The Partitioned Optical Passive Stars (POPS) Topology. *

G. Gravenstreter, R.G. Melhem, D.M. Chiarulli, S.P. Levitan, J.P. Teza

University of Pittsburgh, Pittsburgh, PA 15260

Abstract

This paper presents and analyzes a topological approach to providing multiple data channels using current technologies. The Partitioned Optical Passive Stars (POPS) topology is an all-optical interconnection architecture that uses multiple non-hierarchical couplers. POPS topologies provide powerful configurability for optimization of system complexity, network throughput, power budget, and control overhead. It is shown that high performance and utilization is achieved for random communication patterns. POPS topologies maintain these characteristics even when scaled to large systems.

1 Introduction.

The evolution of computer systems has, for the most part, been driven by an ever-increasing need for network throughput. Although the bandwidth wasted to control pure-electronic networks is not excessive, relative to network capabilities and system loads, these networks are incapable of efficiently accomodating very large volumes of traffic due to several key physical link issues. These include driving power, capacitive/inductive loading, and relatively low transmitter rates implied by sensitivity to noise.

The second era in computer networking incorporated optical fiber link technology into existing multihop network designs [11, 1, 6]. An order of magnitude increase in link throughput was available. Further, the lack of reactive factors and a high noise immunity were well suited to the demands of the faster, larger, and more widespread environments. Such networks, however, still suffer from throughput bottlenecks and high latencies resulting from electronic/optical conversions and processing at intermediate hops. These disadvantages can be avoided in "all-optical" networks, where electronic technology is only present at the beginning and the end of the communications pathway. Certain topologies implemented exclusively with passive optical technology provide enormous potential throughput with very low latencies. Communication protocols incorporating time (TDM) [10], wavelength (WDM) [5], or code (CDM) [8] division multiplexing have been developed to increase utilization and provide more efficient access to this tremendous capacity.

*This work is in part supported by a grant from The Air Force Office of Scientific Research under contract F49620-93-1-0023DEF.

Several studies have been done on innovative topological approaches for "all-optical" networks. Power optimization in multiple star networks, both hierarchical and non-hierarchical, was investigated by Birk [2]. Networks of multiple passive stars, both totally and partially connected, as well as the reconfigurability of WDM systems operating on them, was studied by Ganz [4]. A multi-hop WDM method applied to multiple star topologies was presented by Hluchyj [7].

This paper studies a topological approach to providing multiple physical data channels. The Partitioned Optical Passive Stars (POPS) topology is an interconnection architecture that uses multiple non-hierarchical stars to achieve single-hop networks. It is an "all-optical" topology constructed exclusively with passive optical technology, and benefits from all the corresponding characteristics discussed, i.e., no intermediate electronic/optical conversions, no reactive factors and high noise immunity.

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.

After the introduction, the second section will discuss the components, parameters, and notation for a POPS topology. The third section will discuss the routing and scheduling of messages in a POPS topology. The fourth section will discuss the performance of POPS topologies in detail. The fifth section discusses issues related to scalability for very large system sizes. A conclusion summarizes key points about a POPS topology.

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. 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.

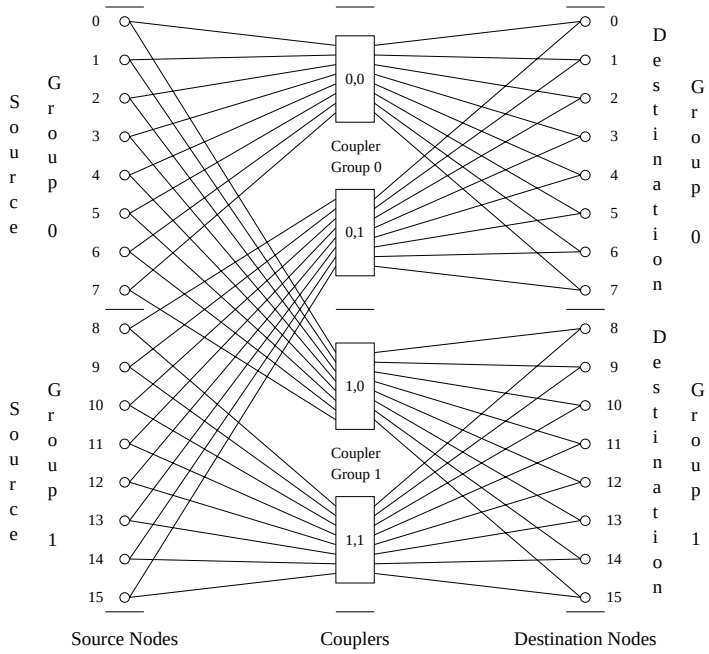

Figure 1: An $n = 16, d = 8$ POPS network ($g = 2$).

