A Photonic Fast Packet Switch for High-Speed Intra-node Routing: Architecture and Delay Analysis

KHALED A. ALY and PATRICK W. DOWD*
Department of Electrical and Computer Engineering
State University of New York at Buffalo
Buffalo, NY 14260

Abstract

A photonic fast packet switch is proposed for low-latency intra-node routing in high-speed optical networks. Applications are demonstrated in telecommunication and parallel computer environments. The switch fabric is based on integrated directional-coupler technology. It has low complexity and is geometrically suitable for real-estate efficient single substrate implementation. The switching architecture relies on time-division multiplexing an optically-coupled tree-structured active routing stage. Both synchronous and asynchronous TDM operation modes are modeled and analyzed to obtain packet buffering and total switching latencies.

1 Introduction

The development of photonic switching architectures is motivated by the recent advances in device technology and the need for high-bandwidth switching to accommodate the increasing offered traffic to a high-speed optical network. In large multi-hop networks, a significant amount of traffic arriving to a node is non-local (trunk traffic) and need to be further routed. In telecommunication networks, this traffic is conventionally handled by the same switch handling local traffic. In parallel computer networks, small-dimension space switches have been considered for intra-node routing in [1]. High-speed multiaccess protocols [2] and routing protocols [3] have been investigated for the same purpose. Figure 1 illustrates the requirement of trunk traffic routing in both environments.

This paper proposes the use of a simple time-division multiplexed photonic fabric to achieve intra-node trunk routing in either telecommunication or parallel computer networks. The fabric consists of an optically-coupled tree routing structure with time-division multiplexing (TDM) scheduling access to the photonic tree router. The architecture is denoted as time-division multiplexed tree routing stage (TDM-TRS). Integrated directional coupler switching devices are considered for the fabric realization. In these devices, surface optical waveguides

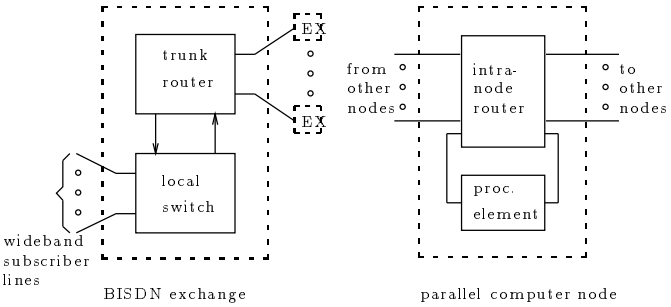

Figure 1: Intra-node *trunk* routing in telecommunication and parallel computer networks

are formed onto a $LiNbO_3$ crystal by Titanium diffusion. $LiNbO_3$ possesses a relatively large linear electrooptic effect that allows some variation of the waveguide's refractive index by means of applying an electric field. Reconfiguration times of $Ti{:}LiNbO_3$ directional couplers have been demonstrated on the order of fractions of a nanosecond [4]. Crossbar switches with dimension of up to 8×8 (64 couplers) have been fabricated [5]. However, topological restrictions are imposed by the physical limit on the number of couplers that can be integrated on a common substrate. The coupling length required to produce sufficient intensity modulation results in a long narrow dimension of directional couplers [6].

The following factors are considered when applying the proposed fabric to intra-node routing in telecommunication and parallel computer network environments:
High bandwidth: due to the photonic switch fabric, no electronic bottleneck is imposed on the wide-band trunk traffic. This is in consistency with the use of optical fibers for serial transmission.
Dimension: The number of nonlocal connections in both telecommunication network nodes and large parallel computer systems is fairly small as determined by the node topological degree. It is practical with current directional-coupler integration technology to fabricate a two-fold of the proposed fabric with dimensions up to

*This work was supported by the National Science Foundation under Grant CCR-9010774.

