A Preliminary Model for Deflection Routing

Qin WEI and Gérard HEBUTERNE

Dep. RST, Institut National des Telecommunications

9, rue charles fourier 91011 evry cedex France

Phone: +33 (0) 1 60 76 46 91 Fax: +33 (0) 1 60 76 42 91

E-mail: Qin.Wei, Gerard.Hebuterne @int-evry.fr

Abstract

Because of development of optical technology, the dream of "All optical networks" has become true. Without conversion between optic et electronic, "All optical networks" are well suited to fulfill the growing demand on bandwidth. One of the popular methods to operate all optical networks is the "deflection" principle, different from recent methods such as diffusion and virtual circuits used in Internet and ATM networks respectively. In this paper, we first compare possible high-bandwidth solutions able to interconnect sources in a metropolitan area, through all optical links. Secondly, we define and study two deflection routing methods, named "simple" method and "configuration" method, based upon the knowledge of network state. Our investigation has been done by using both analytical and simulation approaches.

Keywords: deflection, networking structure, routing, analysis, simulation

Acknowledgements: This study was made in the framework of a FT-CNET grant number 95-3114.

1 Introduction

All optical networks rely on optical techniques for both transmission (fibres) and switching (optical switches). The incredibly large amount of bandwidth optical devices provide is certainly a major component in future networks, and in fact it has already been considered and shown to be well suited to fulfill the growing demand on 1) bandwidth per user, 2) protocol transparency, 3) higher path reliability, and 4) simplified operation and management [6].

Compared to traditional networks making use of electronic devices, one critical issue for such networks is the lack of storage capability. The only way current optical technology can supply the required buffers for storage is through delay

The original version of this chapter was revised: The copyright line was incorrect. This has been corrected. The Erratum to this chapter is available at DOI: [10.1007/978-0-387-35353-1_28](https://doi.org/10.1007/978-0-387-35353-1_28)

D. Kouvatsos (ed.), *Performance Analysis of ATM Networks*

© IFIP International Federation for Information Processing 2000

lines. However, the number of fibre delay lines is very limited. For this reason, new solutions to deal with contention situations and network optimization choices are needed. One of the popular methods allowing to operate all optical networks is "deflection". When two or more packets attempt accessing the same output at the same time, the node lets one packet through and routes others towards idle outputs according to some criteria.

In fact, deflection routing is not only a solution for lack of storage capability but also for congestion control and bandwidth allocation. Deflection networks have no internal congestion as opposed to present ATM or IP networks exhibiting bottlenecks. All packets entering the network will exit, resulting in guaranteed packet delivery and no packet loss. In addition, deflection networks offer the possibility of total bandwidth allocation to a communication between two end-users, so that they are well suited to situations with day-to-day traffic variation or uncertain forecasts.

Most studies in deflection networks [1, 4, 7] are usually based on 2×2 switches. In the present paper, the availability of 4×4 switches is assumed. First, we address some simple arguments to choose between basic architectural alternatives in the framework of Metropolitan Area Networks (MAN), through all optical links using the deflection method. A comparison is given between the "Manhattan Street" architecture and the "star" architecture. The comparison attempts to address such aspects as the optimization of the total length of fibre involved or the traffic capability through simple dimensioning arguments.

The comparison favors the "Manhattan Street" architecture, which is thus considered in the sequel, devoted to a study of deflection routing. The complete definition of the routing algorithm allows a great flexibility. Especially, the choice of the "favored" packet (the one being selected in case of contention) may be done based upon various arguments (such as the "age" of conflicting packets, or the distance to the destination, or more generally the knowledge the nodes have about their neighbourhood). We define and study two routing methods based upon the (partial) knowledge the nodes have about the network state. The methods, named "simple" method and "configuration" method, differ by the degree of knowledge.

The study yields the throughput of a connectionless network using deflection routing. It shows that the "configuration" method is better than the "simple" method. This is probably because the "configuration" method depends on the network-wide information which is available at the node.

An analytical model is proposed under the assumptions of "independence" between nodes and uniformity of the traffics. The model is validated through a simulation experiment.

