Design and Analysis of a Large-Scale Multicast Output Buffered ATM Switch

H. Jonathan Chao, Member, IEEE, and Byeong-Seog Choe

Abstract—We propose and analyze a recursive modular architecture for implementing a large-scale multicast output buffered ATM switch (MOBAS). A multicast knockout principle, an extension of the generalized knockout principle, is applied in constructing the MOBAS in order to reduce the hardware complexity (e.g., the number of switch elements and interconnection wires) by almost one order of magnitude. In our proposed switch architecture, four major functions of designing a multicast switch: cell replication, cell routing, cell contention resolution, and cell addressing, are all performed distributively so that a large switch size is achievable. The architecture of the MOBAS has a regular and uniform structure and, thus, has the advantages of: 1) easy expansion due to the modular structure, 2) high integration density for VLSI implementation, 3) relaxed synchronization for data and clock signals, and 4) building the center switch fabric (i.e., the multicast grouping network) with a single type of chip. A two-stage structure of the multicast output buffered ATM switch (MOBAS) is described. The performance of the switch fabric in cell loss probability is analyzed, and the numerical results are shown. We show that a switch designed to meet the performance requirement for unicast calls will also satisfy multicast calls' performance. A 16 × 16 ATM crosspoint switch chip based on the proposed architecture has been implemented using CMOS 2-μm technology and tested to operate correctly.

I. INTRODUCTION

BROADBAND INTEGRATED SERVICES digital networks (B-ISDN's) are expected to provide efficient and reliable end-to-end transport for various services. The asynchronous transfer mode (ATM) technique [1] has been standardized and widely accepted as a basis for transporting and switching user's traffic in emerging broadband networks, where user's information is packetized and carried in fixed length cells (53 bytes). Recently, many ATM switch architectures have been proposed for point-to-point and/or multicast switching [2], [3].

In the point-to-point switch design, cell-routing and output port contention resolution are the two major issues for implementing a fast packet switch fabric. Preferably, both functions should be performed in a distributed way so that no centralized cell processing will cause a bottleneck to limit the switch size. The output port contention can be resolved by placing buffers at input ports, internal stages, or output ports of the switch.

For the multicast switch design, three more issues need to be considered: 1) how to replicate cells, 2) how to address the multicast cells, and 3) how to reduce the translation table complexity.

Among these three issues, multicast addressing is a fundamental and challenging problem since there are $2^N$ multicast patterns for $N$ destinations, while there are only $N$ unicast patterns for point-to-point communication. In order to distinguish multicast patterns, at least $N(= \log_2 2^N)$ bits information is needed, while for uncasting the required routing information is $\log_2 N$ bytes long. There are three possible schemes for multicast addressing. The first one is to carry the addresses of all output ports in the cell header. The second one is to carry a multicast pattern that is a bit map of all output ports; each bit indicates if the cell is to be accepted at the associated output port. The third scheme often applied to a nonself-routing network uses indirect addressing via a multicast identification that is translated into a bit map of output ports involved in the multicast. The first scheme results in a variable-length header and is undesirable for implementing an ATM switch because of high hardware complexity. All three schemes have a potential problem of having a long cell header for a large switch size, resulting in a higher operation speed required in the switch fabric.

The architectures that have been proposed for multicast ATM switches can be classified into two categories based on their cell-replication methods: 1) multicast tree type [4]–[7], and 2) broadcast tree type [8]–[13] as shown in Fig. 1(a) and (b), respectively. The switches in the first type create a multicast tree in a binary network, such as a banyan or omega network, to replicate cells. They consist of a multicast tree network (often called a copy network) and a routing network, solving the multicast addressing problem by separating cell-replication function from cell-routing function. The copy network uses $O(\log_2 N)$ bits information (copy request number [4], [7] or address intervals [6]) for cell-replication, and the routing network uses $O(\log_2 N)$ bits information (destination address) for routing each copy of multicast cells.

A bit map addressing scheme can be employed for the multicast switches that use a broadcast medium for cell-replication. If a multiple-stage structure is applied to implement the broadcast-typed multicast switches, the number of addressing bits is reduced from $O(N)$ to $O(\sqrt{N})$, where $n$ is the number of stages.

The architectures proposed in [4], [5] use internal buffers at every stage to resolve internal blocking caused by the contention among replicated cells and thus tend to have a
We have proposed a large-scale multicast output buffered ATM switch (MOBAS) that utilizes a recursive structure [14], [15] to achieve a large switch size. The new switch architecture was modified from the one proposed in [16] to cope with the multicast capability. The proposed multicast switch belongs to the second category, broadcast type. It employs: 1) multicast knockout principle, extending the generalized knockout principle to incorporate the multicast function, 2) output queuing with cell duplication capability to achieve the best possible delay/throughput performance [17], 3) distributed processing for cell replication, cell filtering, and cell contention, and 4) multiple stages to overcome the multicast addressing problem. The cell replication is achieved by broadcasting incoming cells to multiple switch modules, each of which consists of a two-dimensional array of switch elements (SWE's). The cell filtering and contention resolution functions are performed distributively by switch elements. In the MOBAS, the output ports are partitioned into a number of groups by switch modules. A routing-link sharing concept (known as channel grouping in [18], [19]) is applied to construct the entire switch, reducing hardware complexity (e.g., number of SWE's and interconnection wires) by almost one order of magnitude when compared to a switch without sharing the routing links.

