(node) = \tau(node - 1) +_g 2J +_g 1$;
$node = node + 1$;
for $r = 0, \ldots, \sqrt{n} - 1$ do /* rotate the assignments of rows $r = 0, \ldots, \sqrt{n} - 1$ */
for $m = 0, \ldots, \sqrt{n} - 1$ do
If $(m - r < 0)$ then $\mu_g\left(\sqrt{n}r + m - r\right) = \tau\left(\sqrt{n}r + m\right)$
else $\mu_g\left(\sqrt{n}(r+1) + m - r\right) = \tau\left(\sqrt{n}r + m\right)$;

For example, consider the embedding of a 256-torus into a POPS with $d = 32$. In this case, $g = 8$, $c = 64$ and, thus, there are four sections, each containing four subsections of 16 nodes each. Here, a subsection spans an entire row. Fig. 10 shows the group assignment for the nodes in any of the four subsections, before and after the row rotation step.

A simpler example is the one for which $n = 16$ and $d = 8$. In this case, there are four sections, each containing a single subsection of four elements. That is, each row of the torus is a subsection and, thus, in each row, the four nodes are assigned to groups 0, 0, 1, and 1, respectively. The assignment in each row, r, $0 \le r < 4$, is thus rotated to the left by r positions, yielding the group assignment 0, 0, 1, 1 for the first row, 0, 1, 1, 0 for the second row, 1, 1, 0, 0 for the third row, and 1, 0, 0, 1 for the last row. The node assignment shown in Fig. 8c follows the above group assignment.

Each node, $\mu_g(u)$, in the POPS contributes one message to the set $\mathcal{M}_h$ of horizontal messages and another message to the set $\mathcal{M}_v$ of vertical messages. Let $\phi_h(u)$ be the offset of the former message and $\phi_v(u)$ be the offset of the latter. As was done for the ring case, an algorithm which specifies the offsets $\tau(u +_n 1) -_g \tau(u)$ for each node may be given. Noting that τ assigns group 0 to the first node in each row, we conclude that $\tau\left(u +_n 1\right) -_g \tau(u) = \tau\left(u - \sqrt{n}\right) -_g \tau(u)$ if u is the last node in a row. Hence, the set of horizontal offsets $\{\phi_h(u) : u = 0, \ldots, n - 1\}$ obtained after rotating the assignments of row r is equal to the set of offsets, $\{\tau(u +_n 1) -_g \tau(u) : u = 0, \ldots, n - 1\}$, obtained before the rotation. In Theorem 1, it was shown that such a set can be delivered in n/c time slots. This proves that the set of horizontal messages in the torus communication structure, $\mathcal{M}_h$, can also be delivered in n/c time slots. In the next theorem, we will show that, if $d \ge 2\sqrt{n}$, the set $\mathcal{M}_v$ can also be delivered in n/c time slots, thus proving the optimality of the alternating-pair embedding for the torus communication structure.

THEOREM 5. *In any POPS topology where n and d are powers of two, if $d \ge 2\sqrt{n}$, then the alternating-pair embedding of a torus communication structure requires a sequence of length $s = 2n/c = 2d^2/n$, which is optimum.*

PROOF. Appropriately, the proof is similar to, though more intricate than, the proof of optimality in the ring embedding. Consider a single torus section. For the c horizontal messages in the section, we have shown that these messages are identical to those in a single section of a ring with the same n and d and, thus, can be achieved in a POPS network in a single time slot. We will prove that the c vertical messages also satisfy the hypothesis of Lemma 2. In order to simplify the discussion, we will define the *rs-section*, $\overline{\mathcal{N}}_i^J$ to be a version of $\mathcal{N}_i^J$ which is rotated to match the row rotation in the alternating-ring embedding. In other words, if $\mathcal{N}_i^J$ contains the nodes $u, \ldots, u + 2g - 1$, in row r, then $\overline{\mathcal{N}}_i^J$ contains the nodes $u - r, \ldots, u + 2g - 1 - r$, where the left rotation wraps around nodes in the same row. In the rest of this proof, we will refer to horizontal offsets by *h-offsets* and to vertical offsets by *v-offsets*. We will also drop the subscripts from $+_g$ and $-_g$, with the understanding that all additions and subtractions on offsets are modulo g.

