STATE UNIVERSITY OF NEW YORK AT STONY BROOK

CEAS TECHNICAL REPORT 603

Information Based Deflection Strategies for the Manhattan Street Network

A.A. Lazar and T.G. Robertazzi

April 24, 1991

Information Based Deflection Strategies for the Manhattan Street Network

Thomas Robertazzi

Dept. of Electrical Engineering,

SUNY at Stony Brook,

Stony Brook NY 11794

Aurel A. Lazar

Dept. of Electrical Engineering &

Center for Telecommunications Research,

Columbia University,

New York NY 10027

Abstract

Alternate deflection routing strategies for a 64 node Manhattan Street network suitable as either a metropolitan area network or as a switch fabric are presented. Results are presented for deflection routing strategies based on using a hop counter, a deflection counter or distance information. It is shown that the use of age or deflection counter based strategies improves delay performance. These results also point to the need for input buffers in a working Manhattan Street network implementation.

1 Introduction

Since the introduction of fiber optics, the way in which the architecture of networks is viewed has changed. It has become apparent that, for the moment, fiber links can operate at throughputs higher than that of nodal electronics. Thus interest has been generated in parallel switching fabrics where simple decision rules implemented locally at the nodes and the use of parallel paths makes the best of the electronics bottleneck.

In 1985 N.F. Maxemchuk of AT&T Bell Laboratories proposed a metropolitan area network (MAN) architecture and routing protocol called the Manhattan Street Network (MSN) [9]. It consists of NxN nodes connected in a toroidal topology [12] using a pattern of uni-directional fiber links, as in Figure 1. The MSN is a synchronous system in which time is slotted. Each node in the MSN has two input links and two output links. It can also hold, at most, two packets in its two packet buffers. During a slot, at most two packets may arrive at a node - one on each input link - and be placed in the node's two packet buffers. Assuming that neither packet is destined for that node, each is assigned (routed) to a separate output link and transmitted during the next slot. The essential routing problem in the MSN is how to assign the packets to output links.

Deflection routing [5,6] can be used to make the assignment of packets to output links. A single packet in a node may either have one preferred (shortest path) output link or both may be equally good. The routing event in which deflection routing comes into play is when there are two packets in a node and both have the same, single preferred output route. Maxemchuk proposed randomly assigning one packet to the preferred output link and "deflecting" the other packet onto the other, non-preferred, output link. We will refer to this

strategy as standard deflection. Unlike in conventional store and forward packet switching, packets do not queue at nodes.

Because of the toroidal topology and relatively simple, light weight, routing protocol, the routing decisions in the MSN can be implemented in hardware to better match the speed of electronics to that of fibers [1]. We note that deflection routing can be seen to be a variation of the older "hot potato routing" [4].

In this paper a number of alternate deflection routing strategies for making the decision as to which packet to deflect are examined. A deflection decision can either be made randomly (standard deflection) or using information concerning a packet's past history, an expectation of it's future transit through the network or a combination of both.

The paper begins in section 2 with a look at standard deflection. Only limited performance data for the MSN has been published previously [5,11]. In section 3.1 the use of the age of a packet and the use of the number of deflections a packet has undergone are examined in the context of making deflection decisions. Generally, deflection strategies using this information achieve a lower maximal delay in transiting the network. In section 3.2 strategies that use distance measures in making deflection decisions are discussed. One set uses the source to destination shortest path distance (static distance strategies) and the other set uses the shortest path distance from the current node to the destination (dynamic distance strategies). A hybrid, predictive path length strategy, appears in section 3.3. A taxonomy of these strategies appears in Fig. 2. A simple relation between mean throughput, mean delay and link utilization is presented in section 4.

Before proceeding, it is well to point out that applications for the MSN include both as MANs and as switch fabrics.

2 Routing without Information: Standard Deflection

2.1 Assumptions

Unless otherwise noted, the data in this paper is for an 8x8 MSN simulated for 20,000 slots with the first 200 slots of data discarded. The data is generally significant to two decimal places. The simulator program, written in C, used table look-up to determine the preferred output link. It was found that the simulator could route approximately 1600 packets/sec on a SPARC 1+ station.

In this paper we assume [2,3,5] that during a slot a single packet may arrive at a node with probability p. This assumption allows an easy parametrization of the data and does not seem overly restrictive. While one could allow two new packets into a node's two packet buffers if they were both empty, this would only make a significant difference at light loads (if p=.5, the probability that a node's two packet buffers are both empty under standard deflection is less than 3%, see Table 3).