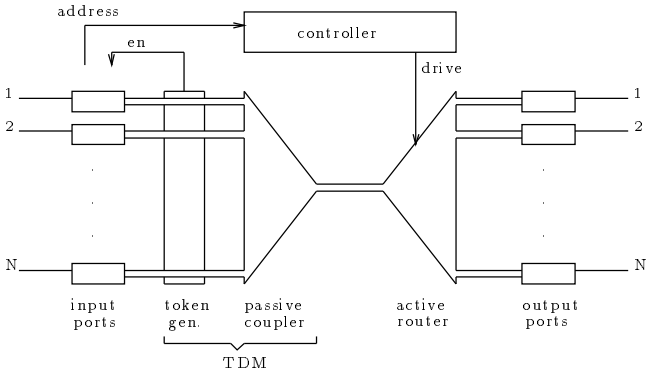

Figure 2: Organization of the TDM-TRS fast packet router

32. Due to TDM, the smaller the dimension, the higher the achievable line rate.

Multiplexing modes: a time-division multiplexed fabric is applicable to the ATM trend in telecommunication networks. In particular, synchronous TDM is suitable for routing highly uniform (pre-multiplexed) trunk traffic. On the other hand, asynchronous TDM is applicable to efficient low-latency routing of bursty computer traffic in optically-interconnected parallel computer systems.

The switch organization is detailed in Section 2. Both synchronous and asynchronous modes of operation are modeled and evaluated respectively in sections 3 and 4. Packet buffering delays and switching latencies are obtained. Results and conclusions are summarized in Section 5.

2 TDM-TRS Architecture

The TDM-TRS organization can be viewed as two subsystems: TDM subsystem and active tree router subsystem, as shown in Figure 2. The TDM subsystem couples the connected lines and drives them at a higher rate into a single aggregate optical channel. The synchronized packets are further fed to a tree-structured active routing network. A simple controller provides address decoding and drive signals at the synchronous switch mini-time slots. For a switch dimension of N, complexity is $O(N)$ in terms of the number of directional couplers and $O(log_2 N)$ in terms of the number of voltage drivers.

Fabric

The fact that the geometries of the coupler and router networks are similar and oppositely situated suggests the fabrication of a two fold switching network onto a common $LiNbO_3$ substrate. It is desirable to have the couplers stacked rather than cascaded on the substrate to take advantage of their physical dimension. Defining the width and height of a $Ti{:}LiNbO_3$ substrate in terms of the number of directional couplers in each dimension, respectively as w and h, the objective is to minimize w. Two possible two-fold configurations are illustrated in Figure 3. Configuration (a) has $w = 3log_2 N$ and $h = \frac{N}{2} + 1$ while configuration (b) has $w = 2log_2 N$ and $h = N$. A one-fold integrated crossbar has $w = 2N - 1$ and $h = N$.

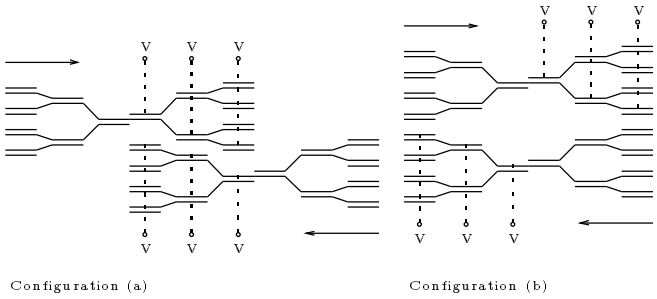

Figure 3: Directional coupler geometry for single-substrate two-fold implementation

I/O Ports

Input/output ports perform bit rate adaptation and provide packet buffering. The bit rates of the switching fabric and line are denoted respectively as R_s and R_l. The packet transmission time (time slots) of the line and fabric are therefore respectively $TS_l = 1/R_l$ and $TS_s = 1/R_s$. The ratio $L = R_s/R_l$ is defined as the switching bandwidth representing the maximum number of packets that could be routed during a line time slot.

In an input port, a received serial optical signal is converted into an L-bit parallel bus using a hybrid package that contains both an optical detector/amplifier and a bit demultiplexer. Incoming packets are held in a parallel (L-bit)-word FIFO buffer. The buffer output data bus is converted back to an optical serial line using another hybrid device that contains both a bit multiplexer and an optical source. Bit parallelism is introduced at the buffer stage to avoid the cost implications of using memory with access time comparable to the switch bit rate. Hybrid mux/demux packages have been fabricated with clock rates of up to 3.2 GHz [7]. This is the only component in the system that is clocked at a rate higher than the line rate. A processor with parallel data bus bit rate R_b and width W_b may directly connect to an L-bit parallel port so long as $R_b \leq R_s/W_b L$ [8]. Output ports perform the complement function. Figure 4 shows the input port organization when $R_l = 0.4$ Gbps, $L = 4$ and consequently $R_s = 1.6$ Gbps. The dashed line at the input to the FIFO buffer represents the option of connecting an 4-bit parallel bus at 100 Mbps.

