Analysis and Comparison of Hot-Potato and Single-Buffer Deflection Routing in Very High Bit Rate Optical Mesh Networks

Fabrizio Forghieri, Alberto Bononi, and Paul R. Prucnal, *Fellow, IEEE*

***Abstract*—The steady state behavior of regular two-connected multihop networks in uniform traffic under hot-potato and a simple single-buffer deflection routing technique is analyzed for very high bit rate optical applications. Manhattan Street Network and ShuffleNet are compared in terms of throughput, delay, deflection probability, and hop distribution both analytically and by simulation. It is analytically verified that this single-buffer deflection routing technique recovers in both networks more than 60% of the throughput loss of hot-potato with respect to store-and-forward when packets are generated with independent destinations. This gain, however, decreases to below 40% when the average message length exceeds 20 packets.**

I. Introduction

MULTIHOP packet-switching networks with regular two-connected mesh topologies, such as Manhattan Street Network (MS) [1] and ShuffleNet (SN) [2], have been proposed for all-optical implementation at very high bit rates [3], [4]. While in electronic networks buffering of hopping packets at intermediate nodes is commonly used with conventional store-and-forward routing, the same is not true of all-optical networks, where the only fast access optical memories available are simple recirculating fiber delay loops which require optical amplification, thus becoming impractical. Deflection routing [5], with its inherent limited-time buffering, can eliminate the need of optical amplifiers in the optical memory [6]. Even more dramatic simplification is obtained with hot-potato [7], which is a special case of deflection routing where buffers are not provided at all.

In these networks an all-optical path is provided between source and destination, without intermediate regeneration of the optical signal. Therefore at very high bit rates the propagation distance, proportional to the number of hops taken by a packet from source to destination, becomes a limiting factor. A limit on the distance-bit rate product is imposed by the optical channel if the packet error rate has to be kept below a fixed threshold. Under deflection routing, the number of hops taken by each packet from source to destination varies randomly with network traffic. In a deflection routing network with equal link lengths, if the average statistics of the number of hops n are known, the packet error rate can be obtained by conditioning on n as

$$P(e) = \sum_{n=1}^{\infty} P(e/n)P(n). \tag{1}$$

For fixed n and fixed link length, the conditional error rate $P(e/n)$ depends on the specific optical channel selected for the network and is a point-to-point communication problem. Knowledge of the distribution of the number of hops $P(n)$ is then necessary in network design to find the maximum bit rate and hence the maximum throughput achievable for a given offered load and physical size of the network.

This paper analyzes the steady state behavior of two-connected mesh networks under deflection routing. The one-packet analytical model appearing in [8], [9] for hot-potato routing is reviewed and extended to the single-buffer memory configuration proposed in [6], which is particularly attractive for optical implementation. Simulation results are provided to confirm the validity of the analytical models and stress the consequences of violating some of the underlying assumptions.

Hop distribution curves $P(n)$ are provided for both MS and SN under various loads, for different network sizes, with no buffers and with the above mentioned single-buffer memory.

Section II reviews deflection routing and summarizes some topological properties of mesh networks whose interplay determines the global network behavior under deflection routing.

Section III describes node operation and provides a detailed analysis of the steady state behavior of two-connected regular mesh networks under both hot-potato and single-buffer deflection routing.

In Section IV, analytical results for MS and SN are discussed and checked against simulation results. These two topologies are compared for 64 node and 400 node sizes and the improvement achievable with single-buffer deflection routing with respect to hot-potato is evaluated. The degradation caused by transmission of long streams of consecutive packets from the same node to a fixed destination is evaluated by simulation.

II. Topological Properties of Importance in Deflection Routing

A two-connected network is one in which each node has two input links and two output links. In this paper the

Paper approved by I. M. Habbab, the Editor for Local Lightwave Networks of the IEEE Communications Society. Manuscript received July 22, 1992; revised April 22, 1993. This paper was presented in part at the IEEE INFOCOM '93, San Francisco, CA, March 28–April 1, 1993.

F. Forghieri was with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 and with the Dipartimento di Ingegneria dell'Informazione, Università di Parma, I-43100 Parma, Italy. He is now with AT&T Bell Laboratories, Crawford Hill Laboratory, Holmdel, NJ 07733 USA.

A. Bononi and P. R. Prucnal are with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA.

IEEE Log Number 9406240.

behavior of two-connected networks using deflection routing is investigated. Deflection routing [5] is a shortest path routing algorithm where buffer overflow is handled without discarding packets.

Assume that a first-in-first-out (FIFO) buffer having N_b one-packet memory elements is provided on each output link. Routing and buffering proceed as in store-and-forward up to the time where one of the two queues overflows. At that point the overflowing packet is deflected onto the other queue. This is possible since the two queues cannot be completely full at the same time, and a single shared output queue turns out to be enough [10]. Deflection routing is thus a variation on store-and-forward where no packet loss occurs and the queueing delay remains bounded by the number N_b of memory elements. A special case of deflection routing is when buffering is not provided at all. This routing is called hot-potato [7].

There are three structural properties whose interplay determines the performance of a multihop network under deflection routing. These are detailed below.

1) *Diameter:* Consider the distance in hops between any pair of nodes along a shortest path connecting them. The diameter of the network $d_{\max}$ is defined as the maximum of this distance over all node pairs in the network. This number is a good indicator of how compact a network is. Starting from a generic node, the smallest diameter two-connected network would ideally have a binary spanning tree reaching 2^i new nodes at each level i of the tree. The Shuffle Exchange Network (SX) is the smallest-diameter implementable two-connected network, where the $N = 2^{n_{\rm SX}}$ nodes are arranged in a single column and connected in perfect shuffle, with an alteration of the perfect shuffle connection at the first and last node of the column [5]. The diameter of SX is $d_{\max} = n_{\rm SX} = \log_2 N$. This network is not perfectly symmetric, i.e., the network configuration, as seen from the node's point of view, is not the same for all nodes.

ShuffleNet is a generalization of Shuffle Exchange in which the $N = n_{\rm SN} 2^{n_{\rm SN}}$ nodes are organized in $n_{\rm SN}$ columns of $2^{n_{\rm SN}}$ nodes each, and each column is connected to the next column in perfect shuffle [2]. Unlike SX, SN is a regular network, in that all nodes are topologically equivalent. The diameter of SN is $d_{\max} = 2n_{\rm SN} - 1 \approx 2\log_2 N$ for large N. Therefore SN has asymptotically minimum diameter as N increases to infinity.

Another regular network considered here is the Manhattan Street Network, in which the $N = n_{\rm MS}^2$ nodes, where $n_{\rm MS}$ is even, are organized in a toroidal grid of $n_{\rm MS}$ rows and $n_{\rm MS}$ columns, with alternating directions like the one-way streets in Manhattan [1]. The diameter of MS is [11] $d_{\max} = n_{\rm MS+1}$ if $n_{\rm MS}/2$ is even, and $d_{\max} = n_{\rm MS}$ if $n_{\rm MS}/2$ is odd. Hence $d_{\max} \approx \sqrt{N}$, much greater than the minimum for large N.