Section II describes an architecture of a two-stage MOBAS. Section III presents the multicast knockout principle and the switch performance analysis. Section IV shows the effectiveness of a cell duplicator at the output port. Section V shows the MOBAS to be "quasifair" by calculating cell loss rates of all input ports. Section VI presents our conclusions.

II. MULTICAST SWITCH ARCHITECTURE

A. Two-Stage Configuration

Fig. 2 shows a two-stage structure of the multicast output buffered ATM switch (MOBAS), which consists of input port controllers (IPC1, IPC2), multicast grouping networks (MGN1, MGN2), and output port controllers (OPC). The IPC's terminate arrived cells, look up necessary information in the translation tables, and attach routing information (e.g., multicast patterns and priority bits) to the front of the cells such that cells can be routed properly in the MGN's. The MGN's replicate multicast cells based on their multicast patterns and send one copy to each output group. The OPC's temporarily store multiple arriving cells destined for that output port in an output buffer, generate multiple copies for multicast cells with a cell duplicator (CD), assign a new VCI obtained from a translation table to each copy, convert the internal cell format to the standardized ATM cell format, and finally send the cells to the next switching node or the final destination. Section II-C contains a detailed description of the translation tables and cell routing formats in the MGN's.

To simplify the explanation of how the MOBAS functions, let us first consider a unicast case and then a multicast case. The key concept used to implement the unicast switch in [16] is the combination of the knockout principle [20] and channel sharing, or the so-called generalized knockout principle [21]. Multiple output ports are bundled in a group so that cells...
Fig. 2. The architecture of the multicast output buffered ATM switch (MOBAS).

As shown in Fig. 2, every M output ports are bundled in a group, and there are a total of K groups (K = N/M) for a switch size of N inputs and N outputs. \(L_1 \times M\) routing links are provided to each group of M output ports, due to cell contention. If there are more than \(L_1 \times M\) cells in one cell time slot destined for the same output group, the excess cells will be discarded and lost. However, we can engineer \(L_1\) (called group expansion ratio) such that the probability of cell loss due to the competition for \(L_1 \times M\) links is lower than that due to the buffer overflow at the output port or bit errors occurring in the cell header. Performance study shows that the larger the M is, the smaller \(L_1\) needs to be to achieve the same cell loss probability. For instance, for a group size of one output port, which is the case in the second stage (MGN2), \(L_2\) needs to be at least 12 to have a cell loss probability of 10^{-10}. But for a group size of 32 output ports, which is the case in the first stage (MGN1), \(L_1\) just needs to be 2 to have the same cell loss probability. Cells from input ports are properly routed in MGN1 to one of the K groups; they are then further routed to a proper output port through MGN2. Up to \(L_2\) cells can arrive simultaneously at each output port. An output buffer is used to store these cells and send out one cell at each cell time slot. Cells that originate from the same traffic source can be arbitrarily routed to any of the \(L_1 \times M\) routing links. Chao presented a solution in [16] to preserve cell sequence.

Now let us consider a multicast situation where a cell is replicated into multiple copies in MGN1, MGN2, or both, and these copies are sent to multiple outputs. Fig. 3 shows an
example to illustrate how a cell is replicated in the MGN's and duplicated in the CD. Suppose a cell arrives at an input port \( i \) and is to be multicast to four output ports: \#1, \#M, \#(M+1), and \#N. The cell is first broadcast to all \( K \) groups in MGN1, but only groups, \#1, \#2, and \#K accept the cell. Note that only one copy of the cell appears in each group, and the replicated cell can appear at any one of the \( L_1 \times M \) links. The copy of the cell at the output of group \#1 is again replicated into two copies at MGN2. There are in total four replicated cells created after MGN2. When each replicated cell arrives at the OPC, it can be further duplicated into multiple copies by the CD as needed. Each duplicated copy at the OPC is updated with a new VCI obtained from a translation table at the OPC before it is sent to the network. For instance, two copies are generated at output port \#1 and three copies at output port \#(M+1). The reason for using the CD is to reduce the output port buffer size by storing only one copy of the multicast cell at each output port instead of storing multiple copies that originate from a traffic source and are multicast to multiple virtual circuits at an output port. Also note that since there are no buffers in both MGN1 and MGN2, the replicated cells from either MGN are aligned in time. But, the final duplicated cells at the output ports can be unaligned in time because they may have different queuing delays in the output buffers.

### B. Multicast Grouping Network (MGN)

Fig. 4 shows a modular structure for the MGN at the first or the second stage. The MGN consists of \( K \) switch modules for the first stage or \( M \) for the second stage. Each switch module contains a switch element (SWE) array, a number of multicast pattern maskers (MPM), and an address broadcaster (AB). Since the structure and the operations for MGN1 and MGN2 are identical, we will only give the explanation for MGN1.

Each switch module in MGN1 has \( N \) horizontal input lines and \( L_1 \times M \) (\( M = N/K \)) vertical routing links. These routing links are shared by the cells that are destined for the same output group of a switch module. Each input line is connected to all switch modules, meaning that cells from each input line are broadcast to all \( K \) switch modules.