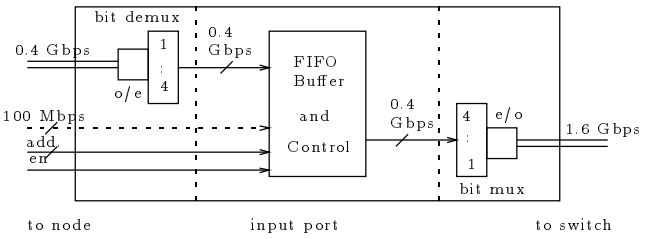

Figure 4: Input port organization

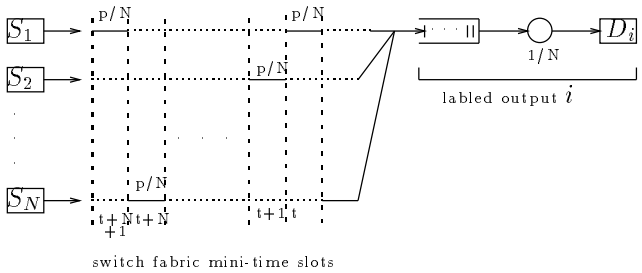

Figure 5: STDM-TRS queueing model (Geom/D/1)

Multiplexing Modes

The TDM subsystem may operate in either synchronous or asynchronous multiplexing modes. The system is denoted respectively STDM-TRS or ATDM-TRS according to the multiplexing mode. TDM is thought of as a cyclic contention resolution scheme. It is implemented by a *token loop* that consists of a closed ring of edge-trigerred delay elements, clocked at the TS_s rate [8]. Only one token is active at any TS_s.

In STDM-TRS a one-packet input buffer is required for synchronization and hold till selection. STDM slightly simplifies the routing control, but it requires a high switching bandwidth ($L = N$). This implies higher complexity of the switch I/O ports in terms of the hybrid mux/demux package. It is only practical to use STDM routing for very uniform loads (eg. high-level trunk switching in telecommunication networks) and when the switch dimension-node bit rate product is implementationally possible.

In ATDM-TRS ports that are not backlogged in a particular round are bypassed by the token loop. Packets are queued on both inputs and outputs. When $L = 1$, queueing is only at inputs, while the opposite holds when $L = N$. ATDM-TRS has a potential application as an internal routing structure in optically-interconnected parallel computer systems with topologies of proportionally high node degree such as the hypercube [1].

3 STDM-TRS Performance Analysis

An exact queueing model for the STDM-TRS is introduced in this section. We obtain the mean waiting time and the total switching latency a packet encounters as it enters the switch enroute to its next destination. It is to be noted that the STDM-TRS performance is similar to that of a nonblocking switch with strict output queueing [9], since it provides guaranteed packet transport within the line time slot TS_l of arrival. An independent model is developed for the STDM-TRS and the output waiting time results are shown to match those of [9].

The system queueing model is shown in Figure 5. The arrival process is assumed to be Poisson and the source nodes are assigned switch mini-time slots in a cyclic manner. The process of synchronized packet arrivals has a corresponding geometric distribution with parameter p. All source and destination nodes are assumed symmetric in terms of packet generation process and likelihood of being addressed. At each TS_s, the probability of a labeled output queue being addressed is p/N (uniform reference model), since each source node will be sampled only once at a $TS_l = N \ TS_s$. The arrival process to the labeled output queue is then a geometric progression of Bernoulli trials with *success* rate equal to p/N. The service time for this Geom/D/1 queue is $N \ TS_s$, which is the packet transmission time at the line rate.

