DIMACS Series in Discrete Mathematics
and Theoretical Computer Science

,

Space-Time-Wavelength Network with Group Communication Locality

Radim Bartoš, Pilar de la Torre, and Rajgopal Kannan

Abstract. This paper proposes an optical space-division interconnection network for distributed memory multiprocessor where nodes are organized in groups of equal size. The network utilizes internally space, wavelength, and time multiplexing to address known scalability constraints associated with any one of these forms of multiplexing. Technologically difficult switching of individual wavelength channels is avoided by rearranging packets within groups in a way that wavelength insensitive switching elements can be used within the switching fabric interconnecting groups. The goal of the design of the proposed network is to support a large number of nodes while maintaining acceptable packet loss performance under inter-group localized and uniform traffic. Under the fully inter-group localized traffic model pairs of groups communicate with each other. The network can be parameterized using three parameters: number of groups, number of available wavelengths, and the maximum number of simultaneous transmissions that can be delivered to a group. Preliminary simulation results indicate that the network exhibits a tradeoff between packet loss rate under uniform as well as inter-group localized traffic and the capability to deliver multiple transmissions. The experiments indicate that acceptable packet loss rate can be achieved with only a small value of the simultaneous arrival threshold.

1. Introduction

One of the areas or research in optical networks that has not yet been fully explored concerns issues related to the design of interconnection networks capable of supporting a large number of nodes and patterns of communication exploited in parallel computing. Methods used to build regular-size networks fail as the size increases. The number of available wavelength channels, tunability range of receivers, transmitters, wavelength filters, and converters limit scalability of wavelength-multiplexed networks. 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 several methods of multiplexing. Such *multidimensional* networks have been recently proposed [**SSL94, DA94, KBLJ97a, CEFS97, CEF97**].

1.1. Assumptions and design goals. This paper proposes a space-division interconnection network capable of supporting a large number of nodes. Multiplexing in space, time, and wavelength utilized internally to alleviate problem associated with a single dimension 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) are organized in g groups of w processors each.

The research of this author was supported in part under NSF Grant CCR-9410592.

©0000 (copyright holder)

Furthermore, optical components are assumed to be capable of handling w independent wavelength channels. It is assumed that the network carries fixed size packets consistent in length with a cache line or an ATM cell. Each packet contains a destination address which consists of destination group identification and in-group node identification. In addition to the destination address, the packet headers also contain an activity bit.

The network is intended to carry traffic load consisting of a combination of uniform and inter-group localized traffic. Under the *uniform traffic* model packets arrive at each input independently with an equal probability and the input load on an input is distributed equally over all the outputs. Under a *fully inter-group localized traffic* load, all packets sent by a group are destined to exactly one group, no two groups transmit to the same group, and the traffic between a communicating pair of groups is uniformly distributed among the members of the groups. Note that a local intra-group traffic is a special case of localized traffic where source and destination groups are the same. Recent research points out that communication primitives often used by algorithms in parallel computing, such as the total-exchange, can be more efficiently implemented in terms of localized traffic [**dlTK96, dlTP97**].

1.2. 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*[1] 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 [**Tam88**]) are readily available and provide sufficient switching rates.

STWnet [**KBLJ97a**], 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 all three dimensions of space, time, and wavelength. STWnet is primarily designed for uniform traffic load. It performs space-time sorting [**KLLJ94**] followed by temporal routing [**KBLJ96**] 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 [**EKY92**], that a small expansion factor is sufficient to maintain a very low probability of packet loss. The packet loss in STWnet occurs at the group level when the number of packets destined to a group exceeds the number of available slots. The central switch of a STWnet is lossless.

[1] WDM packets were independently proposed in [**KBLJ97a**] and [**CEFS97**] (term *photonic slot* is used in [**CEFS97**]).

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 a significant packet loss performance degradation under inter-group localized traffic load. This is a serious drawback since the traffic observed in most networks is often highly non-uniform.