Another feature of the network modeled in this paper is that there is no input buffer at a node. That is, an arriving packet is cleared from the system if it can not immediately be accepted into one of the packet buffers of the node at which it arrives. We also assume that for each packet any node is equally likely as a destination. Successive destinations for arriving packets at a node are independent of each other.

The data and programs that this paper are based on are available in [13-15].

2.2 Performance

Fig. 3 is a plot of the average number of hops a packet undergoes (average delay) in transiting the MSN as a function of packet arrival probability p. Our simulation confirms the simulation curve of Greenberg and Goodman presented in [2,3,5]. Generally, as the offered load increases, deflections become more likely, causing an increase in the average delay in transiting the network.

Fig. 4 is a plot of link utilization (the probability that a link carries a packet during a slot) versus p. Because a single packet may transit several links before reaching its destination, utilization increases steeply at light loads (it is .28 at p=.1).

Fig. 5 shows the average throughput per node versus p. Since at most a single new packet may arrive at a node during a slot, average throughput assumes a value between 0 and 1. Average throughput increases monotonically with increasing p and saturates at about .22. We note that at saturated loads a packet has an average path length (delay) of about 9 (see Fig. 3) and thus may be viewed as having a 1/9 probability of being at its destination each time it reaches a node [2]. The maximal throughput is equal to the probability that a node has one or two empty buffers to accept a new packet and is thus:

 $Thruput = P[buffer\ A\ empty] + P[buffer\ B\ empty] + P[both\ buffers\ empty]\ (2.1)$

$$Thruput = \frac{1}{9} + \frac{1}{9} + \frac{1}{81} = .234$$
 (2.2)

Therefore the simulation data for average delay and average throughput is thus consistent.

While it would seem desirable to operate the network at the smallest value of p that largely achieves saturation, this is not practical without input buffers because of the degradation in the blocking probability presented to new packets. Fig. 6 is a curve of blocking probability vs. p. As arrivals are random, blocking probability is simply the probability that a node has two transiting packets occupying its two buffers. Even at p=.1 it is equal to .09 and rises to totally unacceptable levels under throughput saturation.

For a working, highly utilzed system, arriving packets must be buffered at the input so that the probability of there being a packet at the head of the line of the input buffer is at a value consistent with throughput saturation. Moreover the buffer must be sized so that it presents an acceptable blocking probability to arriving packets. Naturally, the packet arrival probability for an 8x8 network with input buffers can not exceed the saturation throughput level of .22. Thus, if the 8x8 MSN links operate at 1Gbs, the maximum average arrival rate to a node is 220Mbps.

In spite of this, we will continue in the rest of this paper to examine a system without input buffers. One reason is that valuable insights can be gained from this scenario that, we believe, are applicable to systems with input buffers. The network can still be exocised under a variety of loads, though quality of service [17] for individual packets may be poor. Another reason is the scarcity of published performance data concerning the MSN.

2.3 Routing Events

An interesting facet of the MSN's operation is that during a slot there are seven possible routing "events". These are listed in Table 1. For instance, for Event 5, one packet has a single preferred output link and for the other packet both output links are equally good. In

fact, only during Event 3 does deflection occur. We note that the multiplicity of routing events complicates writing routing software or designing routing hardware.

In Table 2 the probabilities of the various events are tabulated vs. p. In Table 3 the probabilities that a node has 0,1 or 2 packets in its buffers are tabulated vs. p. In [9] it is pointed out that if a packet's source/destination pair is equally likely, the probability that a packet will prefer one output link is about .25, the probability it will prefer the other is about .25 and the probability that both output links seem equally good is about .5. At first glance this would seem to imply that the following relationships should hold for the data of Table 2 and Table 3:

$$p(E1) = p(E2) \tag{2.3}$$

$$p(E3) = \frac{1}{8} P(2 \ packets) \tag{2.4}$$

$$p(E4) = \frac{1}{8} P(2 \ packets) \tag{2.5}$$

$$p(E3) = P(E4) \tag{2.6}$$

$$p(E5) = \frac{1}{2} P(2 \ packets) \tag{2.7}$$

$$p(E6) = \frac{1}{4} P(2 \ packets) \tag{2.8}$$

While equation (2.6) is consistent with the data, the other equations are not. An examination of the data indicates that the occurrence of a single output link preference is higher than these equations would indicate. The reason for this is that one actually has to average over all the typical paths taken by packets moving thru the network. The probabilities .25/.25/.50 of [9] are only applicable to the first hop taken by the packets. For the nodes along a typical path the probability that a packet has a single output link

preference is elevated compared to that for the first hop. This elevation is consistent with the routing probability data of [11].

