Proceedings
# of the
1996 Conference
on
Information Science and Systems

Volume II

Department of Electrical Engineering
Princeton University
Princeton, New Jersey 08544-5263

FP-1 Distributed/Decentralized Detection
Chair: Rick S. Blum, Lehigh University

FP-2 Optical Networks
Chair: Paul R. Prucnal, Princeton University

FP-3 Analysis of CDMA Systems
Chair: Laurie Nelson, University of Minnesota

Space-Division Optical Star Networks with Deflection Routing

A. Bononi
Dipartimento di Ingegneria dell'Informazione
Università di Parma, I-43100 Parma, Italy
On leave from Department of Electrical and Computer Engineering, S.U.N.Y. at Buffalo, NY, USA.

Abstract

An optical star network is implemented by an MxM space-division cell switch to which M nodes are connected by dedicated fibers. Each node is equipped with an optical transmitter and an optical receiver. Deflection routing is used to simplify the structure of the central interconnect and the routing of packets within it. Deflected packets delivered to the wrong user are re-routed to the switch. A multi-stage optical implementation of the central interconnect based on Shuffle Exchange stages made of 2x2 crossbar directional couplers is considered. The elementary beta switching elements within the interconnect fabric contain either zero or one optical buffer. The number of stages n is varied from 1 to $log_2 M$. The network topology thus evolves from a pure Shuffle Exchange network, which is multihop, to a potentially single-hop star network when the interconnect is complete ($n = \log M$). Throughput, delay and optical power loss are obtained by simulation and simple approximate closed-forms are obtained for uniform traffic in the case $n = \log M$. It is shown how the drastic reduction of the hop count in the complete interconnect case allows both larger throughput and substantially lower optical power loss as compared to the multihop approach, making such topology an attractive candidate for a transparent optical implementation.

1. Introduction

One of the most serious problems with transparent optical networks is the optical power loss, or attenuation, involved in propagating and switching all-optically signals from source to destination. Even in optically amplified systems, the spontaneous emission noise introduced in the amplification process, proportional to the amplifier gain, ultimately sets stringent limits on the maximum geographical span of the network. If all other transmission impairments were compensated for, such power limit would set the ultimate bound on network size.

We recently considered node structures of reduced functionality in slotted, space-division transparent optical mesh networks in order to locally minimize the optical power loss and still guarantee a reasonable throughput/delay performance [1]. Such networks were based on regular multihop topologies, such as Shufflenet (SN) [2], and employed deflection routing [3],[4] with at most one optical buffer per node [5], since the buffering operation at the optical level, achieved by recirculating fiber-delay loops, introduces large power losses.

In this paper, instead of locally minimizing the attenuation at each optical node, we select a network structure that both reduces the total optical power loss and improves the throughput/delay performance.

We start by observing that a Shufflenet is indeed an Omega network [6] – a well-known multi-stage interconnect – in which the first stage coincides with the last stage, i.e., the outputs are fed back to the inputs to obtain a cylindrically connected structure. Therefore a SN is simply an Omega network in which the passive switches in the interconnect become active sources/sinks of traffic.

In multihop networks, packets have to hop through several intermediate *active* nodes before getting to their destination, thus reducing the opportunity for those nodes of transmitting/receiving their own messages. This causes a quick decrease of the network throughput with increasing network size, for a fixed in/out degree of the nodes. The situation gets worse when deflection routing is employed because of the increased number of intermediate nodes disturbed by each packet.

If we eliminate the active nodes within the switching fabric, leaving the only active nodes at the periphery, the network becomes a central interconnect with feedback. It is thus possible for a packet to reach its destination in a single pass through the interconnect, without disturbing any intermediate node. If the switching fabric cannot deliver the packet in a single pass because of internal blocking, the packet is deflected, i.e., delivered to the wrong node and re-routed to the central interconnect for a second "hop". The process is repeated until the packet is successfully delivered to the intended destination node. Such network is intrinsically single-hop, and it gradually becomes multihop because of deflections.

For a given number of active nodes, the larger throughput obtained in the central-interconnect network is due to the larger number of switching elements with respect to SN, which provide more spatially-disjoint paths from any source to any destination.

