Optical Multistage Interconnection Networks: New Challenges and Approaches

Yi Pan*
Department of Computer Science
University of Dayton, Dayton, OH 45469-2160
Email: pan@cps.udayton.edu

Chunming Qiao†
Department of Computer Science and Engineering
University at Buffalo (SUNY), Buffalo, New York 14260
Email: qiao@computer.org

Yuanyuan Yang‡
Department of Computer Science
University of Vermont, Burlington, VT 05405
Email: yang@emba.uvm.edu

Abstract

Optical interconnections for communication networks and multiprocessor systems have been studied extensively. A basic element of optical switching networks is a directional coupler with two inputs and two outputs (hereafter referred to simply as switching elements or SEs). Depending on the control voltage applied to it, an input optical signal is coupled to either of the two outputs, setting the SE to either the straight *or the* cross *state. A class of topologies that can be used to construct optical networks is multistage interconnection networks (MINs), which interconnect their inputs and outputs via several stages of SEs. Although optical MINs hold great promises and have demonstrated advantages over their electronic counterparts, they also introduce new challenges such as how to deal with the unique problem of avoiding* crosstalk *in the SEs. In this paper, we survey the research carried out, including major challenges encountered and approaches taken, during the past few years on optical MINs.*

1 Introduction

Electronic multistage interconnection networks (MINs) have been studied extensively as an important interconnecting scheme for communication and parallel computing systems [10]. Fiber optic

*Research supported in part by the National Science Foundation (NSF) under Grants CCR-9211621, OSR-9350540 and CCR-9503882, by the Air Force Office of Scientific Research under grant F49620-93-C-0063, by the Air Force Avionics Laboratory, Wright Laboratory, under Grant F33615-C-2218, and by Ohio Board of Regents through the Investment Fund Program and the Research Challenge Grant Program.

†Research supported in part by a grant from NSF Research Initiation Award (RIA) under contract MIP-9409864 and ANIR-9801778.

‡Research supported in part by the U.S. Army Research Office under Grant No. DAAH04-96-1-0234 and by NSF under Grant No. OSR-9350540 and MIP-9522532.

communications with photonic switching promise to meet the increasing demand of communication systems, and received much attention in parallel processing community as well. An optical MIN can be implemented with either free-space optics or guided wave technology [9]. In this paper, we consider optical implementation with guided wave technology. Two types of guided wave optical switching systems can be identified. The first is a hybrid (photonic) approach in which optical signals are switched, but both the switch control and routing decisions are carried out electronically at a speed that can be much lower than the bit rate of the optical signals being switched. The second approach is all-optical switching, which would potentially overcome the speed-mismatch problem associated with the hybrid approach. However, such systems are not likely to become practical in the near future [9], and hence only the hybrid optical MINs are considered in this paper.

In an electronic MIN, it is common to use packet switching [10]. However, in hybrid optical MINs which uses electronically controlled optical switching elements (SEs), such as Lithium Niobate directional couplers, packet switching requires conversions between optical signals and electronic ones, which could be very costly. Compared with the switching speed of the SEs, which can be as fast as hundreds of picoseconds, the process of determining their settings based on the address information in each packet could become a significant overhead.

For these reasons, circuit switching is usually preferred in optical MINs, with which a direct connection between the source and the destination is set up by a network controller (or controllers) before data are sent. Since neither packet processing nor buffering are needed at each SE, circuit switching can be implemented with SEs that are simpler and faster than those required for packet switching. Nevertheless, data transmission can still be packetized in order to utilize the network bandwidth better by letting multiple connections time-share an input/output port of a switch or a link between two switches.

In this paper, we first describe unique characteristics of optical MINs and in particular, the crosstalk introduced by the SEs in the next section. We then discuss general approaches to avoiding crosstalk via network dilation and especially the time domain dilation approach in Section 3. In Section 4, the ability of time domain dilated optical MINs to embed regular structures such as rings, meshes and trees, and to realize permutations is presented. In Section 5, algorithms for establishing an arbitrary set of connections are studied and numerical results from performance analysis and simulation are presented. Finally, we conclude our paper by identifying new research topics in the area of optical MINs.

2 Unique Characteristics of Optical MINs

Wide-band optical signals can be switched under electronic control using directional couplers between Ti:$LiNbO_3$ waveguides on a planar $LiNbO_3$ crystal [1]. The basic SE is a directional coupler with two active inputs and two active outputs. Depending on the amount of voltage at the junction of the two waveguides which carry the two input signals, either of the two inputs can be coupled to either of the two outputs. Many architectures have been proposed to construct an $N \times N$ MIN using the 2×2 directional coupler as the basic component. These architectures are essentially similar to those of electronic MINs.

Much research has been done on electronic MINs in the literature. Some of the analytical methods and results are also applicable to optical MINs. However, optical MINs also hold their own challenges. For example, one problem is *path dependent loss*. In a large MIN, a substantial part of this path dependent loss is directly proportional to the number of couplers along an optical path, which is determined by the architecture used and the network size.