First, it will be shown that the first hypothesis of Lemma 2 is true, that is, each possible v-offset is used g times. The v-offsets for these messages are determined by the group assignments of the source nodes in one rs-section and the destination nodes in another rs-section. These assignments are, in turn, determined by alternating pairs of different h-offsets. For example, the h-offsets in rs-section J are $2J$ and $2J + 1$. Since there are $\sqrt{n}/2g$ rs-sections in each torus row, the h-offsets for the rs-section $J + \sqrt{n}/2g$, which is immediately below rs-section J, are $2J + \sqrt{n}/g$ and $2J + \left(\sqrt{n}/g\right) + 1$. These offsets are depicted in Fig. 11.

These facts, combined with the effect of the row rotation in the assignment algorithm, cause the first v-offset for the Jth rs-section to be $2J + \sqrt{n}/g$, where $0 \le J < g/2$. Further, succeeding offsets within the rs-section alternately (and cumulatively) increase by $\left(\sqrt{n}/g\right) + 1$ and $\left(\sqrt{n}/g\right) - 1$. This pattern of increase in the v-offsets is independent of which rs-section in the section is involved.

Since

$$\left(\sqrt{n}/g\right) + 1 + \left(\sqrt{n}/g\right) - 1 = 2\sqrt{n}/g$$

and

$$\left(2\sqrt{n}/g\right)\left(g^2/2\sqrt{n}\right) = g,$$

the v-offsets repeat after $g^2/2\sqrt{n}$ increments. Thus, the $2g$ v-offsets in a rs-section are divided into $2\sqrt{n}/g$ sets of $g^2/2\sqrt{n}$ v-offsets each. These sets of v-offsets, which we shall call cycles, are identical within an rs-section.

The set of all g^2 v-offsets in a section is, thus, divided into $\sqrt{n}/g$ cycles of $g^2/2\sqrt{n}$ v-offsets each. These cycles can be shown to be disjoint or identical. If the same v-offset occurs both at even or both at odd locations in different cycles (counting from 0 at the cycle beginning), then the same sequence of increments (starting with either $\left(\sqrt{n}/g\right) + 1$ or $\left(\sqrt{n}/g\right) - 1$, respectively) will be applied. These two cycles will consequently be identical. If the same v-offset occurs at an even location in one cycle and an odd location in another cycle, then, since both increments are odd, the first v-offset in one of the cycles is odd. This is a contradiction, since the first v-offset, as previously shown, is always even. Thus, the $\sqrt{n}/g$ cycles are disjoint or identical.

Each of the different cycles can also be shown to occur once in the cycles of the first $\sqrt{n}/g$ rs-sections in any section. Consider the specific v-offset which begins the cycles which occur in rs-section J_1, $0 < J_1 < d/\sqrt{n}$. If this offset occurred in the cycles of a previous rs-section J_2, $0 \le J_2 < J_1 < \sqrt{n}/g$, then

$$\left(\sqrt{n}/g\right) + 2J_1 = \left(\sqrt{n}/g\right) + 2J_2 + l\sqrt{n}/g.$$

Here, l is the location of the v-offset within the cycle. Since even locations have even v-offsets and odd locations have odd v-offsets (the first v-offset is even and the v-offset increments are both odd), the location of the v-offset in rs-section J_2 must be even. Now, $J_1 = J_2 + l\sqrt{n}/2g$, but, since $J_1 < \sqrt{n}/g$, then l must be 0 and $J_1 = J_2$. Thus, the $\sqrt{n}/g$ cycles partition the g possible v-offsets.

A similar argument shows that in each consecutive disjoint set of $\sqrt{n}/g$ rs-sections (two rows of the torus), each of the different cycles occurs once. In the entire section, each cycle occurs in $(g/2)/\left(\sqrt{n}/g\right) = g^2/2\sqrt{n}$ rs-sections. Since a cycle occurs $2\sqrt{n}/g$ times in an rs-section and a specific v-offset occurs once per cycle, a cycle occurs g times per section. This satisfies the first hypothesis of Lemma 2.

