A WDM Interconnection Network for Traffic with Group Locality

Radim Bartoš and Pilar de la Torre

Department of Computer Science, University of New Hampshire, Durham, NH 03824, USA

ABSTRACT

Designing optical communication networks that support large numbers of processors, while maintaining acceptable performance under traffic patterns relevant to parallel-distributed computing, remains an open problem in optical networks research. Central obstacles to its solution appears to hinge on the inadequacy of the standard design methods to overcome the current limitations on the number of available wavelength channels per fiber and the cost and the technologically difficulty of fully-optical switching of wavelength channels.

This paper proposes a WDM optical network architecture design within the limits of current technology. The design goals are to support a large number of nodes and to parametrically configure the network to perform with acceptable packet loss under a variety of offered traffic patterns consistent with communication primitives often used in parallel-distributed computing. Scalability problems due to the limitations of multiplexing in a single dimension (e.g., space, time, and wavelength) are alleviated by an orchestration of the communication traffic according to three design strategies: multidimensional space-time-wavelength multiplexing, spatial reuse of wavelengths, and parametrically driven network reconfiguration as collection groups each made up of processor-clusters where the number of processors per cluster is limited by the number of available wavelengths. The design does not require tunable components nor wavelength sensitive switching. A group-localized traffic model is proposed as a motivation for network design and used for its experimental evaluation.

Keywords: WDM, photonic packet switching, network configuration, locality, parallel architectures

1. INTRODUCTION

One of the ongoing problems in an interconnection network design is the issue of performance under various traffic load conditions. Network designers are faced with the decision either to dimension the network for the worst case scenario, which typically increases the complexity and the cost of the network, or to make an assumption regarding the character of the offered traffic load. Unlike communication networks, interconnection networks used in parallel computers often have the advantage of carrying traffic with characteristics that can be either predicted by an analysis of the computation which is being performed or even controlled by orchestrating the communication among processors from within the program. Since most networks perform well under certain traffic load conditions (e.g., under uniform traffic), the capability to predict and control the offered traffic load in a parallel system can be used to rearrange the network resources or to steer the traffic characteristics towards the areas where the network performs well. The network proposed in this paper exhibits a good performance under several traffic load distributions which are consistent with patterns of communication exploited in parallel computing.

Standard multiplexing methods used to build regular-size optical networks fail as the network sizes grow larger. On one hand, the number of available wavelength channels, tunability range of receivers, transmitters, wavelength filters, and converters limit scalability of wavelength-multiplexed networks, on the other hand the speed of electronic and electro-optic components is the main limiting factor of a time-multiplexed network. A natural solution to this problem is to combine together several methods of multiplexing.[1–5]

Further author information:
Radim Bartoš (correspondence): E-mail: rbartos@cs.unh.edu; Telephone (603) 862-3792; Fax (603) 862-3493. The research of this author was supported in part under UNH Vice President for Research Discretionary Fund Grant.
Pilar de la Torre: E-mail: dltrr@cs.unh.edu; Telephone (603) 862-2682; Fax (603) 862-3493. The research of this author was supported in part under NSF Grant CCR-9410592.

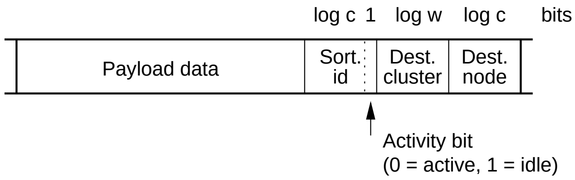


Figure 1. Format of a packet.

1.1. Assumptions and Design Goals

This paper proposes a space-division optical interconnection network capable of supporting a large number of nodes. Multiplexing in space, time, and wavelength is utilized internally to alleviate scalability problems associated with any single form of multiplexing. Wavelength channels are spatially reused throughout the network. The design does not require tunable components nor wavelength sensitive switching.

The proposed interconnection network is intended for distributed memory multiprocessor where nodes (processors), labeled $n_{i,j}$, $0 \le i \le c-1$ and $0 \le j \le w-1$, are organized in c fixed-size clusters of w processors each. The nodes are assigned to clusters in a way that cluster $C_i = \{n_{i,j} : 0 \le j \le w-1\}$, $0 \le i \le c-1$. The WDM components of the network are assumed to be capable of handling at least w independent wavelength channels. Furthermore, it is assumed that c and w are powers of two. The network is expected to carry fixed size packets, consistent in length with a cache line or an ATM cell. Each packet (see Fig. 1) contains a destination address which consists of a destination cluster address (DC) and an in-cluster destination node address (DN). An activity bit distinguishes active packets carrying data (activity bit = 0) from idle packets (activity bit = 1). The packet header also contains a field to carry a *sorting identifier*, a $\log c$-bit address used during routing within a cluster. The activity bit is positioned next to the least significant bit of the sorting identifier. This arrangement allows the two fields to be used together as a single integer value.

1.2. Traffic Assumption: Group-Localized Traffic

This subsection introduces group-localized traffic model which serves as a motivation for the proposed network design and is used for experimental evaluation of the proposed network performance. Recent research points out that communication primitives often used by algorithms in parallel computing can be more efficiently implemented in terms of such group-localized traffic.[6,7]

Groups are formed by partitioning the collection of clusters $\mathcal{C} = \{C_0, C_1, \ldots C_{c-1}\}$ of the proposed network. Let *group configuration* K be a pair of partitions of clusters (k^S, k^D) such that *source partition* $k^S = \{G_0^S, G_1^S, \ldots, G_{|G^S|-1}^S\}$, where group G_i^S, $0 \le i \le |G^S| - 1$, will be refereed to as a *source group*, and *destination partition* $k^D = \{G_0^D, G_1^D, \ldots, G_{|G^D|-1}^D\}$, where group G_i^D, $0 \le i \le |G^D| - 1$, will be referred to as a *destination group*. All nodes in the clusters of a source (destination) group are expected to transmit (receive) traffic with the same distribution of destination (source) addresses.