Another problem in optical MINs is *optical crosstalk*, which occurs when two signal channels interact with each other. There are two ways in which optical signals can interact in a planar switching network. The channels carrying the signals could cross each other in order to embed a particular topology. Alternatively, two paths sharing a SE will experience some undesired coupling from one path to another within a SE.

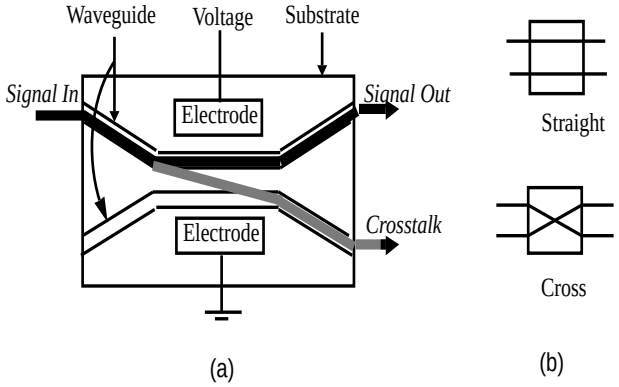

Figure 1: Crosstalk in an electro-optical switching element (SE).

Figure 1 shows an example of crosstalk in a SE. In the figure, the SE is set to straight (Figure 1(b)), the main signal is injected at the upper input as shown in Figure 1(a), and a crosstalk signal having a small fraction of the input signal power may be detected at the lower output. Hence, when a signal passes many SEs, the input signal will be distorted at the output due to the loss and crosstalk introduced on the path. Experimental results [4] show that it is possible to make the crosstalk from passive intersections of optical waveguides negligible by keeping the intersection angles above a certain minimum amount. Studies also indicate that the crosstalk problem is more severe than the path dependent loss problem with current optical technology [1], [9]. Thus, switch crosstalk is the most significant factor which reduces the signal-to-noise ratio and limits the size of a network. This unique characteristics in an optical MIN lead to different design and analysis approaches from those used in its electronic counterpart.

3 Approaches to Avoiding Crosstalk

To reduce the negative effect of crosstalk, various approaches which apply the concept of *dilation* in either the space or time domains have been proposed [4, 5]. With the space domain approach, extra SEs (and links) are used to ensure that at most one input and one output of every SE will be used at any given time. With the time domain approach, the same objective is achieved by treating crosstalk as a conflict, that is, two connections will be established at different times if they use the same SE (even if they use different inputs and outputs of the SE). With the wavelength domain approach, only the wavelength channels that are far apart from each others are assigned to the same SE. The following discussions will concentrate on the first two approaches.

3.1 Space and Time Domain Dilations

As an example of dilating a MIN in the space domain, Figure 2 shows a 2×2 dilated Benes network (DBN) proposed in [4]. The idea is to duplicate hardware to avoid crosstalk. Note that

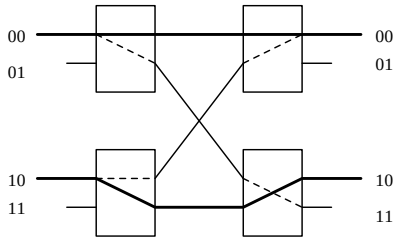

Figure 2: A 2×2 DBN.

although the DBN has four inputs and outputs, only *half* of them will be available to its users. More specifically, only inputs and outputs numbered, in binary, 00 and 10 are used for establishing two connections. This way, the crosstalk signal generated from one connection will not interfere with the signal carried by the other connection. Its construction is recursive in nature. A 2×2 DBN, which is a basic building block, has four inputs and four outputs but can only be used for two connections. An $N \times N$ DBN has $2N$ inputs and outputs although only N of them are used for source and destination connections. The network can be constructed from two $N/2 \times N/2$ subnetworks by putting one on the top of another, and interconnecting the two subnetworks with a front and an end stages as depicted in Figure 3. The number of stages in an $N \times N$ DBN, $M(N)$, satisfies the recursively defined equation, $M(N) = M(N/2) + 2$ where $M(2) = 2$. By solving the equation, we have $M(N) = 2 \log N$, which is referred to simply as M in Figure 3.

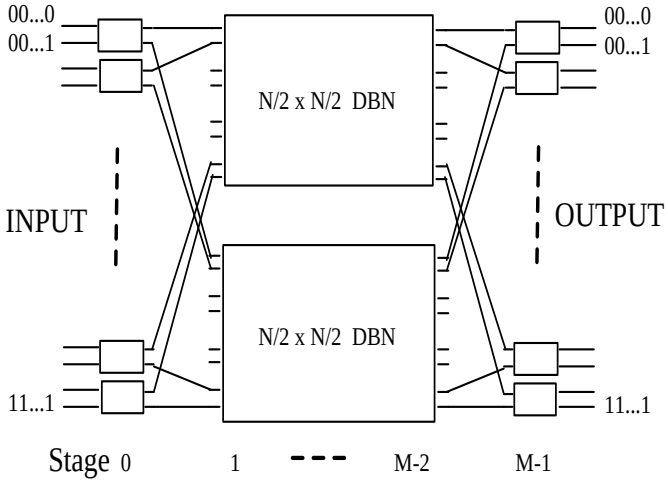

Figure 3: A recursively constructed $N \times N$ DBN.