2. Interconnection Network Architecture

The goal of the design of the proposed network is to support a large number of nodes while maintaining acceptable packet loss performance under inter-group localized and uniform traffic. The proposed network builds upon STWnet architecture. The overall structure of the network is maintained while the internal design of the components was changed to accommodate uniform as well as inter-group localized traffic loads.

The proposed $gw \times gw$ network consists of three main stages: group output interface (GOI), $g \times g$ central switch, and group input interface (GII) (see Fig. 1). Packets sent by the processors are rearranged within a group and transmitted in the form of WDM packets to the wavelength-insensitive central switch. The central switch interconnects groups and routes WDM packets as units to the destination groups. Group input interfaces extract individual packets from WDM packets and route them to processors.

It is assumed that the central switch can handle up to L WDM packets delivered simultaneously to a group. The goal of the design is to have GOIs output WDM packets in a way that probability of having more than L WDM packets destined to a group in one cycle is greatly reduced. Since WDM packets in excess of L delivered to a group are dropped by the central switch, the choice of parameter L directly affects the packet loss performance of the proposed network. GOIs do not introduce any packet loss which occurs only within the central switch when, during a transmission cycle, more than L WDM packets destined to the same group are generated. 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 the following section indicate that for a value of L as low as 3 the network exhibits good packet loss performance.

Let $A(g, w, L)$ represent the proposed network with g groups, w nodes per group, built using components capable of supporting w wavelengths, with a central switch capable of delivering L packet per output.

2.1. Group Output Interface (GOI). The main function of a Group Output Interface is to form WDM packets. Packets destined to the same group are collected together and transmitted simultaneously using an array of fixed wavelength transmitters. In order to ensure that there is enough packets destined to individual groups to fill corresponding WDM packets, a train of g packets from each one of w nodes (instead of just one packet per node) is considered at a time. WDM packet transmissions of individual GOIs are orchestrated in such a way that packet loss caused by conflicts in the central switching fabric is avoided with a high probability.

A GOI (Fig. 2) operates in stages. In the first stage, buffers receive packets from processors and output them in a form $w \times g$ matrix. This matrix is expanded

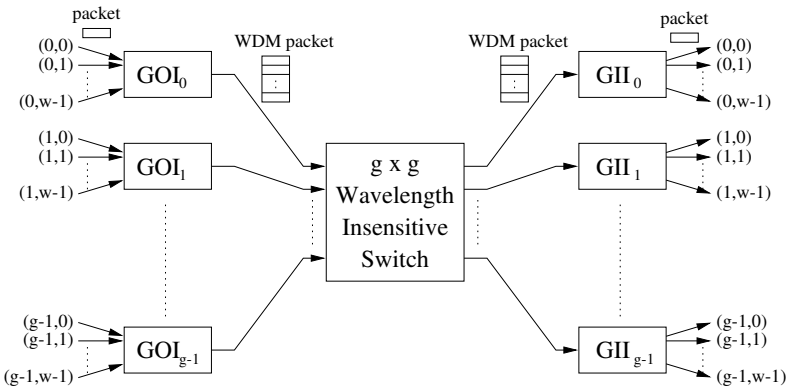

FIGURE 1. The proposed space-time-wavelength interconnection network (GOI: Group Output Interface; GII: Group Input Interface).

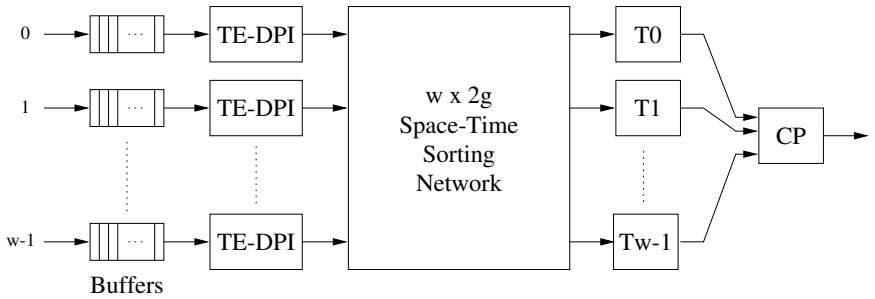

FIGURE 2. Group Output Interface (TE-IPI: Time Expander/Idle Packet Injector; Ti: fixed wavelength transmitter tuned to λ_i; CP: coupler).

to $w \times 2g$ 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. Individual columns of the sorted matrix are transmitted by the array of fixed tuned transmitters in the last stage of a GOI.