Second, it will be shown that the second hypothesis of Lemma 2 is true. That is, for each of the possible v-offsets, exactly one message originates in each source node-group. Consider one of the g possible v-offsets, say f. As a v-offset, f occurs in $g^2/2\sqrt{n}$ rs-sections, and occurs once in each of the $2\sqrt{n}/g$ cycles within each of these rs-sections. Define a collection of nodes with v-offset f to be a matrix of nodes which consists of $g^2/2\sqrt{n}$ rows. The nodes in each row are those nodes that belong to the same rs-section and have f as a v-offset. Hence, a collection is an $g^2/2\sqrt{n}$ by $2\sqrt{n}/g$ matrix. We will show that any two nodes in a collection have different group assignments.

Denote by $n_{r,c}$ the node in row r and column c of a collection and recall that this node is assigned to group $\mu_g(n_{r,c})$. Consider two nodes n_{r_1,c_1} and n_{r_2,c_2} in the same collection. The relationship between both nodes and node n_{r_1,c_2} will be used to demonstrate that

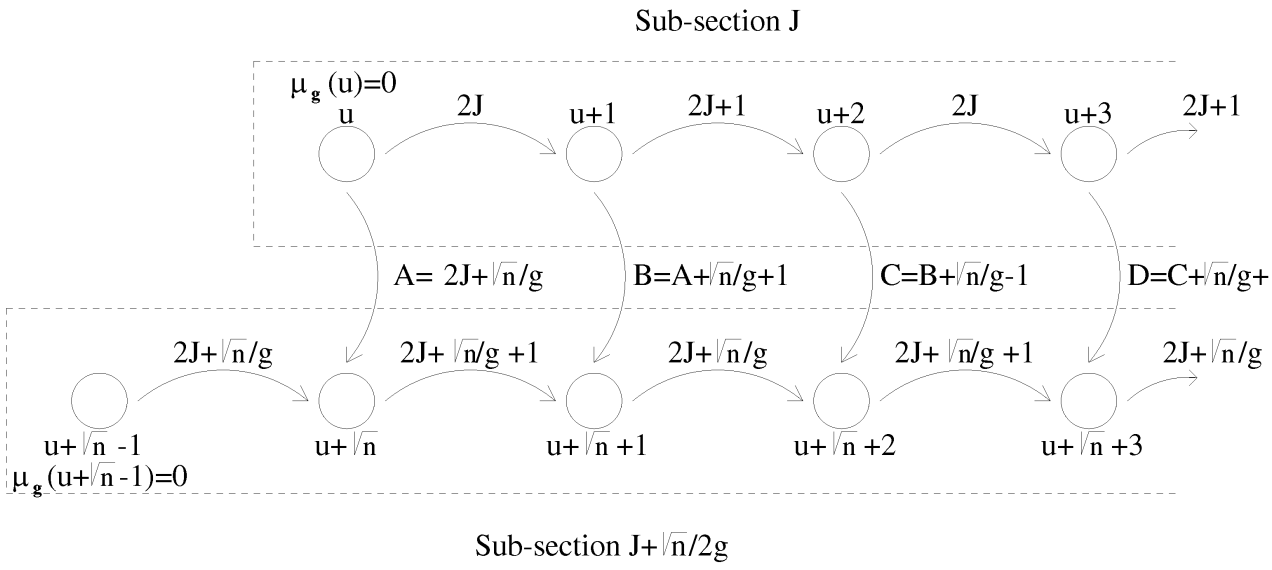


Fig. 11. The relation between v-offsets and h-offsets in rs-sections.

$\mu_g(n_{r_1,c_1}) \neq \mu_g(n_{r_2,c_2})$. If $\mu_g(n_{r_1,c_1}) = \mu_g(n_{r_2,c_2})$, it must be that $\mu_g(n_{r_1,c_1}) - \mu_g(n_{r_1,c_2}) = \mu_g(n_{r_2,c_2}) - \mu_g(n_{r_1,c_2})$. Thus, a specification of the difference between group assignments for collection nodes in the same row and nodes in the same column is needed.

In a collection, the difference between group assignments for adjacent nodes in the same row is a function of n , g, the number of the rs-section, J_0, that contributes the first node in the entire collection, and the number of the collection row r that contains the nodes. All of these, except r, are constant for all nodes in the collection. Specifically, it can be shown that the group assignment difference

$$\mu_g(n_{r,c+1}) - \mu_g(n_{r,c}) = \left(g^2/2\sqrt{n}\right)\left(4r\sqrt{n}/g + 4J_0 + 1\right).$$

This reduces to $\left(g^2/2\sqrt{n}\right)(4J_0 + 1)$, since $2rg = 0$ *mod* g.