The packet hold time at the input stage T_h, until allocated a mini-time slot, is uniformly distributed and bound by $(N-1) \ TS_s$: $P[T_h = k] = \frac{1}{N}, \quad k = 0, 1, ..., N-1$, so

$$\begin{aligned} E[T_h] &= \sum_{k=0}^{N-1} \frac{k}{N} \ TS_s \\ &= \left[\frac{N-1}{2}\right] \ TS_s \\ &= \left[\frac{N-1}{2N}\right] \ TS_l \end{aligned} \tag{1}$$

The hold time approaches $0.5 \ TS_l$ for large N. The following result of discrete-time queueing theory [10] is used to obtain the mean waiting time in the output queue:

$$E[W_o] = \frac{S_0 \rho (1 - \delta t / S_0)}{2(1-\rho)} \tag{2}$$

where S_0 is the fixed service time, δt is the unit "embedding" time and ρ is the output trunk utilization in Erlangs. Using $S_0 = N$, $\delta t = 1$ and $\rho = \frac{p/N}{1/N} = p$, the output buffer mean waiting time is:

$$\begin{aligned} \overline{W}_o &= \frac{p(N-1)}{2(1-p)} \ TS_s \\ &= \left[\frac{N-1}{N}\right] \left[\frac{p}{2(1-p)}\right] \ TS_l \end{aligned} \tag{3}$$

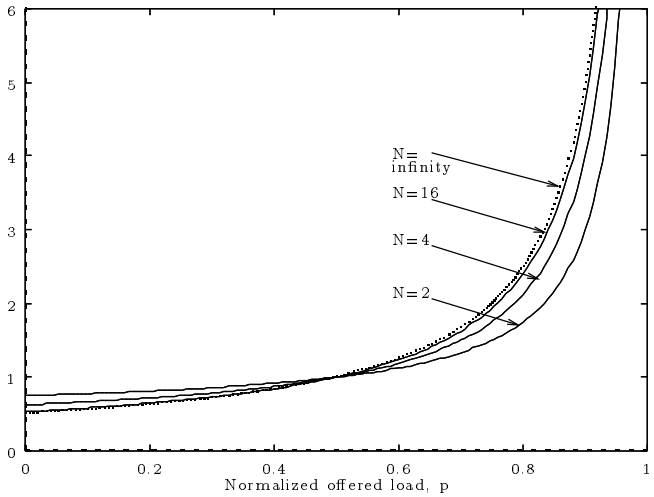

Figure 6: STDM-TRS total switching latency in line time slots

The total system time T_{sys} (referred to alternately as total switching latency) consists of three components: hold time T_h, fabric transmission time $T_x = 1\ TS_s$, and output buffer waiting time W_o. From the linearity property of the expectation of a summed random variable:

$$\begin{aligned} \overline{T}_{sys} &= \overline{T}_h + T_x + \overline{W}_o \\ &= \frac{N + (1-2p)}{2N(1-p)}\ TS_l \end{aligned} \tag{4}$$

For an arbitrarily large switch dimension

$$\overline{T}_{sys}|_{N\to\infty} = \frac{1}{2(1-p)}\ TS_l \tag{5}$$

The total system time is plotted in Figure 6 for $N \in \{2, 4, 16, \infty\}$. The normalized switch throughput: $\rho_o = p$. The results have been validated by simulation [11].

It is interesting to note that the result of Eqn. (3) agrees with the result obtained in [9] for the mean output queueing time in the multicast-select space-division architecture of switching bandwidth equal to N^2. The queueing model of [9] assumed batch arrivals to multiple parallel buffers at output and random selection to exit the queue. The mean waiting time is found to be equivalent to our case where arrivals are scheduled to enter the output queue on cyclic basis at mini-time slots, each equal to $(1/N)\ TS_l$. In addition to the output buffer time, the cyclic scheduling requires selection time at the input stage $T_h = [(N-1)/2N]\ TS_l$ and transmission time $T_x = (1/N)\ TS_l$. The multicast-select policy requires transmission time of $1\ TS_l$ since the fabric operates at the same bit rate as the lines. The cyclic scheduling therefore incorporates less total switching latency. The difference is $T_h = [(N-1)/2N]\ TS_l$ which approaches $0.5\ TS_l$ when $N >> 1$.