The routing information carried in front of each arriving cell is a multicast pattern, which is a bit map of all the outputs in the MGN. Each bit indicates if the cell is to be sent to the associated output group. For instance, let us consider a multicast switch with 1024 inputs and 1024 outputs and the number of groups in MGN1 and MGN2, \( K \) and \( M \), are both chosen to be 32. Thus, the multicast pattern in both MGN1 and MGN2 has 32 bits. For a unicast cell, the multicast pattern is basically a flattened output address (i.e., a decoded output address) in which only one bit is set to "1" and all other 31 bits are set to "0." For a multicast cell, there are more than
one bit in the multicast pattern set to “1.” For instance, if a cell, X, is multicast to switch modules i and j, the ith and jth bit in the multicast pattern are set to “1.”

The MPM performs a logic AND function for the multicast pattern with a fixed 32 b pattern in which only the ith bit, corresponding to switch module i, is set to “1” and all other 31 bits are set to “0.” So, after cell X passes through the MPM in switch module i, its multicast pattern becomes a flattened output address where only the ith bit is set to “1.”

Each empty cell that is transmitted from the address broadcaster (AB) is attached, in the front, a flattened output address with only one bit set to “1.” For example, empty cells from the AB in switch module i have only the ith bit set to “1” in their flattened address. Cells from horizontal inputs will be properly routed to different switch modules based on the result of comparing their multicast patterns with empty cells’ flattened addresses. For cell X, since its ith and jth bit in the multicast pattern are both set to “1,” it matches with the flattened addresses of empty cells from the AB’s in switch modules i and j. Thus, cell X will be routed to the outputs of these two switch modules.

The SWE has two states, cross state and toggled state, as shown in Fig. 5. The state of the SWE depends on the comparison result of the flattened addresses and the priority fields in cell headers. The priority is used for cell contention resolution. The SWE is at cross state initially, i.e., cells from the north side are routed to the south side, and cells from the west side are routed to the east side. When the flattened address of a cell from the west (FA_w) is matched with the flattened address of a cell from the north (FA_n), and when the west’s priority level (P_w) is higher than the north’s (P_n), the SWE’s state is toggled; the cell from the west side is routed to the south side, and the cell from the north is routed to the east. In other words, any unmatched or lower-priority (including the same priority) cells are always routed to the east side. Each SWE introduces a 1 b delay as the cell passes it in either direction. The cells from MPM’s and AB are skewed by 1 b before they are sent to each SWE array, due to the timing alignment requirement [16].

Fig. 6 shows an example of how cells are routed in a switch module. Cells U, V, W, X, Y, and Z arrive at inputs 1 to 6, respectively, and are to be routed in switch module #3. In the cell header, there is a 3 b multicast pattern (m_3 m_2 m_1) and a 2 b priority filed (p_1 p_0). If a cell is to be sent to an output of this switch module, its m_3 b will be set to “1.” Among these six cells, cells U, V, and X are for unicast, where only one bit in the multicast pattern is set to “1.” The other three cells are for multicast, more than one bit in the multicast pattern is set to “1.” We assume a smaller priority value has a higher priority level. For instance, cell Z has the highest priority level (“00”) and empty cells transmitted from the address broadcaster have the lowest priority level (“11”). The MPM performs a logic AND function for each cell’s multicast pattern with a fixed pattern of “100.” For instance, after cell W passes through the MPM, its multicast pattern (“110”) becomes “100” (a_3 a_2 a_1), which has only one bit set to “1” and is denoted as a flattened address. When cells are routed in the SWE array, their routing paths are determined by the state of SWE’s, which are controlled
according to the rules in Fig. 5. Since cells V and X are not destined for this group, the SWE's they pass remain in a cross state. Consequently, they are routed to the right side of the module and are discarded. Since there are only three routing links in this example, while there are four cells destined to this switch module, the one with the lowest priority (i.e., cell U) loses the contention with the other three and is discarded.

Since the crossbar structure inherits the characteristics of identical and short connection wires between switch elements, the timing alignment for the signals at each SWE is much easier than that of other types of interconnection network, such as the binary network, the Clos network, and so on. The unequal length of the interconnection wires increases the difficulty of synchronizing data signals, and, consequently limits the switch fabric's size, such as in the Batcher-banyan network. The SWE's in the grouping networks only communicate locally with their neighbors, so as the chips that contain the SWE's. Therefore, chips do not need to drive long wires to others on the same printed circuit board. To ensure that data signals are aligned at the SWE's at the following stages, the interconnection wires between the SWE's in each column or in each row should have the same length. However, wires between columns and rows do not have to be the same length. Furthermore, the requirement of synchronizing data signals at each SWE is confined in each switch module instead of in the entire switch fabric.

C. IPC and OPC Translation Tables

The translation tables in the IPC1, IPC2, and OPC contain necessary information for properly routing cells in the switch modules, MGN1 and MGN2, and for translating old VCI values into new VCI values. As mentioned above, cells routed in the MGN's depend on the multicast pattern and priority values. To reduce the translation table's complexity, the table contents and the information attached to the front of cells are different for the unicast and multicast calls.

In a point-to-point ATM switch, an arrived cell's VCI at an input line can be identical with other cells' VCI's on other input lines. Since the VCI translation table is associated with the IPC at each input line, the same VCI values can be repeatedly used for different virtual circuits at different input lines without causing any ambiguity. But cells that are from different virtual circuits and destined for the same output port must have different translated VCI's.