3 Routing with Information

3.1 Age and Conflict Deflection

The solid curve in Fig. 7 shows the delay probability density function for the standard deflection protocol with p=.5 (the data of Figures 7 and 9-12 is based on 1.5 x 10⁶ slot simulations). A significant feature of it is a heavy tail, with 1% of the packets taking more than 28 hops to arrive at their destination. To trim this maximal delay, one could consider making deflection decisions based on the "age" of the packets. Here the age of a packet is defined to be the number of hops a packet has undergone so far. That is, when two packets in a node prefer the same output link, the older packet is given its preference. Situations involving ties in age are settled randomly. We refer to this strategy as "age deflection". It has appeared previously in [7,8]. Implementation involves the inclusion of a hop counter field in the packet header that is incremented for each transit of a link.

The delay probability density function for age deflection with p=.5 is also shown in Fig. 7. The tail now decreases much more steeply. The main mode of the density is shifted to the right, compared to that of the standard deflection protocol, as a consequence of giving older packets priority is that younger packets lose priority. The oldest packet now takes 27 slots to reach its destination, compared to 100 slots for the oldest standard deflection packet.

What happens, in fact, with age deflection is that a packet's deflection probability decreases as it ages. Deflection probability as a function of packet age is plotted in Fig. 8.

It can be seen to approach zero as a packet ages (for age deflection, deflection probability is the probability that a packet of age i is competing with a packet of age i+1,i+2... plus one half of the probability that it competes with a packet of age i). In age deflection, packets beyond a certain age are assured of getting their preference during deflection events and thus a shortest path to their ultimate destination.

Table 4 shows the probability density of the number of deflections experienced by packets for standard deflection and age deflection. The age density has support over 0-5 deflections compared to a support of 0-24 deflections for standard deflection.

An alternative to age deflection is "conflict" deflection. Here deflection decisions are based on the number of deflections a packet has undergone [16]. This necessitates a field in the header to record this information. Basically, a packet's deflection probability decreases as it undergoes more deflections.

Fig. 9 compares the delay density for age and conflict routing. The variance of the density for conflict routing is somewhat larger than that for age routing and the tail is somewhat heavier but the two strategies are, roughly, comparable in performance.

Table 4 suggests that conflict deflection can be implemented with a deflection counter field with fewer bits than the hop counter field of age deflection (i.e. packets generally undergo fewer deflections than hops). This may be a consideration for short packet applications.

Table 5 shows the 50th, 90th, 95th, 99th and 100th percentile of the delay density for age and conflict deflection based on 1.5 x 10⁶ slots. That is, the table entry for the 95th percentile is the age that 95% of the packets are less than. Note that the the 100th percentile, or oldest packet age, is a sensitive function of simulation time. There is an evident

improvement for age and conflict deflection with respect to the tail of the delay density, or the maximally delayed packets, with respect to standard deflection.

Finally, Table 6 shows comparable throughputs under various loads for standard, age and conflict deflection.

3.2 Distance Based Deflection

In this section several distance based deflection routings strategies are considered. In a static distance strategy a new packet arriving at a node is assigned a priority equal to the distance from its point of origination to its destination. This priority is not changed as the packet transits the network. During a deflection event the packet with the highest priority is assigned to the preferred output link. Figure 10 shows the delay probability density function when p=.5 and priority is given to the packet with the further source-destination distance, to the packet with the closer source-destination distance and, by way of comparison, age deflection.

The static distance deflection strategies are inferior in shaping the tail of the delay density compared to age deflection or even the standard deflection protocol. For the further distance strategy 1% of the packets took more than 35 hops to reach their destination and for the closer distance strategy 1% of the packets took more than 39 hops to reach their destination. We note that the multiple modes in the density tails appear to be structural features of these density functions.

In a dynamic distance deflection strategy, during a deflection event, the distance for a packet from its current node location to its destination is used in making deflection decisions. Fig. 11 illustrates the delay probability density function when the further packet is given its preference, when the closer packet is given its preference [16] and, by way of comparison, age deflection. The tails for these dynamic strategies are not as heavy as those of the static strategies but they are still inferior to those of the age and conflict strategies. When further packets are given their preference 1% of the packets require more than 35 hops to reach their destination and when closer packets are given their preference 1% of the packets require more than 26 hops to reach their destination.

From table 5 it can be seen that the distance based deflection strategies have tails which are longer or approximately the same as that of standard deflection.