The set of n source nodes are partitioned into $g = n/d$ equal-sized source groups. Similarily, 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 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. These g groups of couplers are denoted by $GC_0, GC_1, ..., GC_{g-1}$. The g couplers in coupler group GC_i are denoted by $C_{i,0}, C_{i,1}, ..., C_{i,g-1}$, where $0 \leq i < g$. The d inputs for a given coupler, $C_{i,j}$, are connected to the d nodes in source group GN_j, and the d outputs of $C_{i,j}$, are connected to the d nodes in destination 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 destination group GN_j. Similarily, 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 source group GN_j.

The optical links in a POPS network fall into one of two classes. In the first class, called source links, each link connects a transmitter at a source node to an input of a coupler. For each coupler, at most one source link should be active at a given time to prevent collisions. In the second class, called destination links, each link connects an output of a coupler to a receiver at a destination node. For each coupler, all destination links are either simultaneously active or inactive.

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. The example POPS topology shown in Figure 1 has a fairly large coupler degree. In Figure 2, another example POPS topology, with the same number of nodes as in Figure 1, but a smaller coupler degree, is given for comparison.

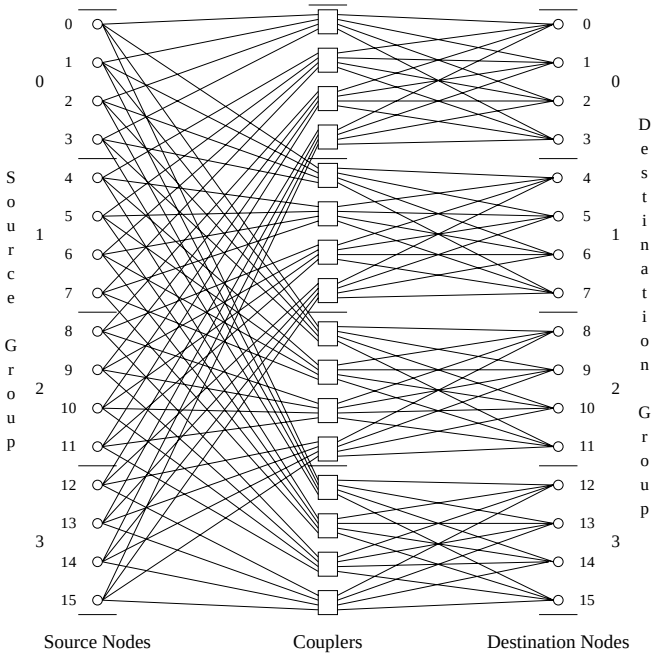

Figure 2: An $n = 16, d = 4$ POPS network ($g = 4$).

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 group, and the destination node determines the appropriate receiver for communication with the required source group. Specifically, consider a single message, denoted $M_{x,y}$, which originates at source node x in source group $GN_{x'}$, where $0 \leq x < n$ and $x' = \lfloor x/d \rfloor$. This message terminates at destination node y in destination 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.

Lemma 1 *In a given POPS network, any specific*

message $M_{x,y}$ has a unique path composed of a transmitter, a coupler, and a receiver, $(T_{x,y'}, C_{y',x'}, R_{y,x'})$.

Lemma 2 *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. No conflict occurs if either $x'_i \neq x'_j$ or $y'_i \neq y'_j$.*

Thus, the conflicts during delivery of a set of messages in a POPS can be determined. Subsequently, the minimal schedule of delivery for a set of messages can also be readily calculated as described next.

Definition 1 *A* **Message Set** *is a set of m messages, denoted by $\mathcal{M} = \{M_{x_0,y_0}, M_{x_1,y_1}, ..., M_{x_{m-1},y_{m-1}}\}$, 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.*

Given a message set $\mathcal{M}$, let a coupler's usage, denoted by u_i , where $0 \leq i < c$, be the number of individual messages that will use the i^{th} coupler. That is, $C_{\lfloor i/g \rfloor, i \bmod g}$.

