Cost of implementing Banyan networks for use in ATM switching fabrics

Christos Bouras* — **Christos Gkantsidis****

** Computer Engineering and Informatics Department, University of Patras,*
26500 Rion (Patras), Greece, and
Computer Technology Institute, Riga Feraiou 61,
26221 Patras, Greece

bouras@cti.gr

*** Computer Engineering and Informatics Department, University of Patras,*
26500 Rion (Patras), Greece, and
Computer Technology Institute, Riga Feraiou 61,
26221 Patras, Greece

gantsich@cti.gr

ABSTRACT. *In this paper, we present a model for computing the cost of implementing Banyan networks. We limit our interest in Banyan networks which are used in ATM switching fabrics and are build in VLSI. The cost is given as a function of the characteristics of the network (i.e. length of buffers, speed of links, etc.). It is well-known that the implementation cost must be related to the performance of the network. The choices, that the designer may have, impact both the performance and the cost. We demonstrate the case where a slight increase in performance implies a great increase in cost (in that case it is not cost effective to build a better switching network), and of course the reverse, i.e., the case where a decrease in cost implies a degradation of the performance of the switch.*

RÉSUMÉ. *En cet article, nous présentons un modèle pour calculer le coût de mettre en application des réseaux Banyan. Nous limitons notre intérêt pour les réseaux Banyan qui sont utilisés dans des ATM tissus de commutation et sont construits avec VLSI. Le coût est indiqué en fonction des caractéristiques du réseau (c' est à dire, du longueur des mémoires tampons, de la vitesse des liens, etc.). Il est bien connu que le coût de mettre en application doive dépendre de l'efficacité du réseau. Les choix, cela que le créateur peut avoir, effectue l'efficacité et le coût. Nous montrerons le cas où une légère augmentation d'efficacité implique une grande augmentation de coût (dans ce cas il n'est pas rentable pour établir un meilleur réseau de commutation), et, bien sûr, l'inverse, c' est à dire le cas où une diminution en coût implique une dégradation de l'efficacité du commutateur.*

KEYWORDS: *Banyan, switching fabrics, ATM, VLSI, cost evaluation, performance evaluation.*

MOTS-CLÉS : *Banyan, tissus de commutation, ATM, VLSI, évaluation du coût, évaluation d'efficacité.*

1. Introduction

The growing need for new services, like video on demand and many others, have increased the demand for new networks that can support a large number of users and a variety of services, which include both traditional bandwidth-hungry data services and others that require a guaranteed quality of service (QoS) from the network. One solution to this problem is the development of Broadband Integrated Services Digital Networks (B-ISDN). Because ATM is the transport protocol of choice for this kind of networks, a need has arisen to build ATM networks that support many customers and demanding services. The core of an ATM network is the ATM switch and thus the afore-mentioned needs must be fulfilled by this device. Large ATM Switches must be build and this must happen in a cost effective way, meaning that we must compromise the performance with the cost. The core of the ATM Switch, which influence in a dominant way both the performance and the cost, is its switching fabric. In this work, we will concentrate on a special architecture for building switching fabrics which is based on Banyan networks. This kind of networks are very efficient in supporting a large number of customers and have been used extensively in the past for this purpose. Our results, techniques and methodology can be applied easily to other architectures with slight modifications.

In this paper, we try to identify the factors that affect the design of a Banyan network. Some of these factors are the existence or not of buffers, the length of them if they exist, the presence or not of a sorting network that precedes the Banyan, the speedup of the network and many others. The designer can experiment with them to determinate which network suits better the design goals. The literature is rich with works that examine the impact of these factors in the performance of the network using both analytical techniques and experimental results. We are using some of them to evaluate the impact of some choices on the performance of the Banyan network.

But, except for the performance of the switching fabric there is another critical parameter that affect the design and this is the cost. A choice which can increase the performance slightly, but will have a significant impact on the cost is not wise, except in rare cases where the only goal of the design is the performance. On the other hand, a small increase in cost which results in a great advancement in performance should be made. Assuming that the switching fabric will be implemented in VLSI, a major parameter of the cost is the area of the chip. This area affects the probability of producing a working chip because the probability of a fault is constant in a unit area and depends only on the technology being used. Thus, a small chip has greater probability to be produced without a fault. Increasing the area, by adding for example buffers in order to improve the performance of the switch, result in an increase of the probability of producing a faulty chip. Chip area is not the only factor that affect cost; another one is the speed up factor which shows how faster will be the links of the switching fabric than the input and output links. In this work we are trying to relate the cost with these parameters. Having found a way to express the cost with the parameters of the network, we compare the cost and the performance and suggest a method to find the optimal cost performance design.

Many papers found in the literature, like [COP 93], [FRA 81] and [BOU 98], are the basis of this work and many of our results depend on these papers. Although the key points of them are mentioned in this paper, the interesting reader must consult them for full proofs and explanations.

In section 2 we describe Banyan networks. We give various kinds of these networks (buffered, unbuffered, sort-Banyans, dilated and replicated) and discuss briefly the factors that restrict their performance. Also, in the same section we discuss some techniques which were developed by [FRA 81] and [COP 93] and can help us find the cost of the switching fabric. In section 3, we derive the function which gives the cost of implementation of Banyan networks and discuss some key-points of it. In section 4 we put together the results of the previous one and the well known results about performance of these kind of networks and discuss about what is optimal in a cost performance sense. Finally, in section 5 we give the conclusions and some ideas for future work.

2. The model

2.1. *Banyan networks*

Banyan networks belong to the class of Multistage Interconnection Networks (MINs). They were defined in [GOK 73] and are characterized by the property that there is exactly one path from any input to any output. A simple example of a 8×8 banyan interconnection network (ie. which has 8 inputs and 8 outputs) is drawn in figure 1. This network consists of nodes and links. Every node is a 2×2 switch, which can receive packets at each of its two input ports and send them through each of its two output ports. Generally, switches may have k input ports and k output ports [1] ($k \times k$ switches). Switches are grouped in stages. This means that every switch in stage j is connected to switches in stages $j-1$ and $j+1$. Of particular interest are networks in which all switches are identical, except perhaps the switches in the first and last stages. The other building element of a MIN is links. Links are used to connect switches of successive stages.

So far, we have made some assumptions about the switches. Now we will concentrate on the network. We assume that the problem is to interconnect N inputs to N outputs, thus we want to construct an $N \times N$ MIN which has a banyan topology. Using $k \times k$ switches, we need a network with $\log_k N$ stages and $\frac{N}{k} \log_k N$ switches. We, also, assume that we want to transport packets from inputs to outputs without preestablishing a path; our transport network will be connectionless.