2 Network Architectures

2.1 Some Networking Architectures

Most research on deflection routing is based upon regular topologies such as the "Manhattan Street" networks with bidirectional links (torus, see Figure 1, or grid, in which the "backplane" links are suppressed), and the "Shuffle-exchange" networks since they offer the same number of inputs and outputs at each node [3, 9]. Among these architectures, the "Manhattan Street" network seems more adequate for the following reasons:

- (1) Most often, there exists two paths with the same shortest distance to the destination.
- (2) The cost of a deflection is constant (2 additional links).

Figure 1: A bi-directional "Manhattan Street" torus

Two variants of the "Manhattan Street" architecture have been considered, namely the *torus* and the *grid*. Figure 1 displays the diagram of the torus, characterized by N^2 identical nodes with 4 input links and 4 output links, and an additional local I/O access – referred to as a 4×4 switch. The end nodes wrap around, row (or column) edges are connected, leading to identical switch nodes in terms of the number of inputs and outputs.

On the other hand, the "Manhattan Street" grid has no connections between the end points (row as well as column). Nodes are not equivalent, and three node matrices (the corner nodes, the edge nodes and the general nodes) must be considered to specify the "Manhattan street" grid architecture.

The *star* architecture provides a simpler solution to connect nodes in a metropolitan environment [10]. It offers many advantages at the physical level of all optical networks ([8]). This raises the point of choosing, among "Manhattan Street" network and "star" network, the best suited for Metropolitan Area Networks (MAN) architecture. In this Section, a comparison of these alternatives is given in terms of the total fibre length and of traffic carried.

Also, a comparison between "Manhattan Street" torus and grid has been studied. To summarise shortly, a "Manhattan Street" torus is far better than the grid version. The detail of the comparison is presented in [10].

2.2 The grid solution

2.2.1 Total Fibre Length

We use the following notations and assumptions.

- L : the "diameter" of the network, i.e. the length of each row / column of the Manhattant Street network;
- N : the number of columns or rows.
- K : the total number of sources.
- c : the packet speed in the fibre.
- m : the time needed to send one packet, so that cm is the spatial extension of the packet (the total fibre length a packet occupies).
- Λ the total arrival rate of packets in the network, and λ the arrival rate per node ($\lambda = \Lambda/N$ for uniformly distributed load).

We assume that each link is composed of two fibres to ensure bi-directional connectivity (see the structure of the node, on Figure 1. The total fibre length could be written as follows:

$$\text{length} = 4NL + \frac{KL}{N}, \quad (1)$$

the first term represents the total fibre length of the network ($2N$ rows, each bidirectional), and the second term is the total fibre length of the distribution network (connection of source to its nearest neighbour node, assuming sources are uniformly scattered in the area; the length is around half the elementary square side, and 2 fibres are provided). The optimization problem can be stated as follows : assuming a given number K of traffic sources, how many rows/columns should be used ? Taking derivative of equation (1), the number of rows or columns N may be optimally chosen as :

$$N = \frac{\sqrt{K}}{2} \quad (2)$$

Using this value as N in equation (1), it is shown that the total fibre length is equally shared between the core network and the access network, and that each node serves 4 sources.

2.2.2 Maximum Carried Traffic in Network and Maximum Traffic per Source

Let us assume that the order of magnitude of a connection length is L , so that the mean packets delay in network is L/c . Each packet occupies a length on the fibre equal to cm . The total fibre length in transit network is $4NL$, so $4NL/cm$ packets may be simultaneously stored in the network. According to *Little formula*, the maximum carried flow Λ of the network is as follows:

$$\Lambda \leq \frac{\frac{4NL}{cm}}{\frac{d}{c}} r = \frac{4Nr}{m} \quad (3)$$

where $r(\geq 1)$ is a factor which takes account of the additional cost (in link utilization) of the "deflection" method. Equation (3) shows that the throughput depends only on N, r and m but not on L . Assuming the sources offer identical traffics, this yields the maximum packet flow intensity per source. Expressed in terms of the admissible load, one gets:

$$\text{Load/Source} = \frac{4Nr}{K} \quad (4)$$

Choosing for N the optimum value (2), and taking for r the ideal value $r = 1$, one gets an upper bound for the admissible load per source:

$$\text{LoadMax} = \frac{1}{N} \quad (5)$$

Note that the above equation means that the maximum load per source (and also the maximum load per node) decreases as network size increases. This result is discussed in Section 6.

