Contribution to: GLOBECOM'88, Hollywood Florida,
November 28 - December 1, 1988

Buffering Concepts for ATM Switching Networks

Erwin P. Rathgeb, Thomas H. Theimer, Manfred N. Huber

University of Stuttgart
Institute of Communications Switching and Data Technics
Seidenstraße 36, D-7000 Stuttgart 1

Abstract

High speed switching nodes with high capacity suited for the Asynchronous Transfer Mode (ATM) require a highly parallel architecture in order to meet the requirements with respect to throughput and transfer delay. Therefore, multistage networks of the Banyan type seem to be a possible choice for the design of the switching fabric, because they can be operated without central control for the single data packets. To avoid congestion and packet loss in these single path networks, some buffering has to be done. In this paper we analyze different buffering strategies, where the buffers are located in the individual switching elements. The influence of the different strategies on the performance of the network is evaluated under equivalent conditions to allow a comparison of the results.

1 Introduction

A major design goal for future Integrated Broadband Communication Networks (IBCNs) is to provide efficient support for all the different services envisaged today with their various, sometimes conflicting requirements. In addition to that, they should also offer enough flexibility to satisfy the needs and service requirements that may arise in the future in order to justify the vast investments needed to introduce these networks.

To fulfill the requirements mentioned above, packet (or cell) based *Asynchronous Transfer Mode (ATM)* networks seem to be a suitable choice.

In order to support real-time critical services like voice or video in a packet oriented environment, the ATM switching nodes must provide a considerable capacity that may only be achieved using a distributed and highly parallel architecture. This applies especially to the interconnection networks used as switching fabric. Therefore, in several publications [4,11,15,16] interconnection networks of the *Banyan* type, which is known from multiprocessor computer applications, have been proposed.

To avoid the problem of internal congestion in these single path networks and to increase their maximum throughput there exists a variety of choices, like e.g. internal speed-up, the application of multiple switching planes or the addition of supplementary stages to the network. Another possibility is the insertion of buffers, either at the inputs or outputs of the switching network or within the individual switching elements. In this paper we shall discuss some variations of the latter strategy and compare them with respect to their performance characteristics.

2 Banyan Networks

The property that defines the Banyan class is, that there is exactly one path between any input and any output of the switching network [5]. Adding some constraints to this rather general classification, the class of Delta Networks has been defined [12]. In this class, a network with n stages can be constructed from several $(n-1)$-stage networks by adding an extra stage and connecting it to the $(n-1)$-stage networks in a regular manner (recursive construction). A network built of switching elements which have the same number of inputs and outputs (b×b switching elements) is called Delta-b Network. Bidelta Networks (Bidirectional Delta Networks) have the special topological property that they remain Delta Networks even if the network input links are interpreted as output links and vice versa [3].

The use of the self routing (digit controlled routing) mechanism is also included in the definition of Delta networks [12]. This mechanism uses the fact that there is a unique path between any input/output pair which is completely specified by the running number of the output. Based on this number of base b with n digits according to the number of stages, a packet can find its way through the network without central control. This simple routing algorithm can easily be implemented in hardware, because it only involves the analysis of one digit per stage.

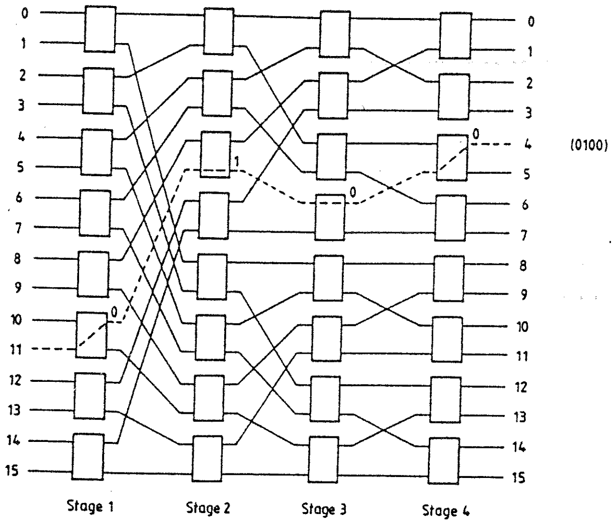

Figure 1: Delta-2 network with 4 stages and 16 input terminals

An example of a Delta-2 Network with 4 stages and a *Baseline* structure [17] is shown in Figure 1. The dashed line indicates the path of a packet from input 11 to output 4, which has the binary address '0100'. Other structures falling into the Bidelta class are e.g. Reverse Baseline [17], Flip [1], Omega [10], Indirect Binary n-Cube [13], Bit-Reversal and Inverse Bit-Reversal networks. They all have the same basic topology and can be transformed into one another by relabelling the links and the switching elements.