This paper considers group configurations where all source and destination groups are of equal size and the total number of groups g is a power of two. Since the number of clusters in the network c is a power of two and $c \ge g$, the number of clusters per group is also a power of two. A non-power-of-two group configurations are possible and a subject of future research. Let U_g be a group configuration with g groups, where each source and destination group contains clusters with consecutive indices. Group configuration U_1, where groups contain exactly one cluster, will be referred to as *discrete configuration.*

DEFINITION 1.1. *An offered traffic load is said to be* group-localized *with respect to group configuration K (K-group-localized) if:*

- *Cluster destination addresses of transmissions sent by nodes in the clusters of a source group are all within one destination group.*
- *Nodes in the clusters of a destination group receive traffic from nodes in exactly one source group.*
- *The traffic between a source group and a destination group is uniformly distributed, i.e., nodes in the clusters of a source group generate packets independently with an equal probability and the packet destination addresses are distributed equally over all the nodes in the destination group.*

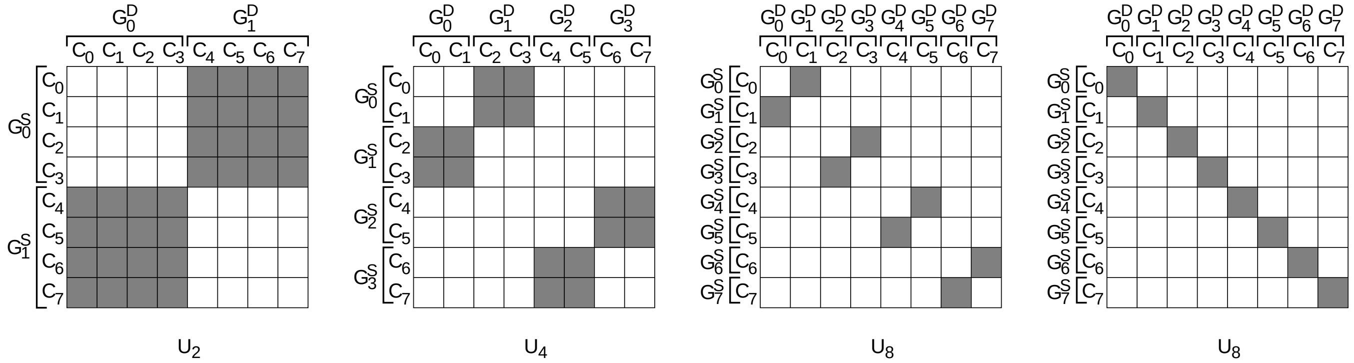


Figure 2. Group-localized traffic generated during an all-to-all communication operation among nodes in a network with $c = 8$ (shaded areas denote exchanged traffic).

Note that group-localized traffic becomes standard uniform traffic[8] when the group configuration consists of only one source and one destination group. The traffic that follows a group configuration where each group contains exactly one cluster will be referred to as *cluster-localized.*

Fig. 2 shows and example how group-localized traffic can be used to implement an all-to-all communication operation among clusters: The clusters in the network are first divided into two equal-sized groups. Then the nodes in one group transmit their data to the other group and vice versa. Since groups received all the data from the other group, no further communication between the two groups is required. In the following steps the whole process recursively repeats within each group. In the end, each group contains one cluster and the communication takes place within that cluster. Note that in the last two steps the traffic turns into cluster-localized since the communication takes place between pairs of clusters.

1.3. Previous Work

One of the methods used in designs of large networks to overcome the limited number of available wavelength channels is wavelength switching. Switching of individual wavelengths can be achieved using a wavelength demultiplexer (which converts the transmissions from the wavelength to the spatial dimension) followed by a spatial switch and an array of transmitters. Such a solution is expensive and requires o/e and e/o conversion. General fully-optical switching of wavelength channels is technologically difficult. The proposed network avoids wavelength switching by utilizing the concept of *WDM packet** where individual packets destined to the same group are combined together and transmitted simultaneously on different wavelength channels. A WDM packet can be switched as a unit since all its packets share the same destination group. Wavelength insensitive switching elements (e.g., Lithium-Niobate directional couplers[9]) are readily available and provide sufficient switching rates.

STWnet,[3] a recently introduced interconnection network architecture, which utilizes the concept of WDM packets to provide connectivity to groups of processors (nodes), utilizes internally multiplexing in space, time, and wavelength. STWnet was designed for uniform traffic load. It performs space-time sorting[10] followed by temporal routing[11] at the group level, which permits the use of a regular reverse Banyan network as the group interconnect and the use of a fixed routing schedule. Expansion in either temporal (speedup) or space dimension is used to accommodate the possibility of having more packets destined to a group than there are nodes. Since the whole group and not an individual node is considered, it can be concluded using probabilistic analysis based on Group Knockout Principle,[12] that a small expansion factor is sufficient to maintain a very low probability of packet loss. STWnet has many desirable features. It is a self routing network that can support a large number of nodes without stretching the limits of the available technology. However, STWnet exhibits significant packet loss performance degradation under a group-localized traffic load.

A preliminary version of the proposed network architecture was presented in.[13] Since the design goal was to support uniform and cluster-localized traffic, the network[13] does not perform well under general group-localized traffic. The architecture of the network[13] is similar to the network presented in this paper but it does not provide any means for reconfiguration.

*WDM packets were independently proposed in[3] and[4] (term *photonic slot* is used in[4]).

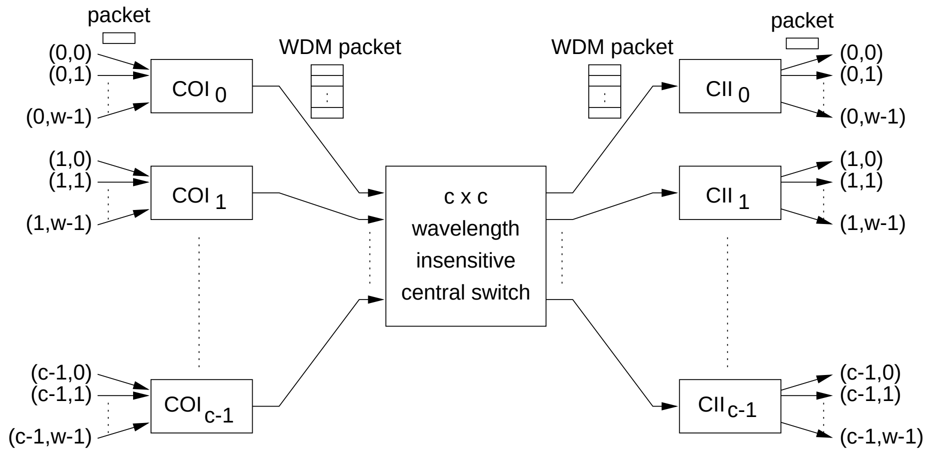


Figure 3. The proposed space-time-wavelength interconnection network (COI: Cluster Output Interface; CII: Cluster Input Interface).

2. INTERCONNECTION NETWORK ARCHITECTURE

The $cw \times cw$ network under consideration consists of three main stages: cluster output interface (COI) stage, $c \times c$ central switch, and cluster input interface (CII) stage (see Fig. 3). For each cluster, packets sent by the processors are rearranged and transmitted in the form of WDM packets to the wavelength-insensitive central switch. The central switch interconnects the clusters and routes WDM packets as units to the destination clusters. The cluster input interfaces extract individual packets from WDM packets and route them to the processors.