2) *Deflection Cost:* The maximum increase in path length, expressed in number of hops, due to a single deflection is called the deflection cost $c_{\rm def}$ of a network. This number is a good indicator of how much network performance degrades under heavy load. In SX, deflections at a node whose distance to destination is k hops result in an increase of $n_{\rm SX} - k+1$ hops in path length. Hence $c_{\rm def} = n_{\rm SX} = \log_2 N$. In SN all deflections result in an increase of $n_{\rm SN}$ hops. Therefore $c_{\rm def} \approx \log_2 N$, increasing with the network size N like in SX. On the contrary, in MS the maximum cost is $c_{\rm def} = 4$ for every N, since it is always possible to "walk around the block" and get back to the point where deflection occurred in four hops. This is the main reason why the tails of the average hop distribution in MS decay much faster than the tails in a SN or SX of similar size, as it is shown in Section IV.

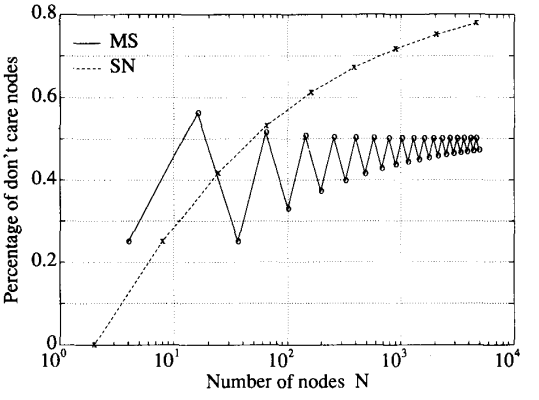


Fig. 1. Percentage of don't care nodes in MS and SN versus network size N.

3) *Don't Care Pattern:* For a given destination node, any other node in the network is *don't care* if both its output links lie on a shortest path to destination. The presence of a high percentage of don't care nodes helps keep the number of deflections to a low level even at high loads. The percentage of don't care nodes $DC\%$ in MS and SN is plotted in Fig. 1 as a function of network size N. For MS the percentage converges oscillating to 50% for large N. Note that, although when $n_{\rm MS}/2$ is odd the network is more compact, the don't care percentage is always higher when $n_{\rm MS}/2$ is even. For SN this percentage tends to 1. The percentage formula for SN is $DC\% = 1 - (2/n_{\rm SN})(1 - 1/2^{n_{\rm SN}})$, and is simply found by looking at the regular don't care pattern. Finally, note that SX has a negligible number of don't care nodes, and these are generated by the alteration of the original perfect shuffle connection at the first and last node of the column. Otherwise it would have none.

A fair comparison between these networks is difficult, since they exist for distinct sets of number of nodes, whose intersection contains only a few cases. However, many networks with number of nodes close to within less than 10% can be found and compared.

MS and SX under deflection routing have already been compared in [10] by simulation for large network sizes. The results show that at high loads MS has larger throughput than SX when no buffers or only one buffer are provided. As the number of buffers is increased, the throughput in SX eventually exceeds that in MS as it should, owing to its minimum diameter property. This shows that, with one or no buffers, the increased number of deflections at high loads completely erodes the initial advantage of SX due to its low diameter. This is caused by the higher deflection cost and the absence of don't care nodes is SX.

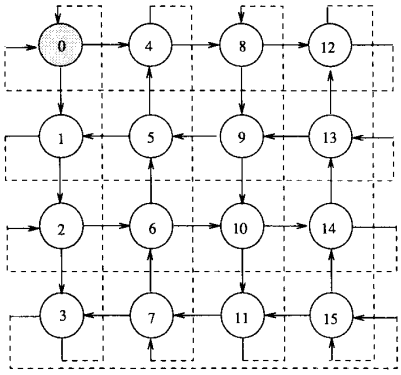


Fig. 2. 16-node Manhattan Street Network.

MS and SN have been compared in [12] under deflection routing for 64 nodes, the only manageable size at which the two networks exist exactly for the same number of nodes. It has been shown there that even without buffers SN has throughput larger than MS at every load. The initial advantage in diameter of SN over MS is thus preserved under heavy traffic by the high percentage of don't care nodes. Note that, for 64 nodes, the two networks have the same deflection cost $c_{\text{def}} = 4$ and the same don't care percentage.

Now, SN and SX are structurally very similar. However, in SX almost every node is *care*, and at full load traffic streams destined to different recipients are all mixed, and congestion quickly builds up. On the contrary, in SN only those nodes whose minimum distance from destination is less than n_{SN} are care nodes. All other nodes are don't care. For given destination node j, outside the "ball" of radius n_{SN} centered around j, traffic destined to node j does not interfere with routing of other packets. For a growing network size this ball is relatively smaller and smaller, and thus even at full load the network does not suffer congestion. This structural property allows SN to outperform SX when deflection routing with a small number of buffers is used.

Since this work is motivated by very high bit rate optical applications of deflection routing, results for a very limited number of optical buffers are of interest. For this reason, in the next sections only MS and SN will be considered.

III. Network Operation and Steady-State Analysis

Consider a two-connected regular mesh network, such as the 16 node MS (MS16) shown in Fig. 2. A common clock is distributed to all nodes, so that node operations are performed in fixed length time slots, and the time axis is discrete. The logical structure of the node is shown in Fig. 3. During each slot, each node performs the following operations:

1) *Absorption:* Incoming packets destined to the node are absorbed. It is assumed that reception (RX) can be performed on both links at the same time.

2) *Generation:* If a new packet is ready for transmission (TX), and if after the absorption block at least one of the two links is free, the new packet is inserted for transmission. It is assumed that only one new packet can be inserted per slot at the node.

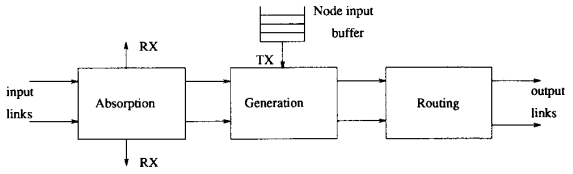


Fig. 3. Logical node structure.

3) *Routing:* Transiting and locally generated packets are routed to the output links or possibly buffered when buffering is provided in this block.

Note that the slotted system allows polite access to the network. A new packet is not inserted if the input links are occupied by transiting packets. This provides an automatic form of flow control.

Assume the network has N nodes, so that the total number of links is $2N$. The steady state behavior will now be analyzed. New arrivals at each node are collected in an input buffer, waiting to be injected in the network. Arrivals are assumed to occur at the same rate and independently at each node. It is assumed that at each node the destination of new packets is chosen independently of other nodes and independently of previously admitted packets, and is drawn from a distribution that is uniform on all other nodes. The reasoning behind these assumptions is that this destination pattern helps the routing algorithm share the load evenly among all links.

With this traffic homogeneity assumption, the local input queues are evenly served. Let g be the probability, equal for all nodes, that the node input buffer has at least one queued packet per slot. Thus g is the probability that a new packet at each node is ready for transmission at every clock. It will be referred to as the generation probability per slot.

Let λ be the network throughput, that is the average number of packets inserted/absorbed per slot in the network at equilibrium.

Define W as the ratio of the link length to the spatial length of a slot, i.e., the number of slots in flight on each link at any time. All links are assumed to have the same length, and W is assumed to be an integer number, which means that the propagation delay on each link is an exact multiple of the slot time. In optical links, W is given by

$$W = \frac{l}{c/n}\frac{R}{M} \approx 10R[\text{Gb/s}]l[\text{km}]$$

where l is the link length, c/n the light speed in optical fibers of refraction index $n = 1.5$, R is the bit rate and M is the packet size, and the numerical value is obtained for the asynchronous transfer mode (ATM) packet size of 424 bits. In very high bit rate optical networks this ratio W is very high and the propagation delay dominates the queueing delay at intermediate nodes when W is much larger than the buffer size.