In a multicast switch, since a cell is replicated into multiple copies that are likely to be transmitted on the same routing links inside a switch fabric, the switch must use another identifier, BCN, to uniquely identify each multicast call. In other words, the BCN of a multicast call at an input line has to be different from that of other multicast calls at different input lines, which is unlike the VCI in the point-to-point switch, where a VCI value can be repeatedly used for different calls at different input lines.

For the unicast situation, upon a cell's arrival, its VCI value is used as an index to access the necessary information in the IPC1's translation table, such as the output addresses in MGN1 and MGN2, a contention priority value, and a new VCI, as shown in Fig. 7(a). The output address of MGN1, A1, is first decoded into a flattened address, which has K bits and is put into the MP1 field in the cell header, as shown in Fig. 8(a). The MP1 and P are used for routing cells in the MGN1, and A2 is used for routing cells in the MGN2. The "1" bit is the multicast indication bit, which is set to "0" for unicast and "1" for multicast. When a unicast cell arrives at IPC2, the A2 field is simply decoded into a flattened address and put into the MP2 field as the routing information in MGN2. Thus, no translation table in IPC2 is required for unicast cells. Note that A2 is not decoded into a flattened address until it arrives at IPC2. This saves some bits in the cell header (e.g., this saves 27 bits for the above example of a 1024×1024 switch) and thus reduces the required operation speed in the switch fabric. The unicast cell routing format in MGN2 is shown in Fig. 8(b).

For the multicast situation, besides the routing information of MP1, MP2, and P, an important piece of information, so-called broadcast channel number (BCN), is used to identify cells that are routed to the same output group of a switch module. Similar to the unicast case, an arrived cell's VCI is first extracted to look up the information in the translation table in IPC1, as shown in Fig. 7(b). After a cell has been routed through MGN1, the MP1 is no longer used. The BCN is used to look up the next routing information, MP2, in the IPC2's translation table, as shown in Fig. 7(c). The cell formats for a multicast call in MGN1 and MGN2 are shown in Fig. 8(a) and
III. MULTICAST KNOCKOUT PRINCIPLE

A new multicast knockout principle, an extension of the generalized knockout principle, has been applied to our two-stage multicast output buffered ATM switch (MOBAS) to cope with the multicasting capability. We will show the cell loss performance of the MOBAS and the numerical results for the first and the second stages (MGN1, MGN2). In order to analyze the performance of MOBAS, we redraw Fig. 2 to show its logical connections. In Fig. 9 each trapezoid corresponds to each switch module (SM) in the MGN. Since the SM performs a concentration function (e.g., N to L1 × M), it is also called a concentrator in this paper. Each concentrator has N input lines, because every incoming cell is broadcast to all concentrators. The concentrator has L1 × M outputs, meaning that up to L1 × M cells can be accepted in each concentrator.

A. Cell Loss Rate in MGN1

In the analysis, we assume that the traffic at each input port of the MOBAS is independent of the other inputs’ and replicated cells are uniformly delivered to all output groups. The average cell arrival rate, \( \rho \), is the probability that a cell arrives at an input port in a given cell time slot. We assume that the average cell-replication in MGN1 is \( E[F_1] \), the average cell-replication in MGN2 is \( E[F_2] \), the average cell-duplication in OPC is \( E[D] \), and that these random variables, \( F_1, F_2, \) and \( D \), are independent of each other.

Every incoming cell is broadcast to all concentrators (SM’s), and is properly filtered at each concentrator according to the multicast pattern in the cell header. The average cell arriving rate, \( \rho \), at each input of a concentrator is considered as

\[
p = \frac{\rho E[F_1]}{K}
\]

where \( K = N/M \) is the number of concentrators in MGN1. The probability \( (A_k) \) that \( k \) cells are destined for a specific concentrator of MGN1 in a given time slot is

\[
A_k = \binom{N}{k} p^k (1-p)^{N-k}
\]

\[
= \frac{N!}{k!(N-k)!} (\frac{\rho E[F_1] \cdot M}{N})^k (1 - \frac{\rho E[F_1] \cdot M}{N})^{N-k}
\]

\[0 \leq k \leq N
\]

(1)

where \( \frac{\rho E[F_1] \cdot M}{N} \) is the probability of a cell arriving at an input of a specific concentrator in MGN1.

As \( N \to \infty \), (1) becomes

\[
A_k = e^{-\rho E[F_1] \cdot M} \cdot (\frac{\rho E[F_1] \cdot M}{k})^k
\]

(2)
where \( \rho \) should satisfy the following condition for a stable system.

\[
\rho \cdot E[F_1] \cdot E[F_2] \cdot E[D] < 1.
\]

Since there are only \( L_1 \times M \) routing links available for each output group, if more than \( L_1 \times M \) cells are destined for this output group in a cell time slot, excess cells will be discarded and lost. The cell loss rate in MGN1, \( P_1 \), is

\[
P_1 = \frac{\sum_{k=0}^{N} (k - L_1 \times M) \cdot A_k \cdot E[F_2] \cdot E[D]}{N \cdot p \cdot E[F_2] \cdot E[D]} \tag{3}
\]

\[
= \frac{1}{\rho E[F_1] \cdot M} \sum_{k=L_1 \times M+1}^{N} (k - L_1 \times M) \left( \frac{\rho E[F_2] \cdot M}{N} \right)^k \left( 1 - \frac{\rho E[F_1] \cdot M}{N} \right)^{N-k} \tag{4}
\]