Since signals go to only one of the two inputs of the SEs at the first and the last stages, crosstalk-free is guaranteed at these two stages. In addition, it is possible to realize any permutation by setting the SEs at each stage such that only one input per SE will be active at any given time, or in other words, crosstalk will be avoided at all stages.

A blocking MIN such as the Banyan may also be dilated in a similar way (see Figure 4(a) and (b)). A dilated Benes (or Banyan) network can realize the same set of permutations crosstalk-free as a regular Benes (or Banyan) can with crosstalk. In either case, the number of SEs (and links) in a dilated network, which is slightly larger than twice of the regular one, may be considered to represent the space cost for crosstalk avoidance. There are two variations of spatially dilated

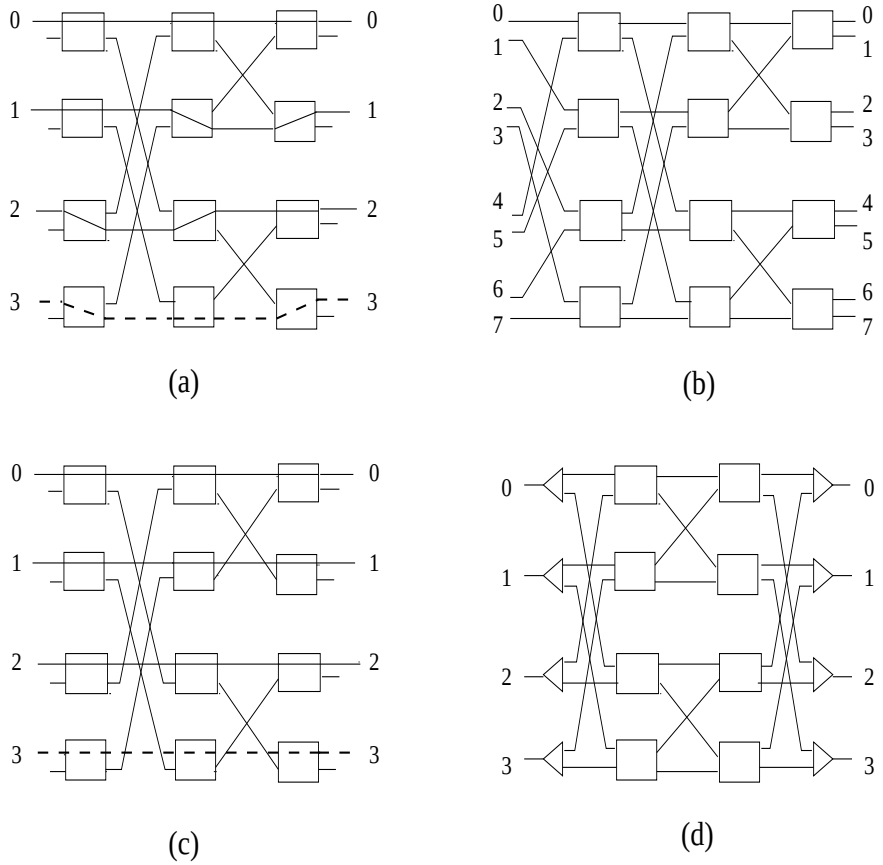

Figure 4: Four switching networks: (a) A 4×4 dilated Banyan. (b) An 8×8 Banyan. (c) A 4×4 Dilated Slipped Banyan (DSB) and (d) A two-plane 4×4 Banyan.

blocking networks: one is the Dilated Slipped Banyan (DSB) [7] shown in Figure 4(c), and the other is a two-plane Banyan shown in Figure 4(d). In these space domain dilated networks, the crosstalk problem is avoided using more hardware.

It should be obvious that no permutation can be realized crosstalk-free in a regular MIN since it requires that at most one input and one output of any SE can be active at a time. However, using the time domain dilation approach [5], which is a special instance of a connection paradigm called Reconfiguration with TDM (or RTDM) proposed by Qiao, et al., a permutation may be partitioned into two (or more) sub-permutations such that each can be realized in a MIN crosstalk-free. In this way, a permutation can be realized in a regular MIN crosstalk-free in several rounds, one for each such sub-permutation, which will be referred to as a crosstalk-and-conflict-free (CF) mapping. The number of rounds (or CF-mappings) needed for any permutation, which is at least two, may be considered to represent the time cost for crosstalk avoidance. Similar approaches for avoiding crosstalk by using more hardware have also been studied in the design of photonic MINs capable of interchanging time slots in [7, 2, 3].

3.2 Space-Time Trade-offs and Implications

As to be discussed in Section 4.2, it was shown in [5] that the number of permutations that can be partitioned into two CF-mappings in a regular Banyan exceeds the number of permutations that can be realized crosstalk-free (in one round) in a dilated Banyan. This seems to indicate that if the cost in time (in terms of the number of rounds) and in space (in terms of the number of SEs and links) were interchangeable, then the time domain dilation is more effective in realizing permutations than the space domain dilation. In addition, if hardware complexity is not a concern, then a two-plane Banyan may be used instead of the dilated Banyan since both have approximately the same space cost, but the former can be more powerful.