The COI stage does not introduce any packet loss which occurs only within the central switch as a result of an output conflict. Due to the probabilistic nature of the offered traffic load and the cost factor, it is impractical to provide an absolute guarantee of no packet loss. The goal of the design is to have COIs output WDM packets in a way that the probability of having multiple WDM packets destined to a cluster in one cycle is greatly reduced. It is assumed that the central switch can handle up to L WDM packets delivered simultaneously to a cluster. WDM packets in excess of L delivered to a cluster are dropped by the central switch and therefore the choice of parameter L directly affects the packet loss performance of the proposed network. Note that L is a design parameter that permits building a network for any desired packet loss rate. However, since the cost of the central switch increases with L, the network may not be practical for a large value of L. Simulation results outlined in Sect. 4 indicate that for a value of L as low as 2 the network exhibits good packet loss performance.

Let $\mathrm{A}(c, w, K, L)$ represent the proposed network with c clusters, w nodes per cluster, built using components capable of supporting w wavelengths, with clusters logically divided into groups according to group configuration K, and with a central switch capable of delivering L packet per output.

2.1. Cluster Output Interface (COI)

The main function of a Cluster Output Interface is to form WDM packets. Packets destined to the same cluster are collected together and transmitted simultaneously using an array of fixed wavelength transmitters. In order to ensure that there is enough packets destined to a cluster to fill the corresponding WDM packet, a train of c packets from each one of w nodes (instead of just one packet per node) is considered at a time. The operation of the COIs is orchestrated using a heuristic based on the current group configuration K to ensure a low packet loss caused by output conflicts in the central switch.

A COI (Fig. 4) operates in stages. In the first stage, buffers receive packets from processors and output them in a form $w \times c$ matrix. This matrix is expanded to $w \times 2c$ by a column of Temporal Expander - Idle Packet Injector modules (TE-IPIs) and sorted in a column-major order by the space-time sorter in the next stage. Expansion together with sorting ensures that the columns of the matrix contain only packets destined to a single destination cluster. Columns of the sorted matrix are transmitted by the array of fixed tuned transmitters in the last stage of a COI. Processing of the complete matrix will be referred to as a *round of communication*.

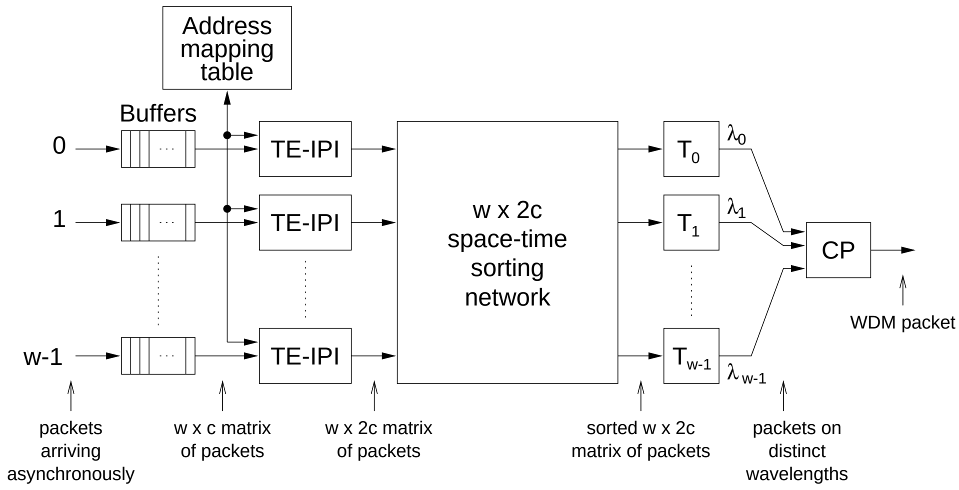


Figure 4. Cluster Output Interface (TE-IPI: Time Expander/Idle Packet Injector; T_i: fixed wavelength transmitter tuned to λ_i; CP: coupler).

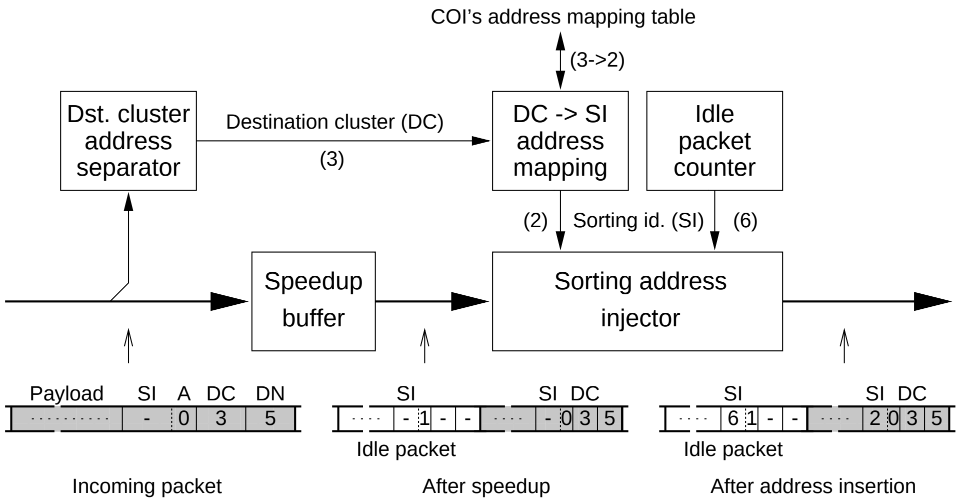


Figure 5. Time Expander/Idle Packet Injector (TE-IPI).

The purpose of the buffer stage is to ensure the temporal alignment of packets and to eliminate the need for a precise mutual synchronization of nodes. Since the buffer stage can only change the temporal position of a packet by less than a packet transmission time, the head-of-line blocking problem does not occur. The buffer stage outputs $w \times c$ matrix of slots (c time-slots on w spatial links) where each slot is capable of holding a packet. Depending on the offered traffic load, the slots may contain active or idle packets.

The next stage rearranges the matrix of slots in a way that individual columns will contain only packets destined to a same cluster and are ready to be transmitted as WDM packets. The order in which WDM packets are generated by COIs is designed to avoid the situation where two or more COIs transmit to the same destination cluster. First, the Time Expander/Idle Packet Injector (TE-IPI) stage (see Fig. 5) introduces additional time-slots by a twofold internal speedup of the network. Incoming packets are retransmitted at double the bit rate and thus created empty slots are filled with idle packets. The desired permutation of the WDM packets is achieved by assigning to each an appropriate integer *sorting identifier* in the interval $[0, c-1]$. Using these identifiers as keys, the sorting network generates the rearranged packets in an increasing order of sorting identifiers.

