Embedding Rings and Meshes in Partitioned Optical Passive Stars Networks.

G. Gravenstreter and R.G. Melhem
Department of Computer Science
The University of Pittsburgh
Pittsburgh, PA 15260

Abstract

We consider the problem of embedding ring and torus communication structures onto the all-optical Partitioned Optical Passive Stars (POPS) topology. We show that the natural embedding, which maps the nodes in the ring or torus directly onto the POPS, does not result in an optimal embedding, and thus does not efficiently utilize the communication capabilities of the POPS. For the ring structure, we present an alternating-pair embedding technique which distributes the communication load uniformly in the POPS. We then apply a slightly modified version of the alternating-pair technique to the embedding of tori onto POPS networks. We prove that both embeddings are optimal, in the sense that they minimize the time required to deliver the messages from each node to its neighbors.

1 Introduction.

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 [4, 9]. A large amount of research has been accomplished on WDM-based systems. 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] is a topological approach to providing multiple physical data channels. It is an interconnection architecture that uses multiple non-hierarchical stars to achieve single-hop networks [2, 3]. It is an all-optical topology constructed exclusively with passive optical technology. 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 TDM or 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 tradeoff. This design flexibility provides for a customized optimization between lower total system complexity versus the combination of higher system throughput with both 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. An embedding is a mapping of the nodes and edges in a graph representing the communication requirements of an application onto the nodes and edges of the graph representing the physical communication network. A key measure of a topology's value as an interconnection network is the dilation and utilization achieved when embedding several standard regular structures. The linear array, or the associated ring, is an example of one of these regular structures. Another important structure is the mesh, or the torus. Several of the most important multi-processor applications generate a mesh type of communication requirement. These 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, 10, 11].

This paper discusses the capabilities of POPS networks to support ring and torus 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 embedding communication structures onto POPS networks. The fourth section examines embeddings of a ring communication structure in a POPS topology. The fifth section examines the embedding of a two dimensional torus in a POPS topology. A conclusion summarizes key points about our embeddings.

2 Description of The POPS Topology.

A small POPS network is shown in Figure 1. Source nodes relay messages through optical links to passive optical couplers. These couplers send the message 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. In order to prevent collisions, at most one input to a given coupler should carry a signal at any one time.

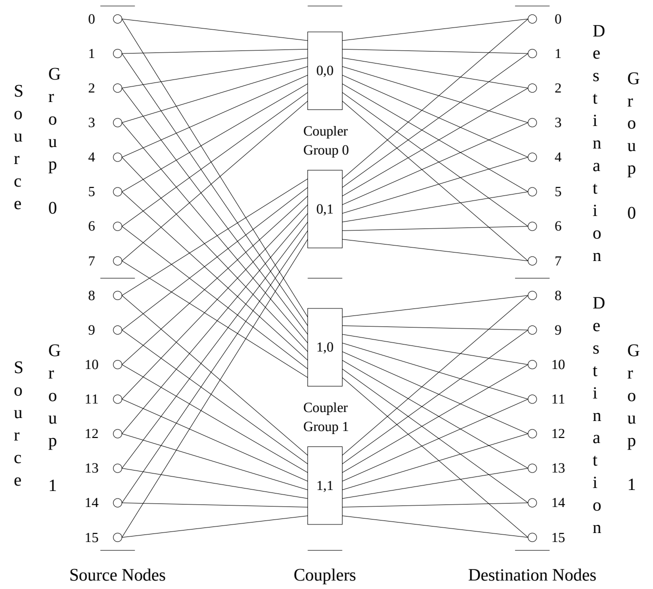

Figure 1: An $n = 16, d = 8$ POPS network ($g = 2$).

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. Typically, a source node and the corresponding destination node are the same processing element. These n nodes are denoted $N_0, N_1, ..., N_{n-1}$. Consequently, the source groups and the corresponding destination groups contain the same set of nodes. These g groups of nodes are are called node-groups and are denoted $GN_0, GN_1, ..., GN_{g-1}$.