Banyan networks have many advantages. First of all they require only $\frac{N}{k} \log_k N$ switches and $N \log_k N$ connections, as opposed to the crossbar network which requires $O\left(N^2\right)$ switches and links. Also, they have the property of self-routing. This means that every switch that accepts a packet in one of its inputs can decide in which

1. We concentrate on switches that have the same number of inputs as outputs

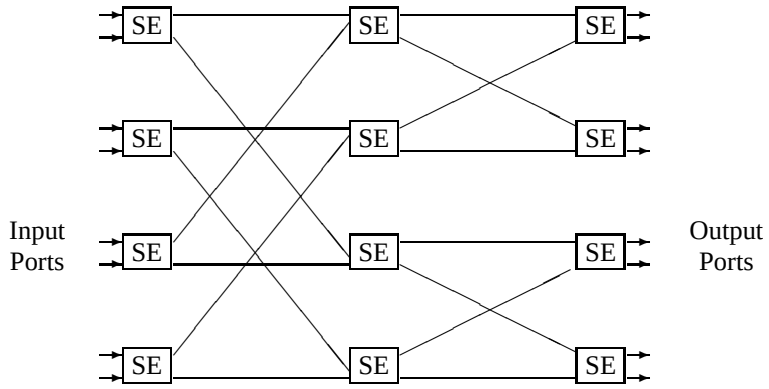

Figure 1. *An* 8×8 *Banyan network*

of its outputs to forward this packet depending only on the destination address of the packet. For example, in the Delta-2 networks, which are a subclass of Banyan networks consisting of 2×2 switches, every switch of stage j can decide in which output port to forward a packet based on the j bit of the destination address. If this bit is 0 then the packet is forwarded to the "upper" output port, and if it is 1, it is forwarded to the "lower" output port. Delta networks were defined and described in [PAT 79], [PAT 81]. The property of self-routing is very important because it means that the implementation of a complex routing algorithm, either centralized or distributed, is not needed. Another advantage of Banyan networks is their regularity and interconnection patterns which are very attractive for VLSI implementation ([AHM 88])

But, these advantages come with a cost. It is well known that Banyan networks are blocking. This means that it is possible for two or more packets to content for the same link somewhere in the network. Only one of them will win this contention and will be transmitted. The others must be buffered and try again in a later time (we assume that the network must not discard packets if there are enough buffers). So, a choice that a designer may have is where to place the buffers. If the buffers are placed in the input ports, which means that a packet stays there until it is finally transported to an output port, then due to head-of-line blocking [2] the throughput of the network will be much lower than $2 - \sqrt{2} \approx 0.586$. We will not prove this limit or discuss the assumptions, which are close to the ones we have made so far; the interested reader may refer to [PRY 95] or [KAR 87]. Another option is to place buffers inside the switching fabric of the switches. A lot of research has been done concerning this matter and a lot of results, both theoretical and from simulations, have appeared (see for example [BOU 98], [DIA 81], [KRU 83]). It is interesting to note that these results are independent of the link pattern which is used to construct the network. Generally, the performance of Banyan networks is independent of the

2. The packet in the head of the queue will prevent the other packets from reaching an output port even if they can find a path in the network

specific topology if the traffic is equally distributed among the input and output ports (one of our assumptions).

So far, we have seen that the designer has two choices. The first is where to place the buffers and the second is the length of them (ie. how many packets they can hold). Both of them can greatly affect the performance of the switching fabric. Another way to increase the performance is by increasing the speed of the internal links. In the previous discussion we have assumed that the speed of the internal links is equal to the maximum speed of the input ports. Thus, if we assume that the network works in cycles, where in one cycle the packets which are in the head of the input queues move through the network and some of them reach the output ports, then it can deliver at most one packet to a given output in one cycle. If the internal links operate s times as fast as the input and output ports (which we assume that operate in the same speed), then at most s packets can be delivered to an output port in a cycle. In the extreme case where $s = N$ then the transport network is internally non-blocking. For practical designs s should be small. [OIE 89] reports that for $s = 2$ and $s = 3$ and for nonblocking architectures (Banyan networks are blocking and an upper limit in their performance is that of nonblocking) the maximum throughput can be as large as 0.8845 and 0.9755 respectively. In this case, where $s > 1$ and generally when more than one packet can be delivered to the same output in the same cycle, buffers must be used in the output ports.

Another way to improve the performance is by using dilated or replicated networks. Unbuffered d-dilated and d-replicated Banyan networks were introduced in [KRU 83]. D-dilated networks have the property that every switch is connected to its successor switch at the next stage with d links. Thus, at most d packets, which compete for the same internal link, can move from one stage to the next. D-replicated networks consist of d parallel Banyan networks. Thus a packet must decide which of these will use. A simple algorithm is to choose one in random. Both of these variances of Banyan networks resemble the method of increasing the performance by speeding up the internal links.

All of the previous methods were basically extensions of the simple Banyan network. This means that all of them are internally blocking. The effects of this phenomenon are reduced using the one or the other technique, but in principle the probability that two packets destined to different outputs collide in one of the intermediate nodes is not zero. One way to remove the internal blocking property is by using a sorting network in front of the Banyan. In this way, packets will be sorted based on their destination addresses and then routed through the Banyan network. Contention for an internal link can not happen in this case and therefore it is not needed to put buffers in the switches. But, buffers must be used in this case in the input ports because sort-banyan networks can route packets with distinct destinations. If two or more packets have the same destination, only one of them will be routed and all the others must stay in the input queues (remember that this may cause the head-of-line syndrome). A very popular sorting network is the Batcher bitonic sort network, which is very similar to the Banyan network. It was introduced in [BAT 68] and consists

of very simple elements (2×2 switches). To construct a network which will sort N packets, $\frac{N}{4}\left((\log_2 N)^2 + \log_2 N\right)$ switching elements must be used (see [AHM 88]). The price that must be paid in this case is that the number of stages that a packet must traverse, in order to get from an input port to an output, has been increased. Fortunately, due to the lack of internal buffers, the time to traverse the sort-banyan network is predictable and equal to the number of stages of the combined networks times the delay in each stage. So, to evaluate the delay in the switching fabric, only the time spent in the input queue must be found.

2.2. *Factors that affect the cost of the networks*

From the description of Banyan network it is easily derived that the number of required switching elements (SE for short) is $O(N \cdot \log N)$ and that the number of stages is $O(\log N)$, where N is the number of input/output ports in a $N \times N$ transport network. This means that the total chip area that must be used is $O(N \cdot \log N)$ when the design is based on SSI technology. But when VLSI is used this doesn't happen as Franklin points out in [FRA 81].