Definition 2 *A* **Coupler Profile** *for a given message set $\mathcal{M}$ is an ordered c-tuple of coupler usages, denoted $C(\mathcal{M}) = (u_0, u_1, ..., u_{c-1})$.*

Lemma 1 states that the coupler used by a given message is a function of the message's source and destination. Applied to the elements of a message set, Lemma 1 yields a unique coupler profile.

Theorem 1 *In a given POPS network, for any given message set, the scheduling of any specific coupler is independent of all other couplers.*

Proof: Any given message set yields a unique coupler profile. From Lemma 2, it follows that the subset of messages requiring a given coupler all have conflicting paths. However, no element of this subset of the messages conflicts with any other message. Thus, the subset can be independently scheduled. ◇

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. When multiple messages require the same coupler, a series of consecutive time slots, called a sequence, is required. During each time slot in a sequence, some non-zero number of couplers will deliver one message each.

A scheduling algorithm chooses, for each coupler, the order and timing of delivery for the pending messages requiring that coupler. A greedy scheduling algorithm is defined to be one that, by some selection criteria in each time slot, assigns one message to each coupler where messages are pending. For example, a random greedy scheduling randomly selects one of the messages pending for a specific coupler. The next result follows directly from Theorem 1.

Corollary 1 *Any greedy scheduling algorithm has the minimum possible sequence length, $u_{max} = max\{u_0, u_1, ..., u_{c-1}\}$.*

Network traffic can be either static or dynamic. A static communications pattern is known apriori, while a dynamic pattern is not. In a POPS network, as seen in Lemma 1, route determination is trivial. Further, Corollary 1 demonstrates that, as a result of the independent nature of the couplers, scheduling is both easy and efficient. For dynamic message traffic, it is still necessary to specify an arbitration protocol on a per coupler basis. Established wavelength-division multiplexing protocols, such as those surveyed in [9], might be adaptable. In [3], a time-division multiplexing protocol is presented for the POPS topology.

The focus of this paper is the study of performance in a POPS network for static communications patterns. That is, a predetermined message set , $\mathcal{M}$, of size m is generated and delivered. Only random message sets that are permutations or subsets of permutations are studied.

4 Permutation Capabilities of POPS.

Certain key aspects concerning the performance of a POPS network are examined. First, we define permutation-based random message sets as follows:

Definition 3 *A* **Permutation-Based Message Set** *is a message set of m messages, $1 \leq m \leq n$, where the m source nodes $x_0, x_1, ..., x_{m-1}$ are distinct, and the m destination nodes $y_0, y_1, ..., y_{m-1}$ are also distinct.*

4.1 Absolute Bounds for Message Sets.

Assume a fixed POPS topology and a specific message set $\mathcal{M}$ of size m, with coupler profile $C(\mathcal{M})$, and a maximum coupler usage of u_{max}. If s is the sequence length that a greedy algorithm yields, then, from Corollary 1, $s = u_{max}$. Clearly, for a given $\mathcal{M}$, u_{max} depends on m and on the message distribution in $\mathcal{M}$. For a given m, however, the value of u_{max} is bounded as defined below.

Definition 4 *The* **Greatest Lower Bound** *for a given message set size m, denoted by $glb(m)$, is the minimum value of s over the sample space of all permutation-based message sets of size m. Similarily, the* **Least Upper Bound** *for a given message set size m, denoted by $lub(m)$, is the maximum value of s over the sample space of all permutation-based message sets of size m.*

The greatest lower and least upper bounds represent absolute limits to the sequence length for a greedy algorithm. These absolute bounds to the sequence length are a function of the message set size and the POPS topology parameters, as specified by the following theorem whose proof is omitted.

Theorem 2 *For a given POPS network and message set $\mathcal{M}$ containing m messages, $glb(m) = \lfloor (m-1)/c \rfloor + 1$, and $lub(m) = min\{m, d\}$.*

For any given message set $\mathcal{M}$ of size ranging from $m = 1$ (a single message) to $m = n$ (a permutation message set), possible sequence lengths are represented by integral values of s that lie in the range $glb(m) \leq s \leq lub(m)$. The probability distribution of s in that range is an important measure of the performance of a POPS network. It determines the average and variance of the communications delay for sets of messages of size m.