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, $0 \leq i < g$, are denoted by $C_{i,0}, C_{i,1}, ..., C_{i,g-1}$. The d inputs for a given coupler, $C_{i,j}$, are connected to the d nodes in node-group GN_j, and the d outputs of $C_{i,j}$, are connected to the d nodes in node-group GN_i.

Each source node, N_i, has g transmitters, denoted $T_{i,0}, T_{i,1}, ..., T_{i,g-1}$, where $0 \leq i < n$. For a given source node, N_i, each transmitter, $T_{i,j}$, is connected to a coupler in GC_j, and thus is used to communicate with nodes in node-group GN_j. Similarly, each destination node N_i has g receivers, denoted $R_{i,0}, R_{i,1}, ..., R_{i,g-1}$, and receiver $R_{i,j}$ is connected to a coupler which provides communication with nodes in node-group GN_j.

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.

3 Communication Using POPS

In a POPS network, the determination of the route for a specified message is a simple process. The source node determines the appropriate transmitter for communication with the required destination node-group, and the destination node determines the appropriate receiver for communication with the required source node-group. Specifically, consider a single message, denoted $M_{x,y}$, which originates at source node x in node-group $GN_{x'}$, where $0 \leq x < n$ and $x' = \lfloor x/d \rfloor$. This message terminates at destination node y in node-group $GN_{y'}$, where $0 \leq y < n$ and $y' = \lfloor y/d \rfloor$. Source node x uses transmitter $T_{x,y'}$ to send the message. Coupler $C_{y',x'}$ transports the message, and destination node y uses receiver $R_{y,x'}$ to get the message. Thus, any specific message has a unique path composed of a transmitter, a coupler, and a receiver, $(T_{x,y'}, C_{y',x'}, R_{y,x'})$. 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 non-zero number of couplers will deliver one message each.

The merit of an interconnection network can be studied by analyzing its capability for delivering different communication patterns. In general. a communication structure can be defined as a set, $\mathcal{M} = \{M_{x_0,y_0}, M_{x_1,y_1}, ..., M_{x_{m-1},y_{m-1}}\}$ of m messages, where $x_0, x_1, ..., x_{m-1}$ are the m message source nodes, and $y_0, y_1, ..., y_{m-1}$ are the m message destination nodes.

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, specifically the ring and two dimensional torus structures. Assignment algorithms for embedding (mapping) the ring or torus logical nodes into the POPS physical nodes will be given. Proofs are provided that these assignments yield optimal embeddings, that is, where coupler utilization is 100% during every time slot in the sequence and which results in the minimum theoretical sequence length. The following definition and lemmas will simplify subsequent discussions. In this definition and in the remainder of this paper, we will use $+_g$ and $-_g$ to denote the addition and subtraction *modulo g* operation. That is, $a +_g b = (a + b) \ mod \ g$ and $a -_g b = (a - b) \ mod \ g$.

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

Any message originating in node-group x' , with an offset of i , requires the x'^{th} coupler in coupler group $x' +_g i$. Further, the subset of g couplers that consists of the x'^{th} coupler in each coupler group is required for delivery of g messages originating in node-group x' when each of the messages have different offsets. Hence, we have the following lemma:

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

Considering g of these sets, each originating at a different node-group, the above lemma 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) every g messages originating at the same node-group have different offsets. The following lemma states these two conditions in a slightly different form.

Lemma 2: *A set, $\mathcal{M}$, of $c = g^2$ messages can be delivered in one time slot if and only if:*

1. *for each f, $0 \leq f < g$, exactly g messages in $\mathcal{M}$ have offset f,*
2. *the g messages that have 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. Short sequence lengths are achieved when communication in each time slot emphasizes both inter-group 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 communications in one time slot.

