Distributed routing protocols for ATM extended banyan networks

Martino De Marco and Achille Pattavina
Dept. of Electronics and Information
Politecnico di Milano/CEFRIEL
20133 Milano, Italy

Abstract

It is well known that a multistage banyan network, which is a single-path blocking structure, becomes rearrangeable non-blocking in a circuit switching environment if the number of its stages is increased so as to obtain a Benes network. Banyan networks, provided with a shared queue in each switching element, have often been proposed as the core of an interconnection network for an ATM packet switching environment. In this scenario, if the classical interstage backpressure protocols are used, adding stages to a banyan network can even degrade the banyan network performance, in spite of the multipath capability given by the additional stages. A class of new simple interstage protocols is here defined to operate in the added stages of the banyan network so that a sort of sharing of the queueing capability in each added stage is accomplished. Large improvements in the traffic performance of these extended banyan networks are obtained, especially in the region of offered loads providing a low packet loss probability.

1 Introduction

The use of banyan networks, that is multistage arrangements of small switching elements (SEs) providing one path per I/O couple, as the core of interconnection networks for ATM packet switching has been investigated for several years. As is well known, their feature of packet self-routing through each SE makes the banyan network a very powerful routing device with a high degree of parallel processing (all the SEs in a stage route concurrently the packets received) that is compatible with the very high packet arrival rates typical of ATM networks.

Owing to the blocking nature of these networks caused by internal conflicts (different I/O paths can share one or more interstage links), basically two different approaches have been proposed to obtain an interconnection network with satisfying traffic performance: adopting a sorting network in front of the unbuffered banyan network, or providing the SEs in the banyan network with buffers. In the former case the overall network, referred to as Batcher-banyan network, becomes non-blocking; however additional devices are required to guarantee the absence of external conflicts, that is the event of two or more packets addressing the same network output [1]. In the latter case both internal and external conflicts are avoided by the buffers internal to each SE. A buffer can be associated to each inlet [2] or outlet of the SEs [3], or can be shared among all the SE inlets and outlets [4].

Performance studies of banyan networks with these buffer arrangements [5], [6] show that the head-of-the-line blocking phenomenon severely limits the throughput performance of input queuing, whereas output and shared queuing are characterized by an optimal throughput-delay performance. Shared queuing however gives better packet loss performance for the same SE queueing capacity owing to its complete buffer sharing. The drawback of shared queuing compared to output queuing is the concurrent read operation of the shared buffer required to feed all the SE outlets. Since the advances in VLSI technology we have witnessed in the last years have shown that even 32 × 32 SEs with a shared buffer can be implemented [7], our attention will be focused only on this kind of queuing for the SE.

As already mentioned, a banyan network, which includes n stages of SEs, provides a single path per inlet/outlet pair. An extended banyan network is obtained by adding extra stages to the basic banyan topology, so that multiple I/O paths are provided. It is well known that by adding n − 1 stages to the baseline banyan topology the resulting network becomes rearrangeable non-blocking (it is actually a Benes network) by applying a suitable centralized routing
scheme in the added stages. Thus we can say that adding stages to a banyan network improves the performance of a circuit-switching network, where centralized routing scheme are likely to be applied.

Our aim here is to study how an extended banyan network behaves when loaded by a packet traffic in an ATM environment. Now several packets entering the network in a slot can address the same network outlet and centralized routing schemes cannot be used as incompatible with the very high packet arrival rate. Thus a packet routing that is either random or alternates deterministically between the SE outlets is in general used in the SEs of the added stages [8] and SEs are provided with a shared queue to minimize packet losses. Surprisingly enough, unlike a circuit switching network, adding extra stages to a banyan network with shared queueing in an ATM environment does not improve and sometimes degrades (depending on the internal protocols) the network loss performance. Thus a Benes network gives in general a worse packet loss performance than a less expensive banyan network. Our aim here is the definition of distributed protocols, compatible with the ATM packet arrival rate, to control the packet transfer in the added stages, so as to take full advantage of the added queueing capability compared to a banyan network. We will show that these protocols provide a loss performance that improves as the number of stages grows, so that a Benes network outperforms a banyan network.