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 a buffer 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 $g \times w$ matrix (w spatial links and g time-slots) of slots, each capable of holding a packet. Depending on the offered traffic load, the slots may contain active or idle packets.

The following stage rearrange the matrix of packets in a way that individual columns will contain only packets destined to a same group and are ready to be transmitted as WDM packets. First, the Time Expander/Idle Packet Injector (TE-IPI) stage (see Fig. 3) introduces additional time-slots by a twofold internal speedup of the network. Incoming packets are retransmitted at double the bit rate and empty slots that follow the retransmitted packets are filled with idle packets with dummy group destination addresses assigned by the the idle packet generator. Dummy

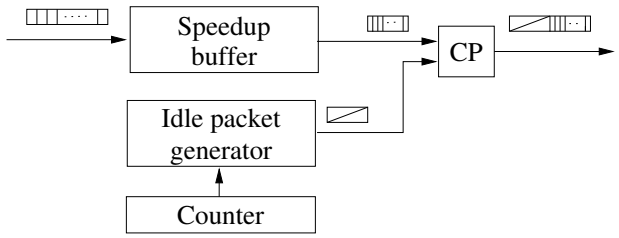

FIGURE 3. Time Expander/Idle Packet Injector (TE-IPI).

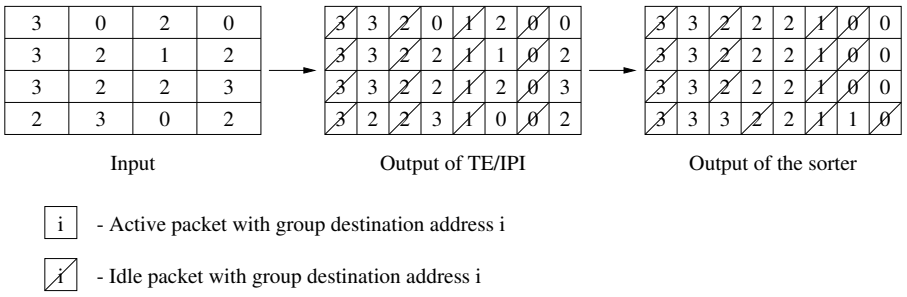

FIGURE 4. Example of routing in a GOI ($g = w = 4$).

destination addresses are generated according to the current value of a mod g counter that is incremented each time an idle packet is generated. Note that in a complete transmission cycle each TE-IPIs inserts g idle packets (one per destination group). There is w TE-IPI modules in a GOI which which brings the total number of inserted idle packets to gw and the complete $2g \times w$ matrix at the outputs of TE-IPIs contains w idle packet for every destination group. The space-time sorting network [**KLLJ94**] sorts $2g \times w$ matrix of packets in a column major order. It is built using 2×2 directional couplers and feedback delay lines. The activity bit is prepended as the least significant bit to the group destination address. Such an expanded address is used by the comparators in the sorting network. Assuming that an activity bit set to 1 denotes an idle packet and conversely 0 marks an active packet, after sorting all active packets destined to group i will be followed by exactly w idle packets with dummy destination address i, followed by active packets destined to group $i + 1$. Since the matrix column height is w, idle packets make it impossible to have a column of packets with active packets destined to more than one group. Fig. 4 shows an example of routing in a GOI of a network with $g = w = 4$. The input matrix contains 16 active packets and no idle packets. The size of the matrix is doubled by TE-IPI and 16 idle packets are inserted. After sorting, column 0 (the rightmost) contains packets to form a WDM packet destined to group 0 followed by WDM packets with group destination addresses 1, b, 2, 2, 3, 3, and b (b denotes a column with no active packets and therefore without a group destination address.