Note in both Fig. 10 and Fig. 11 that strategies that give preference to closer packets have an earlier delay peak because they work well for packets traveling between source-destination pairs that are close together.

The problem with the use of static distance deflection strategies is that they route some packets efficiently (further or closer) at the expense of others (closer or further). With dynamic distance deflection strategies packets are given efficient routes during one part of their journey thru the network (further or closer) at the expense of receiving poorer routes during another part of their journey (closer or further).

Hajek and Krishna [16] have suggested that because of the efficacy of giving priority to packets close to their destination for evacuation problems such a strategy may work well in steady state. Our experience with the dynamic distance strategy giving preference to the closer packet for a Manhattan street network indicates that it is not always safe to make such an extrapolation. However, we did find that when ties in conflict deflection are settled in favor of the packet closer to the destination (as in [16]) then there is a modest improvement in the delay density.

3.3 Predictive Deflection

Consider a hybrid strategy where, at its entry to the network, each packet is assigned a priority equal to its source-destination shortest path distance. Every time a packet is deflected its priority is incremented by four. Preference in a deflection situation is given to the packet with the largest priority. This strategy will be referred to as predictive deflection.

The motivation behind this strategy is that if a packet's priority represents its predicted path length, one would like to give preference in a deflection event to the packet with the longer path length in order to reduce the extent to which packets are maximally delayed. Thus the priority is initialized with the source-destination shortest path length. The increase by four during a deflection is based on an observation by Maxemchuk [9] that when a packet undergoes a *single* deflection its path length is increased by four. Note that this strategy is approximate, for instance after m *multiple* deflections the increase in path length may not be exactly equal to 4m.

The delay density of predictive deflection is shown in Fig. 12 along with that of age deflection. From this and from Table 5 it can be seen that the tail of the delay density is comparable to that of age and conflict deflection. The network throughput vs. various loads for predictive deflection is shown in Table 6.

4 A Simple Relation

One unifying relationship that held with surprising accuracy for all of the deflection strategies discussed in this paper is:

$$Mean\ No.\ of\ Hops = \frac{2 \cdot Link\ Utilization}{Nodal\ Throughput} \tag{4.1}$$

This equation divides the raw total output rate of a node by the actual throughput (which equals the arrival rate) to yield the mean number of hops experienced by packets. With normalization and a homogeneous network, nodal throughput is equal to network throughput. For simulations with 1.5 x 10⁶ slots the relation is accurate to four or five significant decimal places. It would seem reasonable that similar relations (differing only in the constant) would hold for fabrics with different numbers of output links per node.

5 Conclusion

The numerical results presented here indicate that the age, conflict and predictive deflection are superior to standard deflection in controlling and making consistent the maximal delay experienced by packets in the Manhattan Street Network. We believe that these strategies would be useful in a variety of switching fabrics other than the Manhattan Street network. This work also demonstrates a need for input buffers in a practical Manhattan Street Net implementation.

6 Acknowledgement

The research in this paper was supported by the National Science Foundation under grant no. NCR-8703689 and by the SDIO/IST and managed by the Office of Naval Research under grant no. N00014-85-K0610.

References

- [1]H. Abu-Amara, T. Barzilai and Y. Yemini, "PSI: A Silicon Compiler for Very Fast Protocol Processing", Columbia University Dept. of Computer Science Technical Report CUCS-477-89.
- [2]E. Ayanoğlu, "Signal Flow Graphs for Path Enumeration and Deflection Routing Analysis in Multihop Networks", Proceedings of IEEE Globecom '89, pp. 1022-1029.
- [3]E. Ayanoğlu and R.J. Caballero, "Signal Flow Graphs for the Delay Analysis of Deflection Routing in Multihop Networks", submitted to IEEE Transactions on Communications.
- [4]P. Baran, "On Distributed Communication Networks", IEEE Transactions on Communication Systems, Vol. CS-12, March 1964, pp. 1-9.
- [5]A.G. Greenberg and J. Goodman, "Sharp Approximate Models of Deflection Routing in Mesh Networks", accepted for the IEEE Transactions on Communications.
- [6] A.G. Greenberg and B. Hajek, "Deflection Routing in Hypercube Networks", unpublished manuscript.
- [7]A. Iyengar and M. El Zarki, "Switching Prioritized Packets", Proceedings of Globecom '89, pp. 1181-1186.
- [8]Y. Kang and J.S.K. Wong, "Distributed and Fail-Safe Routing Algorithms in Toroidal-Based Metropolitan Area Networks", accepted for Computer Networks & ISDN Systems.
- [9]N.F. Maxemchuk, "The Manhattan Street Network", Proceedings of IEEE Globecom'85, New Orleans, La., 1985, pp. 255-261.