The paper is organized as follows. Section 2 introduces the central-interconnect deflection routing network

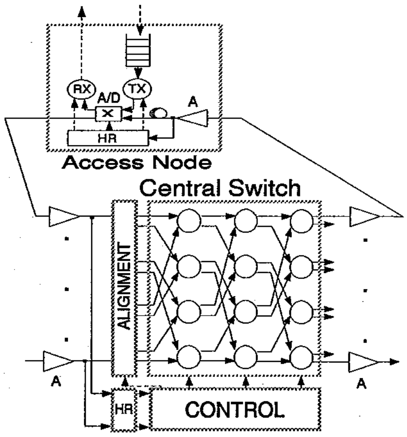


Figure 1: Central-interconnect transparent optical network.

for M nodes. When the switch is a rearrangeable non-blocking MxM cell switch, the network throughput is derived in uniform traffic. In section 3 we choose simpler interconnection networks, starting from a Shuffle Exchange (SX) single-stage interconnect [7] based on 2x2 elementary (beta) switching elements, and we progressively add SX stages, thus increasing the set of allowed input/output connections. When the number of stages is $\log_2 M$ the interconnect is complete, i.e., it is possible to directly connect an input to every output. For this special case, an approximate analytical model is proposed that gives closed-form throughput and delay expressions in uniform traffic. Section 4 introduces simple optically buffered beta elements and finds closed-form analytical expressions for the network throughput. Finally, section 5 contains the conclusions of this study.

2. Central-switch deflection routing networks

Consider an MxM central-switch network, like the one shown in Fig. 1. M active nodes communicate through an MxM optical interconnect (Central Switch). Two dedicated unidirectional fiber links provide bidirectional communications between each node and the central switch. Each access node has one optical transmitter (TX), one optical receiver (RX), and can access the optical fiber bus through a lithium niobate directional coupler (Add/Drop crossbar switch).

The network is a collection of M slotted rings, with a single active node per ring. The slots are aligned at the central interconnect. Fixed-length packets, or cells, are embedded within the slots.

A copy of each cell header is read at the access node (HR) and at the central switch, to set the appropriate switch configuration.

New packets arrive at the optical transmitter at each node uniformly destined to all other nodes, independently slot by slot. They are collected in electronic input buffers, waiting to be injected in the optical transport layer. A polite-access technique is adopted that gives priority to recirculating (deflected) packets over new packets.

Let u be the probability that a slot at the interconnect contains a packet (slot $u$tilization). By the regularity of the network, the uniform traffic assumption and the routing at the interconnect, u is the same for all links.

Since a single link enters each node, a simple application of Little's law gives a relation between throughput per node T and average number of (active-node-to-active-node) hops H:

$$T = \frac{u}{H}. \qquad (1)$$

Such relation is valid for any interconnect structure.

Suppose now the central switch is an MxM rearrangeable non-blocking switch (NB). It could be implemented by a Benes network, which is a multistage structure with $2\log_2 M - 1$ stages and a centralized control [8].

Packets reach the central interconnect, aligned in an array of M input slots per clock. Since the switch is non-blocking, the only conflicts among packets are destination conflicts. If m packets at the input slots compete for the same destination, one of them reaches such destination, while the remaining $(m-1)$ loop back and try again.

If all roundtrip delays are equal, the same $(m-1)$ packets compete again after a roundtrip. The network therefore behaves exactly as an input buffered packet switch with FIFO discipline, the head-of-lines (HOLs) being the recirculating packets.

If instead all feedback delays are different, deflected packets always meet new packets at the interconnect.

The throughput can be easily found in the assumption that the feedback mechanism **(a)** guarantees independence of the slots at the interconnect, which is valid for different roundtrip delays, and **(b)** preserves the uniform destination distribution of packets at the interconnect. This is a good assumption for the ideal non-blocking interconnect, where deflected packets are assigned to idle ports completely randomly. The uniform distribution is not preserved when the internal structure and the control algorithm of the central interconnect do not allow a completely random assignment of deflected packets to the idle ports.