Although specifics of the sorting-identifier mapping schemes are postponed until Sect. 3, we now outline the ideas behind them and their role in the proposed network overall design. The sorting identifier of an active packet is obtained as a result of mapping its destination cluster address. Such mapping depends on the cluster index and the group configuration. The mapping in a COI_i, $0 \le i \le c-1$, under group configuration K can be expressed as a c-tuple $M_i^K = (s_{i,0}^K, s_{i,1}^K, \ldots, s_{i,c-1}^K)$ of sorting identifiers where $s_{i,j}^K$, $0 \le j \le c-1$, is the sorting identifier assigned to packets with destination cluster address j. For example, a COI in a network with $c = 4$ with mapping $(3, 1, 0, 2)$ will first output all WDM packets destined to cluster 2 followed by all WDM packets to 1, followed by 3 and 0. It is easy to see that an arbitrary ordering of destination cluster addresses of generated WDM packets can be achieved by varying the mapping. The next section describes the mapping used to configure network $\mathrm{A}(c, w, K, L)$ when aimed at supporting a K-group-localized traffic load with a low packet loss.

The mapping function is implemented using a simple lookup table with c entries of $\log c$ bits each, located in every COI and shared by all its TE-IPIs. Destination cluster address is extracted from the packet by the cluster address separator and translated to a sorting identifier using the address mapping table. The resulting sorting identifier is inserted into the packet by the sorting address injector.

Idle packets received from the input buffers in the case when the network is not fully loaded are assigned sorting identifier equal COI's index modulo c. The intuition behind this heuristic is to ensure that additional idle packets affect different COIs in a different way and thus further reduce the probability of a collision in the central switch. The sorting identifiers of the speedup-generated idle packets are assigned according to the current value of a mod c counter which is incremented every time an idle packet is generated. Note that in a complete transmission cycle a TE-IPI inserts c idle packets, each with a unique sorting identifier value. Overall, the w TE-IPI modules in a COI generate cw idle packets, w for every sorting identifier value.

The space-time sorting network[10] sorts $w \times 2c$ matrix of packets in a column major order. It is built using 2×2 directional couplers and feedback delay lines. The sorting network performs two tasks in a COI: ensures that WDM packets contain only transmissions destined to one of the destination groups and changes the order in which WDM packets are transmitted. The activity bit is prepended as the least significant bit to the sorting identifier and the whole extended address is used by the comparators in the sorting network. Recall that an activity bit set to 1 denotes an idle packet and conversely 0 marks an active packet. Then, after sorting one round of communication, all active packets with sorting identifier i will be followed by exactly w idle packets with sorting identifier i, followed by active packets with sorting identifier $i+1$. One column of packets contains active packets destined to a single destination cluster since the number of idle packets separating the active packets with different destination cluster addresses is equal to the number of rows of the matrix.

In the last stage of a COI, the outputs of the sorting network are connected to an array of fixed tuned transmitters, each of which is tuned to a unique wavelength. A protocol proposed in[14,15] is used to facilitate routing of packets within a destination cluster. The transmitters examine the packet destination and generate the protocol header while discarding the idle packets. The protocol provides a conflict-free routing between a source and a destination clusters.

COIs are fully pipelined and only a bounded delay is introduced. There is no packet loss introduced by the architecture of a COI.

2.2. Central Switch

The central switch is a $c \times c$ space-division switching network that routes WDM packets between COIs and CIIs. Since WDM packets are handled as units, the central switch must be built using wavelength insensitive switching elements. Any space division switching network capable of simultaneous delivery of L packets to an output (e.g., Knockout switch,[8] SXmin[16]) can be used as a central switch.

2.3. Cluster Input Interface (CII)

The function of a CII (see Fig. 6) is to receive WDM packets delivered by the central switch, separate individual packets and route them to the nodes in the cluster. The incoming WDM signal from the central switch is split by a passive optical splitter and distributed to individual receivers. A CII is equipped with w diffraction-based receivers capable of concurrent reception on all transmitting wavelengths. Receivers analyze the protocol information[14,15] and discard the packets which are not destined to the corresponding nodes. Buffers perform concentration in the case of simultaneous delivery of multiple packets on different wavelengths and traffic smoothing.

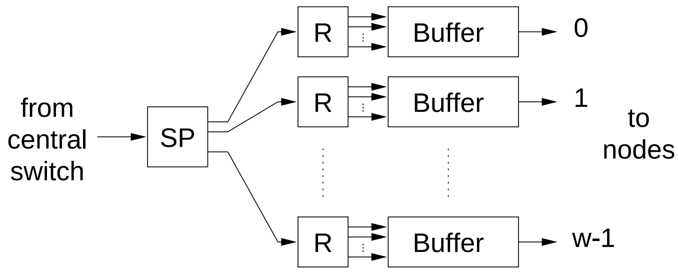


Figure 6. Cluster Input Interface (SP: splitter; R: diffraction-based parallel receiver).

3. NETWORK CONFIGURATION

The mapping of destination cluster addresses onto sorting identifiers in COIs is used to configure the proposed parameterized network $A(c, w, K, L)$ for a given c, w, and group configuration K. Such mapping is a one-to-one projection from the set of cluster destination addresses onto sorting identifiers in the range $[0, c-1]$. Since the sorting identifiers are used as keys within the sorting network, the mapping provides a simple way to control the order in which WDM packets are produced by a COI. The mapping function proposed in this paper is based on the following two heuristic conditions:

1. Destination addresses of group members should be mapped to sorting identifiers as far apart from each other as possible. This requirement reduces the probability of a collision among transmissions sent by members of a group.

2. No two COIs should map a destination address to the same sorting identifier. This condition reduces the probability of multiple clusters simultaneously producing WDM packets with the same cluster destination address.

We now describe a procedure that, given c, w, and K, produces the desired mapping. First a special case of group configuration U_g will be considered and then formulas for a general network configuration K will be given. For purposes of describing its construction, we view the mapping as composed of two separate operations: A permutation, identical for all COIs in the network, that shuffles members of groups according to the first condition (*core mapping*) and a *circular shift* of the permutation in individual COIs to satisfy the second condition. Let $M^K = (m_0^K, m_1^K, \ldots m_{c-1}^K)$ be the core mapping under group configuration K and let $R^K = (r_0^K, r_1^K, \ldots r_{c-1}^K)$ represent the circular shifts of COIs under group configuration K. Then the sorting identifier assigned to packets with cluster destination address j in COI_i under group configuration K, $s_{i,j}^K$, is given by ($\oplus$ denotes addition modulo c):

$$s_{i,j}^K = m_j^K \oplus r_i^K.$$

The purpose of core mapping is to ensure that COIs output WDM packets with destination cluster addresses within a group as far apart from each other as possible. While core mapping is the same for every COI in the network, it depends on the current group configuration. For example, under group configuration U_g, where group members occupy COIs with consecutive indices, j-th element of the core mapping

$$m_j^{U_g} = \left\lfloor \frac{jg}{c} \right\rfloor + g\,(j \bmod \frac{c}{g}),$$

where factor $\lfloor \frac{jg}{c} \rfloor$ corresponds to the group index and factor $j \bmod \frac{c}{g}$ gives the in-group index of the cluster.