Time domain dilation is not only useful for avoiding crosstalk in realizing permutations in blocking MINs, but also for establishing an arbitrary set of connections that would normally cause conflicts in blocking or nonblocking MINs (dilated or not) by partitioning the set into several CF-mappings. Let T_d and T_u be the average time cost of a set of arbitrary connections in a dilated Banyan and an undilated (i.e. regular) Banyan, respectively. Analysis to be presented later in Section 5 has shown that $T_u < 2T_d$. Since the space cost of the former, denoted by S_d, is more than twice of that of the latter, denoted by S_u, we have $S_u \cdot T_u < S_d \cdot T_d$, which implies that the time domain dilation may be more cost-effective provided that the cost in time and in space were interchangeable. An intuitive explanation is that the time domain approach integrates the solutions to the crosstalk and path conflict problems while space domain dilation only deals with crosstalk avoidance and still relies on time multiplexing techniques to resolve path conflicts.

The above discussions have the following implications on the implementation of optical MINs. First, whenever the limit on the network size is reached, the time domain approach may be used as a viable way to trade the maximal bandwidth available to each individual input and output pair for increased connectivity. Secondly, it is useful when future technology allows the transmission rate to scale up faster than the network size, or in other words, when the cost of increasing the bandwidth of each connection becomes as "cheap" as (or even cheaper than) the cost of constructing a network of twice of its original size.

For example, assume that one has a 16×16 dilated Banyan or a 32×32 Banyan, which has essentially the same hardware complexity. To connect 32 sources with 32 destinations using the former in the space domain dilation approach, two sources (and destinations) need to be multiplexed onto a single input (and output, respectively). An alternative is to use the latter as in the time

domain dilation approach, thereby letting each source and destination connect to its own input and output, but of course, only one of the two sources (destinations) sharing the same SE can transmit (receive) at a time. Assume that the channel bandwidth (as determined by the transceiver rate) is 2.5 Gb/s, then theoretically speaking, in both cases, the bandwidth available to each source and destination pair would be 1.25 Gb/s when traffic is evenly distributed among all inputs and outputs. Based on the above cost-effective analysis, however, using the time domain dilation approach may achieve a better bandwidth utilization, or in other words, results in a higher effective bandwidth between a source-destination pair.

Continue the above example and assume that there is a need for network evolution in which the bandwidth for a source-destination pair can be increased to 2.5 Gb/s. Instead of trying to build a 32×32 dilated Banyan (or equivalently a 64×64 Banyan), which may not be feasible, the time domain dilation approach can be used, which deploys 5Gb/s transceivers but still uses the 32×32 Banyan. Of course, it is possible that increasing the network size is feasible and hence, the space domain approach is more desirable. However, if hardware complexity size is not a concern, a two-plane Banyan may be used instead as mentioned above.

4 Establishing Connections with Regular Patterns

To help close the gap between the relatively slow electronic processing speed and the high bandwidth of an optical MIN, the RTDM paradigm, which is a generalization of the time domain approach for avoiding crosstalk, was proposed by Qiao, et al. Specifically, when the set of connections cannot be established in a MIN due to conflicts, it is desirable to partition it into a *minimum* number of CF-mappings. Once these CF-mappings and corresponding network configurations are determined, a sequence of control signals needed to set each SE appropriately can be stored in a cyclic shift-register. At run time, the MIN can simply go through a sequence of configurations under global synchronization without incurring much overhead that would have been introduced had complex electronic processing is involved.

In this section, we examine the ability of optical MINs, especially Banyans, to emulate (or embed) regular structures such as rings, meshes and trees, and to realize permutations.

4.1 Emulation of Common Structures

There are two basic control modes applicable to MINs. The first is called *switch control*, under which the switches in the network are controlled independently of each other, and hence can be set to different states at any given time. The second is called *stage (or column) control* under which, the switches at the same stage are controlled by one common signal, and hence have to be in the same state at any given time.

The DSB network (shown in Figure 4(c)) was proposed to facilitate *stage control* of a dilated Banyan, thus requiring only one control bit (e.g. [0] for straight and [1] for cross) per stage, and more importantly, one electronic driver circuit per stage. An $N \times N$ DSB can emulate a fully-connected network of the same size by applying the time domain approach (along with the space domain approach). Specifically, the set of $N(N-1)$ connections in the fully-connected network can be partitioned into N CF-mappings as follows: the k-th CF-mapping, where $0 \leq k \leq N-1$, contains connections from input i to output j as long as $i \oplus j = k$, where $\oplus$ is the bit-wise exclusive-OR operation. The control word (CW) to set the corresponding configuration in the DSB is $CW = [0]k \oplus [k]0$. By going through a sequence of N configurations corresponding to the

N CF-mappings in a time division multiplexed fashion, the DSB provides the equivalent of the full-connectivity.

While full-connectivity may be reasonable for certain applications such as using the optical MIN as a hub in a local-area network environment, many programs in a parallel and distributed computing environment, however, exhibit communication locality and regularity and do not require the full connectivity. Accordingly, we can reduce the number of CF-mappings needed and thus increase the network bandwidth utilization in most cases.