With assumptions (a) and (b), u/M indicates the probability that an input port has a packet for the tagged output, so that the throughput is

$$T(u) = \left(1 - \left(1 - \frac{u}{M}\right)^M\right) \qquad (2)$$

since no packet is assigned to the tagged port it if no packet destined to it is present at the input. This converges to $T(u) = 1 - e^{-u}$ for large M, giving a saturation throughput of 0.63 at u=1 [9].

In the following section we will try to simplify the central interconnect by selecting blocking structures that give a satisfactory throughput/control-complexity tradeoff.

3. Multistage shuffle-exchange central switch (SX-CS)

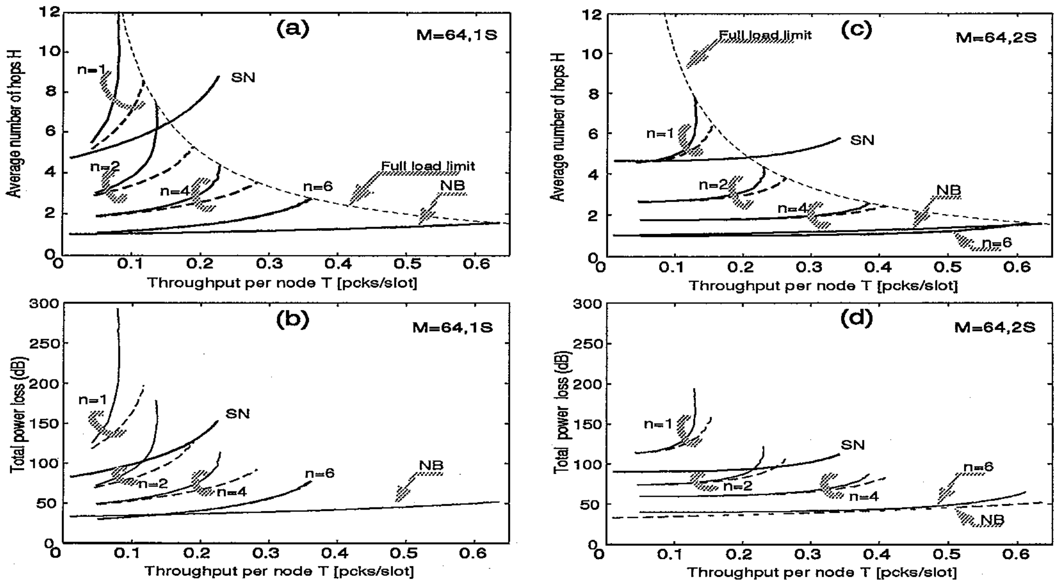


Figure 2: Number of hops and power loss vs throughput in SX-CS networks with n stages for M=64 nodes, beta elements without buffers (1S) and single-buffer (2S), random (R, solid line) and distance priority (DP, dashed line) contention resolution. Curves for two-connected Shufflenet (SN) and Benes interconnect (NB) are given for comparison. Alignment loss of 10 dB and 10 km fiber links are assumed.

Suppose the central switch consists of a number n of Shuffle exchange stages, as shown in Fig. 1. For $n = 1$ this corresponds to a Shuffle Exchange (SX) multihop network, which is known to be quite inefficient with deflection routing [4]. By adding more SX stages, we enrich the set of allowed input/output connections, thereby decreasing the average number of hops, H. A distributed control of the beta elements, shown with circles in Fig. 1, allows to massively parallelize the routing operations. The problem is now to check the throughput/delay improvement as we add more and more SX stages to the central switch.

We will first present some simulation results, and then introduce a simple approximate analytical model.

First consider the case in which each beta element consists of a single 2x2 crossbar with inputs I1 and I2, and outputs H (high) and L (low), which we call 1S. The routing at each beta element makes use of a hop-distance table, by which the controller establishes the preferred output (H or L) for packets at the two inputs. The controller first tries to send a packet along its preferred (shortest-path) output. In case of contention, a fair coin is tossed to assign a packet to its desired output, while the other packet is deflected. This is known as hot-potato routing with random (R) contention resolution (CR). There are cases in which a packet at the beta element can take either output, since it can choose between two equivalent paths to reach its destination. Such a packet is called *don't care* (DC), and the presence of don't cares is exploited to reduce the deflection probability of other (i.e., *care high* (CH) and *care low* (CL)) packets, since a DC never forces a decision on the setting of the beta element and is thus completely equivalent to an empty slot (E). The controller need not distinguish between E and DC packets.

