Broadband Packet Switches Based on Dilated Interconnection Networks

Tony T. Lee, *Senior Member, IEEE,* and Soung C. Liew, *Senior Member, IEEE*

***Abstract*— A theoretical foundation for evaluation and comparison of a very broad spectrum of fast packet-switching techniques is developed in this paper. Based on this framework, we investigate the complexity of various packet switch designs, and demonstrate the advantage of dilation as a switch-design technique. Packet switches are classified either as loss systems or waiting systems, according to whether packets losing contention are dropped or queued. In a loss system, the packet loss probability can be made arbitrary small by providing enough paths between inputs and outputs. We focus on the question: How does the switch complexity grow as a function of switch size for a given loss probability requirement? A uniform approach to this problem is developed here. We show that for an $N \times N$ switch, the required number of switch elements for both the parallel-banyan network and the tandem-banyan network is of order $N(\log N)^2$, whereas the complexity of a dilated-banyan network is of order $N \log N(\log\log N)$. Within the class of waiting systems, we show that the parallel banyan networks in a Batcher-parallel-banyan network can be replaced by a dilated-banyan network without sacrificing the nonblocking property. Thus, as with parallelization, dilation can also be used to increase the throughput of a waiting system. In addition, we also explore the application of dilation in a large modular switch design which is realized by an interconnection structure consisting of Batcher-dilated-banyan networks and statistical multiplexers.**

I. Introduction

ASYNCHRONOUS Transfer Mode (ATM) has been widely accepted as a basis for packet transmission in future broadband communication networks [1]. Besides packet transmission systems, high-speed packet switches are essential elements in high-performance integrated communication networks for providing multimedia services. Various packet switches proposed in the literature are based on interconnection networks, originally intended for multiprocessor interconnects in highly parallel computer systems [2,3,4]. These switches make use of interconnection of many small switch elements in their overall architectures. An attractive feature of these switches is their regular topological interconnection pattern, which can be easily implemented by VLSI technology.

This paper develops a theoretical foundation for evaluation and comparison of a very broad spectrum of fast packet-switching techniques within the framework of performance and complexity studies. The goal is to provide insight into the design of very large switches. Within this framework, we investigate the complexity of various packet-switch designs proposed to date and demonstrate the advantage of dilation as a design technique in reducing switch complexity. The previous investigations of dilated networks [2,3] have not included an order-of-complexity study, and we believe comparison switches according to their fundamental complexities are essential to a complete understanding. Some switch design issues are clarified in the following to put things in the proper context.

Packet contention is one of the fundamental problems that must be overcome in designing packet switches. For switches based on interconnection of small switch elements, we must be concerned with two kinds of contention: output-port conflicts for the overall switch due to multiple input packets destined for the same output address, and internal collisions due to packets simultaneously routed to the same outgoing link at an individual switch element. Obviously, both contention problems can be solved completely by methods that allow all packets to reach their desired destinations. This could be achieved in an $N \times N$ switch by speeding up the switch operation by N times or providing N direct paths from each input to all outputs, where N is the number of input or output ports. In either case, even if all packets were to have the same destination, they could be switched within the same packet cycle. Buffering at the output port is still needed since the output trunk may not be able to handle all arriving packets at once.

It should be emphasized that the output buffers are not the reason why the contention problems are solved; rather, contention is eliminated by allowing up to N packets to reach the same destination. Buffering is required at each output port because of the limited transmission capacity of the output trunk, and each output port behaves essentially like a statistical multiplexer.

When the switch dimensions are large, switching mechanisms that let N packets reach the same destination increase design complexity and become impractical. As long as only fewer than N packets can reach the same output address, the contention problems remain, and they must be dealt with in other ways. There are only two alternative solutions: the switch can drop excess packets that cannot be switched, or it can buffer them for output access in the next time slot. Accordingly, switches based on interconnection networks can be classified either as loss systems or waiting systems, depending on how contention is resolved.

In all the loss systems under consideration here, packets are either dropped or switched to the outputs immedi-

Paper approved by M. S. Goodman, the Editor for Optical Switching of the IEEE Communications Society. Manuscript received July 13, 1991; revised January 22, 1992. This paper was presented in part at IEEE ICC '92, Chicago, June 1992.

The authors are with Department of Information Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong. This work was done while the authors were at Bellcore, New Jersey, U.S.A.

IEEE Log Number 9400960.

ately. By providing sufficient paths from inputs to outputs, packet-loss probability can be made as small as desired. The maximum number of packets that can be received by an output address in one time slot, or packet cycle, will be called the *group size*. Since switch complexity generally increases with group size, one would look for the minimum group size to meet the packet loss probability requirement in practice. In addition, if the group size is more than 1, packet buffers are needed at outputs, since the output port may not be able to transmit all arriving packets at once.

For waiting systems with the nonblocking property, such as the Batcher-banyan switch [5,6], there is no internal conflict, and excess packets are buffered at the input ports. Waiting systems that are blocking, such as the buffered-banyan switch [2,3], require internal buffers at the switch elements where conflict arises. Internal buffers bring about undesirable traffic-management complexity for communication networks that utilizes these switches at their nodes, since more queueing stages must then be considered in order to meet the desired grades of service for an end-to-end connection; that is, multiple queueing stages are associated with each node of the communication networks. Therefore, we will only focus on internally-unbuffered switches as far as waiting systems are concerned. In these systems, output conflicts are resolved before packets are allowed to enter the switch fabric [5,6,7]. Packets may be buffered only at inputs or at both inputs and outputs (in the case where the group size is more than one), and will not be lost except through buffer overflow. The throughput of a waiting system is limited by head-of-line (HOL) blocking, caused when packets waiting at the heads of input queues prevent subsequent packets from output access, even if the subsequent packets were destined for idle outputs [6,8,9]. This limitation can be relaxed by increasing the group size so that more HOL packets can access their destination outputs simultaneously [6,10,11,12,13,14] .

To summarize, we propose the following definitions to characterize precisely two classes of internally-unbuffered switching systems:

- Loss system
  - The switch fabric has no input or internal buffers; packets may be queued at outputs if group size is more than one.
  - Packets may be dropped internally or at outputs due to contention. The loss probability can be made arbitrarily small by adjusting the group size or some related switch design parameters.
- Waiting system
  - Output conflicts are resolved by some contention-resolution mechanism before packets are switched. Packets may be queued only at inputs or at both inputs and outputs, but not internally.
  - The throughput of the switch can be made arbitrarily close to 100size or other design parameters.

This paper focuses on various packet switches which make use of dilated-banyan networks [2,3] to improve performance. The basic idea of the dilated-banyan network is to expand the internal link bandwidth in order to reduce packet-loss probability in the case of loss systems, and to increase switch throughput in the case of waiting systems.

For comparison purposes, we also examine the Knockout switch [15], the parallel-banyan network, and the tandem-banyan network [16] as other known representatives of the class of loss systems. We focus on the question: *How does the switch complexity, in terms of the number of the fundamental switch elements required, grow as a function of switch size for a given loss probability requirement?* The order of complexity of the Knockout switch is known to be N^2. This paper shows that the order of complexity of both the parallel and tandem-banyan networks is $N(\log N)^2$, and that the order of complexity of the dilated-banyan network is $N \log N(\log \log N)$. To our best knowledge, the dilated-banyan network has the lowest order of complexity among all the loss systems proposed to date. Furthermore, because the factor $\log \log N$ grows very slowly with N, the dilated-banyan network is very close to meeting the $N \log N$ Shannon's lower bound on switch complexity [17]. As a side note, we have recently discovered a dual shuffle-exchange network [18] that meets the $N \log N$ bound. However, since our purpose in this paper is to classify various proposed fast packet-switching techniques according to a uniform measure of switch complexity, we will detail the $N \log N$ switch in a separate paper.

Perhaps the best-known waiting system is the Batcher-banyan switch, which consists of a Batcher sorting network followed by a banyan network [6,19,20,21]. To our knowledge, it has the lowest order of complexity, $N(\log N)^2$, among all known waiting systems with self-routing and nonblocking properties. Although the Benes switch has a lower order of complexity, $N \log N$, it is not self-routing in that an external routing algorithm is needed to set up the states of individual switch elements before packets are allowed to enter the switch fabric. The Batcher-banyan switch design is based on the fact that the banyan network is internally nonblocking if the input packets are sorted according to their destination addresses [6,7,10,19,20]. Using multiple banyan networks in parallel after the Batcher network [12], we can generalize the self-routing and nonblocking properties of the regular banyan network to allow multiple packets to access the same output address. We show in this paper that the parallel banyan networks can be replaced by a dilated-banyan network without sacrificing the nonblocking property, thus demonstrating that dilation can also be employed in a waiting system to improve switch throughput.