3 Buffering Concepts

There are various possibilities for implementing the switching elements which are the basic modules of the switch fabric. Each switching element of a buffered Delta Network consists of a switching matrix and several buffers to store the packets. In this section we will focus on the position of the buffers with respect to the switching matrix.

The buffers of a switching element can be located in front of the switching matrix, behind the switching matrix or at each crosspoint of the matrix. The performance of a single switching element with infinite buffers at the inputs or at the outputs of the switch has been investigated by Karol, Hluchyi and Morgan [7]. As the hardware complexity of the network is mainly dominated by the buffers, we restrict our considerations to switching elements with finite buffer capacities.

If a network is constructed using such switching elements, packet loss would occur if a buffer is full and a packet arrives. Therefore, a *backpressure mechanism* is employed to avoid packet loss within the switch fabric. If a buffer in stage n is full, all switching elements of stage $n-1$ which are trying to send packets to this buffer are informed about the blocking situation. This mechanism might propagate through the preceeding stages to the input of the switch fabric, where blocked packets are stored in a larger buffer included in the packet processor unit.

Kumar and Jump [9] have studied the performance of buffered Delta Networks with finite capacity buffers located at the inputs of the switching elements or at the crosspoints of the switching matrix. However, their investigations were based on the assumption that a buffer is able to accept a packet only if there is enough space at the *beginning* of a clock cycle. Consequently, a buffer might be blocked although a packet is going to leave so that another packet could be accepted in the same clock cycle. As this assumption gradually decreases the maximum throughput of the network, our studies refer to a different backpressure mechanism which allows a full buffer to accept a packet if there would be enough space at the *end* of the clock cycle. Thus, even a full buffer is able to store a packet if one of the waiting packets can be forwarded simultaneously.

3.1 Switching Elements with Input Buffers

In a switching element with input buffers, each arriving packet is stored in a buffer corresponding to the input of the switching element. All these b buffers are interconnected to the outgoing lines via the switching matrix (Fig. 2). Using *first-in first-out* (FIFO) buffers, a collision occurs if there are packets in several buffers competing simultaneously for the same outgoing link. This disadvantage could be avoided by speeding up the switching matrix as well as the outgoing lines. However, this solution is impractical because the speed of the interconnection links must be multiplied in each stage of the network.

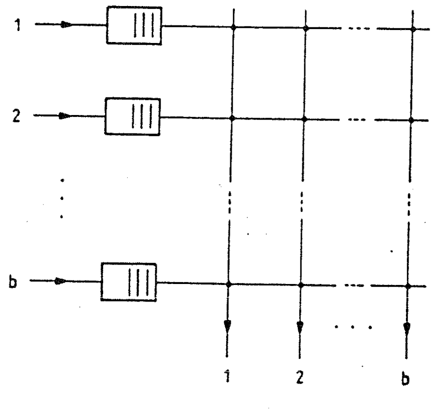

Figure 2: Switching Element with Input Buffers

Another possibility to solve the collision is a *random access* to the input buffers, which means that the FIFO discipline might be violated if the first packet in the queue is unable to move forward. In this case the next packet which is destined for an idle output line will be selected for transmission. Consequently, the total buffer capacity is logically subdivided in a load dependent manner into *b* single FIFO buffers.

The control of the switching matrix chooses one of the internal FIFO's per input buffer in such a way that the number of collisions is minimized and, if possible, the packet with the longest waiting time is transmitted first. However, this strategy needs a complex buffer access control.