In the last stage of a GOI, 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 [**BKLJ95, BKLJ96**] is used to facilitate routing of packets

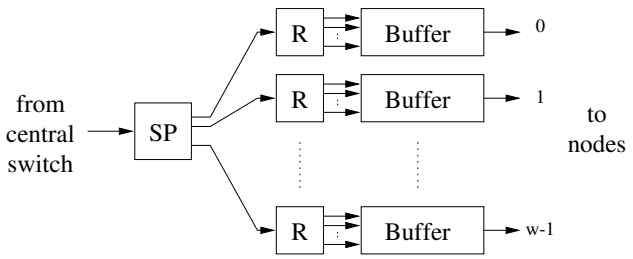

FIGURE 5. Group Input Interface (SP: splitter; R: diffraction-based parallel receiver).

within a destination group. The transmitters examine the packet destination and generate the protocol header while discarding the idle packet.

The operation of a GOI serving group i, $0 \leq i \leq g-1$, is phased back by $2i$ WDM packet transmission cycles. Such an arrangement reduces the probability of having more packets destined to the same group since GOIs outputs WDM packets with destination addresses in an increasing order.

GOIs are fully pipelined and only a bounded latency is introduced. There is no packet loss introduced by the architecture of a GOI.

2.2. Central Switch. The central switch is a $g \times g$ space-division switching network that routes WDM packets among groups. Since WDM packets are handled as units, the central switch can be built using wavelength insensitive switching elements. Any space division switching network capable of delivering L packets to an output (e.g., Knockout switch [**YHA87**], SXmin [**KBLJ97b**]) can be used as a central switch.

2.3. Group Input Interface (GII). The function of a GII is to receive and process WDM packets sent by the central switch. Individual packets in a WDM packet are separated and routed to the individual members of the group. The internal structure of a GII is the same as in STWnet.

The incoming WDM signal from the central switch is split by a passive optical splitter and distributed to individual receivers. A GII is equipped with w diffraction-based receivers capable of concurrent reception on all transmitting wavelengths. Receivers analyze the protocol information [**BKLJ95, BKLJ96**] and discard the packets which are not destined to the corresponding nodes. Buffers perform concentration of packets received on individual wavelengths and traffic smoothing.

3. 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 initial performance evaluation of the proposed network. Packet loss rate was chosen as the criterion to evaluate the performance characteristics. First two experiments evaluate the behavior of the network under a uniform traffic load. The main goal of these experiments was to find out the value of L for which the packet loss performance matches or exceed that of STWnet[2]. The following two experiments study the network under mixed uniform and localized (see 1.1) traffic loads to show that

[2]For the rest of the paper we assume an STWnet with twofold expansion ($lr = 2$).

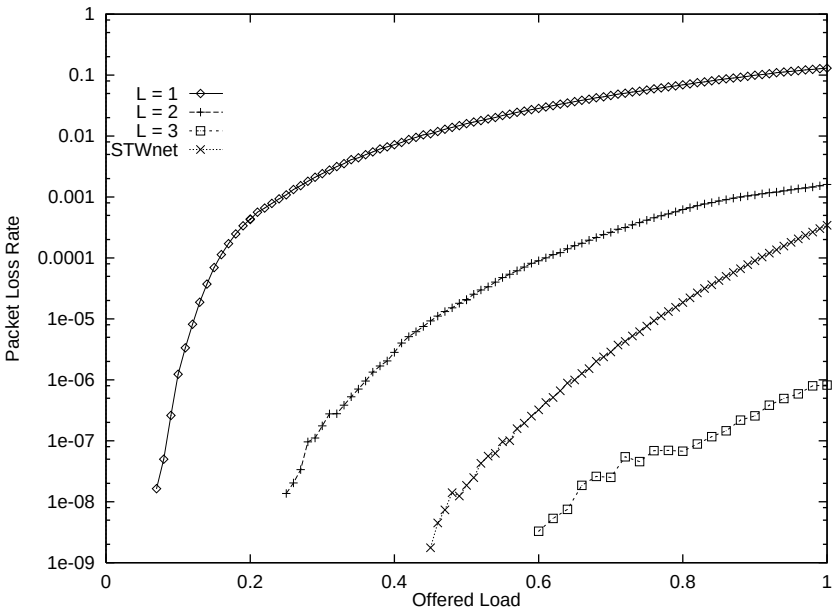

FIGURE 6. Performance of $A(8, 8, L)$ and STWnet under varying uniform traffic load.