- [10]N.F. Maxemchuk, "Routing in the Manhattan Street Network", IEEE Transactions on Communications, Vol. COM-35, No. 5, May 1987, pp. 503-512.
- [11]N.F. Maxemchuk, "Comparison of Deflection and Store and Forward Techniques in the Manhattan Street Network and Shuffle Exchange Networks", Proceedings of IEEE IN-FOCOM'89, Ottawa, Canada, April 1989, pp. 800-809.
- [12]T.G. Robertazzi, "Toroidal Networks", IEEE Communications Magazine, Vol. 26, No. 6, June 1988, pp. 45-50.
- [13]T.G. Robertazzi and A.A. Lazar, "Simulation Data for Standard, Age and Conflict Deflection Routing for a Manhattan Street Network", SUNY at Stony Brook College of Engineering and Applied Science Tech. Rep. 585, July 31, 1990. Available from T. Robertazzi.
- [14]T.G. Robertazzi and A.A. Lazar, "Long Run Simulation Data for Deflection Routing in a Manhattan Street Network", SUNY at Stony Brook College of Engineering and Applied Science Tech. Rep. 593, Oct. 1990, Revised March 13, 1991. Available from T. Robertazzi.
- [15]T.G. Robertazzi and A.A. Lazar, "Simulation Programs for Distance Based Deflection Routing in a Manhattan Street Network", SUNY at Stony Brook College of Engineering and Applied Science Tech. Rep. 594, Oct. 1990. Available from T. Robertazzi.
- [16]B. Hajek and A. Krishna, "Performance of Shuffle-Like Switching Networks with Deflection", Proceedings of IEEE INFOCOM'90, San Francisco, June 1990, pp. 473-480.

- [17] A.A. Lazar, A. Temple and R. Gidron, "An Architecture for Integrated Networks that Guarantees Quality of Service", International Journal of Analog and Digital Communication Systems, Vol. 3, No. 2, 1990.
- [18]T.G. Robertazzi and A.A. Lazar, "Deflection Strategies for the Manhattan Street Network", Proceedings of IEEE International Conference on Communications 91, Denver CO, June 1991. Due to a software error in the simulator, the delay density data between the 99% and 100% percentiles takes on erroneously large values in this paper. The corrected data appears in this manuscript.

Table 1: Routing Events

Event 0	No packets				
Event 1	1 packet, 1 preferred output link				
Event 2	1 packet, 2 equally good output links				
Event 3	2 packets, 1 preferred output link each, same link				
Event 4	2 packets, 1 preferred output link each, different links				
Event 5	2 packets, 1 preferred link for one packet, 2 links for other				
Event 6	2 packets, 2 equally good output links for each packet				

Table 2: Routing Event Probabilities (Standard Deflection)

р	E0	E1	E2	E3	E4	E5	E6
.1	.527	.278	.104	.0228	.0235	.0373	.00716
.2	.226	.321	.134	.0777	.0779	.135	.0286
.3	.101	.258	.116	.124	.124	.226	.0513
.4	.0510	.191	.0902	.155	.153	.291	.0686
.5	.0276	.138	.0676	.176	.174	.336	.0808
.6	.0151	.0982	.0499	.191	.188	.367	.0904
.7	.00841	.0667	.0356	.202	.199	.391	.0971
.8	.00430	.0430	.0244	.209	.207	.410	.102
.9	.00174	.0231	.0151	.216	.213	.424	.107
1.0	.00000	.00665	.00725	.221	.218	.437	.110

Table 3: Nodal Buffer Occupancy Probabilities (Standard Deflection)

р	P(0 packets)	P(1 packet)	P(2 packets)
.1	.527	.382	.0909
.2	.226	.455	.319
.3	.101	.374	.525
.4	0510	.282	.668
.5	.0276	.206	.767
.6	.0151	.148	.837
.7	.00841	.102	.889
.8	.00430	.0674	.928
.9	.00174	.0381	.960
1.0	.00000	.0139	.986

Table 4: Number of Deflections Density (p=.5)

# Deflections	Standard	Age	Conflict	Predictive
0	.556	.446	.395	.404
1	.233	.358	.392	.374
2	.110	.165	.188	.189
3	.0525	.0296	.0245	.0325
4	.0255	.00161	.000683	.00151
5	.0123	.000022	.000005	.000017
6	.00598		1 packet	2 packets
7	.00292			
8	.00145			
9	.000691			
10	.000344			
11	.000161			
12	.000078			
13	275 packets			
14	129 packets			
15	53 packets			
16	27 packets			
17	14 packets			
18	6 packets			
19	6 packets			
20	2 packets			
24	1 packet			