Figure 2: A bi-directional "star" structure

2.3 The "Star Network" Solution

The "star" network is built according to Figure 2 in which each connection has a length around $L/2$. In this case, the total fibre length could be equal to LK ,

while the maximum flow carried is given by Equation (3). Since the maximum flow carried is the same in "Manhattan Street" and in "star" architecture, the comparison is restricted to the total fibre length.

2.4 Comparison in Total Fibre Length

One may compare the equations giving the total length in the two architectures. Equation (6) below gives the condition under which a preference is given to "Manhattan Street" network in term of total fibre length.

$$4N^2 + K < KN \quad (6)$$

This expression may be transformed to give the condition in terms of the number of sources. In the general case, the condition under which the "Manhattan street" network should be preferred is:

$$K > \frac{4N^2}{N-1} \quad (7)$$

In the case where N is chosen optimally according to (2), it is immediate to show that the "Manhattan street" network uses less fibre length as soon as $K > 16$.

2.5 The Choice of an Architecture

The conclusions we get from this short analysis confirm feeling one gets from the literature survey, namely that the "Manhattan Street" architecture seems to be the most valuable one. So in the sequel the analysis is restricted to this case.

3 Node Operation

Without any buffering capability, the basic node consists of five inputs and five outputs according to the right-hand part of Figure 1. Four inputs (or outputs) are interconnected to their neighboring nodes, and one input (or output) links the local source. As mentioned earlier, for a "Manhattan Street" torus all nodes have the same structure.

To analyze performance, the deflection network is assumed to be packet switched and time slotted, with fixed slots. As a consequence, the network is synchronous. New packets are generated independently at each source according to a Bernoulli process, thus providing a uniformly distributed traffic density. Each node receives arriving packets at the beginning of a time slot, and sends them at the end of the time slot.

Top priority is assigned to packets in transit rather than admitting new packets. A new packet could enter the network if at least one of the transit outputs of the associated node is idle. The goal is to protect packets in transit, which have already consumed bandwidth.

In the model, destinations of new packet are chosen *independently* and *uniformly* at random from all nodes not coinciding with their sources.

4 Deflection Routing

4.1 Different Approaches

The choice of the packet to be deflected has an influence on the overall performance. Here, we define and compare two methods which differ by the degree of knowledge a node has of its neighborhood.

The *simple* method works as follows: each packet is preferably routed to its destination according to the "shortest path" criterion. A packet could be deflected if its preferred link is seized by another packet which has been processed previously (either the node has a fixed scanning policy so that packets are processed in a given order, or packets are processed according to some random phenomenon).

In the second method the node processes all incoming packets in a single operation. Among all routing combinations, the node chooses the one which minimizes the sum of the distances of packets to their destinations. We refer to this algorithm as the *configuration* method.

Other methods are applicable, which make the "age" of the packets as a preference criterion. The comparison between our approaches and these methods is beyond the scope of the present study.

The increased complexity of the latter method prompts to verify the gain which is to be expected (note that the high speed at which packet must be processed makes the complexity of the routing operation a key factor).

4.2 Comparison of the Methods by Simulations

By simulations, we compare the two routing methods in terms of throughput (see Figure 3). The configuration is with $N = 6$.

Briefly, it is clear that network throughput (packets/slot) increases as the load increases from 0 packets/slot to 1 packet/slot per node. The maximum network throughput by "configuration" method reaches 0.83 packets/slot against 0.57 packets/slot for the "simple" method. Figure 5 shows clearly the advantage of the "configuration" method over the "simple" one. The explanation is simple: the "simple" method allocates resources on the basis of a more local information, while the "configuration" method incorporates a more global information in its decision making process.

The same kind of results could be derived from the other parameters of interest. For instance, the average end-to-end delay is significantly shorter for the "configuration" method (the curves are omitted here, being of little interest: see below for the delay in the "configuration" method).

Consequently, under weak traffic conditions, there is no use to look for an optimized approach. However, for medium to large loads, the "configuration"

algorithm is worth considering. The gain, in terms of throughput is quite obvious.