it performs well under a wider range of offered loads than STWnet. Experiments were carried out for cases where $g = w$. Further experiments are required to study the tradeoffs in cases where $g \neq w$. The preliminary results presented in this paper do not consider packet loss due to output buffer overflow.

It should be noted that the latency of the proposed network consists of two components: constant delay introduced by the GOIs and the central switch and variable delay introduced by the output buffering in the GIIs. The analysis of the output queueing delays has been widely explored (e.g., [**KHM87**]) and is not repeated here.

3.1. Performance under uniform traffic. Fig. 6 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 $L \geq 3$.

The experiments shows (Fig. 7) that under uniform traffic the packet loss rate slightly increases as the network size grows larger. It can be seen that the rate of the increase is getting smaller for the larger network sizes. Similar behavior was observed in other Knockout Principle-based networks. Analytical models in [**YHA87**] and [**EKY92**] 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.

3.2. Performance under mixed traffic. For the next set of experiments we subjected the proposed network to a traffic load which is composed of inter-group localized and uniform traffic. Let r, $0 \leq r \leq 1$, be the percentage of fully inter-group localized traffic within the overall traffic. The load on an input in a time slot is generated according to localized traffic model with probability r or using

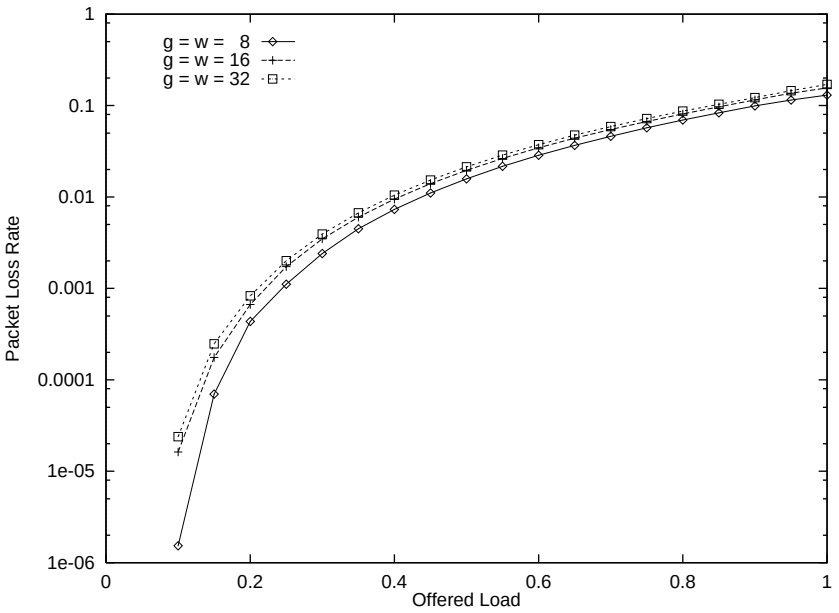

FIGURE 7. Performance of $A(w, g, 1)$ under varying uniform traffic load for various network sizes.

uniform traffic model with complementary probability $1 - r$. For $r = 0$ the traffic load becomes uniform and for $r = 1$ is fully inter-group localized.