In the following of the paper, Section 2 first describes formally an extended banyan network together with the assumed traffic model. Section 3 shows the network behaviour under random routing in the added stages, whereas the new distributed protocols are described in Section 4. Finally Section 5 provides the network performance evaluation of the proposed distributed protocols.

2 Network and traffic model

2.1 Extended banyan networks

A banyan network includes $N$ inlets and $N$ outlets interconnected through $n$ stages of $b \times b$ switching elements (SEs), where $n = \log_b N$, numbered 1 to $n$. This structure is such that in a banyan network only one path exists between any inlet and outlet. Different topologies of banyan networks have been defined among which a functional equivalence applies [9], meaning that one topology can be obtained from the other by applying a suitable permutation at the inlets and/or outlets of the network. It is well known that the ATM traffic performance of a banyan network does not depend on the specific network topology, unless very particular traffic patterns are considered. Thus in the following description we will consider the reverse baseline network as the basic banyan topology.

In order to increase the number of I/O paths, one or more stages are added in front of the network by means of the mirror imaging procedure described in Ref. [10], thus obtaining an extended banyan network. An extended banyan network with $d$ additional stages ($1 \leq d \leq n - 1$) is a network with $n_d = d + n$ stages in which the last $n$ stages represent the basic banyan network and the $d$ first stages are obtained as the mirror image of the last $d$ stages of the basic banyan network including the permutation that precedes stage $n - d$. An example of application of such mirror imaging procedure to a reverse baseline banyan network with $d = 2$ is shown in Fig. 1 for $N = 16$, $b = 2$. For each added stage the number of I/O paths increases by a factor $b$. Note that an extended network with $d = n - 1$ added stages becomes the well-known Benes network. According to Ref. [11], we refer to the last $n$ network stages as the routing stages and the remaining $d$ first stages as the distribution stages.

A $b \times b$ SE is provided with a shared buffer of capacity $B$ cells that is shared by all the SE inlets to hold cells independent of their destination. The buffer is said to include $b$ logical queues, each holding the packets addressed to a specific SE outlet. The SE operation consists in storing the received cell in the logical queue whose associated SE outlet reaches the addressed network outlet. In each slot the SE transmits downstream as many packets as the number of its non-empty logical queues and at the same time stores all the packets received from the upstream SEs within the available buffer capacity. We assume that the packets transmitted downstream make available in the same slot the buffer space they are holding in the SE to packets

Figure 1: Extended banyan network ($N = 16$, $d = 2$).
that can be received from upstream SEs. Such kind of operation corresponds to assuming that each slot is subdivided into two subslots, the first reserved to packet transmissions, the second to packet receptions.

Internal protocols can be adopted in the networks, in that there is an exchange of signalling information between adjacent stages to enable the cell transmission by a SE. We consider the following operation modes.

Queue Loss (QL): no internal protocol is applied to the network, so that each non-empty logical queue always transmits downstream a packet, independent of the buffer occupancy state of the destination SE.

Backpressure (BP): an exchange of information takes place between adjacent stages; each SE of stage $i$ transmits a request signal to the downstream SEs of stage $i+1$ associated to a non-empty logical queue. Each downstream SE enables through an acknowledgement signal to the transmission only a number of cells not exceeding the number of idle places $I$ of its own buffer. There are two kinds of backpressure mechanisms:

- Local Backpressure (LBP), in which $I$ is equal to the number of idle positions in the buffer at the end of the preceding slot;
- Global Backpressure (GBP), in which $I$ is the number of idle positions in the buffer at the end of the preceding slot increased by the number of positions that are going to be freed by downstream cell transmissions in the first half of the current slot.

If the buffer overflows, discarding takes place for the packets in excess which are selected randomly among the received packets. If no internal protocol is applied (QL mode), all the SEs of the network can lose packets, whereas only the SE buffers of the first network stage can overflow with backpressure protocols.