Now, since $g/\left(g^2/2\sqrt{n}\right) = 2\sqrt{n}/g$ is a power of two, and $4J_0 + 1$ is odd, they are relatively prime. Thus, the $2\sqrt{n}/g$ nodes in any given collection row have distinct group assignments and are uniformly distributed among the g possible node assignments. Then, by a rearrangement of the columns in the collection, the group assignment difference becomes $g^2/2\sqrt{n}$. This is the difference for any two adjacent nodes in the rearranged collection, regardless of which row or which columns are involved. The difference between group assignments for two nodes that are in the same row but not adjacent is $\mu_g(n_{r,c_2}) - \mu_g(n_{r,c_1}) = g^2(c_2 - c_1)/2\sqrt{n}$.

In a collection, the difference between group assignments for adjacent nodes in the same column is a function of n , g, the number of the rs-section J_0 that contributes the first node in the entire collection, the position l of the offset within the cycle, and the number of the collection row r that contains the upper node. All of these, except r, are constant for all nodes in the collection. Specifically, it can be shown that the group assignment difference

$$\mu_g(n_{r+1,c}) - \mu_g(n_{r,c}) = \left(2\sqrt{n}/g\right)(l - 4r - 2) - (4J_0 + 1).$$

This difference, which we will denote V_r, is independent of which column in the collection contains the nodes. The difference between group assignments for two nodes that are in the same column but not adjacent is

$$\mu_g(n_{r_2,c}) - \mu_g(n_{r_1,c}) = \left(V_{r_1} - 8\sqrt{n}(r_2 - r_1 - 1)/g\right)(r_2 - r_1).$$

Given nodes n_{r_1,c_1} and n_{r_2,c_2} in a collection, if

$$\mu_g(n_{r_1,c_1}) = \mu_g(n_{r_2,c_2}),$$

then

$$\mu_g(n_{r_1,c_1}) - \mu_g(n_{r_1,c_2}) = \mu_g(n_{r_2,c_2}) - \mu_g(n_{r_1,c_2}).$$

This would imply that

$$\left(g^2/2\sqrt{n}\right)(c_2 - c_1) = \left(V_{r_1} - 8\sqrt{n}(r_2 - r_1 - 1)/g\right)(r_2 - r_1).$$

Since $2\sqrt{n}(l - 4r - 2)/g$ is an even number and $4J + 1$ is odd, then V_{r_1} is odd. Further, since $8\sqrt{n}(r_2 - r_1 - 1)/g$ is even, $V_{r_1} - 8\sqrt{n}(r_2 - r_1 - 1)/g$ is also odd. On the other hand, since $g^2/2\sqrt{n}$ is a power of two, then it must be that $g^2/2\sqrt{n} \leq r_2 - r_1$. This is a contradiction, since $0 \leq r_2 - r_1 < g^2/2\sqrt{n}$. Thus, any two distinct nodes in any collection must have different group assignments, and each of the g possible v-offsets must have distinct source node-group assignments. This satisfies the second hypothesis of Lemma 2 and, thus, proves that that the communication generated in the vertical direction by torus nodes in a single section can be achieved in a single time slot.

TABLE 1
POPS TOPOLOGY SUPPORT

Communications Pattern	Range	Sequence Length
All-to-All Person. Group		1
All-to-All Personalized		d^2
Natural Global Reduction		$(d-1) + (\log_2 g)$
Optimal Global Reduction	$d \le \sqrt{2n}$	$\log n$
	$d = \sqrt{2bn}$, $b > 1$	$\log n + 2(b-1) - \log b$
Natural Unidirectional Ring		$d-1$
Natural Bidirectional Ring		$2d-2$
Optimal Unidirectional Ring		d^2/n
Optimal Bidirectional Ring		$2d^2/n$
Optimal Unidirectional Torus	$d \ge 2\sqrt{n}$	$2d^2/n$
Optimal Bidirectional Torus	$d \ge 2\sqrt{n}$	$4d^2/n$

Hence, the communication generated in the vertical direction in the entire torus requires $p = n/c$ time slots. The result of the theorem follows by adding to this the n/c time slots required for the horizontal communications. The $2n/c$ is the minimum theoretical sequence length to deliver $2n$ messages with c couplers, and implies a 100 percent coupler utilization. □

8 CONCLUSION

The Partitioned Optical Passive Stars (POPS) Topology provides multiple data channels in an all-optical design that utilizes non-hierarchical passive stars. A key design configurability provides for a powerful optimization between lower total system complexity versus the combination of higher system throughput and lower power budgets. This paper has characterized the general communication capability of POPS networks. The group communication capabilities of POPS has been described and used to analyze the efficiency of embedding four of the most common and important regular communication patterns.