Traditional techniques for finding the total area of the chip take for granted that the cost of interconnecting subsequent stages is negligible and this means that they assume that the total area depends only on the number of switching elements. Let's assume that each line, which connects a switching element in one stage to another element in the following stage, has a width of l units of length. Let's also assume that the gap between adjacent lines must be l units of length. Thus, if w parallel lines are needed to connect two SEs (path of length w), then at least $2lw$ units of length are required for this connection. The bad thing about Banyan networks is that connection paths must be mixed up in some stages. Take for example the network in figure 1, where it is clear that the paths between the first and second stages have more points in common than the paths between the second and the third stage. The designer must find a way to avoid these common points given that in VLSI the paths consist of horizontal and vertical segments (ie. there are no diagonal connections as in figure 1). As indicated in figure 2 (which is the same interconnection network as in figure 1, with the exception that the above constraint about communications paths has been taken into account), in order to construct the paths which connect two subsequent stages when there are many points in common, an increase in the horizontal distance of the stages must be made. In a network with more inputs and outputs which is constructed using the pattern of figure 1, the distance between subsequent stages would be a decreasing function. What is most interesting is that the total area used to connect the SEs of states, which is the summation of the products of the vertical times the horizontal distances, is $O\left(N^2\right)$. The interested reader may refer to [FRA 81] for a full proof and a deeper explanation of the above reasoning. Something that we must point out, is that in Fraklin's work it was assumed that the SEs had no buffers. If we put buffers in the switching elements then the area of the Banyan network will be dominated by the total area of the SEs.

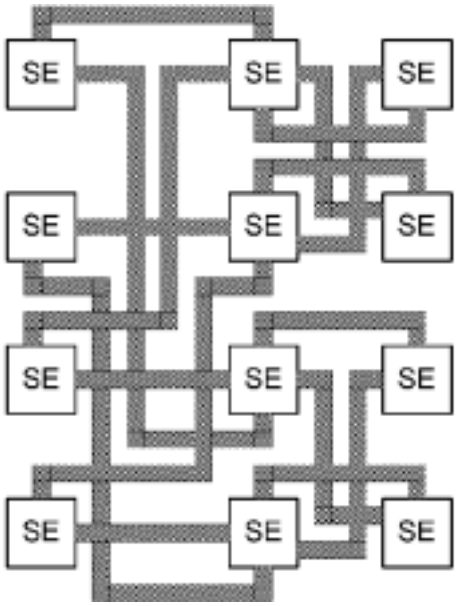

Figure 2. *An* 8×8 *Banyan network implemented in VLSI*

A large part of the previous discussion has been devoted to methods which help us to estimate the required area for the Banyan transport network. We haven't said enough for the importance of this estimation and its relation to the cost function, which is our ultimate goal. It is well known that when producing a number of chips, using any available technique, not all of them will work properly. In fact the vast majority of them will be faulty. The probability of obtaining a working component of unit area in a given technology is called yield and we will denote this by r. Thus if our chip, which is a Banyan network, has area equal to A units, then the expected number of correct chips will be one every r^{-A}. Now, we must try to relate this probability and the area A with the cost of producing the chip. Generally this is a very difficult task. To simplify it, we will adopt the approximation, which is also made in [COP 93], that the cost depends on $A \cdot r^{-A}$.

Except from the area of the chip, we will assume that the cost depends also on the speedup factor. It is evident that this is true for each link; stated in a different way the cost of each link depends on the speedup factor. Also, the cost of each switching element depends on this, because increasing the operational speed of the links means that we must increase the speed of the SE. Another assumption is that the total cost depends on the cost of each link and of each switching element. Although the literature is not very rich in this area, all the above assumptions are usually made (see for example [COP 93]).

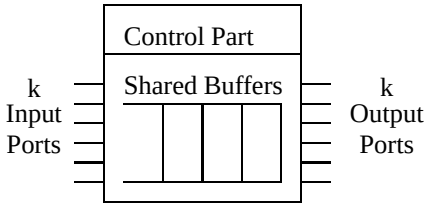

Figure 3. *An abstract structure of a SE*

3. Analytic results

3.1. *Evaluation of the function of the cost*

In this section we try to evaluate the cost of the networks discussed in the previous section. First, we turn our attention to the basic building block of the Banyan network, which is the switching element. Every SE must have the structure of the figure 3. It has k inputs and k output ports. Also, there are at least k buffers in the SE, which hold the packets temporarily before they are transmitted in the next stage. In a non-buffered network each SE will have exactly k buffers. In buffered networks, it is usually assumed that each output port is associated with B buffers. Thus there are $k \cdot B$ buffers in the SE. This doesn't happen in practice. Usually, fewer than $k \cdot B$ buffers are used which are shared between all output ports. By multiplexing all output streams (packets for the same output port) the number of buffers can be reduced without affecting the performance, in terms of cell loss probability or average delay. The penalty for this is an increase in the complexity of the SE and the use of faster, by a factor of at least k, memory. The ratio of the number of buffers when shared memory is not used, which is $k \cdot B$, and the number of buffers in the other case when the same performance is achieved, is called multiplexing gain and depends on the number of inputs and the traffic characteristics. We will make the assumption that traffic is uniformly distributed and thus this ratio will depend only on k. We expect that as k increases then this ratio will also increase because better statistical behavior would be achieved[3]. We denote the reciprocal of this ratio as $m(k)$ and thus the total number of buffers that will be used is $k \cdot m(k) \cdot B$. The size of the area which will be used for the buffers, A_B, depends on the number of them and this means that:

$$A_B = O(k \cdot m(k) \cdot B) \qquad [1]$$

Function $m(k)$ is monotonically decreasing and takes values in $(0, 1]$.

Eq. 1 gives an approximation of the area of the SE which will be used for the buffers. In order to find the total area of the SE, we must approximate the area of the control part (see fig. 3). Clearly, this depends on the number of input and output ports. Also, we expect that if many buffers are used then this area will be much smaller that

3. A fact that is derived from the central limit theorem

the area of the buffers. On the other hand, when no buffers are used then we expect that the area of the control part is comparable, if not bigger, to the area of the buffers. Thus, if we assume that α_B and α_{CP} are constants which depend on the implementation, then the total area of the SE is[4]:

$$A_{SE} = A_B + A_{CP} = \alpha_B \cdot k \cdot m(k) \cdot B + \alpha_{CP} \cdot k$$

An approximation which has been made in [COP 93] and which is also used by us is that the right part of the previous equation is $O(k \cdot m'(k) \cdot B)$. In this way, we use the new function $m'(k)$ to hide the area of the control part. Thus, the total area of the switching element is:

$$A_{SE} = O(k \cdot m'(k) \cdot B) \quad [2]$$

By introducing the constant α_{SE} we can rewrite the previous equation as:

$$A_{SE} = \alpha_{SE} \cdot k \cdot m'(k) \cdot B$$

From now on, $m(k)$ will refer to $m'(k)$.