Given that many algorithms have been proposed for popular structures such as ring, mesh, hypercube and binary tree, optimal emulation of these structures in a Banyan by applying the time domain approach has been considered. For example, while a fully-connected network requires $2N$ CF-mappings, a ring can be emulated with only two CF-mappings, a mesh with only four CF-mappings, a cube-connected-circle (CCC) with three CF-mappings, and a hypercube with only $\log N$ CF-mappings. In addition, a complete binary tree of $N-1$ nodes can be emulated with four CF-mappings using an elaborate procedure involving nested recursions. Such an emulation is shown to be optimal when in-order labeling of the tree nodes is used, but may or may not be improved to involve only three CF-mappings when other labeling orders are used.

We will discuss issues related to realizing permutations in the next subsection and those related to establishing a set of arbitrary connections in Section 5.

4.2 Permutation Capability

As mentioned earlier, a dilated Banyan (and Benes) can realize the same set of permutation crosstalk-free as an Banyan (and Benes, respectively). In [5], a one-to-one but not onto mapping between the set of permutations realizable in a dilated Banyan (or Banyan) and the set of permutations that can be partitioned into two CF-mappings was developed, which showed that the latter contained more permutations than the former.

In this subsection, we approach this problem from a different angle and consider the permutation capability of undilated networks. An interesting question is: What is minimum number of passes required for realizing a permutation in such a network? In other words, we are interested in what types of partial permutations could be possibly realized crosstalk-free in an optical MIN. Recently, Yang, Wang and Pan [11] introduced a concept called *semi-permutation*. A semi-permutation is a partial permutation that ensures that there is only one active link passing through each input SE and output SE, and thus it has the potential to be realized crosstalk-free in an optical MIN. They have shown that any permutation can be decomposed into semi-permutations and the total number of semi-permutations for an even integer N is $2^N \cdot (\frac{N}{2})!$.

A simple algorithm for decomposing a permutation into semi-permutations is also described in [11]. The basic idea is to construct a bipartite graph for the given permutation between N inputs and N outputs of the network. Then, for each connected component of the graph, start from a vertex of this component in the input sets, traverse through an unvisited edge to the neighbor vertex in the output sets, back and forth until we return to the starting vertex. (During the traversing, a visited edge is marked "forward" if the traverse direction on this edge is from V_1 to V_2; and marked "backward" if the direction is opposite.) Finally, take all one-pair mappings corresponding to the edges marked with "forward", to form one semi-permutation; let the remaining one-pair mappings, corresponding to the edges marked with "backward", form another semi-permutation. It is easy to see that the complexity of the above decomposition algorithm is $O(N)$.

Thus, the problem of realizing a permutation in a crosstalk-free network can be transformed into the problem of realizing semi-permutations in the crosstalk-free network. However, it should

be pointed out that introducing the semi-permutation concept in a network composed of 2×2 SEs can only guarantee crosstalk-free in the SEs in the input stage and the output stage of the network. In fact, realizing a semi-permutation in a single pass implies that there is only one active input on each SE in the input stage and only one active output on each SE in the output stage. To ensure the entire network crosstalk-free, we need to know if there exists a proper routing that can eliminate crosstalk in the SEs in the intermediate stages along different active paths. We look into this issue for two different types of networks: Banyan and Benes.

Due to the unique path nature of a Banyan network, a semi-permutation is routed through the network in a fixed switch setting. Consequently, some semi-permutations can be realized in a Banyan network in a single pass while others cannot. Yang, Wang, and Pan [11] showed that the number of semi-permutations that can be realized crosstalk-free in an $N \times N$ Banyan network in a single pass is $2^{\frac{3}{4}N} \cdot N^{\frac{N}{4}}$. By comparing the number of semi-permutations that can be realized in an $N \times N$ Banyan network and the number of all possible semi-permutations for an N-element set, we can see that there are a substantial amount of semi-permutations that cannot be realized in a Banyan network, especially when N gets larger.

A Benes network can be constructed by concatenating a Banyan network and a reverse Banyan network with the center stages overlapped. Electronic Benes networks are well known for being capable of realizing all possible permutations [10]. It is not surprising that Benes networks also have good properties to support permutations in optical networks. Yang, Wang, and Pan [11] have shown that any semi-permutation can be realized crosstalk-free in a Benes network in a single pass. They also gave an efficient algorithm for routing semi-permutations. The routing algorithm for a semi-permutation in an $N \times N$ Benes network is obtained by slightly modifying the decomposition algorithm described earlier in this subsection.

Now we know that any permutation can be decomposed into two semi-permutations and that semi-permutation can be realized in a Benes network in a single pass. Therefore, any permutation can be realized crosstalk-free in a Benes network in two passes. It should be pointed out that a permutation requires at least two passes in any $N \times N$ optical MIN due to the constraint of crosstalk-free in the input stage of SEs. In other words, two is the lower bound on the number of passes for any optical permutation networks under the constraint of crosstalk-free. It indicates that an undilated Benes network reaches this lower bound and realizes permutations optimally.

5 Establishing Random Connections

When emulating rings, meshes and trees or realizing permutations, the communication patterns involved have certain regularity. In many applications, communication patterns are irregular in nature. In this section, we discuss issues related to establishing a set of arbitrary connections in optical MINs, especially photonic Banyans.