4.2 Sequence Length Probability.

Assume a POPS topology with fixed n and d, and consider a set $\mathcal{M}$ of m messages with distinct source nodes and distinct destination nodes. Denote by $\aleph$,

the set of all possible message sets containing m messages. Further, denote by $\aleph_s$, the set of all message sets containing m messages that require a sequence of length s for delivery. Thus, $\aleph = \bigcup_{glb(m) \le s \le lub(m)} \aleph_s$.

Lemma 3 *The probability that a random message set $\mathcal{M}$ of size m will require a sequence of length s for delivery is $\rho(m,s) = |\aleph_s| \,/\, |\aleph|$.*

The number of ways to choose the source nodes for a message set $\mathcal{M}$ is $C_m^n = n!/(n-m)!m!$, the number of unordered ways to select m nodes from n nodes. For each of these choices, the number of ways to select the destination nodes for the message set is $P_m^n = n!/(n-m)!$, the number of ordered ways to select m nodes from n nodes. Thus, $|\aleph| = P_m^n C_m^n = (n!)^2/((n-m)!)^2 m!$.

To determine $\rho(m,s)$, the calculation of $|\aleph_s|$ is still required. It is possible to enumerate all the message sets in $\aleph_s$, but typically, the number of these message sets makes this impractical. A more efficient approach, based on the enumeration of coupler profiles is given.

Denote by $\mathcal{C}(\aleph) = \{C(\mathcal{M}) | \mathcal{M} \in \aleph\}$, the set of all possible coupler profiles for message sets containing m messages. Also, denote by $\mathcal{C}(\aleph_s) = \{C(\mathcal{M}) | \mathcal{M} \in \aleph_s\}$, the set of all possible coupler profiles for message sets containing m messages that require a sequence of length s for delivery. It follows that the POPS topology imposes several constraints on the elements of $\mathcal{C}(\aleph_s)$.

Lemma 4 *Each element, $C(\mathcal{M}) = (u_0, ..., u_{c-1})$, of the set of message sets, $\mathcal{C}(\aleph_s)$, containing m messages and requiring a sequence of length s for delivery, should satisfy the following conditions.*

1. $\forall i, 0 \le i \le c-1$, then $u_i \le s \le d$
2. $\exists i, 0 \le i \le c-1$, such that $u_i = s$
3. $\forall i, 0 \le i \le g-1$, then $\sum_{j=0}^{g-1} u_{ig+j} \le d$
4. $\forall j, 0 \le i \le g-1$, then $\sum_{i=0}^{g-1} u_{ig+j} \le d$
5. $\sum_{i=0}^{g-1} \sum_{j=0}^{g-1} u_{ig+j} = m$.

The first condition says that the number of messages through any coupler is at most s, else a sequence length of s could not deliver that coupler's messages. Further, s is at most d, since only one message in a message set can arrive at each of a coupler's input ports. The second condition says that at least one coupler must handle s messages, else a sequence length of s would not be required for the message set. The third condition says that the total number of messages through couplers connected to any specific source group is at most d, since each node can initiate at most one message. Similarily, the fourth condition says that the total number of messages through couplers connected to any specific destination group is at most d. Finally, the fifth condition says that the total number of messages through all couplers must be m.

In general, each coupler profile $C \in \mathcal{C}(\aleph_s)$ supports more than one message set $\mathcal{M} \in \aleph_s$. For example, consider a message set where two of the messages use a single given coupler. Another message set supported by the same coupler profile, would have the same two destinations reversed. The set of message sets supported by a specific element, C, of the set of coupler profiles, $\mathcal{C}(\aleph_s)$, is denoted $\mathcal{M}_C$. The elements of $\mathcal{C}(\aleph_s)$ yield a partitioning of $\aleph_s$ by the message sets $\mathcal{M}_C$.

The partitioning of $\aleph_s$ implies $\aleph_s = \bigcup_{C \in \mathcal{C}(\aleph_s)} \mathcal{M}_C$. Further, $|\aleph_s| = \left| \bigcup_{C \in \mathcal{C}(\aleph_s)} \mathcal{M}_C \right| = \sum_{C \in \mathcal{C}(\aleph_s)} |\mathcal{M}_C|$. The determination of $|\aleph_s|$, and thus of $\rho(m,s)$, can be achieved by summing the values of $|\mathcal{M}_C|$ during an enumeration of the C in $\mathcal{C}(\aleph_s)$. The value of $|\mathcal{M}_C|$ is now specified.