It is important to emphasize that the intuitive better performance provided by the GBP protocol over the LBP protocol is paid in terms of more stringent hardware requirements. In fact, the exchange of signalling information can take place concurrently between adjacent stages with local backpressure. Thus the required network internal speed is equal to $C(1 + b_{rt})$ bit/s, if $C$ is the inlet/outlet bit rate and $b_{rt}$ is the ratio of the total number of bits of a request and an acknowledgement message to the cell size (propagation times are assumed negligible). Under global backpressure only the request signals can be sent downstream concurrently, while the acknowledgement signals have to backpropagate serially from the last stage to the first stage. So the internal speed is now $C(1 + b_{rt} + (n_1 - 1)b_{ack})$, in which $b_{rt}$ and $b_{ack}$ are the ratios of the lengths of a request and an acknowledgement message to the cell size. Therefore the internal speed-up with GBP is higher than with LBP and grows with the network size.

2.2 Input traffic characterization

The traffic offered to the network is assumed correlated according to the following hypotheses:

- each network inlet is assumed to be loaded by a bursty source, represented by a two-state On/Off source: in the silent period (off state) the source is idle, whereas in the active period (on state) the inlet always receives one cell;
- all the cells belonging to the same burst address the same network outlet;
- cell arrival events at different inlets are mutually independent;
- each burst is destined to one of the network outlets with equal probability.

The two-state Markov chain modeling the external traffic source is specified through its state transition probabilities: $p_{10}$ ($p_{01}$) is the probability that a source moves from state ON (OFF) to state OFF (ON).

The active and silent period of the traffic sources alternate in time: at the end of each slot one source is able to switch its state. The length of the busy period, i.e. the number of consecutive cell arrivals belonging to the same burst, is geometrically distributed with mean value

$$L_b = \frac{1}{p_{10}}$$

The average load $p$ ($0 \leq p \leq 1$), defined as the probability that a network inlet receives a cell in a slot, and the average burst length $L_b$ describe univocally the two-state bursty source. Note that the memoryless Bernoulli source is a subcase of the two-state source with

$$p_{01} = p_{11} = p$$
$$p_{10} = p_{00} = 1 - p$$

Following Ref. [12], we define the burstness factor $k_b$ as the ratio between the average length of a busy period with bursty traffic and the average length of a busy period in the Bernoulli arrival case. Thus we describe here an On/Off bursty source with the two
parameters \( p \) and \( k_b \), so that the transition probabilities of the two-state On/Off model are

\[
\begin{align*}
    p_{01} &= \frac{p}{k_b} \\
    p_{10} &= \frac{1-p}{k_b}
\end{align*}
\]

Hence, the assumption \( k_b = 1 \) means offering the so-called "random traffic" to the network, i.e., each network inlet is loaded by a Bernoulli source.

## 3 Random routing

Adding stages to a banyan network so as to obtain an extended banyan network has two major consequences: the larger number of network stages that one cell must cross and the availability of multiple paths to the addressed network outlet. In order to evaluate the impact of these factors on the network performance, the extended banyan network is assumed to operate according to the following mode.

- Each network inlet assigns to the arriving cells one I/O path, with a random choice among all the possible \( b^d \) I/O paths; the cell crossing the interconnection network is preceded by a path tag, that is the binary information of the assigned cell routing, specifying the SE outlet stage by stage.
- The multipath network operates either without internal protocols (QL) or under the control of a backpressure mechanism, as described in the preceding section.

- The traffic offered to the network is uncorrelated and balanced \((k_b = 1)\): each inlet receives a cell with probability \( p \) in each time slot.

Computer simulation is used here and in the following sections to evaluate the network traffic performance. The 95% confidence interval of the cell loss probabilities in the figures, not shown for readability, is never larger than \( \pm 40\% \) of the plotted values. We consider a \( 16 \times 16 \) network with \( 2 \times 2 \) SEs and a shared buffer capacity \( B = 8 \). Starting from the banyan network \((d = 0)\), we add one stage at a time until we end with the Benes network \((d = 3)\). The cell loss probability of all these networks is now evaluated.

Initially, no internal protocols are applied to the network (QL operation) and the numerical results are plotted in Figure 2. The cell loss probability increases with the number of distribution stages, in spite of the additional storage capability carried by the \( d \) added stages.