Lemma 3: *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 logical communication structure can be formally specified by a graph $G = (V, E)$, where V is a set of n nodes and $E = \{< u, v >: u, v \in V\}$ is a set of edges connecting pairs of nodes from V. An embedding of a logical communication structure, (V, E), onto an n-processor POPS topology is a one-to-one mapping function $\mu : V \rightarrow \{0, \ldots, n-1\}$ which maps each node in V onto a processor 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 $\mathcal{M} = \{M_{\mu(u),\mu(v)} :< u, v >\in E\}$, thus satisfying the requirement of the communication structure in minimal time. More specifically, by partitioning $\mathcal{M}$ into s subsets $\mathcal{M} = \mathcal{M}_1 \cup \ldots \cup \mathcal{M}_s$, such that each subset $\mathcal{M}_i = \{M_{\mu(u),\mu(v)} :< u, v >\in E_i \subseteq E\}$, satisfies the conditions of Lemma 2, all the messages in $\mathcal{M}$ can be delivered in s time slots. An optimal embedding minimizes s.

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 condition necessary to deliver a set of messages in one time slot does not depend on the actual source nodes and destination nodes of the messages, but rather on the source node-groups and the destination node-groups. Thus, the sequence length (number of time slots) needed to deliver the messages in an embedded structure does not depend on μ_n.

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 the group-mapping function μ_g. We will show, however, that a group-mapping function, μ_g, leads to a one-to-one mapping function, μ, by showing that the hypothesis of the following Lemma is satisfied.

Lemma 4: *A mapping function $\mu = d * \mu_g + \mu_n$ is a one-to-one mapping if and only if μ_g maps exactly d nodes of V to each node-group, where d is the size* *of the node-group.*

4 Ring Embeddings in POPS

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 3 provides a means for extending the results to bi-directional 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 $\mathcal{M} = \{M_{k,k+_n1} : k = 0, \ldots, n-1\}$. The following Lemma, which is stated without proof, shows that the natural embedding of the ring onto a POPS network is not optimal.

Lemma 5: *Using the natural embedding of ring nodes onto a POPS network, the ring communication requires a sequence of length $s = d - 1$.*

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 $\mathcal{N}_i$, $\forall 0 \leq i < p$. Since $d \geq \sqrt{n}$ and both d^2 and n are powers of two, then p is a positive integer. 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 $\mathcal{N}_i$, $\forall 0 \leq i < p$, to node-groups in the POPS network. This algorithm yields an identical sequence of group assignments for each section in the ring.

Second, within a given single ring section $\mathcal{N}_i$, partition the c ring elements into $g/2$ subsets, called sub-sections, and denoted $\mathcal{N}_i^J$, $\forall 0 \leq J < g/2$. Since g is a power of two, $g/2$ is a positive integer. Each sub-section, $\mathcal{N}_i^J$, contains $c/(g/2) = 2g$ consecutive ring elements. Figure 2 shows the partitioning of the ring elements into sections and sub-sections.

Third, within a given single sub-section of a given single ring section, the algorithm assigns the first element in the sub-section 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    /* sub-section J */
```

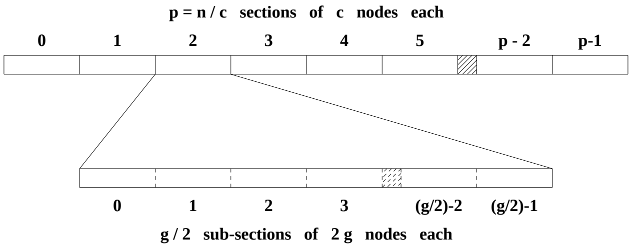

Figure 2: Overview of the Ring Embedding

```
μ_g(node) = 0 ;                /* the first node */
node = node + 1 ;
for i = 1, ... , 2g-1 do
  if i is odd
    then μ_g(node) = μ_g(node − 1) +_g 2J
    else μ_g(node) = μ_g(node − 1) +_g 2J +_g 1 ;
  node = node + 1 ;
```

A different way for 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 /* sub-section J */
    for i = 0, ... , 2g-1 do
      if i is odd
        then φ(node) = 2J +_g 1
        else φ(node) = 2J ;
      node = node + 1 ;
```

Note that if u and v are the first nodes in two consecutive sub-sections, 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.