At any clock time the network links contain $2NW$ slots. Let u be the probability that a spatial slot is occupied by a packet. By the balanced load assumption, u is the same for every slot in the network.

Applying Little's theorem [13] to the whole network, the following balance equation is obtained

$$2NWu = \lambda D_s$$

where D_s is the average propagation delay in number of slots and $2NWu$ the average number of packets in the links at any time at equilibrium. If D indicates the average number of hops, then $D_s = WD$ and Little's formula is simply

$$2Nu = \lambda D. \tag{2}$$

In this analysis the waiting time at the node input queue is not treated. The total delay of a packet, once injected in the network, is the sum of the propagation delay D_s and of the queueing delay D_q at the routing block. For very high bit rate optical networks D_q is small compared to D_s and can thus be neglected.

The probability of packet absorption per slot on a given input link at a node is (3), shown at the bottom of this page, where the last equality is obtained from (2).

To get a steady state equation for the slot occupancy probability u, the approximation that packet arrivals at the two input links at every node are independent events will be introduced. This is a reasonable assumption in homogeneous traffic. The average number of newly transmitted packets per node is obtained as the probability of having a new packet times the probability that at least one of the two inputs is free

$$\frac{\lambda}{N} = g[1 - u^2(1-a)^2]. \tag{4}$$

Equations (2)–(4) yield

$$u = \frac{\sqrt{a^2 + g^2(1-a)^2} - a}{g(1-a)^2}. \tag{5}$$

Note that even for $g = 1$ the value of u is less than 1. The reason is that two packets per slot can be received, but only one new packet can be inserted.

The expected number of hops D noticeably depends on the routing algorithm. For store-and-forward with infinite buffers D is a minimum, since packets always take the shortest path to destination, and is independent of the link load u. Therefore by (2) the throughput is a maximum for a given u. However the queueing delay D_q can diverge to infinity when the network approaches saturation, that is, when g tends to 1. For deflection routing the queueing delay remains bounded, but packets may be deflected to nonoptimal paths and thus D becomes an increasing function of u. The throughput is thus lower than with store-and-forward.

A. Markov Chain Analysis

The objective now is to find the expected number of hops D and the throughput λ as a function of the generation probability g only.

To this aim, the trajectory of a test packet destined to a reference node and generated uniformly at random among all other nodes in the network will be followed [8]. By the assumed regularity of the network the choice of the reference node is arbitrary and node zero will be chosen. Because of the homogeneity of the load, the independence approximation and the fact that the routing is memoryless, the random walk of the test packet towards node zero can be modeled as a homogeneous absorbing Markov chain $n(k)$, representing the node visited by the test packet at the end of its kth hop.

For example, the state transition diagram of this chain for MS16 is drawn in Fig. 4. The labels on the branches represent the transition probabilities. All nodes whose output links are both on a shortest path to node zero are *don't care* for the test packet. The transition probabilities for the test packet at a don't care node are both 1/2. In fact, in the assumption of uniform traffic, a care packet entering a node together with the test packet will prefer either output with probability 1/2, while randomization is applied when this packet is don't care. *Care* nodes are those nodes at which one output provides a shorter path to destination than the other. They are marked by bold circles in the figure. At a care node, p is defined as the probability that the test packet is deflected, so that the transition probability on the preferred branch is $(1-p)$. Since only nonpriority deflection routing in uniform traffic is treated, this quantity p is equal at all care nodes. Note that the destination node zero is the absorbing state of the chain.

Let Π be the $N \times N$ transition matrix whose elements π_{ij} represent the probability that the test packet will move to node i at its $(k+1)$th hop, being at node j at its kth hop. Fig. 4 shows also the matrix Π relative to the transition diagram. Let $\boldsymbol{p}(k)$ be the state vector at time k, whose elements $p_i(k)$ represent the probability that the test packet will arrive at node i at its kth hop. Given the distribution $\boldsymbol{p}(k)$ at time k, the state at time $k+1$ is

$$\boldsymbol{p}(k+1) = \Pi(p)\boldsymbol{p}(k) \tag{6}$$

where the notation stresses the dependence of the matrix Π on p. The state $[1\ 0\ \cdots\ 0]^{\mathrm{T}}$ is the solution to which the chain converges as $k \to \infty$, and in fact it is the eigenvector associated with the eigenvalue $\mu = 1$ of the Markov matrix Π.

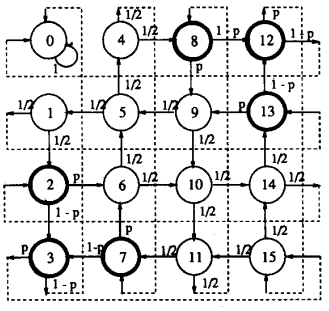
To interpret the information given by the time evolution of the state vector, define

$$I_i(k) \triangleq \begin{cases} 1, & \text{if test packet is at node } i \text{ at time } k \\ 0, & \text{else.} \end{cases}$$

Thus $\{I_i(k);\ k = 0, 1, 2, \cdots\}$ is a stochastic process representing the passage of the test packet through node i. The mean of this process is

$$\mathrm{E}I_i(k) = p_i(k), \qquad k = 0, 1, 2, \cdots.$$

$$a \triangleq \frac{\text{average number of absorbed packets per input link per slot}}{\text{average number of packets per input link per slot}} = \frac{\lambda/2N}{u} = \frac{1}{D} \tag{3}$$



(a)

$$\Pi(p) = \begin{bmatrix}
1 & 0 & 0 & 1-p & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1-p & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 \\
0 & 0 & 1-p & 0 & 0 & 0 & 0 & 1-p & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & p & 0 & 0 & 0 & 0 & p & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p & 0 & 0 & 0 & 0 & p & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 1/2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1-p & 0 & 0 & 0 & 0 & 1-p & 0 & 0 \\
0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 1/2 \\
0 & 0 & 0 & p & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p & 0 & 0 & 0
\end{bmatrix}$$

(b)

Fig. 4. (a) State transition diagram. (b) Transition matrix for MS16.

Now define the random variable V_i as the number of times the test packet visits node i in its travel towards node zero

$$V_i \triangleq \sum_{k=0}^{\infty} I_i(k), \qquad i = 1, 2, \cdots, N-1.$$

Note that when the packet arrives at zero, it remains there forever, so that $I_0(k)$ is a step function jumping from zero to one at the random arrival hop d of the packet. Therefore

$$\sum_{k=1}^{\infty} k[I_0(k) - I_0(k-1)] = \sum_{k=1}^{\infty} k\delta(k-d) = d$$

where $\delta(k)$ is unity at $k = 0$ and zero otherwise. The random variable d represents the total number of hops taken by the test packet in its travel. The expected values of these random variables have interesting interpretations

$$\begin{aligned} \mathrm{E}V_i &= \sum_{k=0}^{\infty} p_i(k) \\ &= \text{average number of times test packet visits node } i, \\ &\qquad i = 1, 2, \cdots, N-1 \end{aligned} \tag{7}$$

$$\begin{aligned} \mathrm{E}d &= \sum_{k=1}^{\infty} k[p_0(k) - p_0(k-1)] \\ &= \text{average number of hops of test packet} \triangleq D. \end{aligned} \tag{8}$$

From this, $p_0(k)$ is seen to be the cumulative distribution function of the number of hops taken by the test packet.