Lemma 5 *Given n and d for a POPS topology, and a coupler profile $C = (u_0, u_1, ..., u_{c-1}) \in \mathcal{C}(\aleph_s)$. The number of message sets supported by C is,*

$$|\mathcal{M}_C| \;=\; \prod_{i=0}^{g-1} \prod_{j=0}^{g-1} \left[C_{u_{ig+j}}^{d-\alpha} P_{u_{ig+j}}^{d-\beta} \right]$$

Where $\alpha = \sum_{l=0}^{i-1} u_{lg+j}$, and $\beta = \sum_{l=0}^{j-1} u_{ig+l}$.

Proof: The enumeration of the message sets by the combinatoric form is derived as follows:

- $\prod_{i=0}^{g-1}$ considers each coupler group in the POPS.
- $\prod_{j=0}^{g-1}$ considers each coupler in the group.
- The large brackets enumerate the contribution of a single coupler.
  - $C_{u_{ig+j}}^{d-\alpha}$ considers the ways to choose the u_{ig+j} inputs from the d source group nodes, minus the total of previous usage of that source group.
  - $P_{u_{ig+j}}^{d-\beta}$ considers the ways to map the u_{ig+j} inputs selected to the d destination group nodes, minus the total of previous usage of that destination group. $\diamond$

Combining the above results, the probability that a random message set of size m will require a sequence of length s for delivery, $\rho(m,s)$, can be specified.

Theorem 3 (The Probability Calculation Theorem)

$$\rho(m,s) \;=\; \frac{\left[\sum_{C \in \mathcal{C}(\aleph_s)} \prod_{i=0}^{g-1} \prod_{j=0}^{g-1} \left[C_{u_{ig+j}}^{d-\alpha} P_{u_{ig+j}}^{d-\beta} \right] \right]}{[P_m^n C_m^n]}$$

Where $\alpha = \sum_{l=0}^{i-1} u_{lg+j}$, and $\beta = \sum_{l=0}^{j-1} u_{ig+l}$. $\diamond$

The coupler profile approach, which enumerates the profiles in $\mathcal{C}(\aleph_s)$, is more efficient than a complete enumeration of all message sets in $\aleph_s$. Moreover, probabilities of the *glb* and the *lub* for certain special values of m can be directly calculated by the coupler profile approach. This is due to various additional restrictions being appropriate at these positions. A couple of these special degenerate cases will be considered.

When the message set size is a multiple of the number of couplers, the greatest lower bound occurs when the messages are evenly distributed among the couplers. In each of these cases, where all coupler usages are identical, only one coupler profile is possible.

This restriction provides for a major simplification of Lemma 5, and consequently, The Probability Calculation Theorem.

Corollary 2 *Given n and d for a POPS topology, $\forall m$, such that $m = xc$, where $1 \leq x \leq n/c$,*

$$\left|\aleph_{glb(m)}\right| = \prod_{i=0}^{g-1} \prod_{j=0}^{g-1} \frac{(d-ix)!(d-jx)!}{(d-ix-x)!(d-jx-x)!x!}$$

For small message sets, $m \leq d$, the least upper bound increases linearly with the message set size. Under these conditions, all the messages in a message set pass through a single coupler. In each of these cases, where only one coupler usage is not zero, c coupler profiles are possible. This restriction provides for another major simplification of Lemma 5, and consequently, The Probability Calculation Theorem.

Corollary 3 *Given n and d for a POPS topology, $\forall m, 1 \leq m \leq d$, then*
$\left|\aleph_{lub(m)}\right| = (n/d)^2 C_m^d P_m^d = n^2(d!)^2/d^2((d-m)!)^2 m!$.

4.3 Probability Experiments.

For POPS topologies of low complexity (relatively small n or large d), a program based on The Probability Calculation Theorem is practical. For a given POPS topology, this calculator program enumerates all coupler profiles of a specific message set size, and is more efficient than an enumeration of all possible message sets. After the profile enumeration, $\rho(m,s), \forall s, glb(m) \leq s \leq lub(m)$ are tabulated.

The exhaustive nature of the calculation-based program is not practical for more complex systems (for example, when $c > 32$). A program which estimates the sequence length probability was also developed. For a given POPS topology, random message sets of a specific size (permutations, or subsets thereof) are generated, then resulting coupler profiles are used to estimate $\rho(m,s), \forall s$ such that $glb(m) \leq s \leq lub(m)$.