This paper is organized as follows. Section II provides some basic results on the throughput of multistage interconnection networks. Based on these results, Section III studies the complexity of various banyan-based loss systems for given loss probabilities, assuming uniform random traffic. We establish that among all loss systems proposed to date, the dilated-banyan network has the lowest complexity measure. Section IV proves that the dilated-banyan

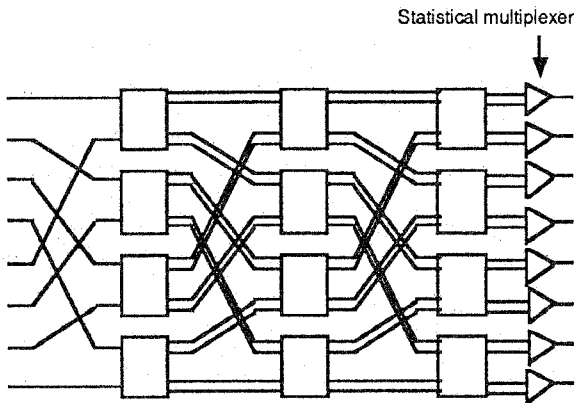

Fig. 1. An 8×8 banyan network with dilation degree 2.

network is nonblocking when the input packets are sorted and the number of packets destined for any output is no more than the dilation degree. In addition, the application of dilated networks in a large modular switch design is also addressed. Finally, we summarize our main results in Section V.

II. Preliminary Throughput and Packet-Loss Probability Equations

In order to establish the relationship between the complexity and capacity of various loss systems based on interconnection networks, throughput and loss probability as functions of switch design parameters are calculated below under the assumption of uniform random traffic. We will use "log" to indicate the logarithm to the base 2, and "ln", the logarithm to the base e. A regular $N \times N$ banyan network is constructed from 2×2 switch elements organized into $n = \log N$ stages, with each stage having $\frac{N}{2}$ switch elements. The switch elements of adjacent stages are interconnected in such a way that there is one and only one path from any input to any output in the overall network. Packets are routed through the network in a decentralized and distributed manner by performing the self-routing algorithm [6,7,10,19,20,22]. Packet loss may occur in an unbuffered network when two packets arriving at a switch element are destined for the same output link.

Intuitively, the loss probability can be reduced if the bandwidth of each link is increased so that more than one packet can be forwarded to each output address. The bandwidth d of an internal link (or the number of packets that can be forwarded to each output address simultaneously) is called the *dilation degree* of the network. An 8×8 banyan network with dilation degree $d = 2$ is shown in Fig. 1. The overall interconnection structure is the same as that in the regular banyan network, except that connected switch elements are linked by a multiplicity of d channels in the d-dilated-banyan network. Thus, the regular banyan network can be considered as a special case of the d- dilated banyan network with $d = 1$. The switch elements are themselves $2d \times 2d$ switches with two output addresses. Each output address has d associated output ports. Consequently, up to a maximum of d packets can be forwarded to the output address in any given time slot. If more than d packets are destined for the same output address, then d packets would be forwarded and the remaining packets dropped from the system. Thus, by making d large, we can achieve arbitrarily small packet loss probability. The drawback, of course, is that the switch becomes complex as d increases.

In the following, we review the exact loss probability calculation. Since this analysis does not yield much insight into the complexity measure of the d- dilated-banyan network, an approximate analysis will then be used to establish the switch complexity. Some results in this section were presented in [2] and [3] originally. Simplified derivations are provided here to make the paper self-contained and consistent.

Let $R_m(j)$ be the probability that j packets are forwarded to an output address of a switch element at stage m, where $0 \le j \le d$. Only d packets are forwarded when more than d packets are destined for the output address. Assuming that packets are equally likely to be targeted for any output (uniform random traffic), the probability of having i packets entering a switch element at stage $m+1$ is

$$S_{m+1}(i) = \sum_{k=0}^{i} R_m(k) R_m(i-k), \tag{1}$$

since $R_m(k)R_m(i-k)$ is the probability that there are k and $i-k$ packets on the upper and lower input-channel groups, respectively. The probability that j of these packets are destined for a particular output address is $\binom{i}{j}2^{-i}$. Thus,

$$R_{m+1}(j) = \begin{cases} \sum_{i=j}^{2d} S_{m+1}(i)\binom{i}{j}2^{-i} & \text{if } j < d \\ \sum_{i=d}^{2d} S_{m+1}(i) \sum_{k=d}^{2d} \binom{i}{k}2^{-i} & \text{if } j = d \end{cases} \tag{2}$$

With the initial condition

$$R_0(j) = \begin{cases} \lambda & \text{if } j = 1 \\ 0 & \text{if } j \neq 1 \end{cases}$$

where λ is the offered load, $R_n(j)$ can be found recursively. The packet-loss probability for the overall switch is simply

$$P_{loss} = 1 - \frac{\sum_{j=0}^{d} j R_n(j)}{\lambda} \tag{3}$$

The result above does not relate P_{loss} to n, d, and λ explicitly, and therefore, it is not amenable to the study of switch complexity. The approximate analysis below will be used in the next section to study how dilation degree d is related to n for given P_{loss} and λ.

There are two groups of d input ports for each switch element. In general, with $d > 1$, input ports of the same group are not independent of each other in the sense that finding a packet on one port is correlated with finding packets on other ports. For analytical tractability, however, we will make the simplifying assumption that the input ports are independent. Let P_m denote the probability that there is a packet at an input channel of a switch element at stage $m+1$. With the independence assumption, the probability

of finding a packet at an output channel of this switch element (or an input channel of a switch element in the $m+2$ stage) is

$$\begin{aligned} P_{m+1} &= \frac{1}{d}\left\{\sum_{k=1}^{d} k\binom{2d}{k}\left(\frac{P_m}{2}\right)^k\left(1-\frac{P_m}{2}\right)^{2d-k} + \right. \\ &\quad \left. d\sum_{k=d+1}^{2d}\binom{2d}{k}\left(\frac{P_m}{2}\right)^k\left(1-\frac{P_m}{2}\right)^{2d-k}\right\} \\ &= P_m - \frac{1}{d}\sum_{k=d+1}^{2d}(k-d)\times \\ &\quad \binom{2d}{k}\left(\frac{P_m}{2}\right)^k\left(1-\frac{P_m}{2}\right)^{2d-k} \end{aligned} \tag{4}$$

Suppose we treat m as a "continuous" variable and expand P_{m+1} as a Taylor series:

$$P_{m+1} = P_m + \frac{dP_m}{dm} + \frac{1}{2!}\frac{d^2P_m}{dm^2} + \cdots$$

If $\frac{d^nP_m}{dm^n}$ is small for $n \geq 2$, then (4) can be viewed as the recursion relation for the following differential equation

$$\frac{dP_m}{dm} = -\frac{1}{d}\sum_{k=d+1}^{2d}(k-d)\binom{2d}{k}\left(\frac{P_m}{2}\right)^k\left(1-\frac{P_m}{2}\right)^{2d-k}. \tag{5}$$

For $d = 1$, the above equation has an exact solution

$$P_m = \frac{1}{\frac{1}{4}m + \frac{1}{P_0}} = \frac{4\lambda}{m\lambda+4}, \tag{6}$$

where $P_0 = \lambda$ is the offered load on each input port. The overall packet-loss probability is then given by

$$P_{loss} = 1 - \frac{P_n}{\lambda} = \frac{n\lambda}{n\lambda+4}. \tag{7}$$

The above simple results, which have been shown in [3] to be excellent approximations, will be used in the next section to derive complexity bounds for parallel and tandem banyan networks. For $d > 1$, (5) can not be so easily solved. However, the complexity of the dilated-banyan network can be estimated from the packet-loss probability for a switch element at stage m defined by

$$\begin{aligned} Q_m &= 1 - \frac{P_m}{P_{m-1}} \\ &= \frac{1}{P_{m-1}d}\sum_{k=d+1}^{2d}(k-d)\binom{2d}{k}\left(\frac{P_{m-1}}{2}\right)^k \\ &\quad \left(1-\frac{P_{m-1}}{2}\right)^{2d-k}; \\ &\quad m > \log d. \end{aligned} \tag{8}$$

Note that $Q_m = 0$ for $m \leq \log d$ because there can be no more than $d/2$ packets at each input group of switch elements at stages 1 through $\log d$. That is, even under full loading conditions, when all input ports of the overall network have a packet, it takes at least $\log d + 1$ stages before contentions between packets may occur. Now, intuitively, the sequence of loading $\{P_m\}$ should be monotonically decreasing with respect to stage number m. This intuition can be easily verified by (5), which shows that the derivative of P_m with respect to m is always negative. Taking the derivative of (8), we can show that dQ_m/dm is also negative. Thus, we have

$$Q_1 \geq \cdots \geq Q_n. \tag{9}$$

The above monotonic sequence is easy to interpret; the loss probabilities become smaller and smaller as the loading becomes lighter and lighter. In the next section, the complexity of various switch architectures will be calculated based on the discussion in this section.