The denominator in (3), \( N \cdot p \cdot E[F_2] \cdot E[D] \), is the average number of cells effectively arriving at a specific concentrator during one cell time, and the numerator in (3), \( \sum_{k=0}^{N} (k - L_1 \times M) \cdot A_k \cdot E[F_2] \cdot E[D] \), is the average number of cells effectively lost in the specific concentrator. Both the denominator and the numerator in (3) include the term \( E[F_2] \cdot E[D] \) to cope with the cell replication in MGN2 and OPC.

As \( N \to \infty \), (4) becomes (5) shown at the bottom of the page.

Note that (5) is similar to the one in the generalized knockout principle [21], except that the parameters in the two equations are slightly different due to the cell replication in MGN1 and MGN2, and the cell duplication in the OPC.

Fig. 10(a) shows the plots of the cell loss probability at the MGN1 versus different \( L_1 \) and fanout values for an offered load (\( \rho \cdot E[F_1] \cdot E[F_2] \cdot E[D] \)) of 0.9 at each output port. We notice from these plots that as \( M \) increases (i.e., more outputs are sharing their routing links), the required \( L_1 \) value decreases for a given cell loss rate. The switch design parameters \( L_1 \) and \( M \) are more stringent to the unicast case than to the multicast. Since the load on MGN1 decreases as the product of the average fanouts in MGN2 and OPC increases, the cell loss rate of MGN1 under the multicast case is lower than that of the unicast case. The replicated cells from a multicast call will never contend with each other for the same output group (concentrator) since the MOBAS replicates at most one cell for each output group. In other words, MGN1 that is designed to meet the performance requirement for unicast calls will also satisfy multicast calls' performance requirement.

### B. Cell Loss Rate in MGN2

Special attention is required when analyzing the cell loss rate in MGN2 because the cell arriving pattern at the inputs of MGN2 is determined by the number of cells passing through the corresponding concentrator in MGN1. If there are \( l \) (\( 0 \leq L_1 \times M \)) cells arriving at MGN2, these cells will appear at the upper \( l \) consecutive inputs of MGN2. If more than \( L_1 \times M \) cells are destined for MGN2, only \( L_1 \times M \) cells will arrive at MGN2's inputs (one cell per each input port), while excessive cells are discarded in MGN1.

We assume the cell loss rate of a concentrator in MGN2 to be \( P_2 \), and the probability that \( l \) cells arrive at a specific concentrator in MGN2 to be \( B_l \). Both \( P_2 \) and \( B_l \) depend on the average number of cells arriving at the inputs of MGN2 (i.e., the number of cells passing through the corresponding concentrator in MGN1). This implies that \( P_2 \) is a function of \( A_k \) in (1). In order to calculate \( P_2 \), the probability of \( j \) cells arriving at MGN2, denoted as \( I_j \), is

\[
I_j = \begin{cases} 
A_j & \text{for } j < L_1 \times M \\
\sum_{k=L_1 \times M}^{N} A_k & \text{for } j = L_1 \times M.
\end{cases}
\]

If \( j \) \((0 \leq L_1 \times M)\) cells arrive at MGN2, they will appear at the upper \( j \) consecutive inputs of MGN2. Since how and where the cells appear at the MGN2's inputs does not affect the cell loss performance, we will simplify our performance analysis by assuming that a cell can appear at any input of the MGN2.

Let us denote the probability that \( l \) cells arrive at the inputs of a specific concentrator in MGN2 for given \( j \) cells arrived at MGN2 to be \( B_{lj} \).

Then,

\[
B_{lj} = \frac{\binom{j}{l} q^l (1-q)^{j-l}}{l} \quad 0 \leq j \leq L_1 \times M, \quad 0 \leq l \leq j
\]

where \( q \) is equal to \( \frac{E[F_2]}{M} \) under the assumption that replicated cells are uniformly delivered to \( M \) concentrators in MGN2.

If less than or equal to \( L_2 \) cells arrive at MGN2 (\( 0 \leq j \leq L_2 \)), no cell-discarding will occur in MGN2, because each concentrator can accept up to \( L_2 \) cells during one cell time slot. If more than \( L_2 \) cells arrive at MGN2 (\( L_2 \leq j \leq L_1 \times M \)),

\[
P_1 = \frac{1 - L_1 \times M}{N \times p} \left[ 1 - \sum_{k=0}^{L_1 \times M} \binom{N \times p}{k} e^{-N \times p} \frac{1}{k!} \right] + \frac{(N \times p)^{L_1 \times M} e^{-N \times p}}{(L_1 \times M)!} \tag{5}
\]

\[
= \frac{1 - L_1 \times M}{\rho E[F_1] \times M} \left[ 1 - \sum_{k=0}^{L_1 \times M} \binom{\rho E[F_1] \cdot M}{k} e^{-\rho E[F_1] \cdot M} \frac{1}{k!} \right] + \frac{(\rho E[F_1] \cdot M)^{L_1 \times M} e^{-\rho E[F_1] \cdot M}}{(L_1 \times M)!}
\]
The cell loss rate in MGN2, $P_2$, is

$$P_2 = \frac{\sum_{l=L_2+1}^{L_1 \times M} (l - L_2) \cdot B_l \cdot E[D]}{N \cdot \rho E[F_1] \cdot (1 - P_1) \cdot E[F_2] \cdot E[D] / M}$$

$$= \frac{\sum_{l=L_2+1}^{L_1 \times M} (l - L_2) \cdot B_l \cdot E[F_1] \cdot (1 - P_1) \cdot E[F_2]}{\rho E[F_1] \cdot E[F_2] \cdot E[D]}.$$