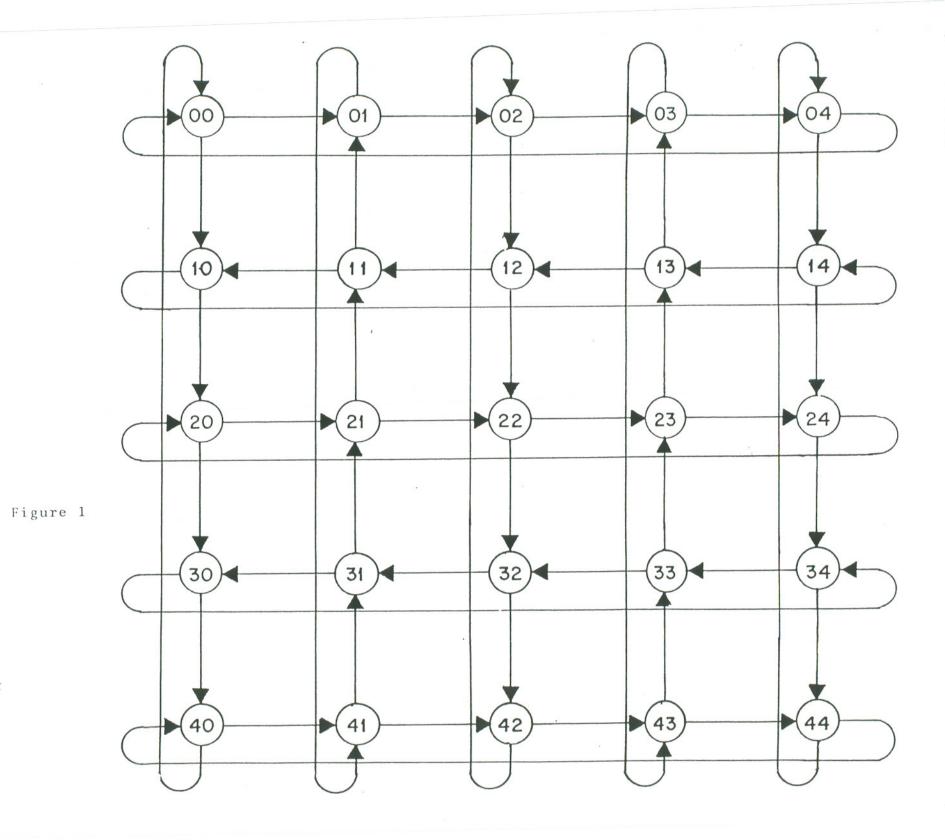
Table 5: Percentiles of Delay Density (p=.5)

Protocol	50	90	95	99	100
Standard Deflection	6	15	19	28	100
Age Deflection	8	13	14	16	27
Conflict Deflection	8	13	15	17	31
Predictive Deflection	8	13	14	16	28
Dynamic Distance	6	14	18	26	85
Preference: Closer					
Dynamic Distance	7	18	23	35	127
Preference: Further					
Static Distance	5	16	22	39	189
Preference: Closer					
Static Distance	6	14	20	35	233
Preference: Further					

 1.5×10^6 slots, appox. 20×10^6 packets

Table 6: Average Throughputs

р	Standard	Age	Conflict	Predictive
.1	.0943	.0945	.0944	.0943
.2	.156	.157	.156	.156
.3	.185	.188	.185	.185
.4	.198	.203	.199	.198
.5	.206	.212	.207	.206
.6	.211	.218	.212	.211
.7	.214	.221	.215	.214
.8	.216	.224	.218	.217
.9	.218	.227	.219	.218
1.0	.220	.228	.221	.220



TOTAL TRANSIT (Predictive, Static Distance)