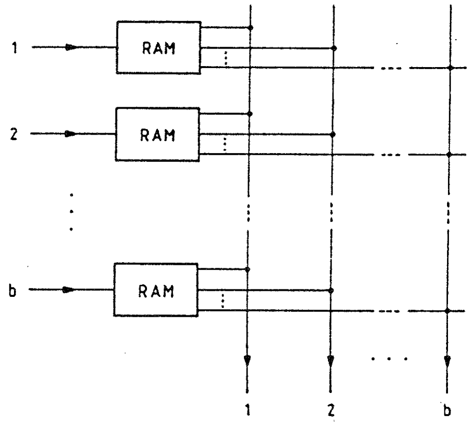

Figure 3: Enhanced Switching Element with Random Access and Multiple Buffer Ports

A further enhancement can be achieved by transferring more than one packet from one input buffer to different outgoing lines in the same clock cycle (Fig. 3). This requires a buffer with multiple output ports or a reduction of the buffer access time, but preserves the same speed on the input and output links [14].

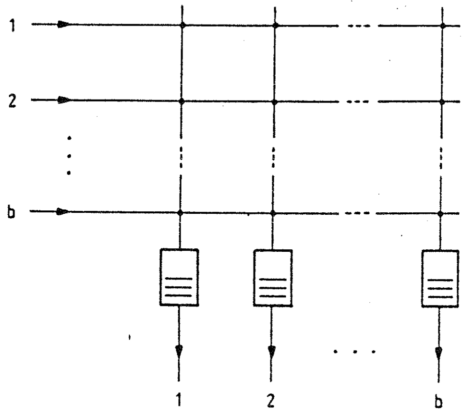

Figure 4: Switching Element with Output Buffers

3.2 Switching Elements with Output Buffers

Exchanging the position of the switching matrix with the buffers results in a switching element with output buffers (Fig. 4). If the buffers and the switching matrix are operating with the same speed as the incoming line, a collision occurs if several arriving packets are hunting simultaneously for the same output buffer. Preserving the speed of the links, this drawback can be compensated by a reduction of the buffer access time and by a speed-up of the switching matrix [7].

3.3 Butterfly Switch

Each input line of the Butterfly switching element [2] has one physically separate FIFO per output line, and the arriving packets are routed to the buffer corresponding to their destination output (Fig. 5). This switching element is similar to the enhanced switching element with input buffers with the difference that the buffers in the Butterfly element have a fixed length. If there are packets in more than one FIFO belonging to the same outgoing line, the control logic of this output has to choose one of these buffers to be served first.

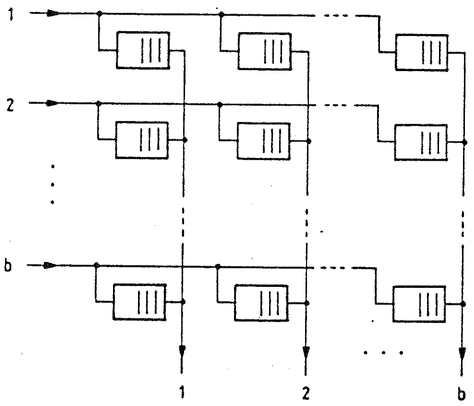

Figure 5: Butterfly Switch

3.4 Performance Comparison

In order to compare the performance of the buffering strategies described above under equivalent conditions, the following assumptions will be made :

1. The network is operated synchronously, that means the packets are transmitted only at the beginning of a time slot given by the packet clock.
2. All packets have a fixed length.
3. Packets arrive independently at the input links of the network and their destination adresses are distributed uniformly (at random) over all output links.
4. Each input link of the network carries the same traffic load.

5. There is no blocking at the output links of the network. This means that the packet processors at the outputs of the network are always able to accept a packet.

Each input of the network is preceeded by a packet processor which performs the protocol handling and generates the local packet labels. The arriving packets are stored in a buffer of the packet processor, which is essential for switching elements with output buffers and no speed-up, because otherwise many packets would be lost at the inputs of the first stage due to collisions. The influence of the size of this buffer on the probability of packet loss will be investigated later on.