The nature of the distribution of sequence length probabilities is extremely consistant in the range of POPS topologies considered. Three characteristics are observed across a wide range of system sizes, coupler degrees, and message set sizes. First, the sequence length extremes, *glb* and *lub*, have extremely low probabilities. Second, all non-trivial probabilities occur in a small range of sequence lengths. Third, sequence lengths with non-trivial probabilities constitute a small fraction of the possible sequence lengths, these being among the very shortest.

Corollary 2 calculates the probability that, given a random permutation-based message set whose size is some multiple of the number of couplers, the minimum possible sequence length for delivery is required. Similarily, Corollary 3 calculates the probability that, given a random permutation-based message set of some fixed size ($m \leq d$), the maximum possible sequence length for delivery is required. A short survey of the probability that the sequence length will be the greatest lower bound or the least upper bound for several POPS topologies indicates that such probabilities are very low.

In general, estimator program data shows that all non-trivial probabilities occur in a very narrow range of sequence lengths. Further, the distribution involves relatively few of the total number of possible sequence lengths, and those involved are the shortest ones. For example, the small POPS network where $n = 32, d = 16$, and $m = 32$, delivers more than 98% of the possible message sets in four sequence lengths (8 through 11). These represent the lowest possible sequence lengths ($glb = 8 \leq s \leq 16 = lub$).

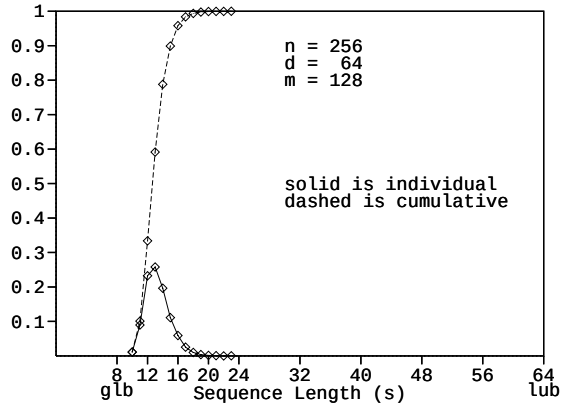

Figure 3: Typical Sequence Length Probabilities with $n = 256, d = 64$ and $m = 128$ ($g = 4, c = 16$).

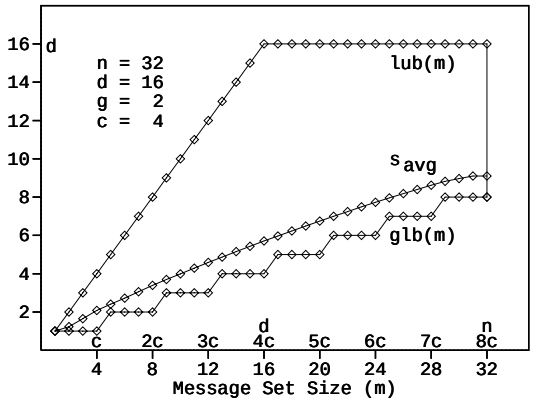

Figure 4: Avg. Seq. Length for $1 \leq m \leq n$.

These distribution characteristics become even more pronounced in higher-complexity POPS networks. Figure 3 shows a typical distribution of sequence length probabilities for a fairly-large POPS topology. For the specified n , d and m, possible sequence lengths range from 8 to 64. The most probable sequence length of 13 delivers over 25% of the possible message sets. In fact, over 88% of the message sets can be delivered by one of only five sequence lengths (11 through 15), more than 98% in ten sequence lengths (8 through 17). This represents the shortest 18% of the valid sequence lengths.

For a given POPS topology and a message set $\mathcal{M}$ of size m, let the average sequence length be the probability-weighted arithmetic mean of all possible sequence lengths, $s_{avg} = \sum_{s=glb(m)}^{lub(m)} s \cdot \rho(m,s)$. The consistent nature of the non-trivial part of the probability distribution as described above, makes the average sequence length a suitable and adequate measure of the performance of a POPS topology.

Figure 4 shows the *glb* , the *lub* , and the average sequence length s_{avg} in the range $m = 1, ..., n$ for a small POPS network. In this example, for any size message set, the average sequence length is always less than two, and often only one, unit larger than the greatest lower bound. Typically, this is only about 10% of the sequence length range. Similar and even more pronounced behaviour is observed in larger systems. Figure 5 shows the average sequence length for seven large POPS networks involving $n = 128, 256, 512, 1024$ and $d = 64, 128$.