The formula for the core mapping under a general group configuration is given in terms of two functions defined to provide the group index and the in-group cluster index of a cluster regardless of the particular group configuration. Let $\mathrm{g}^K(j)$ be an index of a source group such that $C_j \in G^S_{\mathrm{g}^K(j)}$ under group configuration K and let $\mathrm{l}^K(j)$ be the cluster index of cluster j within its source group. Then j-th element of the core mapping under group configuration K is given by:

$$m_j^K = \mathrm{g}^K(j) + g\,\mathrm{l}^K(j).$$

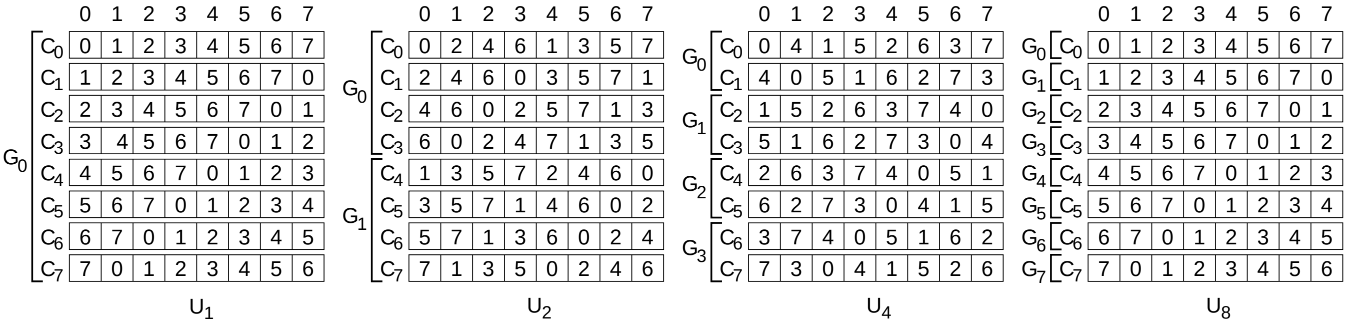


Figure 7. Contents of mapping tables in a network with $c = 8$ under group configuration U_1, U_2, U_4, and U_8, .

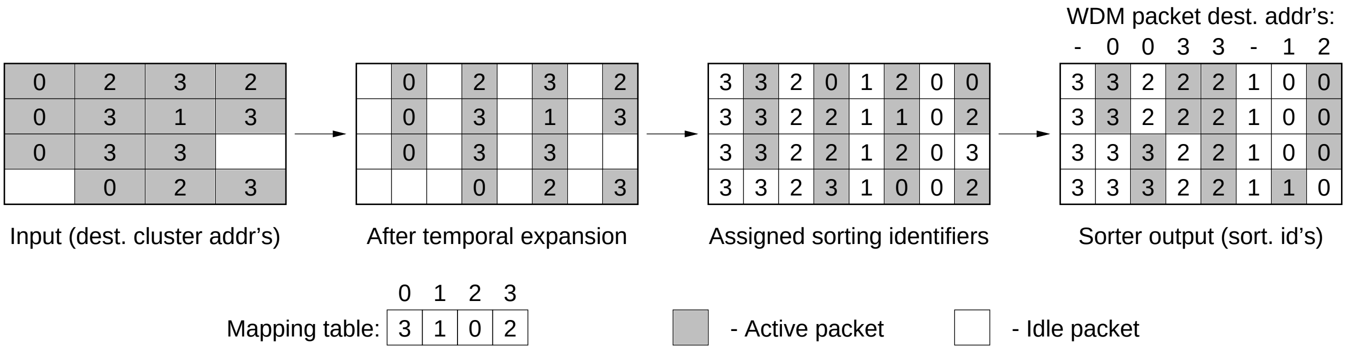


Figure 8. Example of routing in COI_3 of a network with $c = w = 4$.

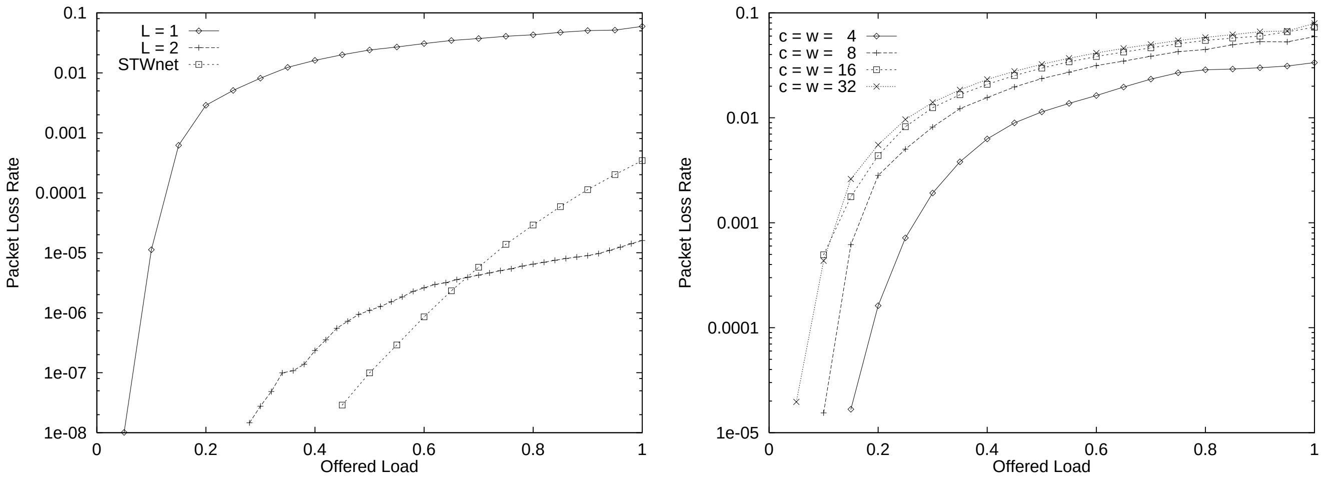
Identical mapping of a cluster destination address would increase the probability of COIs simultaneously outputting conflicting WDM packets. A circular shift is used to ensure that no two COIs in the network use the same mapping for any one of the destination cluster addresses. The proposed circular shift function is similar to core mapping: the operation of members of a group are shifted so that they are as far apart from each other as possible. The network-wide shift is achieved by ensuring that for any two groups there is a also shift between their operation. For example, under group configuration U_g defined above the circular shift for cluster C_i is given by

$$r_i^{U_g} = \left\lfloor \frac{ig}{c} \right\rfloor + g\,(i \bmod \frac{c}{\mathrm{g}}).$$