It was shown that all-to-all personalized communication in a POPS network is optimal when $d \ge \sqrt{n}$. Similarly, it was shown that log n global reduction operation is always possible if the POPS contains enough couplers to support the required message exchange in each step of the reduction. The obvious or "natural" global reductions were shown to be inefficient.

The alternating-pair method of embedding the ring communication pattern in a POPS topology was presented. This embedding was shown to be optimal in any given POPS network. The ring embedding was extended to provide for embedding two-dimensional tori in POPS topologies. The extended embedding was shown to be optimal when $d \ge 2\sqrt{n}$.

The optimal nature of an alternating-pair embedding of the ring or torus communication pattern in a POPS topology was shown to be independent of the unidirectional or bidirectional aspect of the communications. Further, since the linear array is a subset of the ring and the mesh is a subset of the torus, the alternating-pair method is similarly efficient for embedding these patterns into POPS topologies. Additionally, results not presented here demonstrate the applicability of the alternating-pair method to three-dimensional tori embeddings in various POPS topologies. The extension of the embedding algorithm from two- to three-dimensional torus is similar to the extension from the ring to the two-dimensional torus.

The inherent high utilization and high performance of a POPS topology is well matched to common application requirements. Wide ranges of POPS topology parameters have been shown to support these communication structures with 100 percent coupler utilization and/or minimal sequence lengths. Table 1 summarizes the POPS topology support for the various communication patterns discussed assuming that $d \ge \sqrt{n}$. That is, assuming that the number of couplers is not larger than the number of processors.

In [5], it has been shown that an n-node POPS handles random communication patterns efficiently, especially when the coupler degree, d, is proportional to $\sqrt{n}$. Hence, choosing $d = 2\sqrt{n}$ for POPS efficiently supports random traffic and optimally supports the regular communication patterns discussed in this paper. The efficiency of POPS networks to support other regular communication patterns is open for further studies. Another issue that is open for further studies is the comparison of the different ways of adding redundant paths to POPS network for purposes of fault-tolerance or congestion control. Finally, the benefits of using a limited degree of WDM in POPS networks need to be weighted against the hardware and control complexity introduced when both WDM and TDM are combined.

ACKNOWLEDGMENTS

The authors would like to thank Steven Levitan and Donald Chiarulli for helpful feedback during the early part of the research presented in this paper. This work is in part supported by a grant from the U.S. Air Force Office of Scientific Research under contract F49620-93-1-0023DEF. Part of this work was presented at the Second International Conference on Massively Parallel Processing Using Optical Interconnections, San Antonio, Texas, 1995.

REFERENCES

[1] D. Chiarulli, S. Levitan, R. Melhem, J. Teza, and G. Gravenstreter, "Multiprocessor Interconnection Networks Using Partitioned Optical Passive Star (POPS) Topologies and Distributed Control," *Proc. First Int'l Workshop Massively Parallel Processing UsingOptical Interconnections*, pp. 70-80, Apr. 1994.

[2] P. Dowd, "Random Access Protocols for High-Speed Interprocessor Communications Based on an Optical Passive Star Topology," *IEEE J. Lightwave Technology*, vol. 9, no. 6, pp. 799-808, June 1991.

[3] A. Ganz, B. Li, and L. Zenou, "Reconfigurability of Multi-Star Based Lightwave LANs," *Proc. IEEE Globecom '91*, vol. 3, pp. 1,906-1,910, 1992.

[4] M. Goodman, H. Kobrinski, M. Vecchi, R. Bulley, and J. Gimlett, "The LAMBDANET Multiwavelength Network: Architecture, Applications, and Demonstrations," *IEEE J. Selected Areas in Comm.*, vol. 8, no. 6, pp. 995-1,004, Aug. 1990.