When the number of SX stages equals $\log_2 M$ we can use the self-routing property of the Omega network and avoid using the routing table, which gets cumbersome for large networks. In the Omega (or complete) interconnect case, from an input it is possible to reach every output in a single pass by following the complete binary tree rooted at the input. Since there exists a unique path from every input to every output, packets arriving at the first stage are all care, while a deflected packet becomes don't care at the next stage, since, no matter which output it takes, it has to circulate back to the input to find its way.

Since we adopt a distributed control of the beta elements, adding more than $\log_2 M$ stages does not give any throughput improvement. In fact, if we have $K > \log_2 M$ stages, packets at the first $K - \log_2 M$ stages are all don't care. Having DC packets at both inputs, the control operates a random setting of the beta elements at the first $K - \log_2 M$ stages, and such random shuffling does not solve any contention.

Fig. 2(a) shows simulation results for a M=64 node SX-CS network for random (R) CR rule (solid lines). The number of SX stages is increased from 1 (SX network) to $\log_2 M = 6$. The curve for the non-blocking interconnect (NB) is also shown as a reference, as well as the curve for a two-connected Shufflenet (SN) [5]. The figure shows a striking improvement in performance with increasing number of stages. The difference between the complete switch and the rearrangeable non-blocking switch accounts for the deflections caused in the complete switch by the internal blocking. It is seen that $n = 2$ stages already give better performance than SN for throughput values up to 0.14.

Using 4 stages (i.e. using the same number of directional couplers, $3M$, as SN) always gives a better performance.

We tried to improve the throughput by using a distance priority (DP) CR rule [10]. The packet closer to its destination is given priority in a conflict. The curves for DP are shown with dashed lines in Fig. 2. It is seen that DP gives a good improvement when the number of stages is $< \log_2 M$, but no improvement for the complete switch. The reason is that, at all SX stages, packets are either both on their preferred path, thus with same stage-distance to destination (the interconnect output), or are don't care because have been previously deflected. Note that the throughput improvement comes at a slight increase of the control complexity, but does not involve any extra hardware, and hence any extra power loss.

It is interesting to quantify the optical power loss. As the number of stages n increases, the loss per pass increases, but the average number of passes decreases. Fig. 2(b) shows the total loss experienced by a packet from source to destination vs. throughput in a 64-node network with beta elements without buffers (1S). The total loss is $L_T = H\,L_h$, where the loss per hop L_h is obtained by following the path from the input of the access node back to the same point:

$$L_h = l_{HR} + l_{AD} + l_p + l_{HR} + l_A + (2l_c + nl_\beta) + l_p \quad (dB)$$

where $l_{HR} = 1dB$ is the loss due to power tapping to read the packet header, $l_{AD} = 3dB$ is the loss at the add/drop switch, $l_p = 2.5dB$ is the loss in the fiber span from/to the central switch (corresponding to 10 km), $l_A = 10dB$ is the loss in the alignment stages, $l_c = 1dB$ is the coupling loss [8] between fiber and integrated directional couplers at the central switch, $l_\beta = 1dB$ is the waveguide excess loss in the 1S beta elements [8]. For simplicity, we assume that integration of the whole interconnect on a chip is possible. For SN the formula is $L_h = l_{HR} + l_A + (2l_e + 2l_c) + l_p(dB)$, assuming the same propagation span. The loss at the Benes interconnect is assumed to be $(2l_c + (2\log_2 M - 1)l_\beta)(dB)$.

The curves for the SX-CS networks in Fig. 2(b) have been drawn for both R and DP contention resolution rules. It is clear that decreasing the number of hops by increasing the number of stages causes a substantial reduction of the optical power loss. Using the DP rule gives a good gain with a few SX stages, when the deflection probability is large, but minimal power savings when the switch tends to the complete Omega interconnect. Again, SN has larger attenuation than the SX-CS with $n = 4$ stages. Considering a larger loss for alignment gives even more advantage to the centralized networks.