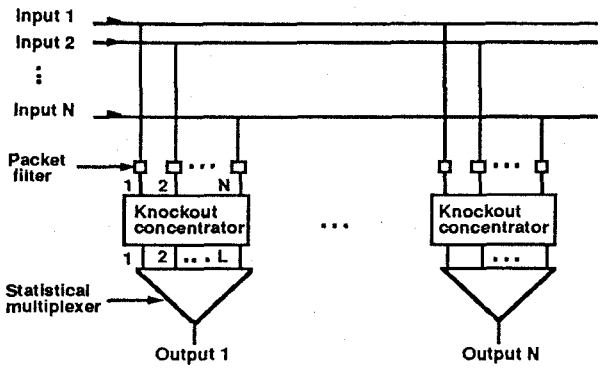

Fig. 2. The knockout switch.

III. Loss Systems

Contention in loss systems is resolved by dropping some of the packets in conflict and allowing the rest to reach their destined outputs without any queueing delay at the inputs or within the switch fabric. Perhaps the most straightforward way of implementing a loss system is with a crossbar switch, in which one and only one packet would be allowed to access a given output in one time slot. For a large crossbar switch, it can be easily shown that the maximum throughput is limited to 63.2% under uniform random traffic [8,9]. Furthermore, the packet-loss probability is likely to be very high, even if the load were far less than 63.2%. For the generic output-buffered switch [8,9], all arriving packets are allowed to access their outputs immediately without dropping any packets. Switch complexity will necessarily be very high in order to achieve zero loss probability this way.

The Knockout switch [15] (Fig. 2) attempts to strike a balance between the two extremes of allowing only one packet and allowing all packets with a common destination to access an output. It makes use of the fact that under uniform random traffic, it is statistically unlikely that more than a certain number packets, say K, will be destined for the same output simultaneously. Thus, by allowing

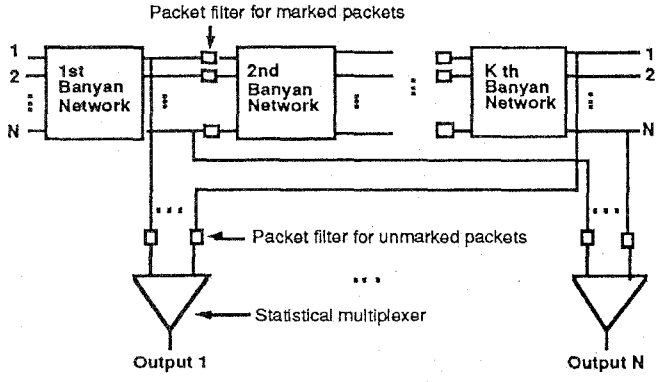

Fig. 3. The tandem-banyan switch.

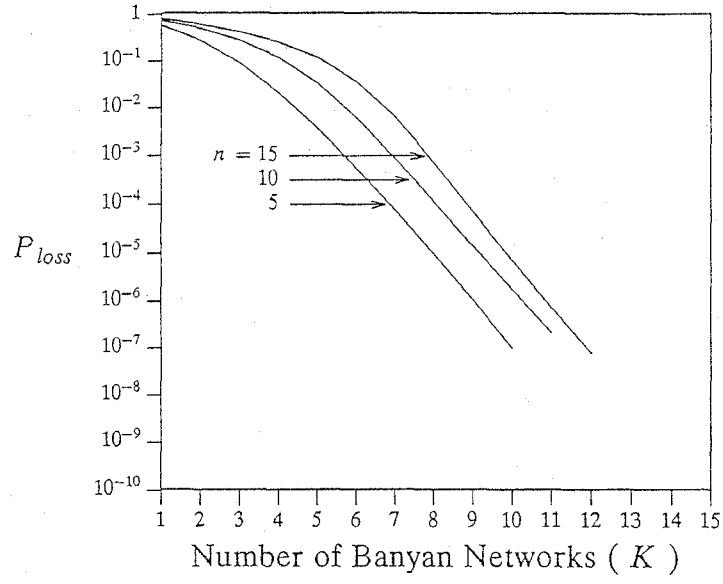

Fig. 4. Packet loss probability versus number of banyan networks K in tandem under full loading: simulation results.

up to a maximum of K arriving packets to access each output, the loss probability can be made very small. For instance, it has been shown that under full loading, $K = 8$ is sufficient to achieve a loss probability of 10^{-6}, regardless of the switch size [15].

The complexity of the Knockout switch architecture in [15] is of order KN^2, where N is the number of ports. In Appendix A, a derivation shows that K is upper-bounded by a quantity K^* which is independent of switch size N for a given loss-probability requirement P_{loss}, agreeing with the results in [15] that K approaches an asymptotic value rather than growing indefinitely as N increases. Since K is largely independent of N for a given loss probability requirement, the complexity as a function of N is of order N^2.

In the following subsections, we address the complexity of a different class of loss systems based on banyan networks. For these networks, contention resolution is distributed and performed internally and throughout the switch fabric. Consequently, packets could be dropped anywhere within the networks. Our focus deals with how switch complexity grows as a function of switch size for a given loss probability requirement.

A. Tandem-Banyan Network

The tandem-banyan switching fabric was originally proposed in [16]. The basic switch structure consists of K banyan networks connected in series (see Fig. 3). Except for the last banyan network, each output of a banyan network is connected to both an input of the subsequent banyan network and a concentrator (statistical multiplexer). With this set-up, a packet would be routed to the concentrator if it reaches the correct output, and to the subsequent banyan network otherwise. Thus, each packet can have up to K attempts to reach its destined output.

Deflection routing is employed within each banyan network; whenever there is a conflict at a 2×2 switch element, one packet would be routed correctly while the other would be marked and routed in the wrong direction. In order to optimize the number of correctly routed packets, the marked packet would have a lower priority than an unmarked one for the rest of its journey within the current banyan network. That is, it is not necessary to route a packet in any particular direction once it is marked because it will reach the wrong output anyway. At the output of this banyan network, a marked packet will be unmarked and forwarded to the next banyan network, and a new attempt to route the packet to its desired output is initiated. A packet is considered lost if it still fails to reach the desired output after passing through all the K banyan networks.

Figure 4 shows our simulation results for packet-loss probability, P_{loss}, versus number of banyan networks, K, under full load for various switch sizes. The curves clearly show that the number of banyan networks required to achieve loss probabilities below a certain threshold increases with the switch size N. To study the functional dependence of K on N for a given P_{loss}, Appendix B applies (7) in successive banyan networks and derives the following result:

For a tandem-banyan switch, assuming packets at successive banyan networks are uncorrelated with each other and that the inputs they occupy are uncorrelated with their destination outputs, the required number of banyan networks K is given by

$$K \approx \frac{\lambda \log N}{4}(1 - P_{loss}) - \ln P_{loss}, \qquad (10)$$

where λ is the offered load and P_{loss} is the required loss probability.

The implication of the above statement is that K is of order $\log N$, and the complexity of the overall tandem banyan switch is of order $N(\log N)^2$, the same as the complexity of the Batcher-banyan switch (see Section IV.B)! It should be noted that the assumption of noncorrelation of input packets is an optimistic one. This can be seen from Fig. 5, where we plot K as a function of $n = \log N$ from simulation results and from analytical results with the noncorrelation

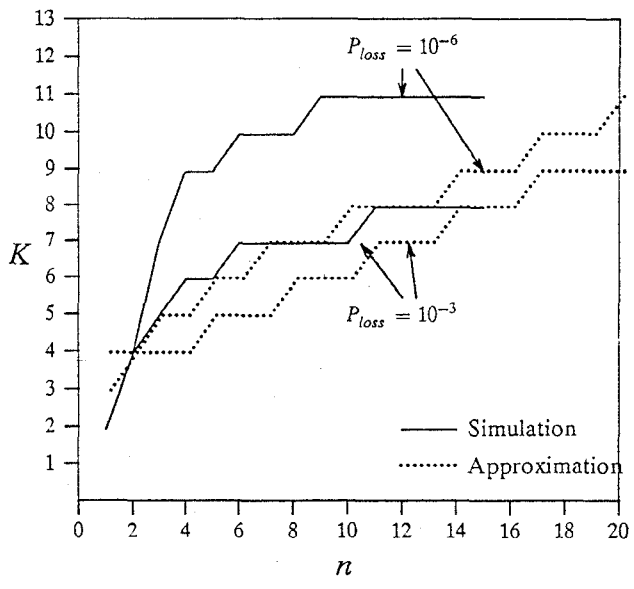

Fig. 5. Number of banyan networks K as a function of $n = \log N$ under full loading

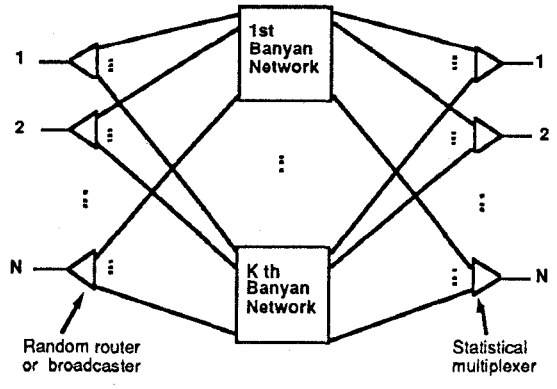

Fig. 6. The parallel-banyan switch.

assumption. Clearly, the simulation results have a higher K value for a given n than the analytical results. Taking this into account, we find that even under the optimistic assumption, the complexity of the tandem-banyan network for a given loss probability is of order $N(\log N)^2$.