The number of SEs of the Banyan network is $\frac{N}{k} \cdot \log_k N$. The consumed area is:

$$A_{total\ SE} = O\left(\frac{N}{k} \log_k N \cdot k \cdot m(k) \cdot B\right) \quad [3]$$

Observe that the above equation gives only the necessary area to implement the functions of the SE (buffering and control). In practice, it is possible to use more area because we must take into account that only a limited number of positions can be used to place the SEs and that the length of the sides of the SEs must be sufficiently large to place the input and output ports. This later approach has been used in [FRA 81], where it is proved that the total area needed for the SEs is $O\left(N^2\right)$. In contrast, eq. 3 says that this area is $O(N \cdot \log N)$. In our analysis we assume that the area of the buffers, when they are used, dominates the total area of the SE. The remaining area, which is asymptotically $O\left(N^2\right)$, is included in the area needed for the connections, which is given below.

The remaining area of the chip is used for the links. Even though the necessary number of links is $N \cdot \log_k N$, the area they use is $O\left(N^2\right)$ (see [FRA 81]). The cost of the links depends also on the speed up factor s. Thus, we can approximate this cost with the function $\alpha_I \cdot N^2 \cdot s$, where α_I is a constant which depends on technology being used to implement the chip. The extra cost which is needed for the reasons discussed above can be included by increasing this constant.

Now, we can give the total cost of a Banyan network which uses buffers. This cost depends on the area used for the SEs and the links, on the speedup factor and on yield and it is:

$$C_{BB} = k_c \cdot s \cdot N \cdot (\alpha_{SE} \log_k N \cdot m(k) \cdot B + \alpha_I \cdot N) \cdot r^{-N \cdot (\alpha_{SE} \cdot \log_k N \cdot m(k) \cdot B + \alpha_I \cdot N)} \quad [4]$$

4. See table 1 for a listing of all parameters and constants used in this work.

N	Number of input and output ports of the switching fabric.
k	Number of input and output ports of the switching elements.
s	Speedup factor.
d	d-replicated or d-dilated networks.
r	Yield. The probability of producing a working chip of unit area.
B	Number of buffers per output port.
$m(k)$	Multiplexing factor of input streams in $k \times k$ SEs.
$m'(k)$	Same as $m(k)$. They are used interchangeable.
A	Area of something.
A_B	Area used for buffers.
A_{CP}	Area used for control part.
A_{SE}	Area of a single SE when buffers are used.
$A_{total\ SE}$	Total area needed for SEs in a buffered Banyan network size N.
C_{BB}	Total cost of buffered banyan networks.
C_{DB}	Total cost of d-dilated or d-replicated banyan networks.
α_B	Constant depending on the implementation of buffers.
α_{CP}	Constant depending on the implementation of the control part.
α_{SE}	Constant depending on the implementation of the SE.
α_I, α'_I and α''_I	Constants depending on the implementation of the links.
k_c, k'_c and k''_c	Constants depending of the technology used to build the chip.

Table 1. *List of symbols and their meaning*

where k_c is a constant which depends on the implementation and the technology used. Constants α_{SE} and α_I must be selected in such way that they accurately give the areas used by SEs and links respectively. Thus, in a buffered network where the area of the SE dominates the area of the chip, α_{SE} must be larger than α_I so that $N \cdot \log_k N \cdot m(k) \cdot B$ is much bigger than N^2 in the previous equation for all practical values of N. When only k buffers per SE are used (no extra buffers, unbuffered network) then the area of the links dominates the area of the chip and this means that α_I is bigger than α_{SE}.

Now, we turn our attention to d-replicated and d-dilated unbuffered Banyan networks. These are simple extensions of the basic network. The cost of d-replicated networks can be easily found from the cost of the unbuffered Banyan. These networks consist of d copies of the simple Banyan network and therefore the chip area they need is d times the area of the unbuffered Banyan plus the area of the control logic which routes the packets to one of the d networks. We will assume that the area of the control logic is negligible. This means that their cost is:

$$C_{DB} = k'_c \cdot d \cdot \alpha'_I \cdot N^2 \cdot r^{-d \cdot \alpha'_I \cdot N^2} \qquad [5]$$

The case of d-dilated networks is somewhat more complicated. These networks have the characteristic that every switching element has $d \cdot k$ input and output ports. Also, the total area of the links is d times larger because d times more links are used. Even though this is a very different situation than the previous, in both cases the area of the network increases by a factor of d. What changes in the case of d-dilated networks are the constants k_c' and α_I', which become k_c'' and α_I'' respectively. These models can be extended using the methods discussed in the buffered Banyan case if the designer chooses to put buffers in the SEs or to increase the speed of operation.

Finally, we examine the Sort-Banyan networks. These networks consist of two subnetworks which are called sorting and banyan. Usually the sorting subnetwork is a Batcher bitonic. In this case the SEs of this subnetwork are 2×2 switches. In practice all SEs are identical and thus the total number of SEs required is

$$\text{Number of SEs} = \frac{N}{4} \cdot \log_2^2 N + \frac{3 \cdot N}{4} \cdot \log_2 N$$

(because $\frac{N}{4} \cdot \left(\log_2^2 N + \log_2 N\right)$ are needed for the Batcher's sorter and $\frac{N}{2} \cdot \log_2 N$ for the Banyan). But, in this case the total area of the chip depends heavily on the links which connect SEs of subsequent stages. As in the case of simple Banyan networks, the total area is $O\left(N^2\right)$. The main difference here is that the chip area needed for this network is two times the area of the simple Banyan, because two subnetworks are used; one of them is Banyan and the other has a similar structure. Doubling the area doesn't mean that the cost will also be double. It would be much more because it increases exponentially with area.

3.2. *Some observations on the functions*

Now, let's look how the functions derived in the previous section can help the designer make some decisions which will reduce the implementation cost of the network.

First of all, a choice must be made about the size of the SEs, ie. the value of k. Increasing k means that fewer buffer per output port will be needed because of statistical multiplexing. But, k can't exceed a maximum value, which depends on the technology used. These two facts can lead to conclusion that k must take its maximum value. Using eq. 4, we can plot the gain in cost of using $k \times k$ switches (ie. $\frac{C_U - C_B}{C_U}$ where C_U is the cost of the switching element when no multiplexing is used and C_B is the cost in the case of shared buffers) versus number of ports. Even though this plot depends on many factors, a general sketch of it can be found in fig. 4. For this sketch, we have used an approximation of $m\left(k\right)$ found in [COP 93][5]; but generally for every valid $m\left(k\right)$ the function have the same shape. A property of this function is that it is increasing very rapidly and its limit is one. Also it is easily observed that there is a value of k before of which the rate of increasing is very rapid and after of which