The strength of the centralized approach lies in the substantial decrease of the number of hops, which allows significant power savings when the per-hop attenuation does not strongly depend on the number of stages, as when for instance alignment and fiber propagation losses dominate.

The lowest power loss is achieved by the Benes network, although integration on a single substrate is not possible for large networks.

Analysis for complete interconnects

A simple analytical formula for the throughput is presented below for the special case of $\log_2 M$ stages. This sets an upper bound for the throughput of the incomplete SX-CS. The key observation is that for the complete interconnect, packets that get deflected at one stage become DC at the following stage. Hence a deflection makes packets temporarily "disappear" within the interconnect since they are routed as empty slots. Therefore we can use the simplified model of [11], [12] for the throughput analysis of open-loop unbuffered delta networks.

We introduce the simplifying approximation that in uniform traffic, at steady-state, the destinations of packets at the input of the interconnect are independent identically distributed (IID) uniform RVs taking values in $\mathcal{D} = \{0, 1, .., M\}$, 0 indicating an empty slot. At each beta element, using the routing table, we sort the M destinations into the appropriate classes E, CH, CL and DC. Let $p_0 \triangleq u$ be the probability of having a packet at an input slot. As noted previously, at the input of the first stage all packets are care, and by symmetry: $P\{CH\} = p_0/2$, $P\{CL\} = p_0/2$, $P\{DC\} = 0$. The probability that a CH packet ends up on H at each beta element of the first stage is thus

$$p_1 = 1 - (1 - p_0/2)^2 \tag{3}$$

since no CH packet ends on H only if no CH packet is present on the two input slots. The same holds for CL packets on L. Hence p_1 is the fraction of packets that get correctly routed on each output at the first stage. Therefore $1 - p_1$ is the fraction of deflected packets at stage 1, which become DC. At the input of the next stage we thus have $P\{CH\} = P\{CL\} = p_1/2$, $P\{DC\} = 1 - p_1$. The process can be repeated at each stage, so that at stage k the fraction of packets correctly routed to an output is

$$p_k = 1 - (1 - p_{k-1}/2)^2 \tag{4}$$

where we used the fact that in uniform traffic the two input slots at each beta element of the kth stage are independent [11].

Equation (4) provides a recursive formula giving the throughput $T_n \triangleq p_n$ of an n stage SX-CS network with $M = 2^n$ nodes, i.e., the fraction of packets correctly delivered at the output of the last stage of the interconnect:

$$\left\{ \begin{array}{lcl} T_0 & = & u \\ T_k & = & 1 - (1 - T_{k-1}/2)^2\,, \quad k = 1, .., \log_2 M \end{array} \right. \tag{5}$$

The accuracy of the formula is shown in Fig. 3.

4. Buffered beta elements

It is interesting to assess how much the throughput can improve by adding optical buffers within the beta elements, and how much the optical power loss decreases.

Fig. 4 shows a single-buffered beta element composed of 2 crossbars and a one-slot fiber-delay loop memory M,

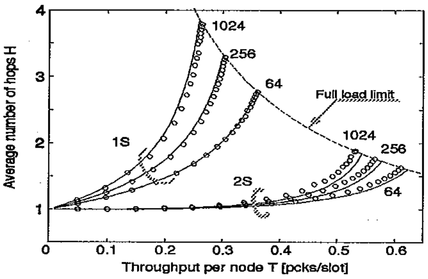


Figure 3: Theoretical curves (solid line) checked against simulation results (circles) for network size 64, 256, 1024 nodes.

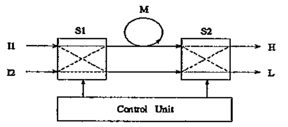


Figure 4: Beta element with single buffer (2S). M is a one-slot fiber delay loop. S1, S2 are crossbar switches [5].

which will be called 2S. This structure has already been studied in space-division meshed transparent optical networks [5], and is particularly attractive for optical implementation because it avoids the use of optical amplifiers within the loop.

The control algorithm for the two crossbars (detailed in [5]) tries to keep the buffer loaded with E/DC packets as often as possible, since deflections occur only when the buffer and the two input slots contain Care conflicting packets. The only inefficiency of the 2S element occurs when two care non-conflicting packets are at the input and the buffer is E/DC. In such a case, one care packet gets stored, thus increasing the deflection probability at the next clock.