Figure 3: Node throughput: "configuration" versus "simple" method

5 An Analytical Model

Previous analytical studies on deflection networks [1, 4, 7] have proposed accurate models to evaluate the behaviour of this routing mechanism. When these approaches are used on larger switching nodes (here, 4×4 nodes, instead of 2×2), they fail in giving accurate results. This had lead to conduct deeper analyses of the phenomenon, allowing building a more sophisticated model which provides reasonable accuracy (a previous model, based on the same assumptions, has given quite poor results see e.g. [5]).

5.1 Notation

Now we use the following additional notations (the "types" are explained in the next section) :

- d : the effective distance between the source of a packet to its destination.
- d_0 : the length of the shortest path from source node to destination.
- D, D_0 : the mean values of d and d_0 .

- u : link utilization; u_1 (resp. u_2) denote the utilization due to type-1 (resp type-2) packets, see below.
- $n_i(k)$: average number of packets of type i (see below) at distance k from their destination, and n_i the total number: $n_i = \sum n_i(k)$.
- p_i^d : deflection probability of a type- i packet.
- $P(j_1, j_2)$: probability that one typical packet meets j_1 type-1 packets, j_2 type-2 packets in a node.
- $\omega_i(k)$: the flow of packets entering the network at distance k from their destination and being of type- i .

5.2 Routing

All packets arriving at the beginning of a time slot should be routed at the end of the time slot. Each node is located according to its coordinates (x, y) from 1 to N .

The algorithm attempts first to route the packets according to their shortest path. If no solution exists, the "configuration" method is run so as to minimize the total number of deflections. Packets to be deflected are taken at random. Note that one deflection at a node induces an increase in length of two additional links.

5.3 Principle of the Analysis

The key point is that 2 kinds of behaviour are to be observed. Packets which are in the same row/column as their destination have only one possible output. On the other hand, all other packets have 2 possible outputs. The latter ones are said *type-2* packets, the former ones being *type-1* packets.

The probability for a packet to be type-1 or type-2, under our uniformity assumptions, depends only on the distance between the packet and its destination. The model consists in tracking the number of type-1 and type-2 packets at distance k from their destination. From this all quantities of interest are derived.

Let n_1, n_2 be the total average number of packets of type 1 and 2, and $n = n_1 + n_2$. Let $D = E(d)$ denote the average sojourn time of a packet (in number of hops). The rate at which packets enter (and exit from) the network is $\Lambda = N^2\lambda$. Little's relation allows to write: $n = \lambda N^2 D$. Now, there are $4N^2$ available slots containing n packets so that $u = n/4N^2$. Eliminating n gives the relation (see [3]):

$$u = \frac{\lambda D}{4} \quad (8)$$

The u_i can be derived from u and the overall population:

$$u_i = \frac{n_i}{n_1 + n_2} u \quad (9)$$

Consider a test packet. According to the Bernoulli assumptions, the probability that it meets j_1, j_2 packets in the next node is simply:

$$P(j_1, j_2) = \binom{3}{j_1} u_1^{j_1} \binom{3-j_1}{j_2} u_2^{j_2} (1-u)^{3-j_1-j_2} \quad (10)$$

The deflection probability of the packet, given its type and given the number (j_1, j_2) of possible conflicting packets may be estimated by a combinatorial analysis. The following table summarizes the results (see [11] for a comprehensive analysis):

j_1	j_2	$p_1^d(j_1, j_2)$	$p_2^d(j_1, j_2)$
0	0	0.0	0.0
0	1	0.0	0.0
0	2	0.073	0.021
0	3	0.223	0.087
1	0	0.125	0.0
1	1	0.172	0.026
1	2	0.279	0.082
2	0	0.229	0.031
2	1	0.297	0.082
3	0	0.316	0.078

Table 1: Deflection probabilities vs configurations

The deflection probability of a type- i packet is given by:

$$p_i^d = \sum_{j_1, j_2} p_i^d(j_1, j_2) P(j_1, j_2) \quad (11)$$

and the average deflection probability (e.g. ratio of the number of observed deflection during a slot to the total number of packets being processed) :

$$P^d = \frac{n_1 p_1^d + n_2 p_2^d}{n_1 + n_2} \quad (12)$$

Since each deflection lengthens the path to the destination by 2 links, the following relation holds:

$$D = D_0 + 2P^d D \quad \text{or} \quad D = \frac{D_0}{1 - 2P^d} \quad (13)$$

D_0 is easy to derive with the uniform assumption: $D_0 \approx N/2$

5.4 Performance Analysis

From the above elements, a set of equations is written which governs the steady-state behaviour of the $n_i(k)$. At each time slot, new packets enter the network,

Figure 4: Building the set of equations 14 at steady-state

at some distance from their destination. At the same time, packets already in the network move, either being deflected (their distance increase by 1) or being successfully transmitted (their distance decreases).