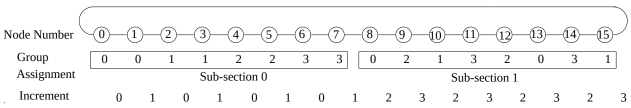

Figure 3: Group Assignments for $n = 16$ and $d = 4$

In Figure 3, we show the application of the algorithm to the embedding of a 16-node ring into the POPS shown in Figure 4. In this case, $g = 4$ and $n = c = 16$. Thus, there is only one section, which is composed of two sub-sections. The first nodes in each sub-section is assigned to node-group 0, and the group assignments of subsequent nodes in sub-section 0 are obtained by incrementing the group assignments by the offsets 0 and 1 alternatively. The group assignments of the nodes in sub-section 1 are obtained by

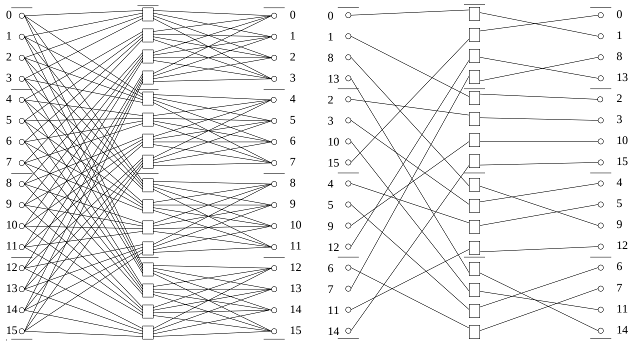

Figure 4: Ring Embedding for $n = 16, d = 4$

incrementing the group assignments by the offsets 2 and 3. The assignment of nodes within groups is done arbitrarily. From Figure 4, 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 1 *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 $g/2$ sub-sections in the section. Within the sub-section 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 sub-section J where a particular offset is used, it 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. It also satisfies the hypothesis of Lemma 4, 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% coupler utilization. $\diamond$

A modified version of the proof of Theorem 1 may be applied to prove that the alternating-pair embedding is optimal even if $d < \sqrt{n}$. In this case, $n < c$ and there is only one section. If we partition this section into $n/2g$ sub-sections 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.

5 Torus Embeddings in POPS

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 3 can extend our result to bi-directional tori. Specifically, we will consider the communication structure $(V, E_h \cup E_v)$, where $V = \{0, \ldots, n-1\}$ and $E_h = \{< u, u+1 >; u \in V \ \& \ u \neq r\sqrt{n} - 1, r = 1, \ldots, \sqrt{n}\} \cup \{< u, u - \sqrt{n} + 1 >; u \in V \ \& \ u = r\sqrt{n} - 1\}$, represents wrapped around horizontal communications, and $E_v = \{< u, u +_n \sqrt{n} >; u \in V\}$ 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.

The alternating-pair embedding, introduced for the ring communication structure and shown to be optimal there, can be extended 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 alternating-pair torus embedding is optimal.

The alternating-pair embedding of the two dimensional torus communication structure in a POPS topology is as follows: First, partition the n torus elements into $p = n/c$ subsets, called sections, and denoted $\mathcal{N}_i$, $\forall 0 \leq i < p$. 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 $\mathcal{N}_i$, $\forall 0 \leq i < p$, 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 sub-sections, and denoted $\mathcal{N}_i^J$, $\forall 0 \leq J < g/2$. Since g is a power of two, $g/2$ is a positive integer. Each sub-section, $\mathcal{N}_i^J$, contains $c/(g/2) = 2g$ consecutive torus elements. It is a requirement that at least one complete sub-section fits within each row of the torus, thus, $2g \leq \sqrt{n}$, and $d \geq 2\sqrt{n}$. Figure 5 shows the partitioning of the torus elements into sections and sub-sections.