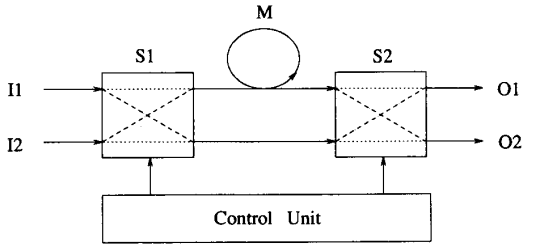


Fig. 5. Optical implementation of the routing block with memory. M is the fiber delay line memory and S1, S2 are exchange-bypass switches.

M	I1	I2	S1	S2	NM
E/DC	E/DC	E/DC	R	R	E/DC
		O1	–	×	E/DC
		O2	–	–	E/DC
	O1	E/DC	×	×	E/DC
		O1	R	×	O1
		O2	R	–/×	O1/O2
	O2	E/DC	×	–	E/DC
		O1	R	×/–	O2/O1
		O2	R	–	O2
O1	E/DC	E/DC	R	–	E/DC
		O1	×	–	O1
		O2	–	–	E/DC
	O1	E/DC	–	–	O1
		O1	R	–	O1
		O2	–	–	O1
	O2	E/DC	×	–	E/DC
		O1	×	–	O1
		O2	R	–	O2
O2	E/DC	E/DC	R	×	E/DC
		O1	–	×	E/DC
		O2	×	×	O2
	O1	E/DC	×	×	E/DC
		O1	R	×	O1
		O2	×	×	O2
	O2	E/DC	–	×	O2
		O1	–	×	O2
		O2	R	×	O2

M State of the memory before switching
I1 State of packet at input I1
I2 State of packet at input I2

S1 State of switch S1
S2 State of switch S2
NM State of the memory after switching

– Bar state
× Cross state
R Random choice between bar and cross states
–/× State of S2 set equal to state of S1
×/– State of S2 set opposite to state of S1

Fig. 6. Truth table of the control unit.

One more observation. Indicating by $\mathcal{DC}$ the set of don't care nodes, the random variable

$$V_{\mathrm{dc}} \triangleq \sum_{i \in \mathcal{DC}} V_i$$

is the total number of times the test packet visits a don't care node in the experiment. Now, d is also the number of times that nodes not coinciding with the destination node are visited in the experiment

$$d = \sum_{i=1}^{N-1} V_i. \tag{9}$$

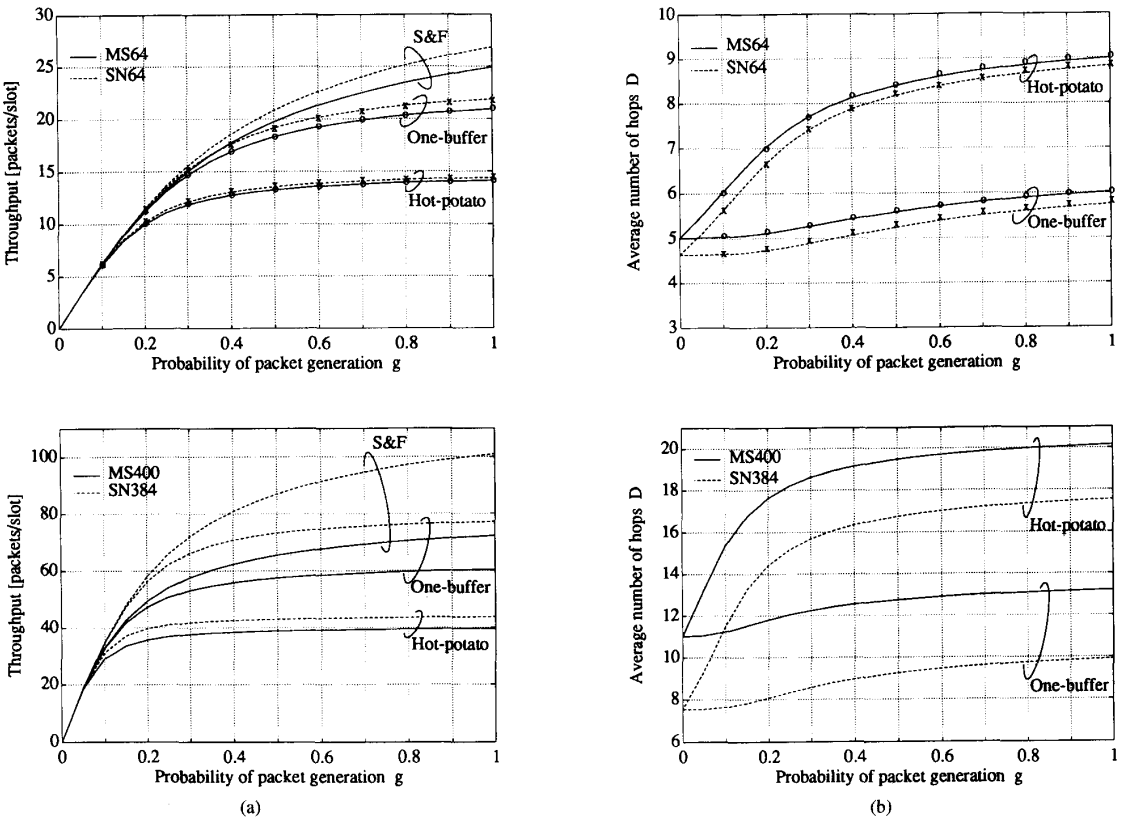


Fig. 7. (a) Aggregate network throughput in MS and SN. Curves for store-and-forward with infinite buffers are provided as a reference. (b) Average number of hops.

Hence the long-run fraction of time the test packet is at a don't care node, referred to as the don't care probability $P_{\rm dc}$, is

$$P_{\rm dc} = \frac{\sum_{i\in\mathcal{DC}} \mathrm{E}V_i}{\sum_{i=1}^{N-1} \mathrm{E}V_i} = \frac{\mathrm{E}V_{\rm dc}}{\mathrm{E}d} = \frac{\sum_{k=0}^{\infty}\sum_{i\in\mathcal{DC}} p_i(k)}{D} \qquad (10)$$

where the last expression on the right-hand side is the operative formula. Since in a homogeneously loaded network the test packet is a typical packet, $P_{\rm dc}$ represents also the probability that a packet entering a node together with the test packet is in a don't care state.

Now, it is possible to express the deflection probability p as a function of the quantities $P_{\rm dc}, a$, and u which all depend on p itself through the transition matrix $\Pi(p)$. For any value of the free parameter g, and initial state vector $\boldsymbol{p}(0) = [0, 1/(N-1), \cdots, 1/(N-1)]^{\rm T}$ to preserve load balance, it is then possible to solve an implicit equation in p, and thus get a curve $p(g)$. From this, the desired curves $D(g)$ and $\lambda(g)$ can be obtained using (8),(5), and (2).