$N \cdot \rho E[F_1]$ is the average number of cells destined for a specific concentrator in MGN1 from the inputs of the MOBAS. $N \cdot \rho E[F_1]$, $(1 - P_1)$ is the average number of cells that have survived in this concentrator, which in turn becomes the average number of cells arriving at the corresponding MGN2. Thus, the denominator in (6), $N \cdot \rho E[F_1] \cdot (1 - P_1) \cdot E[F_2] \cdot E[D] / M$, is the average number of cells effectively arriving at a specific output port. The numerator in (6), $\sum_{l=L_2+1}^{L_1 \times M} (l - L_2) \cdot E[D]$, is the average number of cells effectively lost in a specific concentrator in MGN2 because the lost cell can be a cell that would be duplicated in the OPC.

Fig. 10(b) shows the plots of the cell loss probability at MGN2 versus different $L_2$ values and average duplication values for an effective offered load ($= \rho \cdot E[F_1] \cdot E[F_2] \cdot E[D]$) of 0.9. The average fanout on MGN1 is assumed to be 1.0 ($E[F_1] = 1.0$), the group size to be 32 ($M = 32$), and the group expansion ratio to be 2.0 ($L_1 = 2.0$). Since the traffic load on MGN2 reduces as the average cell duplication ($E[D]$) increases, the cell loss rate in MGN2 decreases as $E[D]$ increases. Therefore, the switch design parameter, $L_2$, is more stringent to the unicast case than to the multicast. Consequently, if MGN2 is designed to meet the performance requirement for unicast calls, it will also satisfy multicast calls' performance requirement.

C. Total Cell Loss Rate in MOBAS

Total cell loss rate in MOBAS is shown in (7) at the bottom of the next page. The first term of numerator in (7), $\frac{N}{M} \sum_{l=L_1 \times M+1}^{L_1 \times M} (k - L_1 \times M) \cdot A_k \cdot E[F_1] \cdot E[D]$, is the average number of cells effectively lost in MGN1. The second term of numerator in (7), $N \sum_{l=L_2+1}^{L_1 \times M} (l - L_2) \cdot E[D]$, is the average number of cells effectively lost in all of MGN2's. The denominator in (7), $N \cdot \rho E[F_1] E[F_2] E[D]$, is the average number of cells effectively offered to the MOBAS.

In (7) $L_1$ and $L_2$ should be in a certain range to guarantee the required cell loss performance in the MOBAS. For example, in order to have the cell loss rate of $10^{-10}$ in the MOBAS with a very large size ($N \geq 1024$) and a group size of M (32), $L_1$ and $L_2$ should be greater than 2 and 12, respectively. If any one of them is less than the required number, the total cell loss rate in the MOBAS cannot be guaranteed since the term of the numerator of (7), which is a function of the smaller number, dominates the total cell loss rate in the MOBAS.
We conclude in this section that a switch that is designed to meet the performance requirement for unicast calls will also satisfy multicast calls' performance requirement because the switch design parameters such as $M$, $L_1$, and $L_2$ are more stringent to the unicast case than to the multicast. The design parameters, $M$, $L_1$ and $L_2$, should be in a certain range to guarantee the required cell loss performance in the the MOBAS.

IV. EFFECTIVENESS OF THE CELL Duplicator

This section discusses the effectiveness of the cell duplicator (CD). Two models for the cell duplication are considered here. Fig. 11(a) shows the model for the OPC in MOBAS, denoted as Model I. Fig. 11(b) shows a multicast switch that consists of input port controllers (IPC's), an expanded copy network, trunk number translators (TNT's), and an output-buffered point-to-point switch. The expanded copy network is similar to the one proposed in [22]. Fig. 11(c) shows the model for the OPC in the multicast switch with an expanded copy network, denoted as Model II.

The mean queue lengths of Model I and Model II are calculated and compared. It is shown that Model I has a smaller queue length than Model II does for the same traffic distributions, and that both models have the same queuing delay.

In this study, the duplication request in Model I is assumed to be a geometric distribution. Based on the analysis of the cell loss rate in the previous section, Model I can be considered as Poisson/Geom/1 [23]. The Poisson arrival assumption is reasonable because we have assumed that traffic at each input port of the MOBAS is independent of the other ports and that the expansion ratios $L_1$ and $L_2$ are sufficiently large to have a negligible cell loss rate in the switch fabric. Model II can be considered as discrete Batch Poisson/D/1 [24]. Here, the batch size of Model II has the same distribution as the service time of Model I. These assumptions are reasonable because the cell duplication that is performed inside the expanded copy network in Fig. 11(b) is performed in the OPC of the MOBAS.