It is always desirable to schedule an arbitrary set of connections in as few rounds as possible. Such a scheduling problem may be transformed into a graph-coloring problem. However, we are not interested in making such a transformation because an optimal algorithm derived that way usually has a time complexity that is exponential to the number of connections to be established, and thus would be of little or no use in high-speed networks. In fact, no optimal scheduling algorithm with a polynomial time complexity is available under switch control. Nevertheless, an optimal algorithm of a polynomial time complexity exists under stage control.

Let us start with a heuristic algorithm for establishing a set of arbitrary (but distinct) connections in a stage control Banyan network, called Odd-Even. In a Banyan, the control word for each

connection is obtained by exclusive-*ORing* its input binary representation with its output binary representation. For example, the control word required for a connection from input $S = 7[111]$ to output $D = 5[101]$ is $W(S, D) = [111] \oplus [101] = [010]$ implying that stage 1 to stage 3 need to be set to straight($w_2 = [0]$), cross($w_1 = [1]$) and straight($w_0 = [0]$), respectively. Under this setting, input 7[111] is first connected to output port $(1 \oplus 0)11$ (i.e. 7[111]) at stage 1, which in turn is connected to output port $1(1 \oplus 1)1$ (i.e. 5[101]) at stage 2, and finally to output $10(1 \oplus 0)$ (i.e. 5[101]) at stage 3.

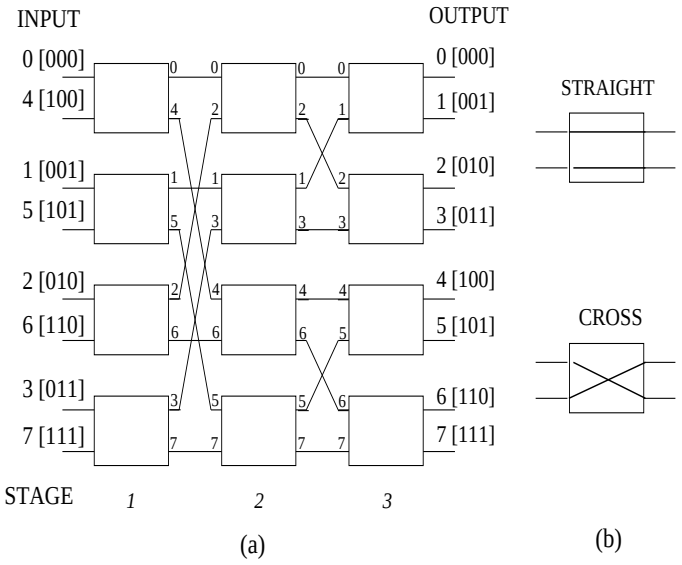

Figure 5: An 8×8 Banyan network and two states of a SE.

From Figure 5, we observe that at every stage, the input and output ports are numbered in the same way. More importantly, the binary representations of the two input (or output) ports of a SE at stage k differ only in bit $n - k$. Accordingly, at stage k, input port $P_{in} = p_{n-1} \ldots p_{(n-k+1)} p_{n-k} p_{(n-k-1)} \ldots p_0$ is connected to output port $P_{out} = p_{n-1} \ldots p_{(n-k+1)} (p_{n-k} \oplus w_{n-k}) p_{(n-k-1)} \ldots p_0$. Accordingly, given two connections (S, D) and (S', D') having the same control word W, they will use a common SE if and only if the *hamming distance* between S and S' is 1. More specifically, they will use a common SE at stage k if and only if (the binary representations of) S and S' differ in bit $n - k$.

For example, the two connections sharing any control word $W[w_2 w_1 w_0]$ that are from $S = 0[000]$ and $S' = 4[100]$, respectively, would use the top SE at stage 1. The two connections from 0[000] and 2[010], respectively, would use the top SE (or the third from the top) at stage 2 if they share control word $[0 w_1 w_0]$ (or $[1 w_1 w_0]$). Similarly, the two from 0[000] and 1[001], respectively, would use the top SE (the second, the third or the bottom SE) at stage 3 if they share control word $[00 w_0]$ ($[01 w_0]$, $[10 w_0]$ or $[11 w_0]$).

A heuristic algorithm called *Odd-Even* works as follows. Partition the connections according to their required control words and their input parities. Specifically, we let $O[i]$ and $E[i]$, where $0 \leq i \leq N - 1$, represent the group of the connections requiring control word i, but whose input parities are odd and even, respectively. The number of connections in $O[i]$ and $E[i]$, ranging from 0 to $N/2$, will be denoted by $|O[i]|$ and $|E[i]|$ respectively.

The heuristic algorithm establishes the connections requiring the same control word in one

round, as long as they have the same input parity. Since the binary representations of the two inputs having the same parity differ by at least two (2) bits, the hamming distance between them is at least 2 (could be 4, 6 and so on). If $|O[i]| > 0$ (or $|E[i]| > 0$), then scheduling the connections in $O[i]$ (or $E[i]$) in one round guarantees that they will be SE-disjoint.