In the diagram, the upper row represents the n_1 's, and the lower row the n_2 's. Remark that all packets at distance larger than $N/2$ are all type-2: they always have two outgoing links initiating a shortest path to their destination. The set of equations (14) is built from the diagram. In the equations, the notation $q_i = 1 - p_i^d$ is used for clarity. The coefficients of the terms account for the possible behaviours. For instance, a type-2 packet which is not deflected may remain type-2 with probability $(k-1)/k$, or may become type-1; a type-2 packet which is deflected remains type-2; etc. We refer to [11] for the whole justifications, which are rather lengthy (especially for limit conditions, see $k = N/2 - 1, N/2$).

$$\begin{aligned}
 n_1(1) &= \omega_1(1) + q_1 n_1(2) + q_2 n_2(2) \\
 n_1(k) &= \omega_1(k) + q_1 n_1(k+1) + \frac{p_1^d}{3} n_1(k-1) + \frac{q_2}{k} n_2(k+1), \\
 n_2(k) &= \omega_2(k) + \frac{q_2(k-1)}{k} n_2(k+1) + p_2^d n_2(k-1) + \frac{2p_1^d}{3} n_1(k-1) \\
 &\quad 2 \leq k < \frac{N}{2} - 1 \\
 n_1(\frac{N}{2} - 1) &= \omega_1(\frac{N}{2} - 1) + \frac{3q_2}{N-1} n_2(\frac{N}{2}) + \frac{p_1^d}{3} n_1(\frac{N}{2} - 2) \\
 n_2(\frac{N}{2} - 1) &= \omega_2(\frac{N}{2} - 1) + \frac{q_2(N-4)}{N-1} n_2(\frac{N}{2}) + p_2^d n_2(\frac{N}{2} - 2) + \\
 &\quad + \frac{2p_1^d}{3} n_1(\frac{N}{2} - 2) \\
 n_2(\frac{N}{2}) &= \omega_2(\frac{N}{2}) + q_2 n_2(\frac{N}{2} + 1) + p_2^d n_2(k-1) + p_1^d n_1(\frac{N}{2} - 1) \\
 n_1(k) &= 0, \\
 n_2(k) &= \omega_2(k) + q_2 n_2(k+1) + p_2^d n_2(k-1) \quad \frac{N}{2} < k < N \\
 n_2(N) &= \omega_2(N) + p_2^d n_2(N-1)
 \end{aligned} \tag{14}$$

Note that the set of equations cannot be solved directly, since the deflection probabilities depend on the link loads u_i which in turn depend on the solution. So, an iterative method has been implemented which begins by choosing an initial distribution for the (n_i) (the distribution without any deflection has been successfully used), from which a first set of (u_i) is drawn, giving a second set of (n_i) , etc. The convergence is quite fast, and the method is thus quite efficient.

6 Validation

Simulation experiments are conducted to compare results with the analytical model. These comparisons are based on the deflection probability, the node throughput and the length of the actual path. We assume a network with $N = 10$ rows and columns.

Figure 5: Comparison of node throughput: simulation versus analysis

The first set of curves displays the throughput obtained as a function of the link load u : Figure 5. The solid line represents results from the model, while the isolated points are simulation results. The precision of the simulation runs, as estimated through the “method of blocks” is better than 10 %. The dotted line indicates the value of the carried traffic, without any deflection. This is the limit performance an infinite buffer system would give rise to. This limit is given by equation (8) with $E(d) = E(d_0)$. An order of magnitude could be obtained by assuming the typical distance as given by $D_O = N/2$, that is:

$$\lambda_{\text{lim}} = \frac{8u}{N} \quad (15)$$

This result has been already given, see e.g. [3]. It shows that the carried traffic increases linearly with N , that is as the square root of the number of nodes. This explains that the traffic per node decreases as the size increases, as already mentioned (Section 2). Actually, the overall admissible load increases, but the efficiency decreases, as for any meshed network. Especially, in the model taken

Figure 6: Comparison of mean length : simulation versus analysis

here, the growth of the network (i.e., of N) is reflected in an increase in D , which corresponds to a very special kind of network evolution.