The deflection probability p depends on the routing technique. Refer to Fig. 3. Define P_c as the probability of having a care packet at the input of the routing block together with the test packet at an intermediate care node. It is

$$P_c = [u(1-a) + uag + (1-u)g](1 - P_{\rm dc}) \qquad (11)$$

since the term in square brackets represents the probability of having a packet at the input of the routing block on the link left free by the test packet. In fact this event occurs if a packet is present at the input link and not absorbed, or is present, absorbed, and a new packet is generated, or the input link is empty but a new packet is generated.

In the case of no buffers (hot-potato routing), the routing block is a simple cross/bar switch. $P_c/2$ is the probability of an output conflict, since in homogeneous traffic the competing packet will wish the same output as the test packet with probability 1/2. Thus the deflection probability is

$$p = \frac{1}{2}\frac{P_{\rm c}(p,\, g)}{2} \qquad (12)$$

where the dependence of P_c on p comes through the matrix $\Pi(p)$. This is the desired implicit equation in p. A different approach must be taken to handle the computation of the deflection probability at the first step of the chain, where the test packet is at its generation node and is trying to access the network. This case is treated in detail in the Appendix.

B. An Optical Single-Buffer Shared Output Memory

This section will derive the implicit equation for p when use is made of the single-buffer optical memory proposed in [6], which lends itself to a simple optical implementation.

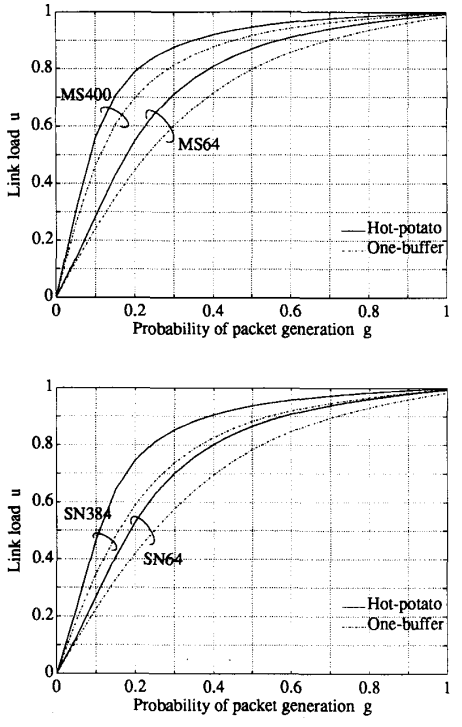


Fig. 8. Link load u versus packet generation probability g.

Description: The scheme of the routing block in this case is shown in Fig. 5. S1 and S2 are two cross/bar switches whose state is controlled by a control unit and the memory element M is a one-packet fiber delay line. This implements a shared output memory, since both inputs can access the buffer M through switch S1, and the buffer can access either output through switch S2. The control unit must know the state of both inputs and of the memory, that is, whether they contain an empty packet (E), or a don't care packet (DC), or a packet wishing to exit on output 1 (O1), or on output 2 (O2). A truth table description of the control unit is given in Fig. 6. It is seen that the empty (E) and don't care (DC) states are collapsed into a single state. This is a simplification with respect to the scheme in [6] that drastically reduces the number of entries in the table and is justified in very high bit rate applications. In fact, it may happen that a don't care packet is stored, while an empty packet is let through, but this extra one-slot delay on don't care packets is negligible compared to the one-hop propagation delay. The simplification has no effect on throughput.

Deflections occur only when a care packet is stored and two packets wishing the same output as the stored packet are present at the inputs. Therefore, if there is no possible conflict between memory and inputs, switch S1 will be set, if possible, to store an empty/don't care packet to avoid a deflection at the next time slot. When a deflection pattern occurs, since statistically it does not make any difference which of the three packets will be deflected, the packet in memory will be given priority over the two inputs and will set the output switch S2, so that one of the two input packets, chosen at random, will be deflected. The only inefficiency of this scheme occurs when two nonconflicting care packets are present at the inputs and the memory is E/DC, in which case one of the two care packets must necessarily be stored, thus increasing the probability of a deflection at the next slot. The addition of an extra switch to access the fiber loop would solve the problem. However, the cost of the extra switch, the increased control complexity, and the fact that in this case the power losses experienced by buffered and unbuffered packets would differ by several dB's make this more complete scheme less attractive for an optical implementation. Moreover, the gain in throughput with respect to the two-switch configuration can be shown to be less than 5% at full load.

Analysis: To find an expression for the deflection probability p, we begin by noting that even in this case $P_c/2$ is the probability of a conflict between the test packet and a competing packet at the other input of the routing block. Now define P_{cm} as the probability of having a care packet in memory at steady state. $P_{cm}/2$ is the probability that this packet collides with the test packet. If the state of the memory is assumed independent of the state of the packets at the input, the deflection probability for the test packet can be evaluated as

$$p = \frac{1}{2}\left[\frac{P_c(p, g)}{2}\frac{P_{cm}(p, g)}{2}\right] \tag{13}$$

since the square bracket indicates the probability of a deflection pattern, and a fair coin is tossed to decide which packet will be deflected between the two input packets. This is the implicit equation in p in this case. Comparing (13) and (12), the reduction in deflection probability with respect to the case of no buffers is seen to be due to the factor $P_{cm}/2$. Keeping low the probability of having care packets in memory will decrease deflections and increase the throughput.

An explicit expression for P_{cm} can be easily obtained by reading off the memory transitions from the truth table in Fig. 6. The probabilities of having a 01 packet and 02 packet in memory are assumed equal, and their value is $P_{cm}/2$. Let $P_u = u(1 - P_{dc})$ be the probability of having a care packet on each input of the routing block. A balance equation on the memory occupancy obtained from Fig. 6 gives after some algebraic manipulation

$$P_{cm} = \frac{P_u^2}{1 - P_u + P_u^2}. \tag{14}$$

This is seen to be a function of only p and g through u and P_{dc}.

IV. Results

In this section, results obtained with the analytical method described in the previous section will be presented for MS and SN. These topologies are compared for 64 nodes (MS64 versus SN64) and for a larger size of about 400 nodes (MS400 versus SN384) where the percentage size difference is less than 5%. All curves will compare hot-potato routing to the single-buffer deflection routing technique analyzed in Section III-B. Simulations for the uniform traffic case have been run for MS64 and SN64 according to the method described in [8] to support the analytical results. Simulation results are shown with circles for MS and crosses for SN in the figures.