5. $m\left(k\right) = \frac{0.35 \cdot k + 2.9}{k + 1.5}$

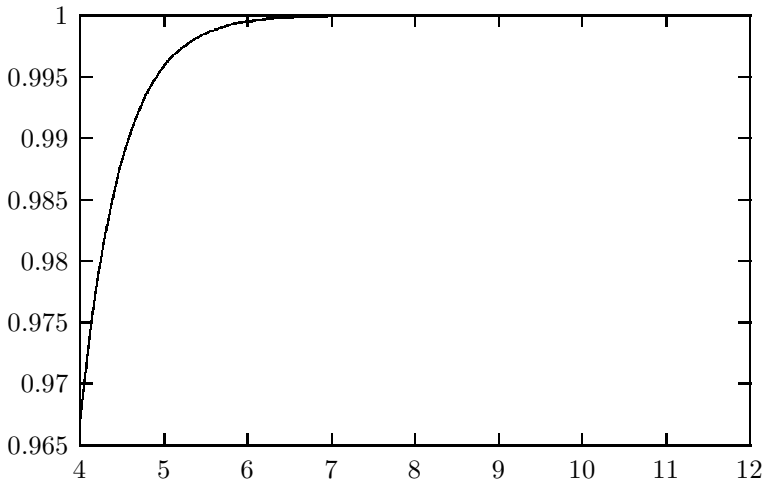

Figure 4. *Gain in cost versus number of ports of SE*

this rate is very slow. Thus, the best choice of k, which reduce the cost and keeps the implementation simple, is this value. For example, for $r = 0.3$ and for several values of N and B the optimum is close to 8, meaning that 8×8 switches must be used.

Some other design parameters, such as length of buffers per port B, d in replicated or dilated networks and N, have the reverse effect in the cost. They increase exponentially the cost of building the network. We will examine B. From eq. 4, if we ignore all other parameters but B, we can easily find that the function $B \cdot r^{-B}$ gives the dependance of cost on number of buffers. A sketch of this function is given in fig. 5. From this figure it is easily understood that there exist a value of B, which prohibits the use of buffers of larger size because the rate of increase of cost after this value is very rapid. The same phenomenon is observed for the parameters d and N. Especially in the case of N the rate of growth is larger (see eq. 4). The only thing the designer can do is to try to increase the critical value, by increasing for example r with the help of more advanced technology.

4. Cost/performance evaluation

In order to study the performance of the switching fabrics, we must define the measures which are used to characterize their performance. One such measure is the probability of a packet being dropped in the network because of buffer overflow. This is particularly important to the end user whose applications depend on the reliable transportation of information. If this probability is large then lost packets will result in a degradation of the quality of applications. Also, this metric is important to the network, because lost packets can trigger retransmissions which can result in congestion. To avoid this problem, large buffers must be used. Another measure of interest

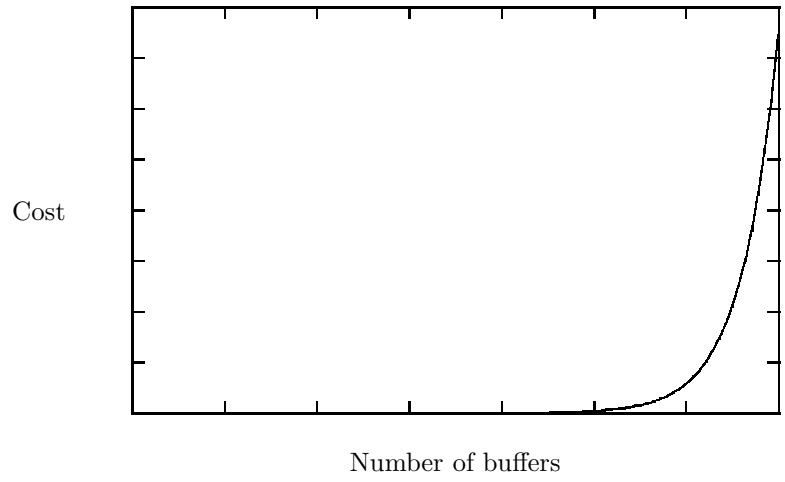

Figure 5. *Cost versus number of buffers per output port*

is the total delay which is needed for the packet to traverse the switching fabric. Since we have assumed the use of ATM networks, in which the length of packets is constant, this measure depends only on the number of packets waiting in queues and on the frequency of operation of the switching fabric. In this paper we focus on the expected time to traverse the fabric and thus we use the mean number of packets waiting in queues. The last measure we use is the throughput of the switching fabric. This is important to the holder of the network who wants to forward as many packets as possible, or, in other words, to make best use of the network.

An interesting thing to study is the effect of using $k \times k$ switches with large k in the probability of dropping a packet. In [BOU 98] it was conjectured that using small k's is preferable when it comes to this probability. Using the algorithm found in that paper we have been able to construct the graphical representation of fig. 6. This figure represents the mean number of packets lost per queue in the first stage of a buffered Banyan network versus k, and it is also valid not only for other stages but for unbuffered networks, too. On the other hand, in the previous section we have found that increasing k has the advantage of using fewer buffers per output port and thus decreasing the total area needed for the SE. In fig. 7 we can see the ratio of the mean number of packets lost per output queue and the gain of using multiplexing (found in the previous section) versus k. In this figure smaller values are better. We can easily observe that increasing k increases slightly this ratio. Thus, it is preferable to use small k.

Another interesting parameter of the buffered-Banyan networks is the number of buffers in each output queue. Increasing this number gives a better (ie. smaller) probability of dropping a packet but also results in an increase in the area needed for the chip and thus in the cost. Fig. 8 gives this probability versus the size of the buffers

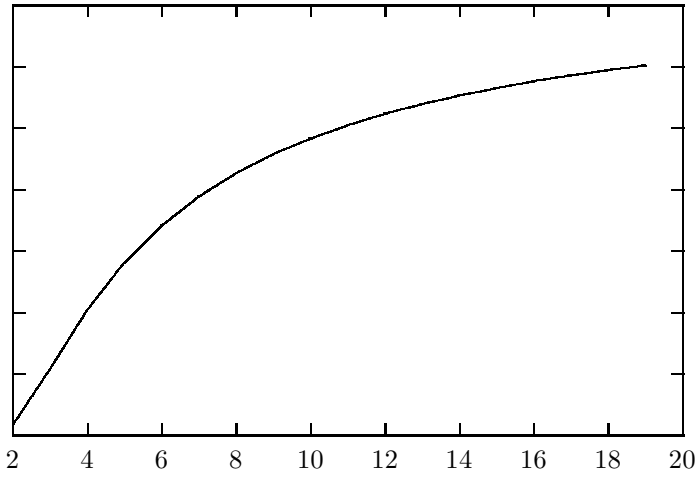

Figure 6. *Mean number of packets being lost per queue in the first stage of a buffered Banyan network versus number of ports of SE*