We now consider internal backpressure and intentionally select the GBP protocol which, although much more demanding in terms of internal speed-up, is expected to provide the best performance. The comparison between basic banyan and extended banyan networks is shown in Figure 3 that plots the cell loss probability for the different \( 16 \times 16 \) multipath networks up to the Benes structure \( d = 3 \). The packet loss probability remains roughly independent of the number of added stages, so that under random routing in the distribution stages the more expensive Benes network is equivalent to the basic banyan network from the loss performance standpoint. Since local backpressure cannot improve the performance of the extended banyan network compared to global backpressure, we can conclude that the added storage capacity carried by the
added stages is substantially useless (or dangerous) under random routing in the distribution stages.

4 Distributed Protocols

In order to improve the performance of the multipath networks, efficient routing algorithms are required. As already observed, a cell in the distribution stage can be routed to any SE outlet, since at least one path to the addressed network outlet crosses each SE outlet. Only in the last n routing stages a single path exists to reach the destination outlet.

We have shown that classical backpressure protocols, although simple, are not able to fully exploit the multipath and storage capability in the added stages: more information must be exchanged within the network. In fact backpressure simply prevents buffer overflow by denying the transmission permit to one or more upstream SEs. What happens is that the random selection of these upstream SEs prevents an optimal use of the overall buffer capacity. On the contrary, if each SE knows the buffer occupancy of its downstream SEs and is capable of coordinating its actions with the SEs in the same stage feeding the same downstream SEs, then a full exploitation of the multiple paths to the required network outlet is possible.

Some distributed protocols based on signalling and reservation are here defined along this line that accomplish the exchange of information about the buffer occupancy between adjacent stages and within the same stage. A basic property of the banyan interconnection network let us apply the signalling protocols to specified subsets of SEs, making it easy the exchange of information between stages. Therefore cells can be routed in the distribution stages so as to maximize the number of packets transmitted downstream in each slot without buffer overflow. In the routing stages a local backpressure protocol is still used.

4.1 Buddy property

In all the topologies of an \( N \times N \) banyan network with \( b \times k \) SEs, a subset of \( b \) SEs has only interconnection with a subset of \( b \) SEs of the downstream stage. We call any such subset of SEs pool of SEs fully interconnected (PFI): PFI\(_{tx}\) denotes the subset of SEs of the upstream stage (that transmits the cells), whereas PFI\(_{rx}\) denotes the subset of SEs of the downstream stage (that receives the cells). Figure 4 shows the PFI in the distribution stages of a 16 \times 16 network with \( b = 2 \) and \( d = 2 \). In the following, a SE of the PFI\(_{tx}\) and a SE of the PFI\(_{rx}\) are referred to as SE\(_{tx}\) and SE\(_{rx}\), respectively.

Figure 4: Example of buddy property.

This important property, know as buddy property, sets a limit to the number of SEs that exchange information with each other about their buffer state: each SE needs only to inform the other SEs in the same PFI.

4.2 Signalling between PFI\(_{tx}\) and PFI\(_{rx}\)

A generic SE of a PFI must transmit its buffer occupancy state to all the SE of the PFI\(_{tx}\) and PFI\(_{rx}\) at the beginning of each time slot. The exchange of such information between a SE\(_{tx}\) and the other SEs in the same PFI\(_{tx}\) would require additional links to connect SEs in the same stage. Rather, we propose a signalling protocol within each PFI that does not require such additional links.

Figure 5 represents the state of the PFI\(_{tx}\) and the connected PFI\(_{rx}\) in a network with \( b = 4 \): the SEs in each PFI are numbered from 0 to \( b-1 \). \( B_k \) is defined as the number of cells that the SE\(_{tx}\)#\( k \) holds in its buffer, and \( I_j \) as the number of idle places in the buffer of the SE\(_{rx}\)#\( j \). The exchange of information between SEs in the same PFI takes place as represented in Figure 6; each SE accomplishes the following operations.

(a) At the end of each time slot, the SE\(_{tx}\)#\( k \) sends the information \( B_k \) to all the SE of the connected PFI\(_{rx}\).

(b) Each SE\(_{rx}\)#\( j \) receives such \( B_1, \ldots, B_b \) from all the SE of the PFI\(_{tx}\).
(c) The SE\textsubscript{tx}\#j collects in a feedback array all the received information, adding its own buffer state \(I_j\), and sends such array of \(b + 1\) elements to all the upstream SEs in the connected PFI\textsubscript{tx}.