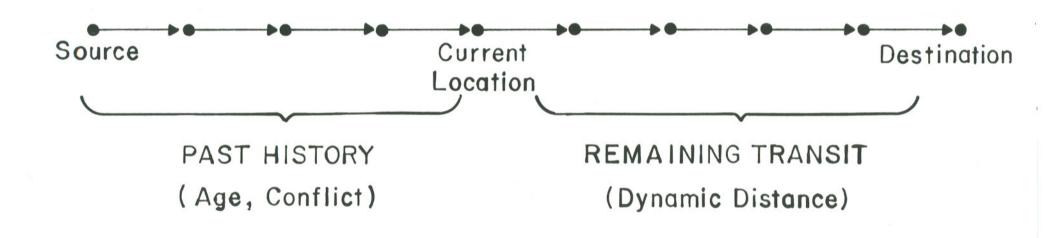


Figure 2

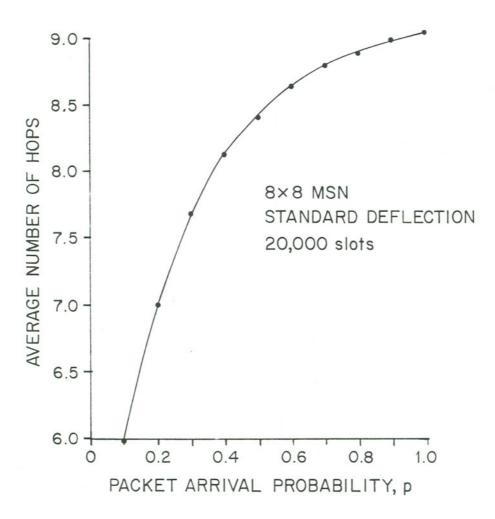


Figure 3

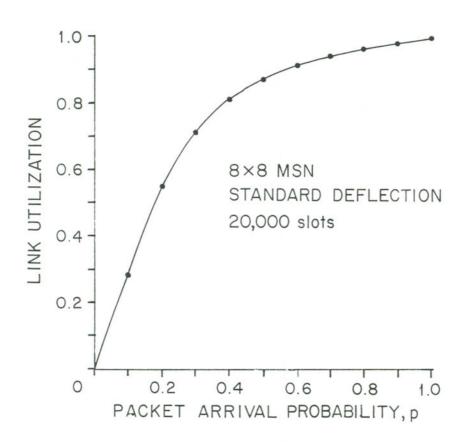


Figure 4

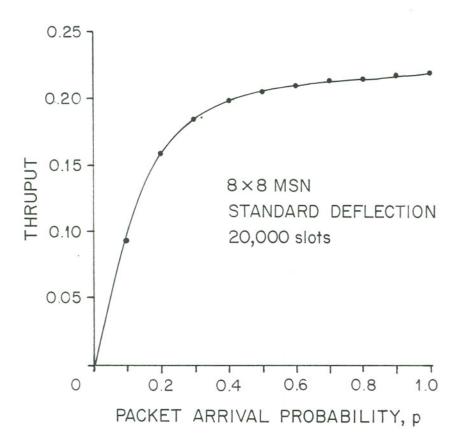


Figure 5

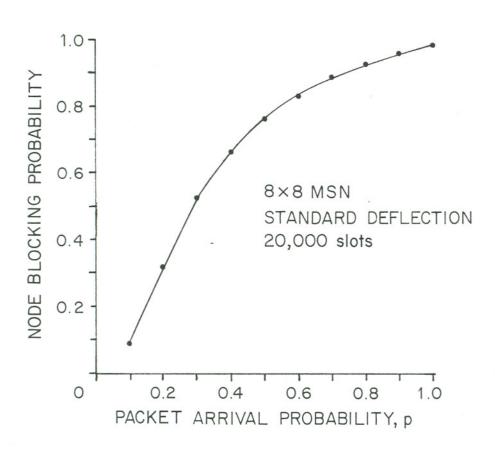


Figure 6