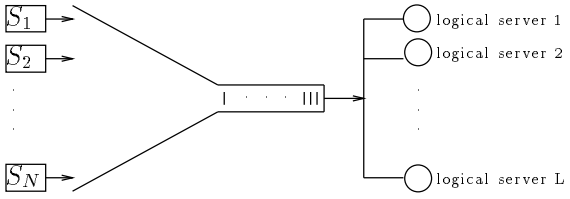

Figure 7: ATDM-TRS queueing mode (Geom/D/L with batch arrivals)

4 ATDM-TRS Performance Analysis

Model assumptions of the previous section are considered holding in this section. An exact descriptive model of the input buffers dynamic behavior under ATDM would require the solution of an array of inter-dependent difference equations, each representing a single input queue [8]. This is due to the dynamic assignment of the fabric's mini-time slots based on the backlog status of input ports throughout the multiplex cycle. A simplified exact model based on Kleinrock's work conservation law for queueing systems [12] is derived.

The work conservation law states that in a work-conservative queueing system, the distribution and consequently the mean, of the waiting time in the queue is independent of the order of service. This is provided that the service priority discipline is not a function of the service time of the customers, or any variation of their service time. The ATDM-TRS is work-conservative since switching time between ports is insignificant and not neglected. As a result a single aggregate queue may be considered instead of the array of inter-dependent queues. Since the service priority is determined by the cyclic policy and is independent of the service time, a FCFS queue is considered instead, without affecting the waiting time results [13]. The resulting queue is a Geom/D/L with batch arrivals, shown in Figure 7. The L servers represent the relative bit rate of the switch fabric (the switching bandwidth). The behavior of this queue is described by

$$X_{m+1} = X_m + A_{m+1} - \Psi_{X_m} \tag{6}$$

where X_n is defined as the buffer occupancy at the epoch of TS_l m, and A_m is the number of packet arrivals during m, now given by the binomial distribution whose PGF is $A(z) = (pz + 1 - p)^N$. The number of transmitted packets during m is:

$$\Psi_{X_m} = \begin{cases} X_m & \text{if } X_m \le L \\ L & \text{if } X_m \ge L \end{cases} \tag{7}$$

The mean buffer occupancy is obtained as $\overline{X} = \lim_{z\to 1} \frac{d}{dz} X(z)$. The standard approach of [14] can be fol-

lowed to obtain $X(z)$. When both sides of (6) are raised to the power of z and the first moment is taken, the following term needs to be evaluated:

$$\begin{aligned} E[z^{X_m - \Psi_{X_m}}] &= \sum_{k=0}^{\infty} P[X_m = k] z^{k-\Psi_k} \\ &= P[X_m = 0] + \ \ldots + P[X_m = L-1] \\ &\quad + \sum_{k=L}^{\infty} P[X_m = k] z^{k-L} \end{aligned}$$

Aside from the difficulty of evaluating the summation on the right hand side, the knowledge of $P[\tilde{X} = 0]$, $P[\tilde{X} = 1]$, . . ., $P[\tilde{X} = L-1]$ is mandatory ($\tilde{X}$ is the steady state queue occupancy). Only $P[\tilde{X} = 0]$ is known and equals $1-\rho$, where ρ is the queue utilization factor [14]. The remaining probability terms may be determined by solving an individual Markov chain for each value of $L > 1$. We choose to solve exactly for $L = 1$ and rather obtain an approximated general solution for $1 < L < N$ which is justified by simulation.

Case 1: $L = 1$

When $L = 1$ both the arrivals and departures of the aggregate queue take place at synchronous and equal time slots. The queueing model has a single server in this case, the service time is equal to 1 TS_l.

$$X(z) = (1-\rho)\frac{A(z)(z-1)}{z - A(z)} \tag{8}$$

The mean buffer occupancy is obtained using $\rho = pN$:

$$\overline{X} = \frac{2Np - N(N+1)p^2}{2(1-Np)} \tag{9}$$

Case 2: $1 < L < N$

Since the steady state probabilities of the queue occupancy being between 1 and $L-1$ are not known, we remain with the solution of normalizing the arrival rate to the switch mini-time slots. Therefore, at each TS_s, a packet arrives with probability p/L. Based on this normalization, the queue is the same Geom/D/1 with batch Bernoulli arrivals and a single server. Substituting for the normalized arrival rate:

$$\overline{X} = \frac{2NLp - N(N+1)p^2}{2L(L-Np)} \tag{10}$$

This solution is an approximate one since breaking down the arrival rate into $1/L$ segments results in loss of independence between successive arrivals. For example, if an arrival took place at TS_s n, it is known in advance that no other arrival may take place until TS_s $n+L$. This dependence loss, as proven by comparison to simulation, has little effect for $L < N$.