Under a general group configuration the circular shift of C_i can be expressed as

$$r_i^K = \mathrm{g}^K(i) + g\,1^K(i).$$

Note that one mapping can serve all possible destination group configurations with all the permutations of the traffic destinations, since it depends only on the source group configuration. Fig. 7 shows four examples of the mappings for a network with 8 clusters, under group configuration U_g for various group sizes g. Fig. 8 shows an example of routing in COI_3 of a network with $c = w = 4$. The input matrix contains 14 active packets and 2 idle packets. The size of the matrix is doubled by the speedup and 16 additional idle packets are inserted. Sorting identifiers of active packets are assigned according to the given mapping of destination cluster addresses. The two received idle packets are assigned sorting identifier 3, equal to the COI's index. Idle packets created by the speedup are assigned sorting identifiers based on their temporal position in the matrix. After sorting, column 0 (the rightmost) contains packets to form a WDM packet destined to cluster 2 (sorting identifier = 0) followed by WDM packets with destination cluster addresses 1, $-$, 3, 3, 0, 0, and $-$ ($-$ denotes a column with no active packets and therefore without a cluster destination address).



(a) A$(8, 8, U_1, L)$. (b) A$(c, w, U_1, 1)$: $c = w$.

Figure 9. Packet loss rate under uniform traffic.

4. EXPERIMENTAL EVALUATION

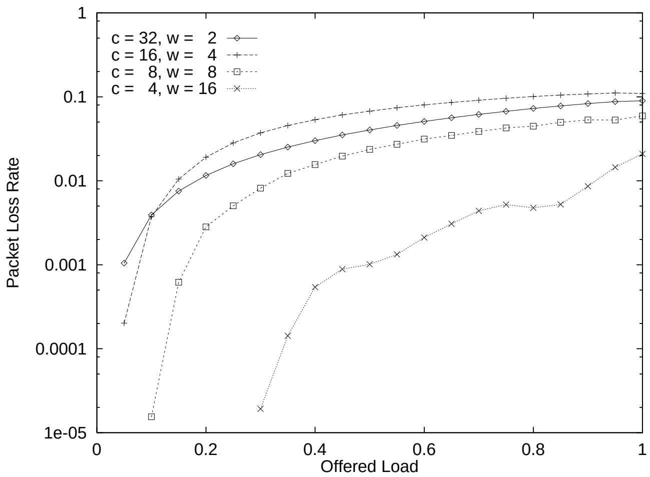
The goal of the experiments was to explore the extent to which the network achieves the design objectives. A simulator was used to carry out the performance evaluation. Packet loss rate was chosen as the measure of performance because of its direct links to the issues addressed by the network design. Two sets of experiments were carried out. First, the network was subjected to a uniform and a general group-localized traffic loads to quantify the performance under conditions directly targeted by the network design. The second set of experiments tested the network performance under mixed traffic consisting of two types of group-localized traffic. The goal of the mixed-traffic experiments was to provide an evidence that the performance of the network does not significantly deteriorate under traffic conditions that may occur during a transition from one network configuration to another. Finally, the data collected from the experiments enabled us to identify the value of L for which the packet loss-based performance of the proposed network matches or surpasses that of STWnet[†]. Moreover, the results of the experiments illustrate the network potential versatility to perform under a wider range of offered traffic loads than its STWnet predecessor.

It should be noted that the latency of the proposed network consists of two components: a bounded delay introduced by the COIs and the central switch and a variable delay introduced by the output buffering in the CIIs. The analysis of the output queueing delays has been widely explored (e.g.,[17]) and it is not repeated here. The results presented in this paper do not consider a packet loss due to output buffer overflow. Nevertheless, this is an important issue, the significance of which in the group-localized traffic setting remains to be systematically investigated.

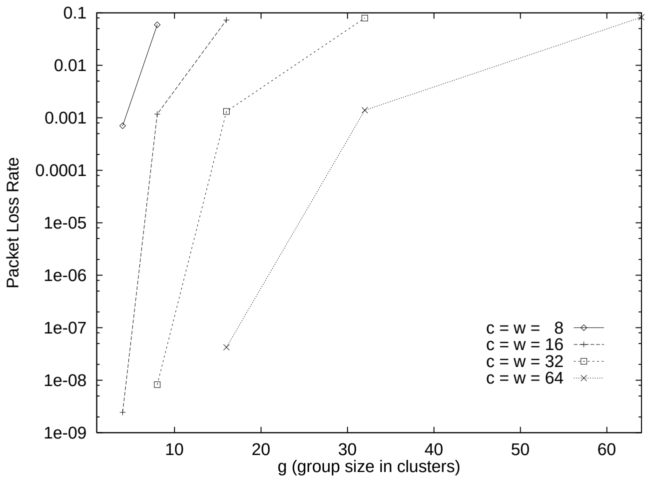
4.1. Performance under Uniform Traffic

Fig. 9 (a) shows packet loss rate under uniform traffic load for varying traffic intensity and value of L. It can be observed that the proposed network exhibits a lower packet loss than STWnet switch for $c = w = 8$, $L = 2$ and the traffic load $\geq 68\%$. The packet loss rate for $L \geq 3$ under uniform traffic is so small that the simulator did not record any packet loss during 10^9 rounds of communication (each round consists of c^2w packets). The next experiment (Fig. 9 (b)) shows that, as the network size grows larger, the packet loss rate increases. However, inspection of the graph suggests that the rate at which the packet loss rate increases appears to decrease. Similar behavior was observed in other Knockout Principle-based networks. Analytical models in[8] and[12] provide tight upper bounds for packet loss rates when the number of nodes approaches infinity. It is expected that systematic analysis of the proposed network will also yield a similar tight upper bound. The last uniform traffic experiment (see Fig. 10 (a)) explored the impact of cluster size variation on the packet loss rate of a fixed-size network. As expected the network performs better in the cases where $w \geq c$ reflecting the lower relative cost of routing within multi-wavelength links.

[†]The experiments assume an STWnet with twofold expansion ($lr = 2$).



(a) $\mathrm{A}(c, w, U_1, 1)$: $cw = 64$, and uniform traffic.



(b) $\mathrm{A}(c, w, U_g, 1)$: load $= 100\%$, and group-localized traffic.

Figure 10. Packet loss rate under uniform traffic and group-localized traffic.