B. Parallel-Banyan Network

We next consider a switch with K parallel banyan networks, as shown in Fig. 6. Suppose that an incoming packet is routed randomly to one of the K networks. Then, the load to each banyan network is reduced by a factor of K, giving rise to a correspondingly lower P_{loss}. Using equation (7), with the load set to $\frac{\lambda}{K}$, we find that

$$P_{loss} = \frac{n\lambda}{n\lambda + 4K}. \tag{11}$$

This yields the following result:

For a parallel-banyan switch with random routing, the number of banyan networks required to achieve a certain loss probability P_{loss} is

$$K = \frac{(P_{loss}^{-1} - 1)n\lambda}{4} \approx \frac{\lambda}{4}\frac{\log N}{P_{loss}} \tag{12}$$

for small P_{loss}.

An alternative routing scheme is to broadcast an incoming packet to all the K parallel banyan networks, and use filters at the outputs to remove redundant packets. With broadcast routing, the load to each banyan network is still λ, but a packet is lost only if all its replicas fail to reach its destination output. For this strategy to work properly, we must adopt a random contention-resolution scheme in each of the parallel banyan networks so that when two packets attempt to access the same output of a 2×2 switch element, the winning packet will be chosen at random. Otherwise, with a fixed contention-resolution scheme (e.g., always choose the packet from the upper input port), packets that are dropped in one banyan network will also be dropped in other banyan networks. Even with the random contention-resolution scheme, the event of a packet being dropped in one banyan network is not independent of the events of its replicas being dropped in the other banyan networks, because all banyan networks have the same set of input packets. For simplicity, if we further make the assumption that the contention-resolution processes in different banyan networks are independent, then

$$P_{loss} = \left(\frac{n\lambda}{n\lambda + 4}\right)^K \tag{13}$$

From this, we get the following result:

For a parallel-banyan switch with broadcast routing, the number of banyan networks required to achieve a certain loss probability P_{loss} is

$$K = \frac{\log P_{loss}}{\log\left(1 - \frac{4}{n\lambda+4}\right)} \approx \frac{(4 + \lambda \log N)}{4}(-\ln P_{loss}). \tag{14}$$

The number of parallel banyan networks needed for broadcast routing is less than the number needed for random routing, and it is close to the upper bound obtained for the tandem-banyan structure. In any case, as with the tandem-banyan switch, the complexity of the parallel-banyan switch is of order $N(\log N)^2$ with either routing scheme. By comparing (10) and (14), we also notice that the parallel-banyan network is uniformly worse than the tandem-banyan network.

C. Dilated-Banyan Network

We now investigate the performance and complexity of the dilated-banyan network. Figure 7 plots P_{loss} versus d for dilation networks of various dimensions, based on the exact analytical calculation given in Section II and assuming 100% offered load. Compared with the results of the tandem-banyan network, P_{loss} is smaller in the dilated-banyan network if $d = K$. We also see that the dilation

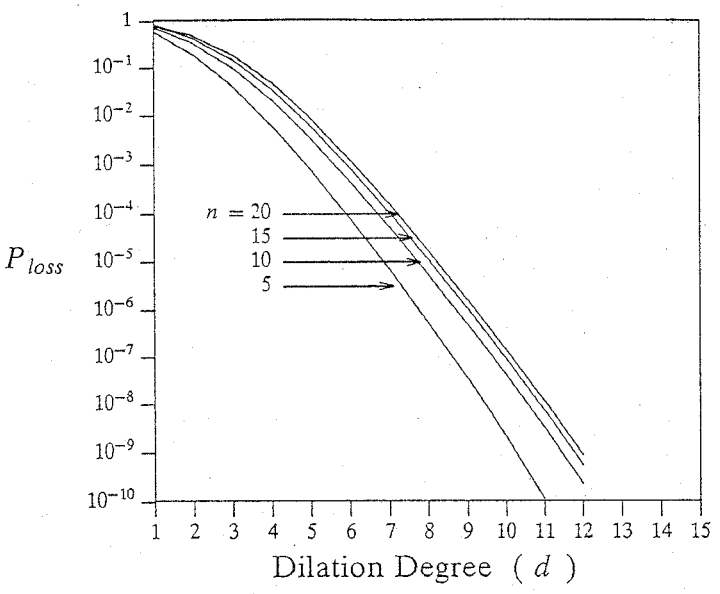

Fig. 7. Packet loss probability versus dilation degree based on Eqn. (3) under full loading

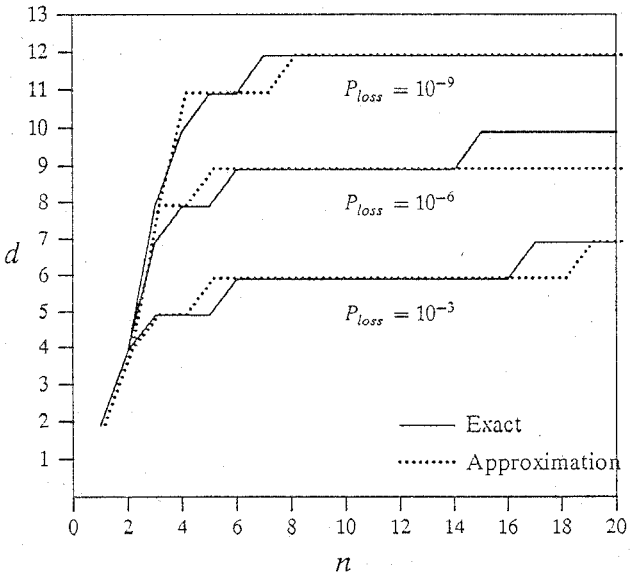

Fig. 8. Dilation Degree versus $n = \log N$ for various fixed loss probabilities under full loading

degree d required for a fixed P_{loss} is not a strong function of n. In fact, the curves indicate that d increases much less than linearly with $n = \log N$ for a fixed P_{loss}. This can be seen more clearly from Fig. 8, where we plot d versus P_{loss} (the solid lines) for various values of required P_{loss} under full loading.

It is interesting to compare the performance of the the dilated-banyan network with that of the Knockout switch [15]. Intuitively, the P_{loss} of the Knockout switch would be smaller than that of the dilated-banyan network if $K = d$. This is because no packet would be lost in the Knockout switch if K or fewer packets were destined for any output address, whereas the possibility of internal conflicts in the dilated-banyan network implies that packets could be lost even if no more than $d = K$ packets were destined for any external output address. Setting $K = d$, however, results in an unfair comparison because the corresponding complexity of the dilated network is lower than that of the Knockout switch. Figure 8 shows that for $n = 10$ ($N = 1024$) and full loading, $d = 9$ for $P_{loss} = 10^{-6}$ and $d = 12$ for $P_{loss} = 10^{-9}$. This compares with $K = 8$ and $K = 11$ in the Knockout switch [15] for the same P_{loss}, respectively. Thus, we see that d in the dilated-banyan network does not need to be much higher than K in the Knockout switch to achieve the same loss probability.

We now examine the complexity of the dilated-banyan network. In Appendix C, The following result is established under the assumption of uniform random traffic:

For a dilated-banyan switch, assuming that the packets on the input ports of each $2d \times 2d$ switch element are independent,

$$\frac{(\log N - \log d)}{4}\left(\frac{\lambda}{d}\right)^d\left(1-\frac{\lambda}{4d}\right)^{d-1} \leq P_{loss} \leq \frac{(\log N - \log d)}{4}\left(\frac{\lambda}{d}\right)^d\left(1-\frac{\lambda}{4d}\right)^{d-1}\frac{2^d}{\sqrt{\pi d}}, \quad (15)$$

where d is the dilation degree required to meet loss probability P_{loss}.