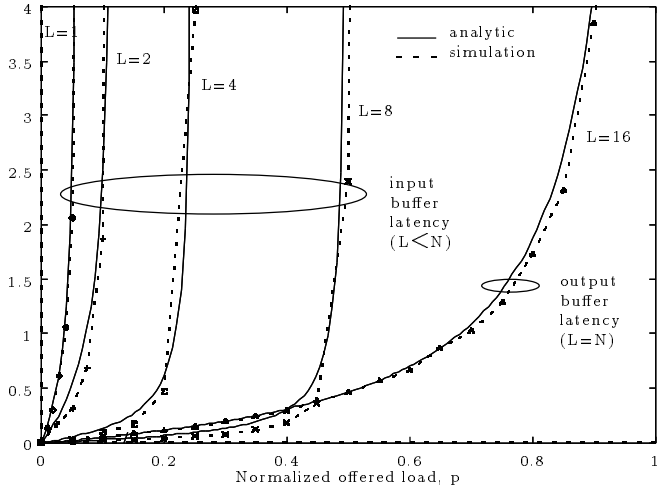

Figure 8: ATDM-TRS input/output buffer time latence (analysis vs. simulation) in line time slots; N=16

In the two previous cases with buffering on inputs (i.e. $1 \leq L < N$), the mean waiting time at input buffer W_i is given by the mean buffer occupancy multiplied by the deterministic service time

$$\overline{W_i} = \frac{\overline{X}}{L}, \qquad 1 \leq L < N \tag{11}$$

Case 3: $L = N$

In this case simulation has predicted the insignificance of input buffering time [11]. Reference can be made to the STDM-TRS case when the server was N times faster than the nodes. Due to the asynchronous server being dynamic, it may appear that packets are moved to the output buffers in a faster manner than by a synchronous server. However, due to the work conservation law, the ATDM server at steady state will not do more work than the work pending in the system. Therefore, output buffering time of ATDM-TRS with $L = N$ and that of STDM-TRS are essentially the same. This conclusion is verified by simulation.

Model Results

Figure 8 shows the buffering latencies for $N = 16$ and $L \in \{1, 2, 4, 8, 16\}$. When $L < N$, buffering takes place at the input stage with no significant output buffering delay, as was verified by simulation. An exact model is used for $L = 1$, while $L = 2$, 4 and 8 are obtained by normalizing the arrival rate. Input buffering latency is insignificant when $L = N$, and the corresponding curve represents output buffer latency as given by Eqn. (10). Simulation curves are plotted for all cases to verify the model accuracy.

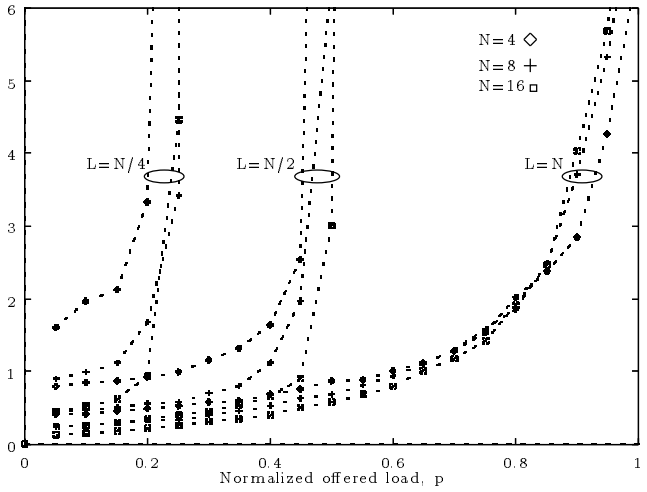

Figure 9: ATDM-TRS total switching latency (simulation) in line time slots

The total switching latency is evaluated by simulation and plotted in Figure 9 for $N \in \{4, 8, 16\}$ and $L \in \{N/4, N/2, N\}$. The plots show that for low offered loads the transmission time dominates when L is small. It appears in the graphs of Figures 8 and 9 that the switch always saturates at an offered load $p = \frac{L}{N}$, which is the buffer delay singularity expected from Eqn.s (9), (10) and (3).