As expected, the proposed network does not exhibit any packet loss under fully inter-group localized traffic since GOIs generate WDM packets destined to distinct output groups and therefore there are no output conflicts in the central switch to cause a packet loss. Fig. 8 shows the performance of the proposed network under mixed localized and uniform traffic. The outcome of the experiment shows that the network performance improves as one type of traffic becomes prevalent 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. 8 also shows the performance of STWnet switch. As expected, it performs well under uniform traffic but the packet loss rate rapidly increases as the traffic becomes more localized. The packet loss rate for a fully localized load is $1 - \frac{2}{g}$ due to the fixed nature of the transmission schedule of the central switch which provides slots for $2g$ packets per complete transmission cycle.

The effect of the network size on the packet loss performance under mixed localized/uniform traffic is shown in Fig. 9. As in the case of fully 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 all localized traffic becomes narrower as the network size increases.

4. Conclusions

The proposed self-routing network combines advantages of space and temporal switching with benefits of wavelength-parallel data transmissions. It can support a large number of nodes without stretching limits of the available technology. Tunable

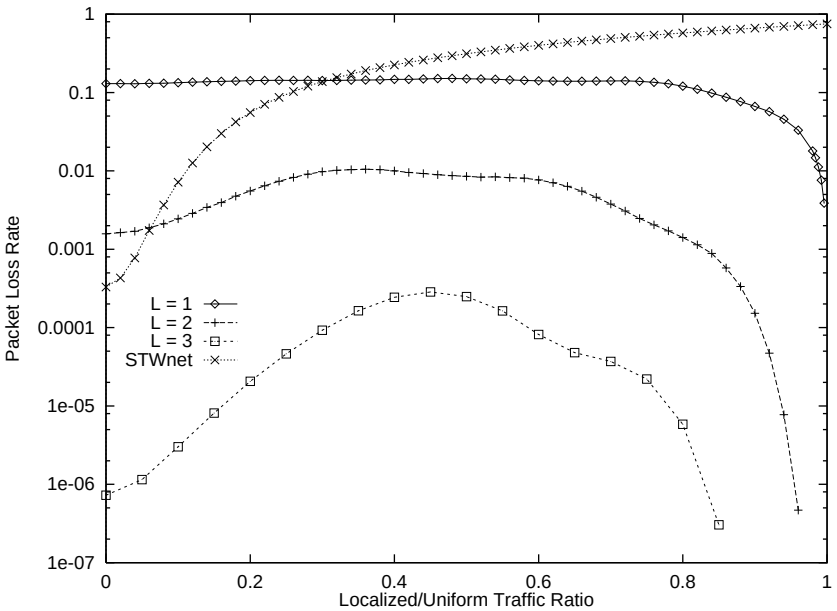

FIGURE 8. Performance of $A(8, 8, L)$ and STWnet under mixed uniform and localized traffic.

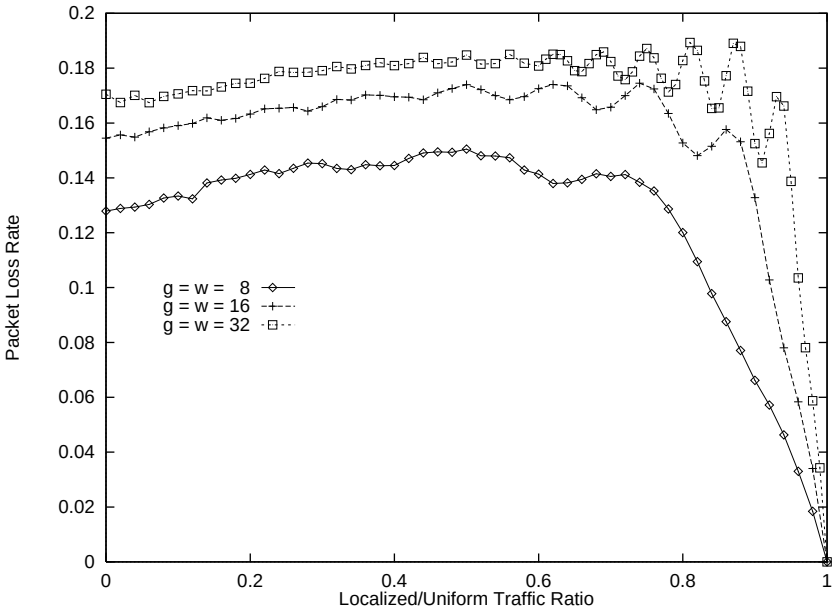

FIGURE 9. Performance of $A(g, w, 1)$ under mixed uniform and localized traffic for various network sizes.