Because in a Banyan whose SEs are already set under stage control, there are exactly $N/2$ connections having even (or odd) input parity. Based on the description of the algorithm, all of the $N/2$ connections can be established in one round. Hence, the maximum number of SE-disjoint connections that can be established in one round under stage-control, $N/2$, can be reached using the Odd-Even algorithm. The time complexity of the algorithm is clearly polynomial of the number of connections to be established.

The Odd-Even algorithm just described may be too conservative since having the same parity is sufficient but not necessary for connections requiring the same control word to be SE-disjoint. For example, it is possible that connections having different input parities, such as $(0, 2)$ and $(7, 5)$, are SE-disjoint (in addition to requiring the same control word), and thus can be established in one round.

The optimal algorithm is an improved version of the Odd-Even algorithm through merging some connections in one pass. The optimal algorithm tries to reduce the number of rounds resulted from using the Odd-Even algorithm by merging $O[i]$ and $E[i]$ into one round. For each control word i ($0 \leq i \leq N-1$), if both $O[i]$ and $E[i]$ are not empty, an attempt is made to establish the connections in the two groups in one round. Specifically, $O[i]$ and $E[i]$ are merged if (and only if) every connection in $O[i]$ is found to be SE-disjoint with every connection in $E[i]$. The attempt to merge $O[i]$ and $E[i]$ is aborted as soon as a connection in $O[i]$ is found to share a SE with another connection in $E[i]$.

Note that merging $O[i]$ and $E[i]$ is the only possible way to reduce the number of rounds because given two different control words i and j, one cannot merge $O[i]$ with $O[j]$, nor $O[i]$ with $E[j]$, nor $E[i]$ with $E[j]$. Hence, it is obvious that the algorithm will result in a minimum number of rounds after the merging takes place and is optimal. Its time complexity is also polynomial of the number of connections to be established.

The average schedule lengths using each of the two algorithms in a Banyan are shown in Figure 6. For comparison purposes, the schedule length using a heuristic *"Greedy"* algorithm, which considers connections to be established randomly and tries to schedule as many connections as possible in each existing round before using a new round, is also shown. These results are for $N = 32$ but the results for other sizes are similar.

As can be seen from Figure 6, when the number of connections to be established, R, is small (i.e. ≤ 100), the Odd-Even algorithm performs the worst and the Greedy algorithm performs nearly as well as the Optimal algorithm. In addition, the number of rounds needed in a DSB is smaller than that needed in a Banyan. However, when R is large (between 100 and 1000), the Odd-Even algorithm performs nearly as well as the Optimal algorithm since merging becomes rare and both algorithms result in a schedule length which approaches the maximum (which is 64 for $N = 32$). However, the Greedy algorithm performs poorly. Specifically, we note that the schedule length using the Greedy algorithm will be about 1.8 times longer when $N = 32$. This is because the Greedy algorithm allows several connections having one parity to be established in one round with several other connections having the opposite parity. This results in fewer than $N/2$ connections in one round and accordingly, more than $2N$ rounds for N^2 connections. Since the Optimal algorithm has a polynomial time complexity and is only a little more complicated than either the Greedy or the Odd-Even algorithm, it should be used when the load condition (i.e. the number of connections

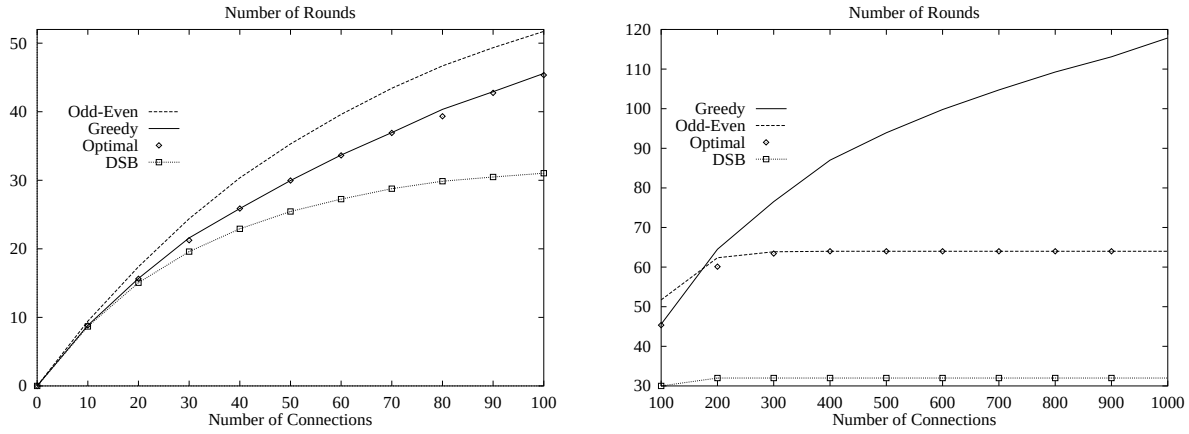

Figure 6: Number of rounds under stage control.

to be established) is unknown or varies greatly.

From Figure 6, we also observe that the schedule length in a DSB reaches its maximum of 32 at about the same time when the schedule length in a Banyan of the same size reaches its maximum of 64. Prior to that point, the latter is less than twice as long as the former for a wide range of R values.