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 sub-section is assigned to node-group 0, then

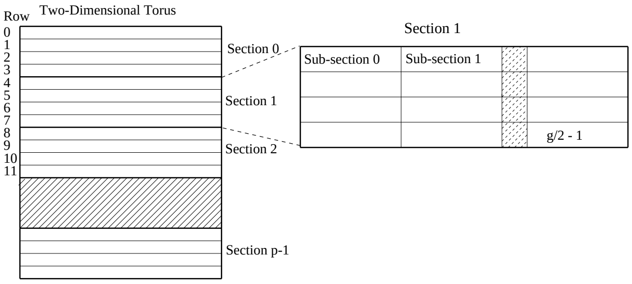

Figure 5: Overview of the Torus Embedding

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

$node = 0$;
for $J = 0, \ldots, g/2 - 1$ do /* sub-section J */
    $\tau(node) = 0$; /* first node */
    $node = node + 1$;
    for $i = 1, \ldots$, $2g$-1 do
        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 rows */
    for $m = 0, \ldots$, $\sqrt{n}$ -1 do
        If $(m - r < 0)$
            then $\mu_g(\sqrt{n}r + m - r) = \tau(\sqrt{n}r + m)$
            else $\mu_g(\sqrt{n}(r + 1) + m - r) = \tau(\sqrt{n}r + m)$;

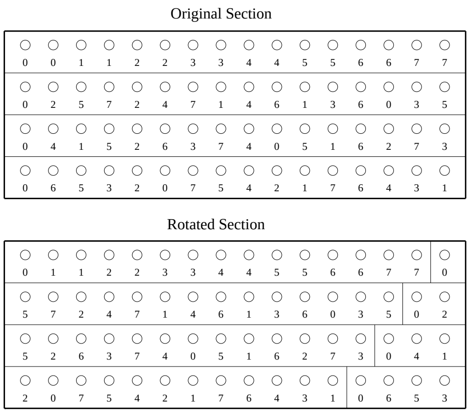

Figure 6: Group Assignments for $n = 256$ and $d = 32$.

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 4 sections, each containing 4 sub-sections of 16 nodes each. Here a sub-section spans an entire row. Figure 6 shows the group assignment for the nodes in any of the 4 sub-sections, before and after the row rotation step.

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 $\tau(u +_n 1) -_g \tau(u) = \tau(u - \sqrt{n}) -_g \tau(u)$ 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 $d \geq 2\sqrt{n}$, 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 2 *In any POPS topology where n and d are powers of two, if $d \geq 2\sqrt{n}$, then the alternating-pair embedding of a torus communication structure requires a sequence of length $s = 2n/c = 2d^2/n$, which is optimal.*

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 $2J + (\sqrt{n}/g) + 1$. These offsets are shown in Figure 7.

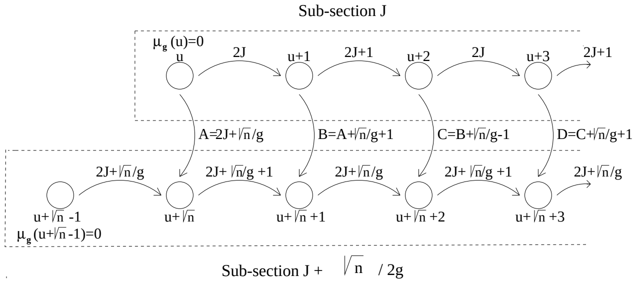

Figure 7: v-offsets and h-offsets in rs-sections.