Figure 6 shows the variation of the effective distance. For the upper limit, the distance is around 7 (in number of links). This implies that on the average a packet incurs one deflection (increasing the shortest path by 2).

Figure [?] summarizes the previous two. The relation between D and P^d makes it redundant; we present it however because deflection is the main phenomenon. The deflection probability remains moderate even for high link utilization.

On the whole, the Figures show that the results of the analytical model are in good agreement with simulations, slightly overestimating the influence of deflection.

The results are also quite encouraging, as far as the deflection method is concerned. They show that, in the configuration under study, deflection remains a rather unfrequent event (on the average, one deflection incurred during the packet's travel). Thus, these results tend to accredit the possibility of using deflection principle for all optical high-speed networks without degrading the performance.

For instance, Figure ?? tells that the maximum achievable carried load for the 10×10 network is around 0.5 per node, which is quite comparable with figures a store-and-forward network would achieve. As a matter of fact, congestion phenomena strongly limit the load in such networks: limits u to values around

Figure 7: Comparison of deflection probability : simulation versus analysis

0.7, resulting in λ 's around 0.56 – that is, only slightly greater.

7 Conclusions

This study has concentrated on the use of deflection routing for bufferless metropolitan area networks. Assuming a "Manhattan Street Network" topology, it is shown that the so-called "configuration" routing algorithm exhibits significantly better behavior than the simpler, sequential, one. The improvement in network performance, exemplified in Figure 3, largely justifies the increase in the algorithmic complexity.

Also, a quite efficient analytical method is proposed, and validated through a simulation study. The numerical example shows clearly the appealing features of deflection routing, as compared with more traditional store-and-forward mechanisms operating in buffered networks. So, deflection routing could appear as a possible solution for electronic-based networks, where buffer availability is not a critical issue.

References

- [1] A.G. Greenberg and J. Goodman : "Sharp Approximate Models of Adaptive Routing in Mesh Networks", *Teletraffic Analysis and Computer Per-*

formance Evaluation, North-Holland, 1986, pp.225-270.

- [2] J.Brassit and R. Cruz "Nonuniform Traffic in the Manhattan Street Network", ICC'91, pp1647-1651, 1991.
- [3] F. Borgonovo, L. Fratta: "Deflection Networks: Architecture for metropolitan and Wide Area Networks". *Computer Networks and ISDN Systems*. North Holland, 1992, pp 171-183.
- [4] A. G.Greenberg and B. Hajek "Deflection Routing in Hypercube Networks", IEEE transaction on communication, VOL.40, N.6, June 1992.
- [5] Q. Wei and G. Hebuterne: *Performance Analysis of Optical Deflection Networks based on "Manhattan Street"*. Montreal (Canada), April 1996.
- [6] Paul E. Green, Jr: "Optical Networking Update", *IEEE Journal on Selected Areas in Communications*. VOL.14, N.5, pp 764-779, June 1996.
- [7] A. Bononi and P. R. Prucnal: "Analytical Evaluation of Improved Access Techniques in Deflecting Routing Networks", *IEEE/ACM Transactions on Networking*, Vol. 4, N.5, October 1996.
- [8] D. Kofman and M. Gagnaire: *Réseaux haut débit*, pp. 407-462, InterEditions/Masson, Paris, France,1996.
- [9] G. Hébuterne and Q. Wei: "Structure and Routing in All Optical Deflection Networks", *NOC'97*, Vol 3, pp 71-78, June 1997, IOS Press.
- [10] Q. Wei and G. Hebuterne : "optical deflection networks based on Manhattan Street: torus or grid ?", *ICCC'97*, Cannes, November 1997.
- [11] Q. Wei and G. Hebuterne: *Performance Analysis of Deflection Networks based on the Manhattan Street Topology*. Research Report RR-981002, Institut National des Télécommunications, October 1998.

PART THREE

Switch and Multiplexer Models