Thus, the number of arriving cells at Model I has a Poisson distribution with a rate of $\lambda$ (i.e., number of cells/one cell time slot),

$$P_r\{k \text{ cells arrive during a cell time slot}\} = e^{-\lambda} \frac{\lambda^k}{k!}$$

where

$$\lambda = \rho E[F_1] E[F_2].$$

The service time has a geometric distribution with an average of $1/q$ (i.e., $E[D]$) cell times,

$$P_r\{\text{service time} = k \text{ cell time slots}\} = (1-q)^{k-1} q, \quad k = 1, 2, \cdots.$$

The number of arriving cell batches at Model II has a Poisson distribution with a rate of $\lambda$ (i.e., number of cell batches/one cell time slot),

$$P_r\{k \text{ batches arrive during a cell time slot}\} = e^{-\lambda} \frac{\lambda^k}{k!}.$$

The batch size has a geometric distribution with an average of $1/q$ cells,

$$P_r\{\text{batch size} = b \text{ cells}\} = (1-q)^{b-1} q, \quad b = 1, 2, \cdots.$$

The distribution of the number of arriving cells at Model I and that of the number of arriving cell batches at Model II are the same. The service time of Model I (the time needed to complete cell duplications and transmissions) has the same distribution as the batch size in Model II (i.e., a geometric distribution).

$$P_T = \frac{\text{Avg. # of cells effectively lost in both MGN1 and MGN2's}}{\text{Avg. # of cells effectively offered to MOBAS}}$$

$$= \frac{\sum_{k=L_1 \times M+1}^{N} (k - L_1 \times M) \cdot A_k \cdot E[F_2] E[D]}{N \cdot \rho E[F_1] E[F_2] E[D]} + \frac{\sum_{l=L_2+1}^{L_1 \times M} (l - L_2) \cdot B_l \cdot E[D]}{N \cdot \rho E[F_1] E[F_2] E[D]}$$

(7)
Let us define \( X \) and \( Y \) be random variables for the number of cells in the buffer in Model I and Model II, at the beginning of a call time slot, respectively. And let us define \( D_I \) and \( D_{II} \) be random variables for the delays experienced by cells in the buffer in Model I and Model II, respectively. From [23] and [24], we can find the average queue lengths of Model I and Model II as follows:

\[
E[X]_{M/Geom, I} = \frac{\lambda (2 - \lambda)}{2(1 - \frac{\lambda}{q})}.
\]

\[
E[Y]_{M/\sigma/D/1} = \frac{\lambda}{2q} + \frac{\lambda}{q} \left( \frac{2 - \lambda}{2(1 - \frac{\lambda}{q})} \right) = \frac{1}{q} \left[ \frac{\lambda (2 - \lambda)}{2(1 - \frac{\lambda}{q})} \right] = \frac{1}{q} E[X]_{M/Geom, I}.
\]

Note that the mean queue length of Model I is smaller than that of Model II by a factor of the mean batch size, \( E[B] = 1/q \). Using Little’s formula [25], the average delays of Model I and Model II are as follows and found to be equal:

\[
E[D_I]_{M/Geom, I} = \frac{E[X]_{M/Geom, I}}{\lambda} = \frac{\lambda (2 - \lambda)}{2(1 - \frac{\lambda}{q})}.
\]

\[
E[D_{II}]_{M/\sigma/D/1} = \frac{E[Y]_{M/\sigma/D/1}}{\lambda q} = \frac{1}{q} \left( \frac{(2 - \lambda)}{2(1 - \frac{\lambda}{q})} \right) = \frac{E[D_I]}{M/Geom, I}.
\]

From the above, the delays experienced by multicast cells are the same in these two models, but the queue length occupied by the multicast cells in the MOBAS is smaller than that in the cascaded type output-buffered multicast switch by a factor of the average cell duplication request in the OPC of the MOBAS, which is equal to the mean batch size in Model II. It implies that for a finite output buffer, the cell loss rates in the MOBAS’s output buffer are smaller than those in the cascaded type multicast switch for any arrival process and any duplication request distribution. The effectiveness of the CD will be more profound when congestion occurs at the output buffer due to bursty arrivals of multicast calls with large number of duplication requests.

V. QUASIFAIRNESS OF MOBAS

In this section, we show that the MOBAS is quasifair by calculating the cell loss rates of all input ports. We conclude that unfairness among the input ports can be neglected by properly choosing switch parameters (e.g., expansion ratios), based on multicast knockout principle, such that the cell loss rate of every input port is less than \( 10^{-10} \).

Fig. 12 depicts an example of unfairness (different cell loss rates) among the input ports. The operation of the SWE (switch element) and how the cells are routed in the SWE array can be found in Section II. In this example, 5 cells, U, V, W, X, and Y are destined for an output port, e.g., #3. Since only four routing links are available in this SWE array, one cell will be discarded. Cells arriving at upper input ports have better service preference than those that have the same priority level but arrive at lower input ports. As shown in Fig. 12, cell Y is discarded because it arrives at the bottom input port.