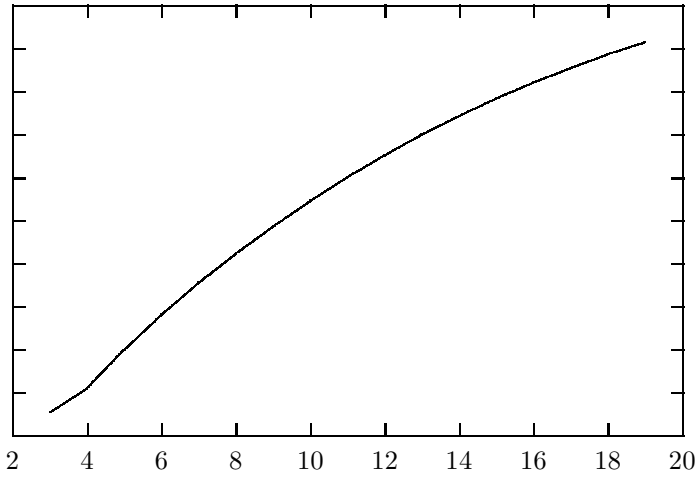

Figure 7. *Mean number of packets lost per queue over gain of multiplexing versus k*

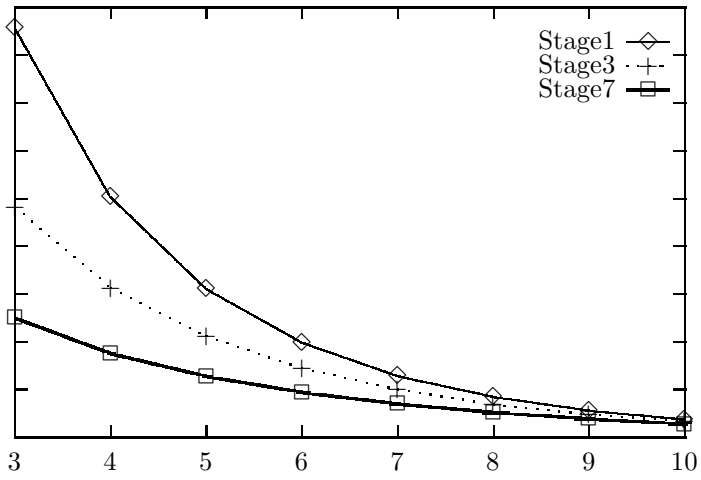

Figure 8. *Mean number of packets being lost per queue of a buffered Banyan network versus the size of the queue*

for some stages of a buffered Banyan switching fabric. This probability is decreasing very fast but, on the other hand, the cost is increasing exponentially, as we have proved in the previous section. The product of this probability and the cost is given in fig. 9 and again we see that small values are better. It is easily observed that this product is increasing quite fast (exponentially) and this means that adding few buffers beyond a point increases dramatically the product and thus it is not a wise choice to increase the buffers.

A parameter that greatly affect the design of the switching fabric is the typical load of the network (traffic load). Switches in heavy-loaded networks must have many buffers to sustain the probability of packet drop below a threshold. For example, fig. 10 shows the necessary number of buffers for networks of typical loads between 0.5 and 0.95 when the probability of a packet being dropped must be below 10^{-10}. It is well known from the queueing theory that this number increases exponentially as it is shown in the figure. The increase in cost, drawn in fig. 11 in a logarithmic scale, is even bigger according to eq. 4 and means that the designer must carefully estimate the traffic characteristics of the network and put no more than the necessary buffers.

Another interesting parameter is the speedup factor s. The effect of speedup has been extensively studied in the literature for the case of non-blocking switches; Batcher-Banyan switches fall in this category. In [CHE 91] and [OIE 89] was shown that the throughput of the switching fabric increases as indicated in fig. 12. We can observe that when s is small enough, the increase in throughput is tremendous. On the other hand choosing a speedup factor of $s + 1$ instead of s will result in a small increase when s is large (practically zero increase for $s > 9$). The price that is paid to increase the throughput is a linear increase in cost. The value of s impacts not only

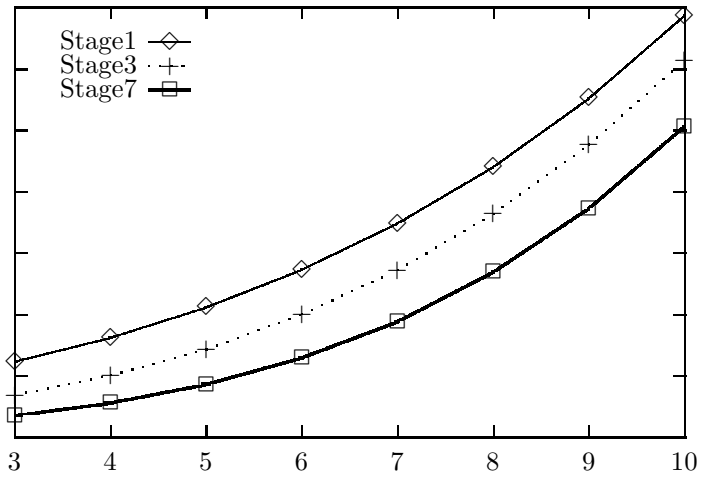

Figure 9. *Mean number of packets lost per output queue times the cost versus k for the first, the third and the seventh stages*

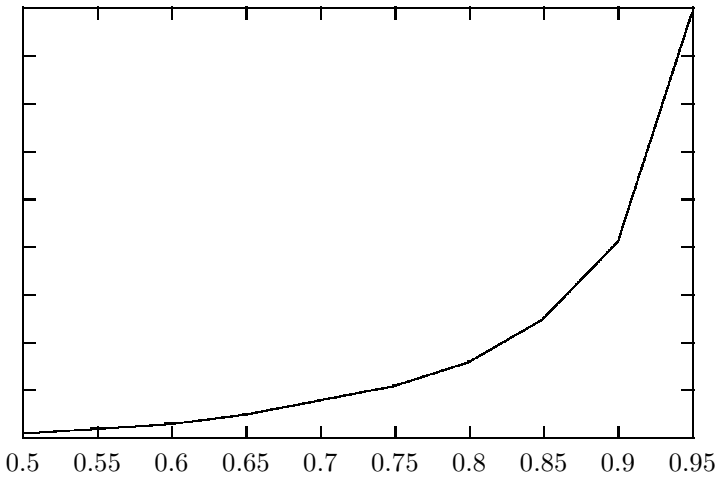

Figure 10. *Number of buffers needed versus load of the network to sustain the probability of packet drop below* 10^{-10}