[5] G. Gravenstreter, R. Melhem, D. Chiarulli, S. Levitan, and J. Teza, "The Partitioned Optical Passive Stars (POPS) Topology," *Proc. Ninth IEEE Int'l Parallel Processing Symp.*, Apr. 1995.

[6] P. Green, "An All-Optical Computer Network: Lessons Learned," *IEEE Network*, pp. 56-60, Mar. 1992.

[7] I. Habbab, M. Kavehrad, and C. Sundberg, "Protocols for Very High-Speed Optical Fiber Local Area Networks Using a Passive Star Topology," *IEEE J. Lightwave Technology*, vol. 5, no. 12, pp. 1,782-1,793, Dec. 1987.

[8] C. Ho and S. Johnsson, "On the Embedding of Arbitrary Meshes in Boolean Cubes with Expansion Two Dilation Two," *Proc. Int'l Conf. Parallel Processing*, pp. 188-191, 1987.

[9] S. Johnsson and C.-T. Ho, "Optimum Broadcasting and Personalized Communication in Hypercubes," *IEEE Trans. Computers*, vol. 38, no. 9, pp. 1,249-1,268, Sept. 1989.

[10] V. Kumar, A. Grama, A. Gupta, and G. Karypis, *Introduction to Parallel Computing*. Redwood City, Calif.: Benjamin/Cummings, 1994.

[11] C. Leiserson, "The Network Architecture of the Connection Machine CM-5," *Proc. Fourth Ann. ACM Symp. Parallel Algorithms and Architectures*, pp. 272-285, 1992.

[12] R. Melhem and G. Hwang, "Embedding Rectangular Grids into Square Grids with Dilation Two," *IEEE Trans. Computers*, vol. 39, no. 12, pp. 1,446-1,455, Dec. 1990.

[13] B. Mukherjee, "WDM-Based Local Lightwave Networks, Part I: Single-Hop Systems," *IEEE Networks*, pp. 12-27, May 1992.

[14] M. Nigam, S. Sahni, and B. Krishnamurthy, "Embedding Hamiltonians and Hypercubes in Star Interconnection Networks," *Proc. Int'l Conf. Parallel Processing*, pp. 340-343, 1990.

[15] S. Ranka and S. Sahni, *Hypercube Algorithms for Image Processing and Pattern Recognition*. New York: Springer-Verlag, 1990.

[16] Y. Saad and M. Schultz, "Topological Properties of Hypercubes," *IEEE Trans. Computers*, vol. 37, no. 7, pp. 867-872, July 1988.

Greg Gravenstreter received the BS degree in mathematics and computer science from the University of Pittsburgh in 1988. He is currently involved in networking and security at the Software Engineering Institute within Carnegie Mellon University. From 1971 to 1985, he did research in the Chemical Physics and Computer Science Departments at Westinghouse Electric Corporation's Research and Development Center. He worked as director of Systems and Operations for Guidance Technologies in Pittsburgh from 1988 to 1992. From 1992 to 1995, he did research in the Computer Science Department at the University of Pittsburgh. His research interests include a wide range of network and security issues, architectural, and systems issues for parallel and distributed systems.

Rami Melhem received a BE in electrical engineering from Cairo University in 1976, an MA in mathematics and an MS in computer science from the University of Pittsburgh in 1981, and a PhD in computer science from the University of Pittsburgh in 1983. He is a professor of computer science at the University of Pittsburgh. Previously, he was an assistant professor at Purdue University and an assistant/associate professor at the University of Pittsburgh. He has published numerous papers in the areas of optical interconnections, fault tolerance, systolic architectures, and high performance computing. He served on program committees of several conferences and workshops and as the general chair for the Third International Conference on Massively Parallel Processing Using Optical Interconnections. He is a member of the editorial board of *Parallel & Distributed Processing* and was on the editorial board of *IEEE Transactions on Computers*. He was a guest editor of a special issue of the *Journal of Parallel and Distributed Computing* on optical computing and interconnection systems. He was on the advisory board of the IEEE Technical Committees on Parallel Processing and Computer Architecture. Dr. Melhem is a member of the IEEE Computer Society, the ACM, and the International Society for Optical Engineering, and a senior member of the IEEE.