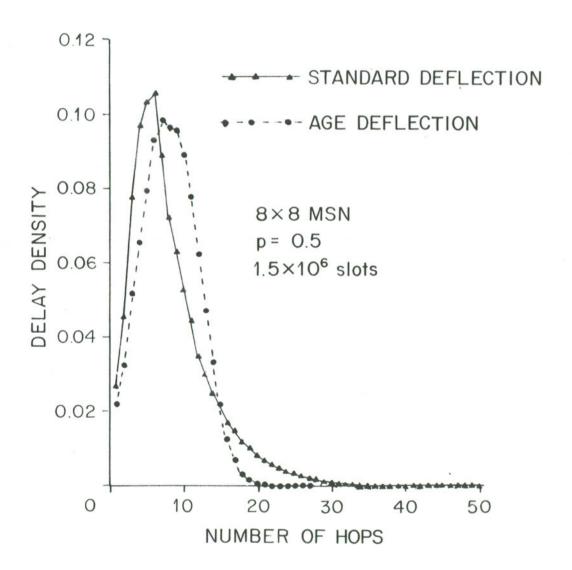


Figure 7



Figure 8

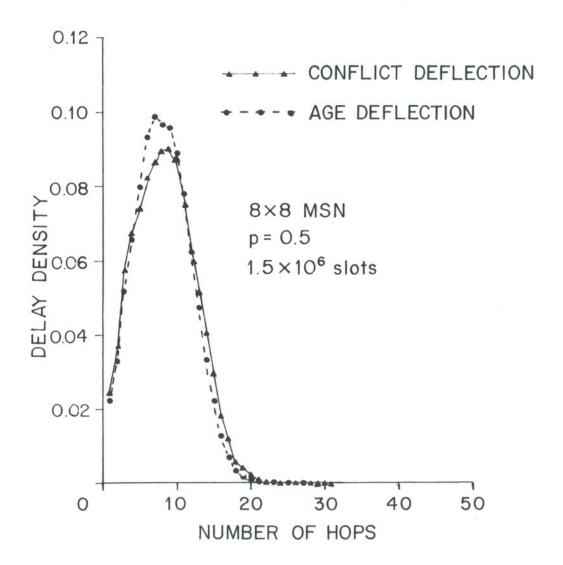


Figure 9

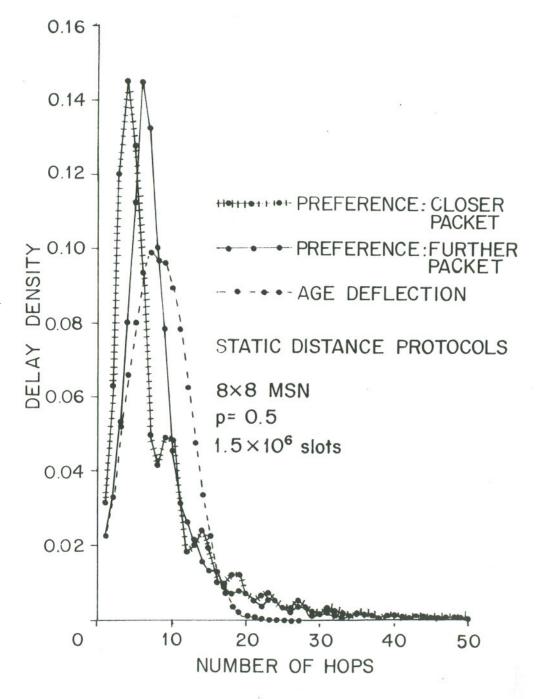


Figure 10

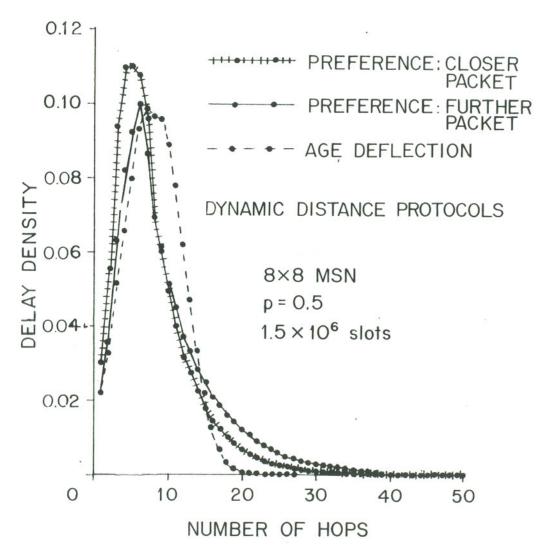


Figure 11

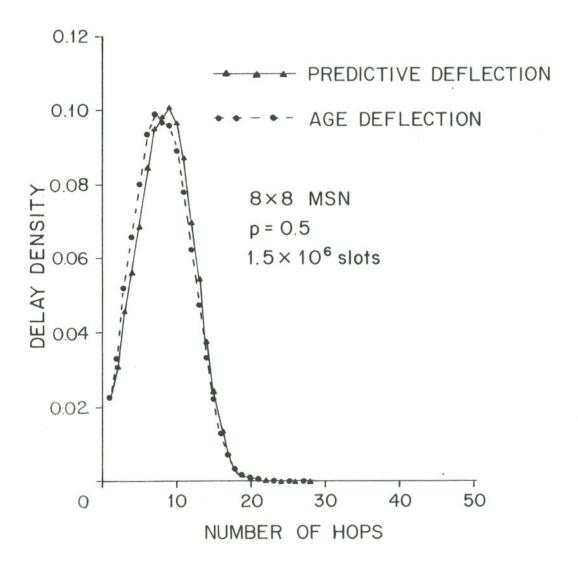


Figure 12