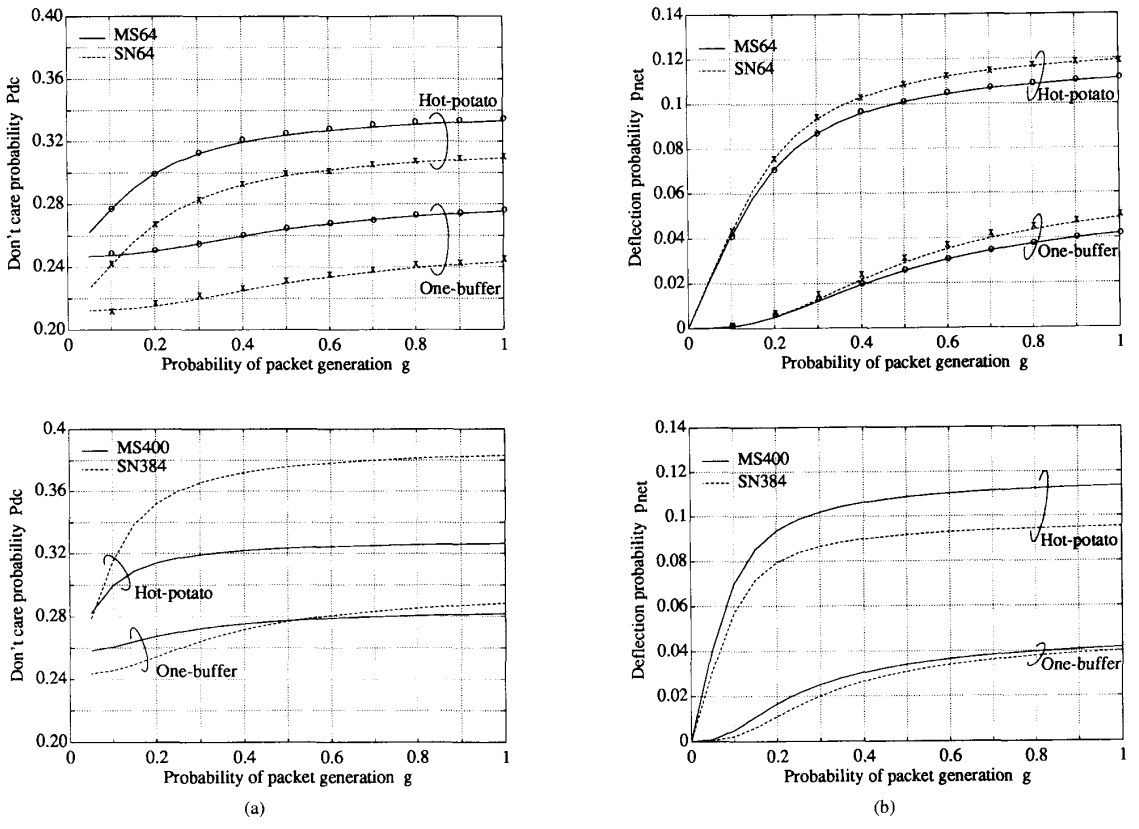


Fig. 9. (a) Don't care probability and (b) network deflection probability in hot-potato and single-buffer deflection routing.

Fig. 7 shows throughput λ and expected number of hops D versus packet generation probability g. Throughput curves for store-and-forward with infinite buffers (shortest path routing) are also provided as a reference. These are readily obtained from Little's formula (2) in which the link load u is evaluated by (5) and the zero-load delay is used. The throughput is higher for SN for all values of g and its gain over MS increases for larger networks. For instance SN384, although smaller than MS400, has much higher aggregate throughput. Note also that one buffer is enough to fill a substantial portion of the throughput gap between store-and-forward and hot-potato, as already shown by simulation in [10] for MS. The portion of the gap at full load recovered by the use of the analyzed single-buffer memory is around 60% in all four networks presented. Curves in Fig. 7 can be read also versus link load u by using relation (5) between g and u plotted in Fig. 8.

To give a better insight of network behavior, Fig. 9 presents curves of don't care probability $P_{\rm dc}$ and network deflection probability $p_{\rm net}$. The analytical curve $P_{\rm dc}$ is obtained from (10), while the curve for $p_{\rm net}$ is obtained as $p_{\rm net} = p(1-P_{\rm dc})$, since p represents the deflection probability at a care node. In the simulations, $P_{\rm dc}$ and $p_{\rm net}$ are obtained as long-run time averages. At every clock cycle t, the number of packets after absorption and generation at all nodes is denoted as $n(t)$. Let $n_{\rm dc}(t)$ and $n_{\rm def}(t)$ be the portion of these packets which are, respectively, don't care and deflected at time t. Then

$$P_{\rm dc} = \frac{\sum_{t=1}^{K} n_{\rm dc}(t)}{\sum_{t=1}^{K} n(t)} \qquad p_{\rm net} = \frac{\sum_{t=1}^{K} n_{\rm def}(t)}{\sum_{t=1}^{K} n(t)} \tag{15}$$

where K is the total number of clock cycles in the simulation. The summations were started after the network had reached steady state and the number K of clock cycles after the transient was chosen to be 10 000. These simulated quantities match very well with the respective analytical values for the test packet, which confirms that the test packet is actually a "typical" packet, and that the network traffic is really homogeneous, so that the independence approximation at the node is accurate. From Fig. 9 one can note the increase in don't care probability in SN with increasing network size due to the increased don't care percentage. The values for MS remain instead almost unchanged since in MS the don't care percentage is approximately 50% for both sizes, as shown in Fig. 1. For both MS and SN the effect of adding one buffer is to route packets more directly toward their destination, thereby reducing the number of don't care visits. Fig. 9 also shows that, for 64 nodes, $p_{\rm net}$ in SN is slightly higher than in MS, while for 400 nodes $p_{\rm net}$ is much lower in SN, due to the increase in $P_{\rm dc}$ observed in the figure. The reduction in $p_{\rm net}$ in both MS and SN using the single-buffer memory is remarkable. The reduction is never less than 60% in both MS

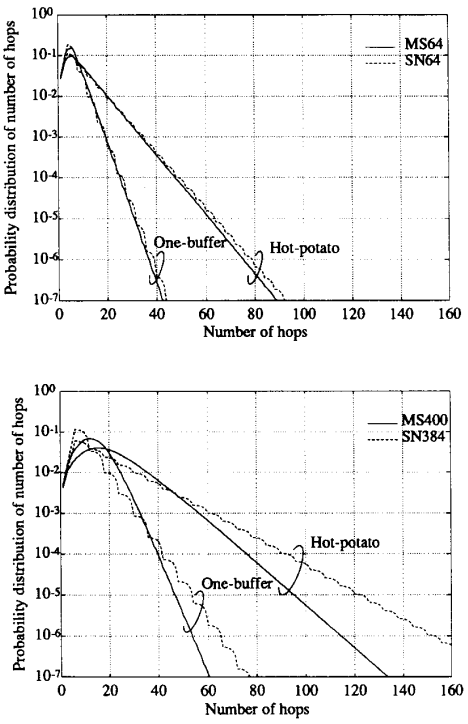


Fig. 10. Hop probability distribution at full load in MS64 and SN64 (top) and MS400 and SN384 (bottom).

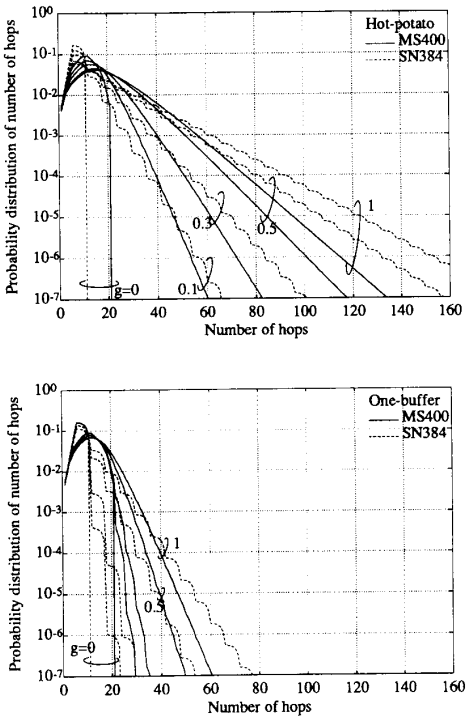


Fig. 11. Hop distribution in MS400 and SN384 for generation probability $g = 0.05, 0.1, 0.3, 1$. Curves at zero load are given as a reference.

and SN, both for 64 and 400 nodes, and is slightly higher for MS when the load approaches one.