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

Since $(\sqrt{n}/g) + 1 + (\sqrt{n}/g) - 1 = 2\sqrt{n}/g$ and $(2\sqrt{n}/g)(g^2/2\sqrt{n}) = g$, 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 $(\sqrt{n}/g)+1$ or $(\sqrt{n}/g)-1$, 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 $J_2, 0 \leq J_2 < J_1 < \sqrt{n}/g$, then $(\sqrt{n}/g)+2J_1 = (\sqrt{n}/g)+2J_2+l\sqrt{n}/g$. 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 $(g/2)/(\sqrt{n}/g) = g^2/2\sqrt{n}$ 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 $\mu_g(n_{r_1,c_1}) \neq \mu_g(n_{r_2,c_2})$. 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 $\mu_g(n_{r,c+1}) - \mu_g(n_{r,c})$ $= (g^2/2\sqrt{n})(4r\sqrt{n}/g + 4J_0 + 1)$. This reduces to $(g^2/2\sqrt{n})(4J_0+1)$, since $2rg = 0 \ mod \ g$.

Now, since $g/(g^2/2\sqrt{n}) = 2\sqrt{n}/g$ 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, if can be shown that the group assignment difference $\mu_g(n_{r+1,c}) - \mu_g(n_{r,c})$ $= (2\sqrt{n}/g)(l - 4r - 2) - (4J_0 + 1)$. 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 $\mu_g(n_{r_2,c}) - \mu_g(n_{r_1,c})$ $= (V_{r_1} - 8\sqrt{n}(r_2 - r_1 - 1)/g)(r_2 - r_1)$.

Given nodes n_{r_1,c_1} and n_{r_2,c_2} in a collection, it can be shown, using the above formulas, that $\mu_g(n_{r_1,c_1}) \neq \mu_g(n_{r_2,c_2})$. 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.

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% coupler utilization. $\diamond$

6 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 a POPS network. The all-to-all personalized (complete) group communication that is possible in a POPS network has been described. This capacity has been used to analyze the efficiency of embedding two of the most common and important regular communication patterns.

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. This alternating-pair method was extended to provide for embedding a two dimensional torus in a POPS topology. This extended embedding was shown to be optimal over a wide range of POPS configurations.

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 uni-directional or bi-directional 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.

Acknowledgment
The authors would like to thank Steven Levitan and Donald Chiarulli for helpful feedback during the early part of this research. This work is, in part, supported by a grant from The Air Force Office of Scientific Research under contract F49620-93-1-0023DEF.

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. In *Proceedings of the First International Workshop on Massively Parallel Processing Using Interconnections*, pages 70–80. IEEE, April 1994.

[2] P. Dowd. Random access protocols for high-speed interprocessor communications based on an optical passive star topology. *Journal of Lightwave Technology*, 9(6):799–808, June 1991.

[3] A. Ganz, B. Li, and L. Zenou. Reconfigurability of multi-star based lightwave LANs. In *Globecom '91*, volume 3, pages 1906–1910. IEEE, 1992.

[4] M. S. Goodman, H. Kobrinski, M. P. Vecchi, R. M. Bulley, and J. L. Gimlett. The LAMBDANET multi-wavelength network: Architecture, applications, and demonstration. *IEEE Journal of Selected Areas in Communications*, 8(6):995–1004, August 1990.

[5] G. Gravenstreter, R.G. Melhem, D.M. Chiarulli, S.P. Levitan, and J.P. Teza. The partitioned optical passive stars (POPS) topology. In *Proceedings of the Ninth International Parallel Processing Symposium*. IEEE, April 1995.

[6] P. Green. An all-optical computer network: Lessons learned. *IEEE Network*, pages 56–60, March 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 Journal of Lightwave Technology*, LT-5(12):1782–1793, December 1987.

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

[9] B. Mukherjee. WDM-Based local lightwave networks, Part I: Single-hop systems. *IEEE Networks*, pages 12–27, May 1992.

[10] M. Nigam, S. Sahni, and B. Krishnamurthy. Embedding hamiltonians and hypercubes in star interconnection networks. In *Proceedings of the International Conference on Parallel Processing*, pages 340–343, 1990.

[11] Y. Saad and M. Schultz. Topological properties of hypercubes. *IEEE Transactions on Computers*, C-37(7):867–872, July 1988.