4.2. Performance under Cluster-Localized Traffic and Group-Localized Traffic

The proposed network does not exhibit any packet loss under cluster-localized traffic since COIs generate WDM packets destined to distinct output groups and therefore no output conflicts occur in the central switch to cause a packet loss.

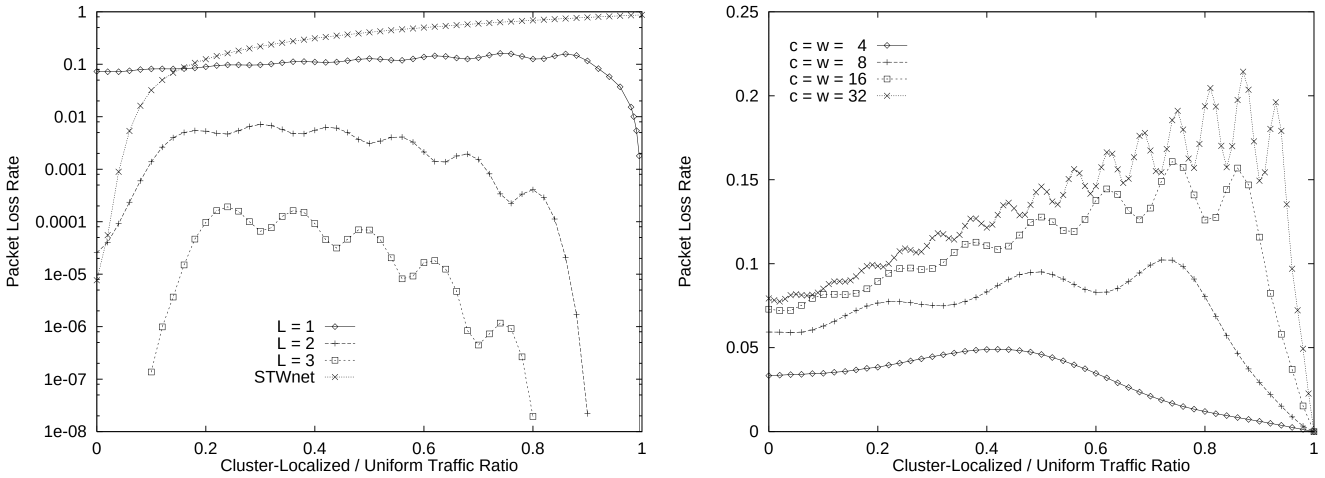
This observation is not true for group-localized traffic where packet loss may occur due to imbalance of the number of packets destined to individual clusters in a group during one round of communication. Experimental results (see Fig. 10 (b)) show that the packet loss rate of a fixed-size network under group-localized traffic rapidly decreases as the groups become smaller.

4.3. Performance under Mixed Traffic

The next set of experiments subjected the proposed network to a traffic load composed of two types of group-localized traffic T_0 and T_1. Let r, $0 \leq r \leq 1$, be the percentage of T_1 within the overall traffic. The load on an input in a time slot is generated according to T_1 with probability r or using T_0 with complementary probability $1 - r$. For $r = 0$ the traffic load follows distribution T_0 and for $r = 1$ distribution T_1.

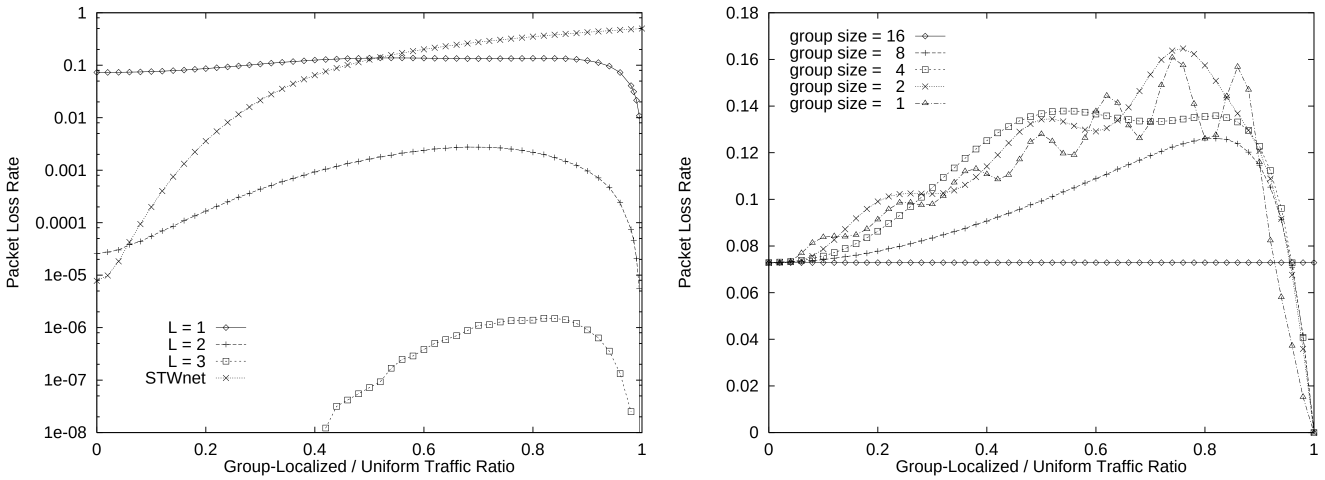
Fig. 11 (a) shows the performance of the proposed network under mixed cluster-localized and uniform traffic. The outcome of the experiment indicates that the network performance improves as one type of traffic becomes dominant and degrades when the two types of traffic mix. Nevertheless, even at the peeks of the curves, the network maintains acceptable packet loss performance. For comparison, Fig. 11 (a) also shows the performance of STWnet switch. STWnet performs well under uniform traffic for which it was designed but the packet loss rate rapidly increases as the traffic becomes more localized. This is as expected since the fixed transmission schedule of the central switch of STWnet provides slots for only $2c$ packets destined to the same cluster per round of communication. The packet loss rate of STWnet under purely cluster-localized traffic is $1 - \frac{2}{c}$. Slots of the central switch in the proposed network are dynamically assigned and therefore this problem does not occur. The effect of the network size on the packet loss performance under mixed cluster-localized/uniform traffic is shown in Fig. 11 (b). As in the case of purely uniform traffic, the packet loss rate slightly increases as the network grows larger. Another observation can be made that the area of good performance for almost purely cluster-localized traffic becomes narrower as the network size increases.