The classical average delay and throughput analysis so far carried out would indicate a superiority of SN over MS, especially for large network sizes, and when buffers are used. The probability distribution of the number of hops, however, provides more information than its average value does, giving more insight of network behavior. This is important when hop distribution tails must be taken into account for optimizing network performance, such as probability of error, as suggested by (1). Based on (1), a comparison between MS and SN in an ultrafast all-optical network has been presented in [4].

Fig. 10 gives the hop distribution curves obtained analytically from (6) at full load. The key observation is that SN has lower mean, but the tails decay more slowly than in MS, both with and without memory. The difference is amplified as the network size increases, as seen in the 400 nodes networks. This behavior reflects the fact that SN is more compact, hence has lower mean, but the cost of each deflection grows higher as network size increases, thereby increasing the probability in the tails. Fig. 11 shows the effect of increasing load on hop distributions. The different behavior of the tails in MS and SN as the load increases to 1 can be well appreciated. For fixed g the tails always cross, the tails in SN eventually exceeding those in MS. At very low loads the cross-point is at very low probability and SN behaves better than MS. However, at high loads the cross-point is up to high probability and SN has much heavier tails.

As a final result, simulations for MS64 and SN64 are presented to check the effect of correlation between destinations of packets generated at the same node, as is the case when a message of several packets has to be transmitted to the same recipient. In the simulations, the message length M_l was chosen as $M_l = X + 1$, where X is a Poisson random variable. Fig. 12 shows throughput and delay curves for SN and MS for an average value $E(M_l) = 1, 5, 20$. The $E(M_l) = 1$ curves are those already given where no correlation between successive packets exists and match the one-packet model curves. At correlation values of 5 and 20 the curves relative to hot-potato show little degradation with respect to the case of no correlation. However, much greater throughput degradation is observed for the single-buffer deflection routing curves. The important result that a single buffer is enough to get a substantial throughput gain over hot-potato was obtained in the assumption of uncorrelation among packet destinations. The potential 60% gain on throughput gap predicted in the absence of correlation can actually decrease to less than 40% in the presence of long messages because one buffer only cannot efficiently handle successive conflicts arising from streams of consecutive packets with the same destination colliding at the node. More buffers are required in this case to substantially improve network performance.

V. Conclusions

A detailed review has been given of the one-packet model used to analyze the steady state behavior of regular multihop

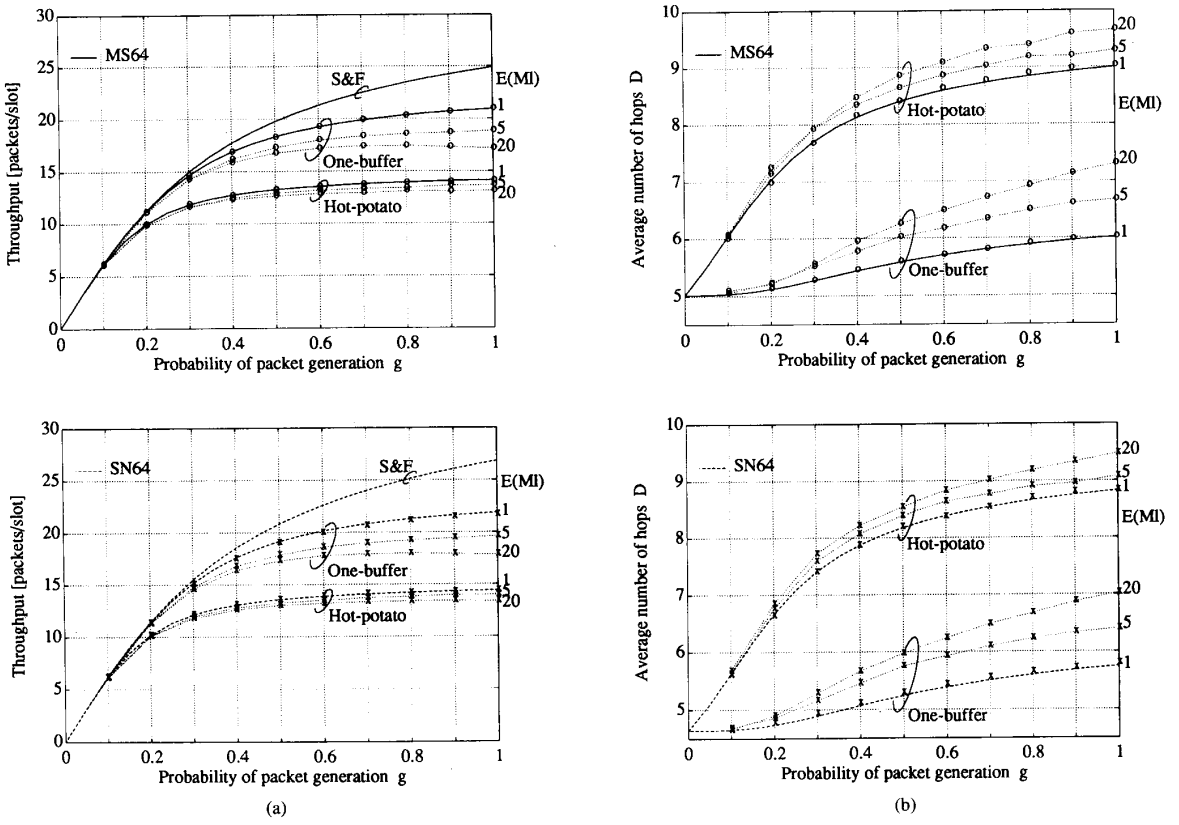


Fig. 12. Simulation results of aggregate network throughput and average number of hops for MS64 and SN64 for average message length $EM_l = 5, 20$ packets. Curves for $EM_l = 1$ (uncorrelated case) are given as a reference.

networks in uniform traffic under hot-potato routing and the method has been extended to include the analytical treatment of a single-buffer deflection routing technique which is particularly attractive, due to its simplicity, for optical implementation in very high bit rate optical networks. The analytical model is applied to MS and SN, which are compared in terms of throughput, delay, don't care probability, deflection probability, and hop distribution, and all results are supported by simulations. The average analysis shows that SN has higher throughput than MS at all loads, and the difference increases with network size. However, the tails of the hop distribution in MS decay much more rapidly than in SN. Therefore, the choice of which regular mesh topology is more appropriate depends on the network parameters to be optimized. The effectiveness of the single buffer is analytically quantified. It is verified that under the assumption of independence of packet destinations, the single-buffer deflection routing recovers more than 60% of the throughput loss of hot-potato with respect to store-and-forward. However, when messages of average length as high as 20 packets are transmitted to the same recipient, consecutive collisions arise and a single buffer cannot efficiently handle them anymore. The achievable gain in this case is reduced to below 40%.