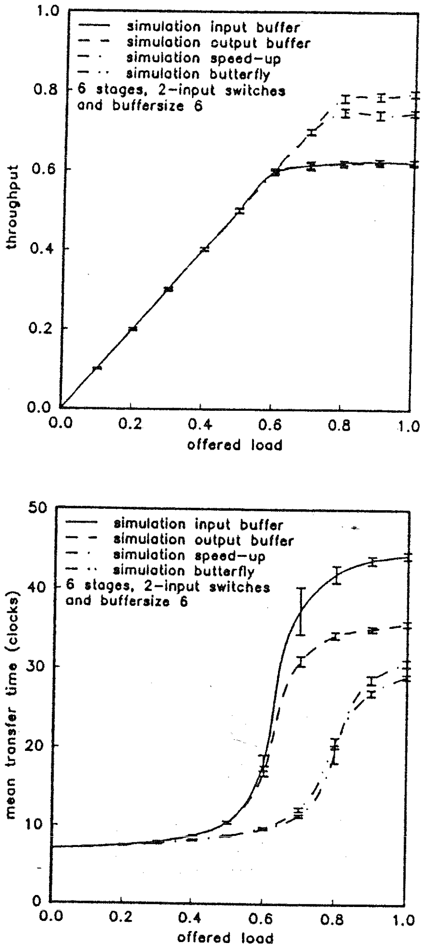

Figure 6: Throughput and delay for various types of switching elements

A very important performance criterion of the switch fabric is the maximum throughput, which is defined as the mean number of packets leaving the network in one clock cycle divided by the total number of network outputs. The maximum throughput depends on various parameters, e.g. the size of the switching elements and their internal buffersize [9,6]. The influence of different buffering strategies on throughput and delay of a Delta-2 Network with 6 stages and an internal buffersize of 6 is shown in figure 6. The results were obtained using simulation techniques and refer to a confidence level of 95%.

It must be pointed out, that the total amount of bufferspace is assumed to be identical for all buffering strategies. Therefore, the internal buffersize of the butterfly switching element is one half of the buffersize of the other types, because in the special case of a 2 × 2 switching element the butterfly architecture requires 4 buffers, whereas 2 buffers are sufficient for all other strategies. Under these conditions, the switching elements with output buffers and internal speed-up achieve the best performance with respect to throughput and delay.

The performance of the architectures 'butterfly' and 'output buffer with speed-up' is superior to the performance of the other strategies, because there are no internal collisions in the switching elements even if two packets are destined for the same output line. Each output buffer of the switching elements with internal speed-up is shared by both input lines in a load dependent manner. This allows more flexibility for the use of the buffers in comparison to the butterfly architecture, where each output buffer is subdivided into two equal sections (one buffer per input line).

It is interesting to note that the switching elements using input buffers and output buffers without speed-up achieve the same maximum throughput. This is due to the fact, that both architectures differ only in the location of the buffers with respect to the interconnection lines of the switching elements. Shifting the output buffers from the output lines of the switching elements to the inputs of the switching elements in the next stage according to the interconnection pattern between the stages leads to a network with buffers at the inputs of the switching elements.

The occurence of packet loss due to buffer overflow at the switching nodes is depending on the traffic load of the switch fabric. In order to minimize the probability of packet loss, the offered traffic load must be kept below the maximum throughput of the network. Especially for the efficient operation of ATM networks it is necessary to avoid any packet loss, because these networks require simplified protocols supporting only end-to-end error recovery mechanisms.

The size of the external buffer which is located in the packet processor has a considerable influence on the probability of packet loss at low and moderate traffic

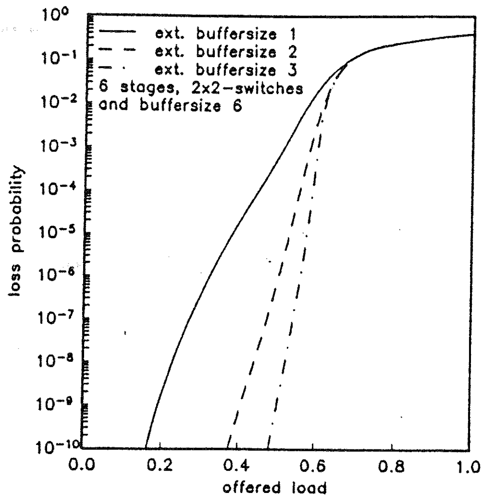

Figure 7: Loss probability for switching elements with input buffers

loads. Figure 7 shows that an external buffersize of 2 leads to a bit error rate (loss probability) of 10^{-6} at a traffic load of 50%. If the size of this buffer is extended to 3, the bit error rate decreases to 10^{-9}.