5 Summary

The paper has introduced and analyzed a photonic fast packet router based on time-division multiplexing a tree-structured routing stage (TDM-TRS). Two fabric geometries have been suggested, based on integrated directional coupler switches for two-fold single-substrate fabrication. Delay analysis has been conducted for synchronous and asynchronous TDM modes.

STDM-TRS is simpler to implement but requires the fabric to operate at a bit rate equal to the sum of the bit rates of the connected lines. It is practical for small dimensions to handle uniform trunk traffic sources. It was shown that the performance of STDM-TRS resembles that of a nonblocking space switch with strict output queueing. Due to its narrow time slot (packet transmission time through the fabric), it results in lower total switching latency than the multicast-select approach. This difference is proportional to the dimension N and approaches half the line time slot period for large N.

ATDM-TRS was modeled using the work conservation law of queueing systems. Exact results were obtained for the cases $L = 1$ and $L = N$ and an approximated result was obtained for $1 < L < N$. Output buffering takes place only when $L = N$, and the switch saturation load was shown to be given by L/N in all cases. This ratio represents the buffer delay singularity. The total switching latency was evaluated by simulation, accounting for the insignificant input/output (depending on L) buffering that has not been analytically evaluated. The domination of transmission time, below saturation, appears when the dimension is small ($N = 4$).

6 References

[1] P. W. Dowd, M. Dowd and K. Jabbour, "Static Interconnection Network Extensibility based on marginal Performance/Cost Analysis", *IEE Proc.,* Vol. 136, Pt. E, No. 1, pp. 9-15, Jan. 1989.

[2] P. W. Dowd, "High Performance Interprocessor Communication Through Optical Wavelength Division Multipleaccess Channels", *Proc. 18th Int. Symp. Comput. Arch.,* Toronto, Canada, May 1991.

[3] J. R. Sauer, "Multi-Gbps Optoelectronic Interconnection System", *SPIE,* Vol. 1178, *Optical Interconnects in The Computer Environment,* pp. 36-45, 1989.

[4] H. S. Hinton, "Photonic Switching Using Directional Couplers", *IEEE Commun. Mag.,* Vol. 25, N0. 5, pp. 16-26, May 1987.

[5] R. C. Alferness, "Waveguide Electrooptic Switch Arrays", *IEEE J. Sel. Areas Commun.,* Vol. 6, No. 7, pp. 1117-1130, Aug. 1988.

[6] R. V. Schmidt and R. C. Alferness, "Directional Coupler Switches, Modulators, and Filters Using Alternating $\Delta\beta$ Techniques", *IEEE Trans. Circuits and Systems,* Vol. CAS-26, No. 12, pp. 1099-1108, Dec.1979.

[7] M. K. Kilocyne et Al., "Optical Signal Interconnection Between GaAs Integrated Circuit Chips", *SPIE* Vol. 703, *Integration and Packaging of Optoelectronic Devices,* pp. 148-155, 1986.

[8] K. A. Aly, "Fast Packet-Switched Photonic Interconnection Networks: Architecture and Performance Evaluation", *M.S. Thesis,* Dept. Elec. and Comp. Engr., State University of New York at Buffalo, Jan. 1991.

[9] M. J. Karol, M. G. Hluchyj, and S. P. Morgan, "Input Versus Output Queueing on a Space-Division Packet Switch", *IEEE Trans. Commun.,* Vol. COM-35, No. 12, pp. 1347-1356, Dec. 1987.

[10] T. Meisling, "Discrete-Time Queueing Theory", *Oper. Res.,* pp. 97-105, Feb. 1957.

[11] K. A. Aly and P. W. Dowd, "Simulation of Fast Packet-switched Photonic Networks for Interprocessor Communications", *Int. J. Comput. Sim.,* to appear, 1992.

[12] L. Kleinrock, *Queueing Systems, Vol. II: Computer Applications,* John Wiley and Sons, 1976.

[13] O. J. Boxma and W. P. Groenendijk, "Waiting Times in Discrete-Time Cyclic-Service Systems", *IEEE Trans. Commun.,* Vol. 36, No. 2, pp. 164-170, Feb. 1988.

[14] L. Kleinrock, *Queueing Systems, Vol. I: Theory,* John Wiley and Sons, 1975.