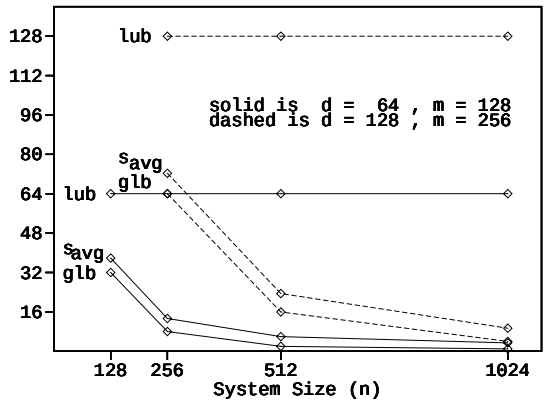

Figure 5: Avg. Seq. Length for Various POPS.

As an example of high performance in a very large POPS topology, consider a system with 1024 nodes and a coupler degree of 64 (16 groups, 256 couplers), and message sets containing 512 messages. In this system, message sets require sequence lengths of at least 2 and at most 64. Simulation results showed the most probable sequence length to be 7. This sequence length delivers 45.1% of the possible message sets.

5 Scalability Issues for POPS.

Given a fixed system size n, the choice of d in a POPS topology determines the maximum network throughput available. Lower d yields more couplers and a higher potential application throughput, but at the cost of higher system complexity (more couplers and links).

Typically, for any fixed n, many POPS topologies may be possible through the choice of d. The method of selecting d, as system size increases, is called a scaling rule. Three varied scaling rules are presented, and their effects on system complexity, network performance, and efficiency are studied.

The first scaling rule, called fixed-g, maintains a fixed number of node groups. This results in a fixed node degree (number of transmitters and receivers) and increases the coupler degree as system size increases. The second scaling rule, called fixed-d, maintains a fixed coupler degree. This results in both an increasing number of groups and an increasing node degree as system size increases. The third and final scaling rule, called the root-n rule, increases the coupler degree as the square root of the increasing system size. This also results in both an increasing number of groups and an increasing node degree. The following table summarizes these characteristics for the three scaling rules.

Characteristic	Fixed-g	Fixed-d	Root-n
# Couplers	*fixed*	$O(n^2)$	$O(n)$
Coupler Deg.	$O(n)$	*fixed*	$O(\sqrt{n})$
Node Deg.	*fixed*	$O(n)$	$O(\sqrt{n})$

Figure 6 illustrates the performance, for permutation message sets ($m = n$), of POPS topologies which follow one of the three scaling rules to determine the coupler degree. For each scaling rule, two examples are shown, one by a solid line and one by a dotted line. These curves represent the average sequence length for the scaling rules over a wide range of fairly large system sizes. Figure 7 illustrates the coupler utilizations for curves 1, 3, and 5 of figure 6.

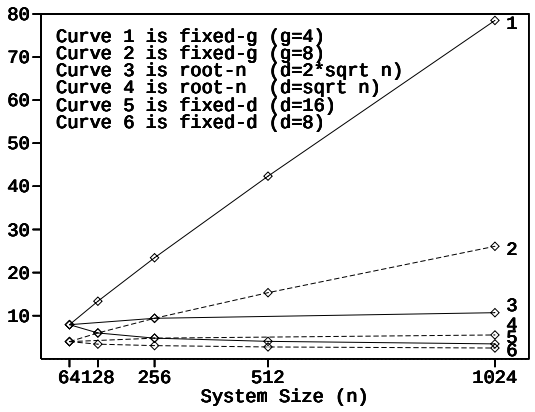

Figure 6: Scaling Rules Performances (Sequence Length Average).

Under the fixed-g scaling rule, as system size becomes very large, the coupler degree becomes impractical (for $n = 1024, d > 100$). Though network complexity is minimized, and Figure 7 shows coupler utilization is very high, practical aspects concerning power dissipation and availability limit this approach. Further, as seen in Figure 6, the fixed-g scaling rule has only modest performance.

Under the fixed-d scaling rule, as system size becomes very large, the number of couplers and links becomes impractical (for $n = 1024, c > 4000$, and # of links $> 130,000$). Commercially-available couplers

may be utilized, but the network complexity becomes prohibitive, and Figure 7 shows that the utilization becomes unacceptably low. Nonetheless, Figure 6 does indicate that the fixed-d rule provides excellent performance.