The dotted lines in Fig. 8 correspond to an approximation based on the upper bound above. The fact that the upper-bound approximation is actually lower than the exact analysis at certain portions of the curves indicates that the independent-input-ports assumption in the approximation is optimistic with respect to the actual situation. As can be seen, however, the d values in the approximation and exact calculation do not differ by more than 1. Appendix C also shows that

$$d\log d = \log\log N - \log P_{loss} + O(d), \quad (16)$$

where $O(d)$ is a function of order d. This implies that for a fixed $Psubloss$ requirement, $d\log d$ is $O(\log\log N)$. Now, as shown in Fig. 9, the $2d \times 2d$ switch elements in the d- dilated network can be implemented by $2d$ 1×2 switch components and two $2d \times d$ concentrators, which in turn can be realized by a running-adder address generator and a reversed-banyan network [12,20,22]. With this design, the order of complexity of each switch element is $O(d\log d)$. Since there are altogether $(N/2)\log N$ switch elements, the order of complexity for the overall switch fabric is $O(N\log N(\log\log N))$. In other words, for a fixed P_{loss}, the dilated-banyan network has a lower order of complexity than the tandem-banyan network and the parallel-banyan network.

Now, the complexity study outlined above is relevant to implementation only if we consider switches with very large dimensions (i.e., large N and d). In practice, if we restrict ourselves to small d, then, in order to make P_{loss} small,

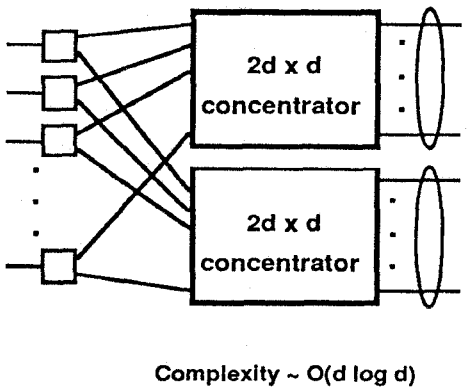

Fig. 9. An implementation of $2d \times 2d$ switch element with order of complexity $d \log d$.

we can cascade several dilated networks in tandem, generalizing the idea of the tandem-banyan network. Upon examining (15), we see that the upper and lower bounds differ by a factor of $\frac{2^d}{\sqrt{\pi d}}$, which is independent of the offered load λ. In fact, roughly speaking, given a fixed d and N, P_{loss} is proportional to the dth power of the offered load λ. This means that reducing the offered load would improve the packet-loss rate more noticeably in a switch with a larger d than one with a smaller d. In a tandem-banyan switch, the offered load is successively reduced in subsequent banyan networks. Replacing the regular banyan networks with the dilated-banyan network will ensure that the load decreases much more quickly and that the required P_{loss} can be achieved with fewer networks.

Before moving onto the next section, it is worth comparing our results with the conclusion of Reference [3], which states: "Although dilated networks provide asymptotically better performance, for practical numbers of processors dilated and replicated (i.e., parallel) networks have similar performance". This assertion has been substantiated in [3] by graphical plots of loss curves for switch size up to $N = 2^{40}$ for dilated and parallel networks of "similar complexity". Closer examination reveals that the above conclusion is based on an unfair comparison of dilated and parallel networks. Specifically, to arrive at the alleged similar-complexity common denominator, Reference [3] assumes the parallel network to consist of generalized banyan networks in which the dimensions of the switch elements are $2d \times 2d$ rather than 2×2. However, as we have already mentioned, the $2d \times 2d$ switch elements in the dilated network, unlike those is the generalized banyan network, has only 2 rather than $2d$ output addresses. Consequently, the $2d \times 2d$ switch elements in the dilated network is less complex than those in the parallel network. This fact has been ignored in Reference [3]. Otherwise, the dilated network would have been shown to be superior to the parallel network, even for moderate switch size.

IV. Waiting Systems

We now turn our attention to the Batcher-banyan switch and its variants as representatives of waiting systems in which switching is accomplished by a combination of sorting and nonblocking routing processes. A contention-resolution mechanism is required to resolve output conflicts before selected packets are allowed to enter the routing stage. The throughput is limited to 58.6% if only one packet can be switched to an output port in one time slot [9]. To improve the throughput, more than one banyan network can be arranged in parallel so that multiple packets can be switched to the same output simultaneously. This enhanced scheme was employed in the Sunshine switch [10]. The group size in this case is equal to the number of parallel banyan networks. There are a number of other architectures that exploit the same concept to increase the throughput [7,12,13,14]. It has been shown in [11] that the overall throughput can be increased to 95% for a group size greater than or equal to 3. In the next subsection, we will show that the set of parallel banyan networks that follows the Batcher network can be replaced by a single dilated-banyan network. That is, when the output space is expanded through dilated internal links, the nonblocking property is still preserved; therefore, the Batcher-parallel-banyan network and Batcher-dilated-banyan network are functionally equivalent.

Another technical challenge associated with the Batcher-banyan switch is scalability. If we attempt to scale up the switch size by interconnecting multiple Batcher-banyan switches in stages, then the overall system is no longer self-routing and nonblocking even though each switch module is. Depending on the actual interconnection structure used, we may face a number of undesirable system problems, such as the need for centrally controlled load-balancing and path-hunting algorithms during call setup. A modular Batcher-banyan switch has been proposed in [7] to avoid these problems. Each switch module constitutes a Batcher network and a set of routing subnetworks, including parallel binary trees and banyan networks. The integration of these routing subnetworks into an equivalent dilated network is also addressed below.

A. Nonblocking Condition of Dilated-banyan Networks

An interconnection network is nonblocking if packet collisions can be completely avoided and if internal buffers are not needed. It is well-known that the banyan network is nonblocking if incoming packets are ordered according to their destination addresses and that there are no output conflicts. The Starlite switch [20], a combination of the Batcher sorting network and banyan routing network, is based on this principle. Intuitively, the nonblocking condition for the dilated-banyan network should be much more relaxed than for the regular banyan network, because of the substantial reduction in the probability of packet collisions at each switch element. Formally, a d- dilated-banyan network is nonblocking if the active inputs (inputs with arriving packets) $x_1, \ldots, x_k$ and the corresponding outputs $y_1, \ldots, y_k$ satisfy the following.

1. (Monotone): $y_1 \leq y_2 \leq \cdots \leq y_k$ or $y_1 \geq y_2 \geq \cdots \geq y_k$.

2. (Dilation): No more than d packets have the same destination address.
3. (Concentration): Any input between two active inputs is also active. That is, $x_i \le w \le x_j$ implies input w is active.

The above is certainly a straightforward generalization of the nonblocking condition for a regular banyan network, and a proof is given below.

Suppose that internal blocking occurs because there are more than d packets destined for the same local output of a $2d \times 2d$ switch element at stage k in an $N \times N$ dilated network. Consider $d+1$ of these packets and denote their external input addresses by $x_1 < x_2 < \cdots < x_{d+1}$ and their corresponding external output addresses by $y_1 \le y_2 \le \cdots \le y_{d+1}$. It has been shown in the Appendix of [22] that

$$(x_j - x_i) \ge 2^{n-k};$$

and

$$(y_j - y_i) \le 2^{n-k} - 1,$$

for any two of the $d+1$ packets, i and j with $j \ge i$. It follows that

$$\begin{aligned}(x_{d+1} - x_1) &= (x_{d+1} - x_d) + (x_d - x_{d-1}) + \cdots + (x_2 - x_1) \\ &\ge d\,2^{n-k}.\end{aligned}$$

Therefore,

$$(y_{d+1} - y_1) \le 2^{n-k} - 1 \le \frac{(x_{d+1} - x_1)}{d} - 1. \qquad (17)$$

Now, consider the input ports between the two input ports occupied by packets $d+1$ and 1, inclusively. By the concentration condition above, each of these input ports is active and has a packet. Therefore, there are altogether $(x_{d+1} - x_1) + 1$ packets among these input ports. Each packet has an associated output destination, and it is easy to see that

$$\begin{aligned}&\textit{the minimum number of distinct destination addresses among the packets} \\ &\ge \frac{(x_{d+1} - x_1) + 1}{d},\end{aligned}$$

since there can be at most d packets with the same destination by assumption. In addition, by the monotone condition,

$$\begin{aligned}&\textit{the maximum number of distinct destination addresses among the packets} \\ &\le (y_{d+1} - y_1) + 1.\end{aligned}$$

Therefore,

$$(y_{d+1} - y_1) + 1 \ge \frac{(x_{d+1} - x_1) + 1}{d}. \qquad (18)$$

It is obvious that (17) and (18) can not hold simultaneously. Thus, the dilated network must be nonblocking under the conditions listed above.