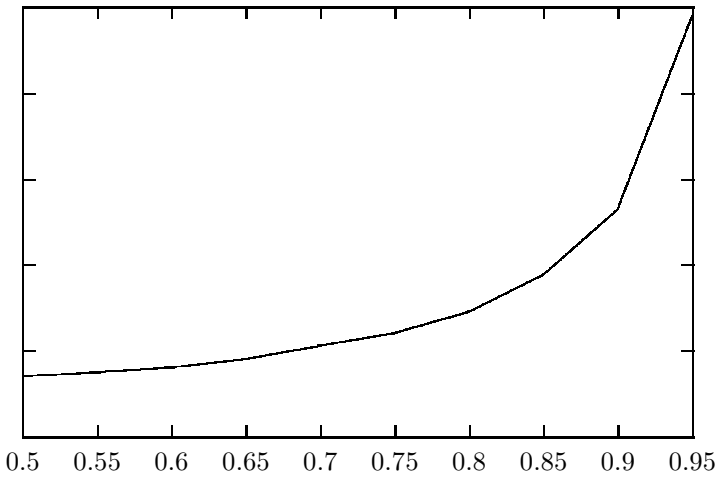

Figure 11. *Cost of implementing a switching fabric when the traffic load is between 0.5 and 0.95 and the probability of dropping a packet must be below* 10^{-10}*, in logarithmic scale.*

the throughput, but also other performance measures such as the average system delay (ie. the average number of queued packets) and the probability of a packet drop due to buffer overflow. Increasing s will result in fewer queued packets and smaller loss probability (see [CHE 91] and [OIE 89]). Thus, the designer must choose to use a speed up factor between 3 and 6, whenever this is possible, because in this way the increase in the performance will overwhelm the increase of cost. Even though the above results apply to nonblocking switching fabrics, such as Batcher-Banyan, we expect that the general conclusions will also apply to buffered Banyan networks.

The designer may also choose to use d parallel Banyan networks or to connect the SEs of subsequent stages with d lines instead of one, in order to increase the performance; in other words use of d-replicated or d-dilated networks. Even though these types of networks have different performance characteristics, the differences can be observed only for impractical N (see [KRU 83]). Thus, we assume that they have the same performance and we concentrate on d-replicated networks, because they are easier to analyze. A simple thought, which we use for the analysis, is that each one of the parallel networks receives a packet with probability d times less than in the case of the simple Banyan network ($p' = p/d$, where p is the probability of a packet arrival in an input link and p' is the probability of a packet arrival in an input link of one of the parallel networks). Selecting in random the network which will be used to transfer a packet is not a very smart choice. More advanced routing algorithms (including the ones found in [LEA 90]), which try to route the packets in order to minimize the packet loss probability inside the switching fabric, result in better performance characteristics. Given this simple observation we can assume that the throughput of the switching fabric can be increased by a factor of at least d when a d-replicated

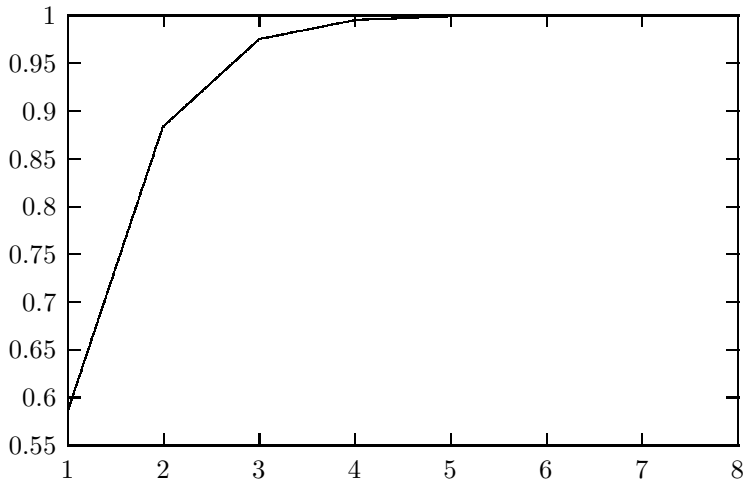

Figure 12. *Throughput versus speed-up factor s in Batcher-Banyan switching fabrics*

network is used. The price for this is an increase in the area of the chip of at least a factor of d (for the parallel networks) and an increase in cost which is even bigger according to eq. 5. Assuming that we can characterize the cost/performance of this network by the ratio of the throughput and the cost for various d, we get the plot of fig. 13. It is easily observed that bigger values are preferable from a cost/performance point of view described previously. Thus, to increase the throughput, the designer must sacrifice cost. If the probability of packet loss is assumed to be the main measure which characterize the performance of the switching fabric, then we can see that the increase of the number of parallel network results in a tremendous decrease of this probability. Fig. 14 gives some experimental results which have been taken using the algorithm of [BOU 98]. Assuming that the measure that characterizes the switching fabric from a cost/performance point of view is the product of the probability of a packet loss and the cost needed to implement the network (see eq. 5), we can see from fig. 15 that it is preferable[6] to use many parallel networks (large b).

So far, we have studied the impact that the parameters of the switching fabric may have both to its cost and performance. Using various cost/performance measures we have seen that a change in a parameter, which results in an increase in performance, in most cases, results in a bigger increase in cost. Also, the cost in these cases increases exponentially, which means that there is a value of the parameter in question which is a threshold. Moving from a point, which is bigger than this threshold, to another point close to it, results in a great increase in cost. On the other hand, moving between points, which are smaller than this threshold result in relatively small increases.

6. Small values are preferable in this figure

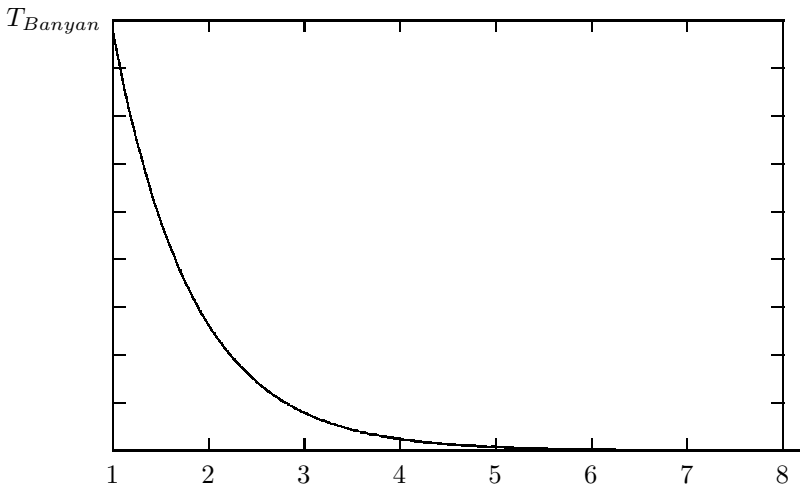

Figure 13. *Ratio of throughput and cost of a d-replicated buffered Banyan network versus the parameter d*