(d) The SE\textsubscript{tx}\#k receives all the feedback arrays building two arrays: \((B_1, \ldots, B_R)\) called Tx-array and \((I_1, \ldots, I_R)\) referred as Rx-array.

In this way, each SE\textsubscript{tx} knows the state of any other SEs in the same PFI\textsubscript{tx} and of all the SEs in the connected PFI\textsubscript{tx}.

The network speed-up needed to exchange the signalling information in the proposed algorithms is now evaluated (see also Section 2.1). Consider that the signal transmission can take place concurrently in the interstage pattern between distribution stages, as enabled by our proposed protocols, and also between routing stages as inherent in the LBP protocol here adopted. For the sake of simplicity we assume: (i) the processing time in the SEs is negligible; (ii) any information about the buffer state requires one byte. Therefore a total number of \(b + 2\) bytes is exchanged between any couple of SEs in the distribution stages and the internal network speed is equal to \(C[1+(b+2)/53]\) bit/s (an ATM cell includes 53 bytes). In the case of 2 \(\times\) 2 SEs the network speed-up is equal to about 7\%, whereas it is about 10\% with 4 \(\times\) 4 SEs.

### 4.3 Reservation sequence

Once the information about the buffer state in the SEs of the same PFI is available, the cells that can be transmitted must be selected. To this extent, a reservation protocol is defined in which a SE\textsubscript{tx} books an idle place into the buffer of a SE\textsubscript{rx}. A reservation sequence must be established to guarantee that all the SE\textsubscript{tx} book in turn the only idle places in the PFI\textsubscript{tx} buffers.

We consider the two tokens TK\textsubscript{tx} and TK\textsubscript{rx}: the SE\textsubscript{tx} holding the TK\textsubscript{tx} can book an idle place in the SE\textsubscript{rx} holding the TK\textsubscript{rx}. The passing sequences and the starting assignment of the two tokens are defined by either of the two following modes.

The Round Robin (rr) mode is very simple: the tokens are assigned at the beginning of each time slot to each SE of the PFI in a circular fashion so that a token comes back to the same SE after \(b\) slots. The reservation sequences are determined by the SE topological position (SE index).

A more complex reservation sequence is accomplished by the Queue Sorting (qs) mode. It consists of sorting the SE in each PFI: the sequence of the SE\textsubscript{tx} in the PFI\textsubscript{tx} is built according to decreasing buffer occupancies, whereas the sequence of the SE\textsubscript{rx} in the PFI\textsubscript{tx} according to the decreasing number of idle places in the buffer. In such way the most busy SE\textsubscript{tx} get a priority in booking cell transmissions, whereas the least busy SE\textsubscript{tx} get a priority in booking cell receptions.

### 4.4 Reservation Protocols

Once we have established the sequence of visits of the two tokens for each slot, we must now select the booking operation of the SE\textsubscript{tx} holding the token TK\textsubscript{tx}. Since all the SE\textsubscript{tx} adopt the same reservation procedure, each SE\textsubscript{tx} knows a priori the reservations al-
Figure 7: Step by step execution of algorithm TPR-rr.

5 Performance evaluation

In this section the traffic performance given by the proposed reservation protocols are presented with reference to multipath networks. The network cell loss probability is first evaluated for random traffic ($k_b = 1$) and then for bursty traffic ($k_b > 1$); finally same considerations about the dimensioning of the resequencing buffer size are given.

In order to compare the performance of different multipath networks, two operation modes have been adopted:

- **Increasing Network Memory (l.NM):** starting from a banyan network, a variable number of stages is added whose SEs have the same buffer as in the initial network.
- **Constant Network Memory (c.NM):** starting from a Benes network, the distribution stages are removed one by one by keeping the same total network buffer capacity.

In the c.NM mode, for example, a $16 \times 16$ Benes network with $b = 2$ and $B = 16$ becomes a network with 6, 5, 4 stages with a SE buffer $B' = 18$, 22, 28, respectively. All of those networks will be denoted as having an equivalent buffer capacity per SE $N_{se} = 16$ as the initial Benes network. The performance of these multipath networks will always be compared with the loss figure of a non-extended banyan network operating under LBP protocol.