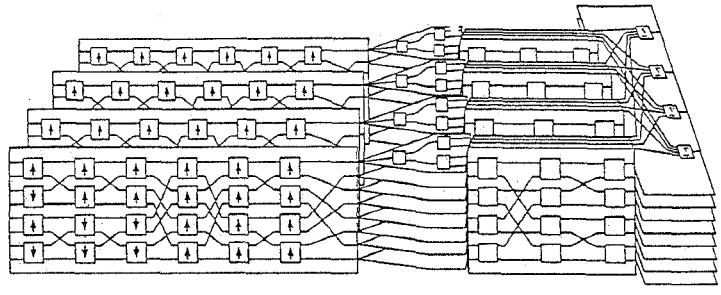

Fig. 10. A 32 x 32 modular Batcher-banyan switch with 4 modules

B. Modular, Dilated Batcher-banyan Networks

A switching node in the public broadband network may require access to as many as 10000 high-speed ports. The sorting performed by the Batcher network requires bit synchronization of all input packets in every time slot. This stringent timing requirement limits the size of a Batcher-banyan network. A modular approach, based on divide-and-conquer, has been proposed in [7] to scale up the switch size while preserving its nonblocking and self-routing properties.

An $N \times N$ modular Batcher-banyan switch consists of K switch modules. Each module is an $M \times N$ packet switch, where $M = N/K$ is called the *base dimension*. A 32×32 switch with 4 modules is shown in Fig. 10. The basic idea of this modular approach is to partition the set of inputs into K subsets. Each subset is sorted by a Batcher network. The sorted sets are then routed to their destinations by an expansion network, which is a combination of M binary trees (each of size $1 \times K$) and K banyan networks (each of size $M \times M$). The $M \times N$ expansion network also performs the self-routing algorithm and it is nonblocking under the same condition as a regular banyan network.

The complexity of a modular Batcher-banyan switch, counting the total number of sorting and routing elements, is by no means optimal. The number of all switch elements of an $N \times N$ switch with K modules is

$$\begin{aligned}&K\left(\frac{M}{4}\log M(1+\log M)\right) + \\ &K\left(M(K-1) + \frac{KM}{2}\log M\right) \\ &= \frac{N}{4}(\log N - \log K)(1 + \log N - \log K) + \\ &\quad N(K-1) + \frac{NK}{2}(\log N - \log K), \qquad (19)\end{aligned}$$

which is monotonically increasing with respect to the number of modules K. The extra cost, however, is compensated by many other advantages that we may gain from this modular approach, such as improvements in the reliability, maintainability and performance [7].

In each switch module, the M binary trees and K banyan networks can be integrated and replaced by an expansion network with dilation degree $d = K$. This point is illus-

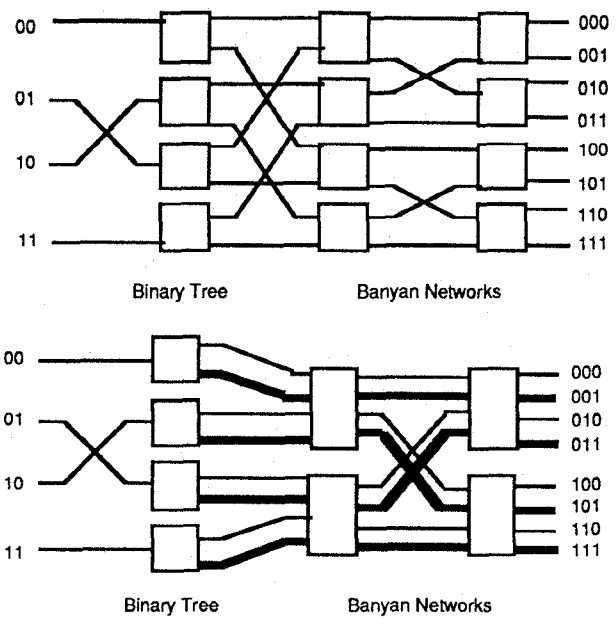

Fig. 11. A 4×8 expansion network and its equivalent dilated expansion network

trated by an example shown in Fig. 11. The 4×8 expansion network, consisting of 4 binary trees and 2 banyan networks, is equivalent to a 4×4 network with dilation 2. It should be noted that the logical paths in the original banyan networks should keep their independence in the equivalent expansion network in order to route the packets to the correct output ports. As illustrated in this example, the thin-line paths and bold-face paths should not be mixed. If $K = d$ is too large to be implemented, then an $M \times N$ expansion network can be replaced by $\frac{K}{d}$ expansion networks, each with dilation degree d. The group size of each output port can also be enlarged in a straightforward manner in a dilated expansion network to improve the overall throughput of the switch further.

V. Conclusion

We have developed a theoretical foundation for evaluation and comparison of a broad spectrum of fast packet switch architectures within the framework of performance and complexity studies. Based on this framework, we have investigated the throughput and complexity of various packet-switching techniques proposed to date, with an emphasis on comparison with designs based on the dilated-banyan network. Specifically, the relationships between dilation degree, internal-link bandwidth, complexity, and switch throughput are established.

The switches under consideration have been classified either as loss systems or waiting systems according to how packet contention is handled. The complexities of various loss systems based on banyan networks have been estimated for given throughput and loss-probability requirements. Our main discovery is that the complexity of dilated-banyan networks is of order $N \log N(\log \log N)$, while the complexity of all the other proposed fast-packet switches, including parallel banyan networks, tandem banyan networks and Batcher-banyan switches, is of order $N(\log N)^2$; the only loss system that has a lower order of complexity is the $N \log N$ dual shuffle-exchange network [18] that we have discovered recently, and this will be reported elsewhere. In addition, we have also established the nonblocking condition of dilated-banyan networks. This result implies that Batcher-dilated-banyan switches, operated as waiting systems, can be constructed to meet any throughput requirement.

On the whole, our work suggests that dilation is a powerful design technique for improving performance and reducing complexity in a large switch. We have argued this from the complexity as well as throughput viewpoints. A far as implementation technology is concerned, since dilation involves multiple, parallel links from one location to another, multiplexing techniques (e.g., wavelength-division multiplexing in the optical domain) is a natural way for reducing the interconnection complexity. This further motivates the use of dilation as a path-diversification technique as opposed to other techniques which do not make use of parallel-running links.

As a final note, this paper concerns dilation at the microscopic level for individual 2×2 switch elements. The use of dilation for interconnecting nonblocking switch modules of larger dimensions to construct a very large overall switching network has been treated in [12]. Since our results are related to switches of very large dimensions (i.e., larger than can be realized with a single VLSI chip), they suggest that dilation should be applied at both microscopic and macroscopic levels in order to make best use of its power. Thus, the nonblocking switch modules in a large switch architecture can be replaced by blocking but dilated switch modules. To reduce the complexity of interconnecting these switch modules, multiplexing techniques can be applied to the parallel links connecting them.

Acknowledgement

Discussion with Ed Arthurs during the initial stage of this work is very much appreciated. We are also grateful to Howard Lemberg, Stu Personick, Paul Shumate, and Mario Vecchi for their comments and suggestions which have improved this paper substantially. We are indebted to one of the anonymous reviewers for pointing out the relative importance of the results putting them into proper perspectives.

Appendices

The derivations of (10), (15), and (16) are given here. The following relations will be useful in our derivations.

$$\begin{aligned} \ln(1+z) &= z - \frac{z^2}{2} + \frac{z^3}{3} + \cdots; \quad |z| < 1, \\ \log(1+z) &\approx (\log e)z \quad |z| \ll 1. \end{aligned} \tag{20}$$

$$n! = n^n e^{-n}\sqrt{2\pi n}(1+\epsilon(n)), \quad \text{(Stirling's Formula)}, \quad (21)$$

where $\epsilon(n) > 0$ is a decreasing function of n.

$$\binom{N-K}{i} \le \binom{N}{K+i} \le \binom{N-K}{i}\binom{N}{K}; \quad 0 \le i \le N-K. \quad (22)$$

Appendix A
Complexity of Knockout Switch

Suppose that P_{loss} is the loss probability of the Knockout switch with group size K. Then

$$\begin{aligned} P_{loss} &= \frac{1}{\lambda}\sum_{i=K+1}^{N}(i-K)\binom{N}{i}\left(\frac{\lambda}{N}\right)^i\left(1-\frac{\lambda}{N}\right)^{N-i} \\ &= \frac{1}{\lambda}\sum_{i'=0}^{N-K} i'\binom{N}{K+i'}\left(\frac{\lambda}{N}\right)^{K+i'}\left(1-\frac{\lambda}{N}\right)^{N-K-i'} \\ &= \frac{1}{\lambda}\left(\frac{\lambda}{N}\right)^K \times \\ &\quad \sum_{i'=0}^{N-K} i'\binom{N}{K+i'}\left(\frac{\lambda}{N}\right)^{i'}\left(1-\frac{\lambda}{N}\right)^{N-K-i'} \quad (A.1) \end{aligned}$$