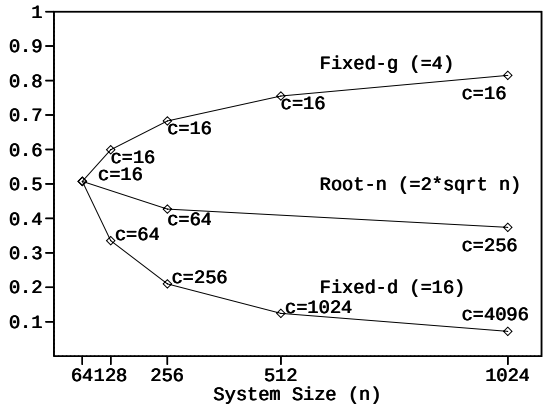

Figure 7: Scaling Rules Utilizations.

Using the root-n scaling rule, the coupler degree is practical even for very large system sizes (for $n = 1024, d \leq 64$), and the number of couplers is linearly related to the number of nodes. Very good performance which approximates the fixed-d rule is observed in Figure 6, and good utilization is maintained even for very large systems, as shown in Figure 7.

In summary, for very large system sizes, the fixed-g scaling rule is limited by practical aspects of the coupler degree, and the fixed-d scaling rule is restricted by network complexity. However, the effect of the root-n scaling rule is an efficient high-performance POPS topology for a wide range of large system sizes. These root-n POPS topologies use commercially-available couplers and have network complexities proportional to their system size.

6 Conclusion.

The POPS topology is an interconnection design suitable for implementation using current optical technology. It is an all-optical non-hierarchical star-based topology. A powerful configurability is provided which allows a critical design tradeoff between low coupler-degree networks, with higher throughput and lower power dissipation, and high coupler-degree networks, with lower system complexity.

In a POPS topology, the route determination is trivial and message scheduling is simple. This latter fact is due to the independence of individual coupler scheduling. These characteristics promote a very low control overhead. Absolute bounds of the sequence length for delivery of a random permutation-based message set can be easily calculated.

For random permutation patterns, results consistantly show a high and narrow sequence length distribution centered relatively close to the theoretical minimum. Surveys of system size, coupler degree, and message set size show that this results in an average sequence length close to the greatest lower bound under almost any conditions. Even when considering very large system sizes, scaling methods exist which provide powerful combinations of system characteristics. Low system complexity, high utilization and high throughput are all possible in a POPS topology.

References

[1] E. Arnould, F. Bitz, E. Cooper, R. Sansom, and P. Steenkiste. The design of nectar: A network backplane for heterogeneous multicomputers. In *Proc. of 3rd Int. Conf. on Arch. Support for Prog. Lang. and Oper. Sys.*, pages 205–216. IEEE, 1989.

[2] Y. Birk. Power-optimal layout of passive, single-hop, fiber-optic interconnections whose capacity increases with the number of stations. In *Infocom '93: 12th Joint Conf. of the Comp. and Comm. Societies*, pages 565–572. IEEE, 1993.

[3] 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 *Proc. of 1st Int. Workshop on Mass. Par. Proc. Using Intercon.*, pages 70–80. IEEE, April 1994.

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

[5] P. Green. *Fiber Optic Networks.* Prentice Hall, Englewood Cliffs, NJ, 1993.

[6] R. Handel and M. Huber. *Integrated Broadband Networks: An introduction to ATM-based Networks.* Addison-Wesley, Reading, MA, 1990.

[7] M. Hluchyj and M. Karol. ShuffleNet: An application of generalized perfect shuffles to multihop lightwave networks. In *Infocom '88*, pages 379–390. IEEE, 1988.

[8] A. Holmes and R. Syms. All-optical CDMA using quasi-prime codes. *Jour. of Lightwave Tech.*, 10(2):279–286, February 1992.

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

[10] C. Qiao and R. Melhem. Reconfiguration with time division multiplexing MINs for multiplexing communications. *IEEE Trans. on Par. and Dist. Sys.*, 5(4):337–352, 1994.

[11] M. Sexton and A. Reid. *Transmission Networking: SONET and the Synchronous Digital Hierarchy.* Artech House, Norwood, MA, 1992.