VI. Appendix

The initial probability p_0 that the test packet be deflected at the injection node will now be found. Recall Fig. 3. The test packet in this case is waiting to access the network at the generation block. At equilibrium, define $\mathcal{A}_0$, $\mathcal{A}_1$, and $\mathcal{A}_2$ as the event of having respectively 0, 1, or 2 packets on the node links after the absorption block, whose probabilities are

$$P(\mathcal{A}_0) = (1-u)^2 + 2u(1-u)a + u^2a^2$$
$$P(\mathcal{A}_1) = 2u(1-u)(1-a) + 2u^2a(1-a)$$
$$P(\mathcal{A}_2) = u^2(1-a)^2$$

and are obtained reasoning as in (11).

Since it is assumed that the test packet *is* injected in the network, and this is possible only if at least one link is free, its deflection probability is actually conditioned on the event $\mathcal{A}_0 \cup \mathcal{A}_1$. Its deflection is possible only if event $\mathcal{A}_1$ occurs, i.e., a transiting packet is present. Thus the probability of having a care packet together with the test packet at the injection node is

$$P_{c_0} = \frac{P(\mathcal{A}_1)}{P(\mathcal{A}_1) + P(\mathcal{A}_0)}(1 - P_{\mathrm{dc}})$$

and from (12) and (13) the initial deflection probability is

$$\begin{cases} p_0 = \frac{1}{4}P_{c_0} & \text{without buffers} \\ p_0 = \frac{1}{8}P_{c_0}P_{cm} & \text{with single buffer.} \end{cases}$$

Therefore (6) describing the state vector at time $k+1$ becomes

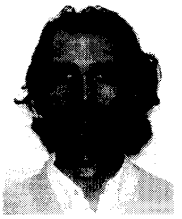
$$\begin{cases} \boldsymbol{p}(k+1) & = \Pi(p)\boldsymbol{p}(k), \quad k \geq 1 \\ \boldsymbol{p}(1) & = \Pi(p_0)\boldsymbol{p}(0). \end{cases}$$

ACKNOWLEDGMENT

The authors are grateful to A. Fumagalli and A. Greenberg for providing preprints of their papers [6] and [8].

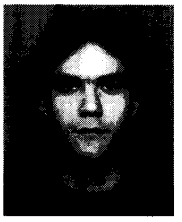
REFERENCES

[1] N. F. Maxemchuk, "The Manhattan Street Network," in *Proc. GLOBECOM '85*, pp. 255–261.
[2] A. S. Acampora, M. J. Karol, and M. G. Hluchyj, "Terabit lightwave networks: The multihop approach," *AT&T Tech. J.*, vol. 66, pp. 21–34, Nov./Dec. 1987.
[3] J. R. Sauer, "An opto-electronic multi-Gb/s packet switching network," Optoelectronic System Center, University of Colorado, Feb. 1989.
[4] A. Bononi, F. Forghieri, and P. R. Prucnal, "Design and channel constraint analysis of ultra-fast multihop all-optical packet switching networks with deflection routing employing solitons," *IEEE J. Lightwave Technol.*, vol. 11, no. 12, pp. 2166–2176, Dec. 1993.
[5] N. F. Maxemchuk, "Regular mesh topologies in local and metropolitan area networks," *AT&T Tech. J.*, vol. 64, no. 7, pp. 1659–1685, Sept. 1985.
[6] I. Chlamtac and A. Fumagalli, "An all-optical switch architecture for Manhattan networks," *IEEE J. Select. Areas Commun.*, vol. 11, pp. 550–559, May 1993.
[7] P. Baran, "On distributed communications networks," *IEEE Trans. Commun. Syst.*, vol. 12, pp. 1–9, Mar. 1964.
[8] A. G. Greenberg and J. B. Goodman, "Sharp approximate models of deflection routing in mesh networks," *IEEE Trans. Commun.*, vol. 41, pp. 210–223, Jan. 1993.
[9] A. S. Acampora and A. Shah, "Multihop lightwave networks: A comparison of store-and-forward and hot-potato routing," *IEEE Trans. Commun.*, vol. 40, pp. 1082–1090, June 1992.
[10] N. F. Maxemchuk, "Comparison of deflection and store-and-forward techniques in the Manhattan Street and Shuffle-Exchange networks," in *Proc. IEEE INFOCOM '89*, Apr. 1989, pp. 800–809.
[11] T. Y. Chung and D. P. Agrawal, "On network characterization of and optimal broadcasting in the Manhattan Street Network," in *Proc. IEEE INFOCOM '90*, 1990, pp. 465–472.
[12] E. Ayanoglu, "Signal flow graph for path enumeration and deflection routing analysis in multihop networks," in *Proc. IEEE GLOBECOM '89*, 1989, pp. 1022–1029.
[13] D. Bertsekas and R. Gallager, *Data Networks*. Englewood Cliffs, NJ: Prentice-Hall, 1987.



Fabrizio Forghieri was born in Modena, Italy, in 1962. He received the Doctor Engineer degree in electrical engineering, summa cum laude, from the University of Pisa, Italy, in 1988, and the M.A. degree in electrical engineering from Princeton University in 1991.

In 1989 he was with the Istituto di Scienze dell'Ingegneria, University of Parma, Italy, working on optical communication systems. From 1990 to 1992 he was with Princeton University, where he worked on ultrafast transparent optical networks. From 1992 to 1993 he was with the Dipartimento di Ingegneria dell'Informazione, University of Parma, Italy, where he worked on soliton networks and taught a course on optical fiber nonlinearities. He is currently with AT&T Bell Laboratories, Crawford Hill Laboratory, working on nonlinear optical effects in fibers and WDM systems.



Alberto Bononi received the Laurea in Ingegneria Elettronica degree from the University of Pisa, Pisa, Italy, in 1988 and the M.A. degree in electrical engineering from Princeton University, NJ, in 1992.

In 1989 he was a Visiting Researcher at the University of Parma, Parma, Italy, working on coherent optical communications. In 1990 he worked at GEC-Marconi Hirst Research Centre in Wembley, UK, on a Marconi S.p.A. Project on coherent FSK systems. He is working toward the Ph.D. degree at Princeton University, where his research interests include system design and performance issues in fast packet switching and high-speed all-optical networks.

Paul R. Prucnal (S'75–M'78–SM'90–F'92) received the A.B. degree, summa cum laude, from Bowdoin College, Brunswick, Me, in 1974, and the M.S., M.Phil., and Ph.D. degrees from Columbia University in 1976, 1978, and 1979, respectively.

He was an Assistant Professor of Electrical Engineering at Columbia University from 1979 to 1984 and an Associate Professor from 1984 to 1987. He was a Member of the Executive Board of Columbia's NSF Engineering Research Center in Telecommunications Research from 1985 to 1987. He is now a full Professor at Princeton University, where he is Director of the Lightwave Communications Research Laboratory, and recently served as Acting Director of the newly established New Jersey Advanced Technology Center in Photonics and Optoelectronic Materials. He has taught courses in the areas of fiber-optic communications systems, quantum electronics, and digital signal processing. He has been a Technical Consultant for Philips Labs, Optical Information Systems Inc., GTE Labs IBM, Dove Electronics, and AT&T Bell Labs. He has published more than 50 journal papers in the areas of optical networks, photonic switching, optical techniques for advanced VLSI/VHSIC interconnections and optical signal processing, and holds three patents.

Dr. Prucnal is a member of OSA and SPIE, and is an Associate Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE CIRCUITS AND DEVICES MAGAZINE, and the IEEE LIGHTWAVE COMMUNICATIONS MAGAZINE.