components and wavelength switching are not required and wavelength channels are spatially reused. Preliminary experimental results indicate that the network performs well under uniform and localized traffic loads and exhibits tradeoff between cost and locality. Both systematic experimental evaluation and formal analysis of the network performance are interesting issues that remain to be investigated.

References

[BKLJ95] R. Bartoš, R. Kannan, K. Y. Lee, and H. F. Jordan, *Multinet: High-performance WDM interconnection architecture*, Proceedings ICC'95, June 1995, pp. 277–281.

[BKLJ96] R. Bartoš, R. Kannan, K. Y. Lee, and H. F. Jordan, *Multinet: High-performance WDM interconnection architecture*, Proceedings ICC'96, June 1996, pp. 277–281.

[CEF97] 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** (1997), no. 3.

[CEFS97] I. Chlamtac, V. Elek, A. Fumagalli, and C. Szabó, *Scalable WDM network architecture based on photonic slot routing and switched delay line*, Proc. of IEEE Infocom, Kobe, Japan, April 1997.

[DA94] P. W. Dowd and K. A. Aly, *Time-space-wavelength networks for low-complexity processor interconnection*, Proc. of Intnl. Conf. on Massively Parall. Proc. with Optical Interconnects, Cancun, April 1994, pp. 281–292.

[dlTK96] P. de la Torre and Clyde Kuskal, *Submachine Locality in the Bulk Synchronous Setting*, Proc. Euro-Par'96 Parallel Processing, vol. 2, 1996, pp. 352–355.

[dlTP97] P. de la Torre and M. D. Plumlee, *Towards a Distributed Realization of Dynamic Virtual Locality in the NOW Setting*, October 1997.

[EKY92] K. Y. Eng, M. Karol, and Y. S. Yeh, *A growable packet switch architecture: Design and applications*, IEEE Trans. on Communications **40** (1992), no. 2, 423–430.

[KBLJ96] R. Kannan, R. Bartoš, K. Y. Lee, and H. F. Jordan, *Optical TDM switching architectures with distributed control*, Proc. of IEEE Milcom, McLean, VA, October 1996, pp. 155–159.

[KBLJ97a] R. Kannan, R. Bartoš, K. Y. Lee, and H. F. Jordan, *STWnet: A high bandwidth space-time-wavelength multiplexed optical switching network*, Proc. of IEEE Infocom, Kobe, Japan, April 1997.

[KBLJ97b] 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** (1997), no. 6, 710–722.

[KHM87] 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** (1987), no. 12, 1347–1356.

[KLLJ94] R. Kannan, D. Lee, K. Y. Lee, and H. F. Jordan, *Optical TDM sorting networks for high-speed packet switching*, Proc. IEEE Globecom'94, December 1994, pp. 1948–1952.

[SSL94] J. Sharony, T. E. Stern, and Y. Li, *The universality of multidimensional switching networks*, IEEE/ACM Trans. on Networking **2** (1994), no. 6, 602–612.

[Tam88] Theodor Tamir (ed.), Guided Wave Optoelectronics, Springer Series in Electronics and Photonics, vol. 26, ch. 4, Springer Series in Electronics and Photonics, Springer-Verlag, 1988.

[YHA87] 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** (1987), no. 8, 1274–1283.

Department of Computer Science, University of New Hampshire, Durham, New Hampshire 03824
E-mail address: rbartos@cs.unh.edu

Department of Computer Science, University of New Hampshire, Durham, New Hampshire 03824
E-mail address: dltrr@cs.unh.edu

Department of Computer and Information Sciences, University of Michigan–Dearborn, Dearborn, Michigan 48124
E-mail address: rkannan@umdsun2.umd.umich.edu