The next two experiments explore the network performance under mixed group-localized traffic and uniform traffic. Fig. 12 (a) shows the performance for the fixed number of groups $g = 4$. As in the case of mixed cluster-localized and uniform traffic, the performance improves when one of the types of traffic becomes dominant and degrades when the traffic is mixed. The main difference is in the narrowness of the area of improved performance when the traffic is almost purely group-localized. The second experiment (see Fig. 12 (b)) explores the impact of a different number of groups in a fixed-size network. The results of the experiment show no major performance variations when the number of groups changes.



(a) $A(16, 16, U_1, L)$: load = 100%. (b) $A(c, w, U_c, 1)$: $c = w$, and load = 100%.

Figure 11. Performance under mixed cluster-localized traffic and uniform traffic.



(a) $A(16, 16, U_4, L)$: load = 100%. (b) $A(16, 16, U_g, 1)$: load = 100%, group size = $\frac{16}{c}$.

Figure 12. Packet loss rate under mixed group-localized traffic and uniform traffic.

The last experiment (see Fig. 13) explores the performance under mixed cluster-localized and group-localized traffic for different numbers of groups in the network. The main performance variation can be observed when traffic is almost purely group-localized. In that region the performance significantly improves as the group size decreases.

5. CONCLUSIONS

The proposed optical network combines advantages of space and temporal switching with benefits of wavelength-parallel data transmissions in order to provide a support for a large number of nodes without stretching limits of the available technology. Tunable components and wavelength switching are avoided and the available wavelength channels are spatially reused. Group-localized traffic model is proposed as a motivation for the network design and is used for experimental evaluation of the network design. In order to explore the extent to which the proposed network design achieves the intended objectives, simulation based experiments were carried out under conditions directly targeted by the design: uniform traffic, group-localize traffic, and a hybrid of such traffic types in various proportions. Data collected from the experiments illustrate the network's potential versatility under a wide range of offered traffic loads, and enable us to quantify how variation on network parameters values affect network performance.

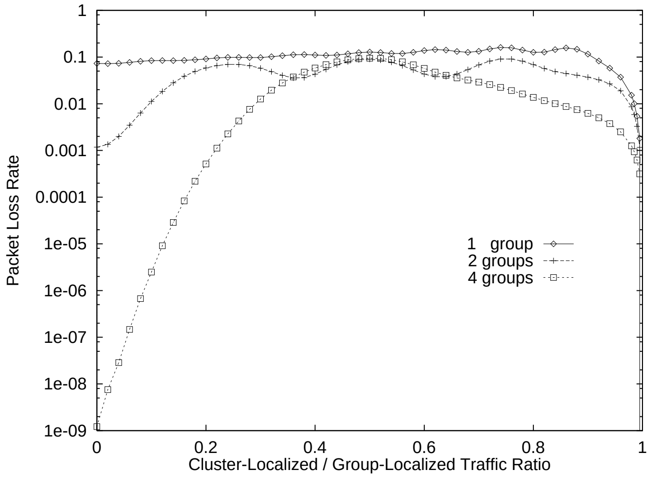


Figure 13. Packet loss rate under mixed cluster-localized traffic and group-localized (A$(16, 16, U_g, 1)$: load = 100%).

REFERENCES

1. J. Sharony, T. E. Stern, and Y. Li, "The universality of multidimensional switching networks," *IEEE/ACM Trans. on Networking* **2**, pp. 602–612, December 1994.
2. P. W. Dowd and K. A. Aly, "Time-space-wavelength networks for low-complexity processor interconnection," in *Proc. Intnl. Conf. on Massively Parall. Proc. with Optical Interconnects, Cancun*, pp. 281–292, April 1994.
3. R. Kannan, R. Bartoš, K. Y. Lee, and H. F. Jordan, "STWnet: A high bandwidth space-time-wavelength multiplexed optical switching network," in *Proc. IEEE Infocom, Kobe, Japan*, April 1997.
4. I. Chlamtac, V. Elek, A. Fumagalli, and C. Szabó, "Scalable WDM network architecture based on photonic slot routing and switched delay line," in *Proc. IEEE Infocom, Kobe, Japan*, April 1997.
5. I. Chlamtac, V. Elek, and A. Fumagalli, "A fair slot routing solution for scalability in all-optical packet switched networks," *Journal of High Speed Networks* **6**(3), 1997.
6. P. de la Torre and C. Kruskal, "Submachine Locality in the Bulk Synchronous Setting," in *Proc. Euro-Par'96 Parallel Processing*, vol. 2, pp. 352–355, 1996.
7. P. de la Torre and M. D. Plumlee, "Towards a Distributed Realization of Dynamic Virtual Locality in the NOW Setting," in *DIMACS Workshop on Networks in Distributed Computing*, October 1997.
8. Y. Yeh, M. G. Hluchyj, and A. S. Acampora, "The Knockout Switch: A simple, modular architecture for high-performance," *IEEE J. on Selected Areas in Communications* **SAC-5**, pp. 1274–1283, October 1987.
9. T. Tamir, ed., *Guided Wave Optoelectronics*, vol. 26 of *Springer Series in Electronics and Photonics*, ch. 4. Springer-Verlag, 1988.
10. R. Kannan, D. Lee, K. Y. Lee, and H. F. Jordan, "Optical TDM sorting networks for high-speed packet switching," in *Proc. IEEE Globecom'94*, pp. 1948–1952, December 1994.
11. R. Kannan, R. Bartoš, K. Y. Lee, and H. F. Jordan, "Optical TDM switching architectures with distributed control," in *Proc. IEEE Milcom'96, McLean, VA*, pp. 155–159, October 1996.
12. K. Y. Eng, M. Karol, and Y. S. Yeh, "A growable packet switch architecture: Design and applications," *IEEE Trans. on Communications* **40**, pp. 423–430, February 1992.
13. R. Bartoš, P. de la Torre, and R. Kannan, "Space-time-wavelength network with group communication locality." to appear in Proc. DIMACS Workshop on Multichannel Optical Networks: Theory and Practice, March 1998.
14. R. Bartoš, R. Kannan, K. Y. Lee, and H. F. Jordan, "Multinet: High-performance WDM interconnection architecture," in *Proc. IEEE ICC'95*, pp. 277–281, June 1995.
15. R. Bartoš, R. Kannan, K. Y. Lee, and H. F. Jordan, "An efficient WDM architecture with a wavelength-distributed synchronization protocol," in *Proc. IEEE ICC'96*, pp. 1788–1792, June 1996.
16. R. Kannan, R. Bartoš, K. Y. Lee, and H. F. Jordan, "SXmin: A self-routing high-performance ATM packet switch based on Group-Knockout principle," *IEEE Trans. on Communications* **45**, pp. 710–722, June 1997.
17. M. J. Karol, M. G. Hluchyj, and S. P. Morgan, "Input versus output queuing on a space-division packet switch," *IEEE Trans. on Communications* **COM-35**, pp. 1347–1356, December 1987.