It follows from (22) that

$$\begin{aligned} P_{loss} &\le \frac{1}{\lambda}\left(\frac{\lambda}{N}\right)^K\binom{N}{K}(N-K)\frac{\lambda}{N} = \left(\frac{\lambda}{N}\right)^K\binom{N-1}{K} \\ &\le \frac{\lambda^K}{K!} \le \frac{\lambda^K}{K^K e^{-K}\sqrt{2\pi K}} \end{aligned}$$

Taking the logarithm of the above inequality, we have

$$\begin{aligned} f(K) &= K(\log K - \log\lambda - \log e) + \\ &\quad \frac{1}{2}\log K + \frac{1}{2}\log 2\pi \\ &\le -\log P_{loss}. \quad (A.2) \end{aligned}$$

Since f is an increasing function with respect to K, the above inequality implies that $K \le K^*$, where the upper bound K^* is the root of $f(K) = -\log P_{loss}$. For example, $K^* = 9.7$ when $\lambda = 1$ and $P_{loss} = 10^{-6}$. This is indeed an upper bound for $K = 8$ needed to achieve $P_{loss} = 10^{-6}$ [15].

Appendix B
Complexity of Tandem-Banyan Switch

Let $L_k = \frac{\lambda_k}{\lambda}$ be the probability that a packet still fails to reach its destination after traveling through k banyan networks, where λ is the initial offered load and λ_k is the load per link offer to the input of $(k+1)^{th}$ banyan network. Let ρ_k be the carried load on each output of the the k^{th} banyan network. It follows from (6) that

$$\begin{aligned} \lambda_{k+1} &= \lambda_k - \rho_{k+1} = \lambda_k - \frac{4\lambda_k}{n\lambda_k + 4} \\ &= \frac{n{\lambda_k}^2}{n\lambda_k + 4}. \quad (B.1) \end{aligned}$$

We then immediately have the recursive formula of the loss probability

$$L_{k+1} = \frac{aL_k^2}{aL_k + 4}, \quad (B.2)$$

where $a = \lambda n = \lambda \log N$. From Eqn. (B.2), we get

$$L_{k+1} - L_k = \frac{-4L_k}{aL_k + 4}. \quad (B.3)$$

Using the Taylor Series approximation technique introduced in Section II, we can transform the above difference equation into the following differential equation:

$$\frac{dL_k}{dk} = \frac{-4L_k}{aL_k + 4}.$$

Simple integration and matching of the boundary conditions, $L_0 = 1$ and $L_K = P_{loss}$, gives

$$K = \frac{\lambda \log N}{4}(1 - P_{loss}) - \ln P_{loss}. \quad (B.4)$$

Appendix C
Complexity of Dilated Banyan Switch

We know from Section II that both the sequence of loading $\{P_m\}$ and the sequence of loss probabilities $\{Q_m\}$ at intermediate stages are monotonically decreasing. Also, we know that if the dilation degree is d then $Q_m = 0$ for $m \le \log d$. By the definition of loss probability, we have

$$\begin{aligned} P_{loss} &= \frac{P_0 - P_n}{P_0} \\ &= \frac{(P_0 - P_1) + (P_1 - P_2) + \cdots + (P_{n-1} - P_n)}{P_0} \\ &= Q_1 + Q_2\frac{P_1}{P_0} + \cdots + Q_{\log d}\frac{P_{\log d - 1}}{P_0} + \\ &\quad Q_{\log d+1}\frac{P_{\log d}}{P_0} + \cdots + Q_n\frac{P_{n-1}}{P_0} \\ &\le (n - \log d)\, Q_{\log d+1}. \quad (C.1) \end{aligned}$$

The above is basically a union bound (i.e., the overall loss probability is upper-bounded by the sum of loss probabilities at different stages), which is likely to be very good when P_{loss} (and therefore Q_m) is small. To find a bound for $Q_{\log d+1}$, substituting $P_{\log d} = \lambda/d$ into (8) gives

$$\begin{aligned} &Q_{\log d+1} \\ &= \frac{1}{\lambda}\sum_{j=d+1}^{2d}(j-d)\binom{2d}{j}\left(\frac{\lambda}{2d}\right)^j\left(1-\frac{\lambda}{2d}\right)^{2d-j} \\ &= \frac{(\lambda/2d)^d}{\lambda}\sum_{i=1}^{d} i\binom{2d}{i+d}\left(\frac{\lambda}{2d}\right)^i\left(1-\frac{\lambda}{2d}\right)^{d-i} \quad (C.2) \end{aligned}$$

where we have made the index change $i = j - d$. Now,

$$\binom{2d}{i+d} = \binom{2d}{d}\binom{d}{i} \times \frac{d!\,i!}{(i+d)!} \le \binom{2d}{d}\binom{d}{i}2^{-i}.$$

Substituting the above into (C.2) and simplifying by Stirling's formula (21), we have

$$Q_{\log d+1} \le \frac{1}{4}\left(\frac{\lambda}{d}\right)^d \left(1-\frac{\lambda}{4d}\right)^{d-1} \frac{2^d}{\sqrt{\pi d}}.$$

Thus,

$$P_{loss} \le (n-\log d)\frac{1}{4}\left(\frac{\lambda}{d}\right)^d \left(1-\frac{\lambda}{4d}\right)^{d-1} \frac{2^d}{\sqrt{\pi d}}, \quad \text{(C.3)}$$

To get a lower bound for P_{loss}, continuing from the first line of (C.1), we have

$$\begin{aligned} P_{loss} &\ge \frac{P_n}{P_0}(Q_{\log d+1} + Q_{\log d+2} + \cdots + Q_n) \\ &\ge (1-P_{loss})(n-\log d)Q_n. \end{aligned} \quad \text{(C.4)}$$

From (8) and with the index change $i = k - d$,

$$Q_n = \frac{(P_{n-1}/2)^d}{P_{n-1}d}\sum_{i=1}^{d} i\binom{2d}{i+d}\left(\frac{P_{n-1}}{2}\right)^i \left(1-\frac{P_{n-1}}{2}\right)^{d-i}. \quad \text{(C.5)}$$

Now,

$$\binom{2d}{i+d} = \binom{d}{i} \times \frac{2d\cdots(d+1)}{(d+i)\cdots(i+1)} \ge \binom{d}{i}2^{d-i}.$$

Substituting into (C.5) and simplifying

$$Q_n \ge \frac{[(\lambda/d)(1-P_{loss})]^d}{4}\left(1-\frac{\lambda}{4d}\right)^{d-1}, \quad \text{(C.6)}$$

where we have made used of the facts that $P_{n-1} \ge P_n = (\lambda/d)(1-P_{loss})$ and $P_{n-1} \le \lambda/d$. Substitution into (C.6) gives

$$\frac{P_{loss}}{(1-P_{loss})^{d+1}} \ge \frac{1}{4}\left(\frac{\lambda}{d}\right)^d \left(1-\frac{\lambda}{4d}\right)^{d-1}(n-\log d). \quad \text{(C.7)}$$

If we are only interested in very small P_{loss}, then

$$\frac{P_{loss}}{(1-P_{loss})^{d+1}} \approx P_{loss}(1+(d+1)P_{loss}) \approx P_{loss}.$$

Making this approximation, we obtain

$$P_{loss} \ge \frac{1}{4}\left(\frac{\lambda}{d}\right)^d \left(1-\frac{\lambda}{4d}\right)^{d-1}(n-\log d).$$

Combining (C.3) and the above, we have

$$\begin{aligned} &\frac{(n-\log d)}{4}\left(\frac{\lambda}{d}\right)^d \left(1-\frac{\lambda}{4d}\right)^{d-1} \le P_{loss} \\ &\le \frac{(n-\log d)}{4}\left(\frac{\lambda}{d}\right)^d \left(1-\frac{\lambda}{4d}\right)^{d-1} \frac{2^d}{\sqrt{\pi d}}. \end{aligned} \quad \text{(C.8)}$$

We now investigate the relationship between d and n for some fixed λ and P_{loss}. Taking the logarithm of (C.3),

$$\begin{aligned} \log P_{loss} &\le \log(n-\log d) - 2 + d\log\lambda - d\log d + \\ &\quad (d-1)\log\left(1-\frac{\lambda}{4d}\right) + d - \frac{1}{2}\log\pi - \frac{1}{2}\log d \\ &\le \log n - 2 + d\log\lambda - d\log d + d - \\ &\quad \frac{1}{2}\log\pi - \frac{1}{2}\log d. \end{aligned}$$