Note that similar results have been obtained even when the connections to be established are random and may contain duplicates [6]. In addition, when switch control (instead of stage control) is used, results from both analysis and simulations show that the number of rounds needed in a Banyan is less than twice that needed in a DSB (see Figure 7).

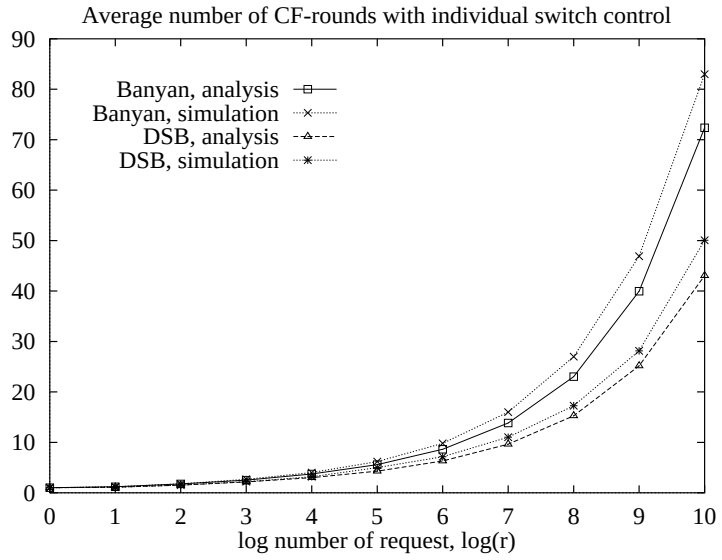

Figure 7: Average number of rounds in a 32×32 Banyan and DSB under switch control.

As mentioned earlier, this means that if we consider the number of SEs (and links) used in a network as a cost in space, and the number of rounds needed as a cost in time (or wavelength), then using a Banyan results in better space-time tradeoffs than using a DSB of the same size because the DSB has at least twice as many SEs and links as the Banyan. A practical implication is that

when space cost is not a concern, one may use a two-plane Banyan instead of a DSB, both of which have approximately the same space cost, to reduce the cost in time (or wavelength) [6].

6 Conclusions

Research in optical MINs, in contrast to its electronic counterpart, is still in its infancy. In this paper, we survey the major challenges encountered and approaches adopted in the research of optical MINs. As optical MINs are more widely adopted in communication and parallel/distributed computing systems [1, 9], many new research issues will emerge and demand for better design and analysis methods. For example, a subject of future research is packet switched photonic MINs with and without internal buffers and the performance of such MINs when traffic can be self-similar (or bursty) in nature. Although some work on diagnosing faulty switching elements generating excessive crosstalk in integrated photonic switching networks has been done, the issues related to engineering fault-tolerant photonic MINs deserve further investigations. Implementing efficient multicast operations in optical MINs is also a new challenge. Finally, how to design efficient parallel algorithms on systems with optical MINs is another interesting research topic.

References

[1] H. Hinton, *An introduction to photonic switching fabrics,* Plenum Press, 1993.

[2] D. Hunter and D. Smith, "New architecture for optical TDM switching", *IEEE/OSA J. Lightwave Technology,* vol. 11, no. 3, pp. 495-511, March 1993.

[3] H. Jordan, D. Lee, K. Lee and S. Ramanan, "Serial array time slot interchangers and optical implementations", *IEEE Trans. Computers,* vol. 43, no. 11, pp. 1309-1318, Nov. 1994.

[4] K. Padmanabhan and A. N. Netravali, "Dilated networks for photonic switching," *IEEE Trans. Communications,* vol. 35, no. 12, pp. 1357-1365, Dec. 1987.

[5] C. Qiao, R. Melhem, D. Chiarulli and S. Levitan, "A time domain approach for avoiding crosstalk in optical blocking multistage interconnection networks," *J. Lightwave Technology,* vol. 12. no. 10, pp. 1854-1862, Oct. 1994.

[6] C. Qiao, "A universal analytic model for photonic Banyan networks", *IEEE Trans. Communications*, vol. 46, no. 10, pp. 1381-1389, Oct. 1998.

[7] R.A. Thompson, "The dilated slipped banyan switching network architecture for use in an all-optical local-area network," *J. Lightwave Technology,* vol. 9. no. 12, pp. 1780-1787, Dec. 1991.

[8] R.A. Thomspon and P.P. Giordano, "An experimental photonic time-slot interchanger using optical fibers as reentrant delay-line memories," *J. Lightwave Technology,* vol. LT-5, no. 1, pp. 154-162, Jan. 1987.

[9] C. Tocci and H.J. Caulfield, *Optical Interconnection – Foundations and Applications*, Artech House Publishers, 1994.

[10] A. Varma and C.S. Raghavendra, *Interconnection Networks for Multiprocessors and Multicomputers: Theory and Practice,* IEEE Computer Society Press, 1994.

[11] Y. Yang, J. Wang and Y. Pan, "Permutation capability of optical multistage interconnection networks," *Proceedings of the 12th IEEE International Parallel Processing Symposium*, Orlando, FL, March 1998, pp. 125-133.