A third crossbar to bypass the buffer is needed to avoid such inefficiency. However, this improved scheme is less attractive for an optical implementation since an optical amplifier would be needed in the fiber delay loop, and gives a negligible performance improvement.

Fig. 2 (c) shows simulated delay/throughput curves for $M = 64$, 2S beta elements, with both random (R) and distance priority (DP) conflict resolution rule. The NB curve is shown as a reference. It is seen that the addition of one buffer gives essentially the same performance as the NB when the SX-CS is complete. Indeed, for a limited range, a throughput larger than that of NB can be obtained, since the buffered beta elements help resolve both internal and destination conflicts, while the rearrangeable non-blocking interconnect does not have internal blocking, but cannot resolve any destination conflict. By using DP, an improvement is achieved for $n < \log_2 M$, but this is only marginal since the deflection probability is already low because of the presence of the buffers [10].

The increase of the number of crossbars in the beta elements implies an increase of the power loss. When alignment and propagation loss dominate, the larger beta losses are offset by the loss decrease due to the decrease of the number of hops. Fig. 2(d) shows the total average loss, which is lower with respect to the 1S case.

Analysis for complete interconnects

It is possible to extend the approximate model to complete interconnects with buffered beta elements and get closed-form expressions for throughput and average number of hops. A third approximation is introduced, namely, the content of the buffer is assumed independent of the input slots at each beta element. This implies independence of consecutive slots from the same input, a condition that is clearly violated because of the time-correlation introduced by the buffer at the previous stage. However, for a single buffer, such correlation is not strong and the approximation, as we will see, is satisfactory.

Control table and memory updates of the 2S elements are detailed in [5]. At the input of a beta element at stage k, let $P_c^{(k)}$ be the probability that an input slot is care (CH and CL having equal probability), and let $P_{cm}^{(k)}$ be the probability that the slot in the buffer is care (CH and CL being equal by the symmetry of the control). The probability that a CH packet ends up on output H, i.e., the fraction of packets correctly routed on each output of the beta element at the k-th stage, is

$$p_k = 1-\left(1-\frac{P_c^{(k)}}{2}\right)^2\left(1-\frac{P_{cm}^{(k)}}{2}\right)-\frac{P_c^{(k)2}}{4}(1-P_{cm}^{(k)}) \quad (6)$$

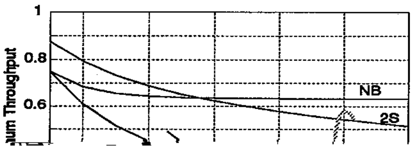
since a CH packet does not exit on H when i) none of the independent RVs I1,I2,M are CH, or ii) when M=E/DC (w.p. $(1-P_{cm}^{(k)})$) and I1,I2 are both care but not in conflict (w.p. $\frac{P_c^{(k)2}}{2}$) and w.p. 1/2 the CH one ends up in the buffer M and not on H. This last condition is due to the fact that in the 2S beta element the memory cannot be bypassed.

Now, $P_c^{(k)}$ is the probability of having a care packet at the input of the k-th stage, i.e., the probability that a packet was correctly routed at the $(k-1)$-st stage (remember that incorrectly routed packets become DC). Hence

$$P_c^{(k)} = p_{k-1}\,. \quad (7)$$

We now need an expression for the buffer content of the beta element at the k-th stage, $P_{cm}^{(k)}$. With the assumption that I1(t), I2(t) and M(t) are independent RVs at the same clock t, the buffer content M(t) can be modeled as a markov chain taking values E, CH, CL, DC, whose transition probabilities can be found from the control table. By the symmetry of the control and of the input-slot distribution, the states CH and CL have the same probability and can be grouped in a care (C) state. The steady state probability of this care state can be found as [5]

$$P_{cm}^{(k)} = \frac{p_{k-1}^2}{1-p_{k-1}+p_{k-1}^2} \quad (8)$$



because of the drastic reduction of the number of hops.

Disadvantages of such a network are:

i) packets can be delivered in scrambled ordered since the routing is dynamic;

ii) the slotted approach needs a global synchronization