Thus,

$$\log n - \log P_{loss} \ge d\log d + f(d), \quad \text{(C.9)}$$

where

$$f(d) = 2 - d\log\lambda - d - \frac{1}{2}\log\pi + \frac{1}{2}\log d = O(d)$$

Taking the logarithm of (C.7),

$$\begin{aligned} &\log P_{loss} - (d+1)\log(1-P_{loss}) \\ &\ge \log(n-\log d) - 2 + d\log\lambda - d\log d + \\ &\quad (d-1)\log\left(1-\frac{\lambda}{4d}\right) + d \\ &\ge \log n + \log\left(1-\frac{\log d}{n}\right) - 2 + d\log\lambda - \\ &\quad d\log d + (d-1)\log\left(1-\frac{\lambda}{4d}\right) + d. \end{aligned}$$

Now,

$$(d-1)\log\left(1-\frac{\lambda}{4d}\right) \ge (d-1)\log\left(1-\frac{\lambda}{4}\right)$$

since $d \ge 1$, and

$$\log\left(1-\frac{\log d}{n}\right) \ge \log\left(1-\frac{\log d}{2^{d\log d+f(d)}}\right)$$

by substitution from (C.9). Hence,

$$\log n - \log P_{loss} \le d\log d + g(d), \quad \text{(C.10)}$$

where

$$\begin{aligned} g(d) &= -\log\left(1-\frac{\log d}{2^{d\log d+f(d)}}\right) + 2 - d\log\lambda - \\ &\quad (d-1)\log\left(1-\frac{\lambda}{4}\right) - d - (d+1)\log(1-P_{loss}) \\ &= O(d) \end{aligned}$$

From (C.9) and (C.10), we conclude that

$$d\log d = \log\log N - \log P_{loss} + O(d), \quad \text{(C.11)}$$

and therefore $d\log d = O(\log\log N)$.

REFERENCES

[1] CCITT, New Draft Recom. I. 150, "B-ISDN ATM Layer Functionality and Specifications," *Committee XVIII*, January 1990.

[2] M. Kumar and J. R. Jump, "Performance of Unbuffered Shuffle-exchange Networks" *IEEE Trans. Computers*, Vol. 35, No. 6, June 1986, pp. 573–577.

[3] C. P. Kruskal and M. Snir, "The Performance of Multistage Interconnection Networks for Multiprocessors" *IEEE Trans. Computers*, Vol. 32, No. 12, December 1983.

[4] J. H. Patel, "Performance of Processor-memory Interconnections for Multiprocessors," *IEEE Trans. Computers*, Vol. 30, October 1981, pp. 771–780.

[5] B. Bingham B. and H. Bussey, "Reservation-Based Contention Resolution Mechanism for Batcher-Banyan Packet Switches," *Electronics Letters*, Vol. 24, No. 13, June 1988, pp. 772–773.
[6] Y. N. J. Hui and E. Arthurs, "A Broadband Packet Switch for Integrated Transport," *IEEE Journal on Selected Areas in Communications*, Vol. 5, No. 8, October 1987, pp. 1264–1273.
[7] T. T. Lee, "A Modular Architecture for Very Large Packet Switches," *IEEE Trans. Commun.*, Vol. 6, No. 9, July 1990, pp. 1455–1467.
[8] M. G. Hluchyj and M. J. Karol, "Queueing in High-Performance Packet Switching," *IEEE J. Select. Areas Commun.*, Vol. 6, No. 9, December 1988, pp. 1587–1597.
[9] M. J. Karol, M. G. Hluchyj, and S. P. Morgan, "Input vs. Output Queuing on a Space Division Packet Switch," *IEEE Trans. Commun.*, Vol. 35, No. 12, December 1987, pp. 1347–1356.
[10] J. Giacopelli, M. Littlewood, and W. D. Sincoskie, "Sunshine: A High Performance Self-routing Broadband Packet Switch Architecture," *Proceedings of ISS '90.*
[11] S. C. Liew and K. W. Lu, "Comparison of Buffering Strategies for Asymmetric Packet Switch Modules," *IEEE J. Select. Areas Commun.*, April 91, pp. 428–438.
[12] S. C. Liew and K. W. Lu, "A 3-stage Interconnection Structure for Very Large Packet Switches," *International J. Digital and Analog Cabled Systems*, Vol. 2, 1989,pp. 303–316.
[13] P. Newman, "A Fast Packet Switch for the Integrated Services Backbone Network," *IEEE J. Select. Areas Commun.*, Vol. 6, No. 9, December 1988, pp. 1468–1479.
[14] A. Pattavina, "Multichannel Bandwidth Allocation in a Broadband Packet Switch," *IEEE J. Select. Areas Commun.*, Vol. 6, No. 9, December 1988, pp. 1489–1499.
[15] Y.-S. Yeh, M. G. Hluchyj, and A. S. Acampora, The Knockout switch: A Simple, Modular Architecture for High-Performance Packet Switching, *IEEE J. Select. Areas Commun.*, Vol. 5, No. 8, October 1987, pp. 1274–1283.
[16] F. A. Tobagi and T. Kwok, "The Tandem Banyan Switching Fabric: A Simple High-Performance Fast Packet Switch," *Proceedings of IEEE INFOCOM '91.*
[17] C. E. Shannon, "Memory Requirements in a Telephone Exchange," *Bell System Technical Journal*, Vol. 29, pp. 343–349, 1950.
[18] S. C. Liew and T. T. Lee, "$N \log N$ Dual Shuffle-Exchange Network with Error-Correcting Routing," *Conference Record of IEEE ICC '92* , June 1992, pp. 262–268.
[19] C. Day, J. Giacopelli, and J. Hickey, "Applications of Self-Routing Switches to LATA fiber Optic Networks," *Proceedings of ISS '87*, March 1987.
[20] A. Huang and S. Knauer, "Starlite: A Wideband Digital Switch," *Proceeding of Globecom '84*, pp. 121–125.
[21] K. E. Batcher, "Sorting Networks and Their Applications," *AFIPS Proceeding of the Spring Joint Computer Conference*, 1968, pp. 307–314.
[22] T. T. Lee, "Non-blocking Copy Networks for Multicast Packet Switching," *IEEE J. Select. Areas Commun.*, Vol. 6, No. 9, December 1988, pp. 1455–1467.

Tony T. Lee received his B.S.E.E. degree from National Cheng Kung University, Taiwan in 1971, and his M.S. and Ph.D degrees in electrical engineering from Polytechnic University in New York, in 1976 and 1977, respectively. Currently, he is a Professor of Information Engineering at the Chinese University of Hong Kong.

He works in areas of broadband packet switch systems, performance analysis, parallel sorting networks, interconnection networks, relational database systems and protocol verification. Before joining Bellcore, he was with AT&T Bell Laboratories, Holmdel, NJ, from 1977 to 1983, He was an adjunct faculty member in the Electrical Engineering Department of Columbia University for the Fall term of the 1989 academic year. He also taught a pilot course on fast packet switches at MIT in 1990. From 1991 to 1993, he was a Professor of Electrical Engineering at Polytechnic University, Brooklyn, New York.

He is a member of Sigma Xi since 1977 and a senior member of IEEE since 1988. He was selected to be a Distinguished Member of Professional Staff by Bellcore since 1988. He is the recipient of the 1988 Leonard G. Abraham prize paper award from the IEEE Communications Society.

Soung C. Liew received the S.B., S.M., E.E., and Ph.D. degrees in electrical engineering from Massachusetts Institute of Technology, Cambridge, in 1984, 1986, 1986, 1988, respectively. From 1984 to 1988, he was a Research Assistant in the Local Communication Networks Group at the M.I.T. Laboratory for Information and Decision Systems, where he investigated fundamental design problems in high-capacity fiber-optic networks. He was also a Teaching Assistant for a graduate course on data communication networks.

In March 1988, he joined Bellcore, Morristown, New Jersey, where he has been a Member of Technical Staff in the Network Systems Research Laboratory. He is currently taking a leave of absence from Bellcore and is Senior Lecturer in the Chinese University of Hong Kong. He has conducted research and published actively in various areas related to broadband communications, including wavelength-division-multiplexed optical networks, high-speed packet-switch designs, system-performance analysis, routing algorithms, network-traffic control, and reliable and survivable networks.

His current research interests include interconnection networks, broadband network control and management, distributed and parallel computing, fault-tolerant networks, and optical networks. Dr. Liew is a senior member of the IEEE and a member of Sigma Xi and Tau Beta Pi.