5.1 Random input traffic

In Section 3 the loss performance of $16 \times 16$ extended banyan networks, with $b = 2$, $B = 8$ has been shown under random routing. We have seen how the network performance becomes worse when distribution stages are added, even though the total buffer capability of the network increases.
Figures 8–9 show the loss performance for the same network in which now the Round Robin sequence with both TPR and TGR protocols are used to control the distribution stages. It is seen that both protocols improve significantly the network performance especially for lower offered loads. Even larger improvements are provided by the TGR-qs protocol (Figure 10) owing to its optimized sharing of the buffer in each stage implied by the sorting of the buffer contents in the same stage. Figure 11 compares the loss performance of different protocols (QL, GBP, TPR-rr, TGR-rr, and TGR-qs), for p = 0.75 and a growing number d of distribution stages. As observed in Section 3, the random routing protocols (QL, GBP) make worse the network performance as d increases, since the additional loss due to the larger number of network stages prevails over the multipath capability. On other hand, the proposed reservation protocols improve the network performance owing to their dynamic sharing of the queueing resources in the distribution stages. As one might expect, the TGR-qs protocol provides the best performance. Therefore only this operation mode will be considered in the following discussion. We point out however that this performance improvement is obtained at the expense of a growth of the total buffer capacity in the network, which is equal to $d \cdot N/2 \cdot B$ cells. For example with offered load $p = 0.75$ the cell loss probability improves by more three order of magnitude from the banyan to the Benes network, whereas the network buffer capacity increase is equal to 192 cells for the Benes network, that is equivalent to about 10 Kbytes.

In order to evaluate the improvement of the network performance by disregarding the effects of the additional buffer capacity, the c.NM operation mode
Figure 12: Loss performance for TGR-q$_5$ (c.NM).

Figure 13: Cell loss under bursty traffic (l.NM).

is now applied to the $16 \times 16$ network with $B_{eq} = 8$
under the control of the TGR-q$_5$ protocol. The internal SE buffer sizes are 9, 11, 14 cells for $d = 2, 1, 0$, respectively. Figure 12 shows the cell loss performance of these networks. It is interesting to note that in the range of offered loads providing very low loss probabilities (those we are interested in for ATM applications) all the extended banyan networks show a smaller loss probability.

5.2 Bursty input traffic

We consider the same network analyzed before which are now loaded by a bursty traffic with $k_4 = 4$. The loss performance of the $16 \times 16$ networks are evaluated with l.NM and c.NM modes. The numerical results regarding a internal buffer size $B = 16$ (or $B_{eq} = 16$) cells are plotted in Figures 13-14, whereas

Figure 14: Cell loss under bursty traffic (c.NM).

Figure 15: Cell loss under bursty traffic (l.NM).

Figure 16: Cell loss under bursty traffic (c.NM).
Figures 15–16 show the loss performance for $B = 32$ (or $B_\text{eq} = 32$): the TGR-qs protocols is always applied. Again the extended banyan networks reveal a better behaviour than the basic network and such improvement is even larger than in the case of random input traffic. In fact, the distribution stages act as a "randomization network" for the cells belonging to the same burst so that the statistical correlation between consecutive cells crossing the following routing stages is decreased. By comparing the two cases $B = 16$ and $B = 32$, we note that the reservation protocols are more effective for smaller buffer sizes.

6 Conclusions

A class of distributed protocols has been defined here to accomplish as much as possible the sharing of queueing resources in the distribution stages of extended banyan networks. These protocols have a very limited cost in terms of additional processing required in each switching element and of internal speed-up due to the exchange of signaling information between adjacent stages slot by slot. We have shown that these distributed protocols under random loading accomplish a decrease of the network loss probability in the range of offered load levels where the packet loss is lower than a given threshold. Such result, which is of relevant interest in an ATM environment where very low loss probabilities are expected, is not only due to the additional buffer carried by the added stages. In fact, it has been shown that an extended banyan network performs better than a single-path banyan network, even if the two networks have the same total queueing capacity. Even larger performance improvements are obtained by the distributed protocols under a more realistic bursty traffic pattern.

References


11a.3.10