These results have been computed using analytical approximations, because simulation techniques are not applicable to yield probabilities in the order of 10^{-10}. A validation of the analytic results at high traffic loads by means of simulations showed a very high accuracy. Figure 7 refers to switching elements with input buffers, because at the moment this is the only buffering strategy for which the appropriate analysis techniques are available. The loss probabilities of the strategies 'butterfly' and 'output buffer with speed-up' will still be less than the depicted values.

4 Conclusion

We have studied various buffering strategies for switch fabrics of the Banyan type, where the buffers are located in the individual switching elements. The influence of the different strategies on the performance of the network has been evaluated under equivalent conditions to allow a comparison of the results and the packet loss probability for networks constructed of switching elements with input buffers has been analyzed.

The investigations indicate, that significant performance improvements may be achieved by applying advanced buffering concepts. The choice of the best buffering strategy depends on the size of the switching elements as well as on economical and technological constraints. Further work is needed to investigate the sensitivity of the results with respect to non-uniform traffic, which is likely to occur in a connection oriented ATM environment characterized by high bandwidth dynamics for the different services. Suitable analytical approximations are also required to overcome the limitations which are inherent to the simulation technique when large switch fabrics are investigated.

References

[1] Batcher K.E., "The Flip Network in STARAN", *Int. Conf. on Parallel Processing*, 1976.

[2] Brooks E. D., "A Butterfly Processor-Memory Interconnection for a Vector Processing Environment", *Parallel Computing 4 (1987)*, North-Holland.

[3] Dias D.M., Kumar M., "Packet Switching in nlogn Multistage Networks", *IEEE GLOBECOM'84*, Atlanta.

[4] Giorcelli S., Demichelis C., Giandonato G., Melen R., "Experimenting with Fast Packet Switching Techniques in First Generation ISDN Environment" *ISS'87*, Phoenix.

[5] Goke L.R., Lipovski G.J. "Banyan Networks for Partitioning Multiprocessor Systems", *First Annual Symposium on Computer Architecture*, 1973.

[6] Huber M. N., Rathgeb E. P., Theimer T. H., "Banyan Networks in an ATM-Environment", *ICCC'88*, Tel Aviv, 1988.

[7] Karol M. J., Hluchyi M., Morgan S. P., "Input Versus Output Queueing on a Space-Division Packet Switch", *IEEE GLOBECOM'86*, Houston.

[8] Kulzer J.J., Montgomery W.A., "Statistical Switching Architectures for Future Services", *ISS'84*, Florence.

[9] Kumar M., Jump J. R., "Performance Enhancement in Buffered Delta Networks Using Crossbar Switches and Multiple Links", *Journal of Parallel and Distributed Computing 1*, 1984.

[10] Lawrie D. H., "Access and Alignment of Data in an Array Processor", *IEEE Transactions on Computers*, Vol. C-24, No. 12, December 1975.

[11] Luderer G. W. R., Mansell J. J., Messerli E. J., Staehler R. E., Vaidya A. K., "Wideband Packet Technology for Switching Systems", *ISS'87*, Phoenix.

[12] Patel J.H., "Performance of Processor-Memory Interconnections for Multiprocessors", *IEEE Transactions on Computers*, Vol. C-30, No. 10, October 1981.

[13] Pease M. C., "The Indirect Binary n-Cube Microprocessor Array", *IEEE Transactions on Computers*, Vol. C-26, No. 5, May 1977.

[14] Schneider H., "Der Asynchrone Übermittlungs-Modus – Ein neues Multiplex- und Vermittlungs-Prinzip für das Breitband-ISDN", *Das ISDN in der Einführung*, Berlin 1988 (in German).

[15] Turner J.S., "Design of an Integrated Services Packet Network", *Ninth Data Communications Symposium*, ACM Sigcomm Computer Communication Review, Vol. 15, No. 4, September 1985.

[16] Turner J.S., "New Directions in Communications", *1986 Int. Zurich Seminar on Digital Communications*, Zurich.

[17] Wu C-L., Feng T-Y., "On a Class of Multistage Interconnection Networks", *IEEE Transactions on Computers*, Vol. C-29, No. 8, August 1980.