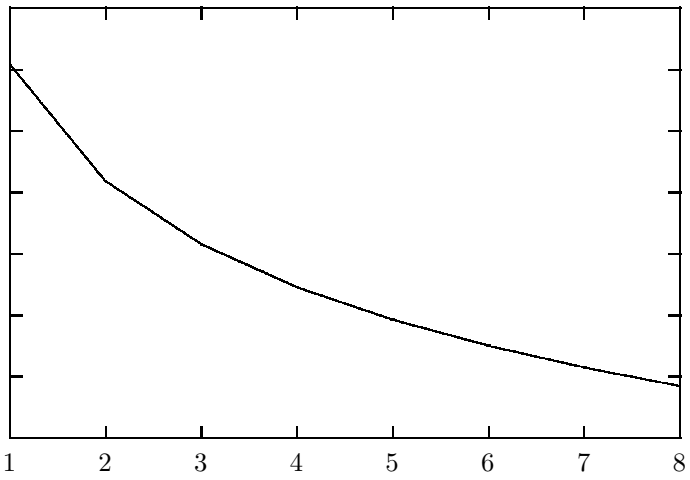

Figure 14. *Mean number of packets being lost per queue of a d-replicated buffered Banyan network versus the parameter d (logarithmic scale)*

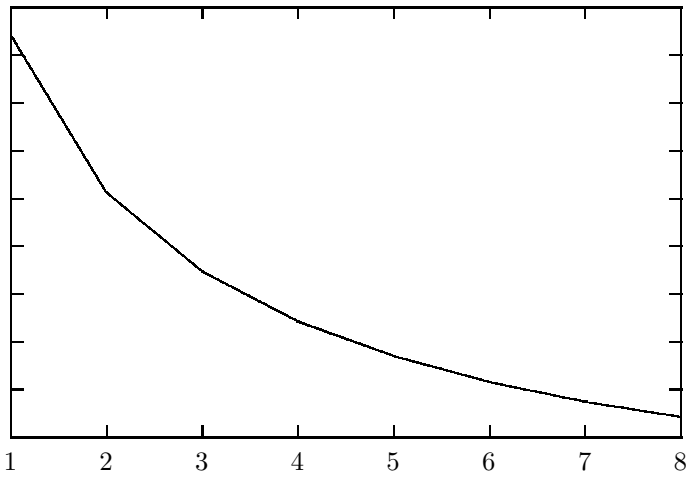

Figure 15. *Mean number of packets being lost per queue times cost of a d-replicated buffered Banyan network versus the parameter d (logarithmic scale)*

5. Conclusions — Future work

In this paper we have concentrated to the analysis of various types and extensions of Banyan network both from performance and cost of implementing in a single chip point of view. We have found that using small SEs is better from a cost/performance perspective, that replicated networks can decrease substantially the probability of packet loss without increasing the cost too much and that usually an increase in performance results in a bigger increase in cost. Also, we have found that in most cases it is possible to increase the performance without increasing the cost by many factors, and this happens when the value of the parameter, which increases the performance, is below a threshold. If the designer keeps these rules, then the design of low-cost switching elements and generally ATM Switches can happen without reducing the performance too much.

Even though we have examined the impact of various parameters of Banyan networks and variations of them from a cost/performance point of view, we feel that a lot of work must be done to understand the strengths and the weaknesses of these networks and their applicability as building blocks for large ATM switches. A very interesting area is the analysis of the behavior of these networks, again from a cost/performance point of view, under more general traffic assumptions (such as traffic with a high degree of correlation which is typical to multimedia applications). Also, the behavior of these networks when multicasting is used must be studied. Another interesting area is to find analytical results for the effect of speedup in buffered and unbuffered banyan networks. This will help us verify the conjectures made in the pre-

vious section. Another parameter which must computed analytically is $m\,(k)$, which is the decrease in buffers needed when k input streams use the same buffers.

6. References

[AHM 88] AHMADI H., DENZEL W. E., "A Survey of Modern High-Performance Switching Techniques", *IEEE J. Selected Areas Commun.*, vol. 7 no. 7, September 1989, pp. 1091–1103.

[BAT 68] BATCHER K. E., "Sorting networks and their applications", *Proc. Spring Joint Comput. Conf. AFIPS*, 1968, pp. 307–314.

[BOU 98] BOURAS C., GAROFALAKIS J., SPIRAKIS P., TRIANTAFILLOU V., "An analytical performance model for multistage interconnection networks with finite, infinite and zero length buffers",*Performance Evaluation*, vol. 34, 1998, pp. 169–182.

[CHE 91] CHEN J. S.-C., STERN T. E., "Throughput Analysis, Optimal Buffer Allocation, and Traffic Imbalance Study of a Generic Nonblocking Packet Switch", *IEEE J. Selected Areas Commun.*, vol. 9 no. 3, April 1991, pp. 439–449.

[COP 93] COPPO P., D'AMBROSIO M., MELEN R., "Optimal Cost/Performance Design of ATM Switches", *IEEE/ACM Transactions on Networking*, vol. 1 no. 5, October 1993.

[DIA 81] DIAS D. M., ROBERT JUMP R., "Analysis and Simulation of Buffered Delta Networks", *IEEE Trans. Comput.*, vol C-30 no. 4, April 1981, pp. 273–282.

[FRA 81] FRANKLIN M. A., "VLSI Performance Comparison of Banyan and Crossbar Communications Networks", *IEEE Trans. on Computers*, vol. C-30 no. 4, April 1981.

[GOK 73] GOKE L. R., LIPOVSKI G. J., "Banyan Networks for partitioning multiprocessor systems", *Proc. 1st Annu. Int. Symp. Comput. Architectures*, December 1973, pp. 21–28.

[KAR 87] KAROL M. J., HLUCHYJ M. G., MORGAN S. P., "Input Versus Output Queueing on a Space-Division Packet Switch", *IEEE Trans. Commun.*, vol. COM-35 no. 12, Dec. 1987, pp. 1347–1356.

[KRU 83] KRUSKAL C. P., SNIR M., "The Performance of Multistage Interconnection Networks for Multiprocessors", *IEEE Trans. Comput.*, vol. C-32 no. 12, December 1983, pp. 1091–1098.

[LEA 90] LEA C.-T., "Multi-Log_2N Networks and Their Applications in High-Speed Electronic and Photonic Switching Systems", *IEEE Trans. Commun.*, vol. 38 no. 10, October 1990, pp. 1740–1749.

[OIE 89] OIE J., MURATA M., KUBOTA K., MIYAHARA H., "Effect of speedup in nonblocking packet switch", *IEEE International Conference on Communications '89*, June 1989, pp. 410–414.

[PAT 79] PATEL J. H., "Processor-memory interconnections for multiprocessors", *Proc. 6th Annu. Int. Symp. Comput. Architecture*, April 1979, pp. 168–177.

[PAT 81] PATEL J. H., "Performance of processor-memory interconnections for multiprocessors", *IEEE Trans. Comput.*, vol. C-30, October 1981, pp. 771–780.

[PRY 95] DE PRYCKER M., *Asynchronous Transfer Mode: Solution for Broadband ISDN*, Third Edition, Prentice-Hall, 1995.