Let us assume only one priority level exists, the switch has \( N \) inputs and \( N \) outputs, and \( L \) routing links are available per output port. \( P_k \) is defined to be the cell loss rate of the \( k \)th input port. The cell loss rates of \( L \) uppermost input ports is zero, meaning that cells from these input ports are delivered in loss free.

\[
P_k = 0 \quad 1 \leq k \leq L
\]

Now let us consider the cell loss rate of the remaining input ports, from the \( (L + 1) \)th to the \( N \)th (shown at the bottom of this page).

Fig. 13 shows the cell loss rate of each input port for a single stage 64 x 64 MOBAS. For \( L = 12 \), the cell loss rate for the bottom input port (the worst case) is less than \( 10^{-10} \), which meets the cell loss rate requirement for existing services.

VI. CONCLUSION

We proposed a recursive modular architecture to implement a large-scale multicast output buffered ATM switch (MOBAS).
3) distributed processing for cell replication, cell filtering, and cell contention, and 4) multiple stages to reduce routing address overhead.

Cell replication is achieved by broadcasting incoming cells to multiple switch modules, which consists of a two-dimensional array of switch elements. In the MOBAS, the output ports are partitioned into a number of groups by the switch modules. The routing-link sharing concept is applied to construct the entire switch, which reduces the hardware complexity (number of switch elements and interconnection wires) by almost one order of magnitude. The routing address used in the MOBAS is a bit map of all the outputs in the first and second stage routing networks. In order to reduce translation table complexity, the contents stored in the tables, including routing addresses, priority levels, and new VCI values, are different for unicast and multicast calls.

Through the analysis based on the multicast knockout principle, we conclude that for a two-stage MOBAS, the switch design parameters, such as routing link expansion ratios at the first and the second stage networks, are more stringent to the unicast case than to the multicast case. This implies that if a switch is designed to meet the performance requirements for unicast calls, it will also satisfy multicast calls’ performance requirement.

The effectiveness of the cell duplicator (CD) has been shown by comparing the mean queue lengths of the output buffer of two models: 1) Model I, the OPC in the MOBAS, and 2) Model II, a multicast switch consisting of an expanded copy network, trunk number translators (TNT’s), and an output-buffered point-to-point switch. The delays experienced by multicast cells are the same in these two models, but the queue length occupied by the multicast cells in the MOBAS is smaller than that in the cascaded type output-buffered multicast switch by a factor of the average cell duplication request in the OPC of the MOBAS. It implies that the cell loss rates in the MOBAS’s output buffer are smaller than those in the cascaded type multicast switch for any arrival process and any duplication request distribution under a finite output buffer condition.

The architecture of the MOBAS has been proven to be quasifair through the calculation of the cell loss rates of all input ports. By properly engineering design parameters (e.g., routing link expansion ratios), the cell loss rate of each input port can be arbitrarily small (e.g., $10^{-12}$) to meet the performance requirement of all broadband services. Thus, there is no need to be concerned about the unfairness issue in the MOBAS.

An ATM crosspoint switch chip with $16 \times 16$ switch elements has been implemented using CMOS 2-μm technology. The chip has been tested and functions correctly at about 50 Mb/s (limited by test equipment’s speed); it was simulated to operate properly at 100 Mb/s. The chip’s photograph is shown in Fig. 14.

ACKNOWLEDGMENT

The authors would like to thank Dr. N. Uzun for laying out and testing the switch chip. They are also grateful to
anonymous reviewers’ comments that improved the quality of this paper.

REFERENCES


Byeong-Seog Choe was born in Seoul, Korea in 1957. He received the B.S.E.E. degree from Seoul National University in 1985, and the M.S.E.E degree from Fairleigh Dickinson University, NJ, in 1987. He received the M.S. and Ph.D. degrees in electrical engineering from Polytechnic University, NY, in 1993 and 1994, respectively.

From 1992 to 1993, he was a Research Fellow in the Center for Advanced Technology in Telecommunications of Polytechnic University. He was involved in the design and performance analysis of ATM switches. Since March 1994, he has been on the faculty of the Department of Electronics Engineering at Myongji University, Korea, where he is an Assistant Professor. His current research interests include congestion control in BISDN and photonic ATM switches.

H. Jonathan Chao (S’82–M’85) received the B.S.E.E. and M.S.E.E. degrees from National Chiayi University, Taiwan, in 1977 and 1980, respectively, and the Ph.D. degree in electrical engineering from The Ohio State University, NY, in 1985.

He is an Associate Professor of Electrical Engineering at Polytechnic University, New York, which he joined in January 1992. His areas of research include large-scale multicast ATM switches, photonic ATM switches, high-speed computer communications, and congestion/flow control in ATM networks. He holds 14 patents and has published over 40 journal and conference papers in the above areas. From 1985 to 1991, he was a Member of Technical Staff at Bellcore, NJ, where he has implemented several high-speed VLSI chips for SONET/ATM-based broadband networks. He has also proposed various architectures, along with their performance studies, for B-ISDN applications. These architectures include an optical customer premises network, a large-scale multicast ATM switch, traffic shaper, and priority queue manager for ATM network congestion control. He played a key role in designing a HIPPI-to-SONET/ATM network interface for the Nectar Gigabit Testbed, one of the five gigabit network testbeds in the US. From 1977 to 1981, he worked at Taiwan Telecommunication Laboratories, where he was engaged in the development of a digital switching system.