MILITARY TECHNICAL COLLEGE
CAIRO-EGYPT

FIRST INTERNATIONAL CONF. ON
ELECTRICAL ENGINEERING

AN ARCHITECTURE FOR SCALABLE MULTICAST ATM SWITCH

Dr. Amani S. Amin* **Eng. Hanaa A. Ibrahim****

ABSTRACT

In this paper an interconnection architecture for a large scale multicast ATM switch [LSMA] is introduced. The proposed architecture permits the modular growth of its size from a small number of ports to very large switch sizes. The expansion of the switch size is implemented in a smoothing way with simple modification of already existing connections. An important feature of LSMA is that the system growth is independent of the number of vertical stages, i.e. the system always contains three stages only. The general description of the input switch module which contains the multicast network is also introduced. Comparison between the number of middle stage modules and the number of overall switch modules for different architectures is presented.

KEY WORDS

ATM switch, Multicast network, Scalable architecture

NOMENCLATURE

h: Number of first stage switch modules in the partition = number of middle stage switch modules in the partition.

h`: Number of output switch modules in the partition.

K: Number of partitions in the system.

L: Number of outputs of the middle stage switch module

m: Number of outputs of the first stage switch module = number of inputs of the second stage switch module.

* Head of the Switching Dep. National Telecommunication Inst. – Nasr City - Cairo – Egypt

** Teacher Assist. – Switching Dep. National Telecommunication Inst. – Nasr City - Cairo – Egypt

N: The required system capacity.
n: Number of inputs of the first stage switch modules.
n`: Number of outputs of the third stage switch module.
r: The number of channels between the switch modules.
S: The number of the required middle stage switch modules within the partition.
X: The number of the middle stage switch modules within the system.
Y: The number of the required switch modules within the system.

I. INTRODUCTION

In the emerging field of high-speed networking, the standard committees as an underlying transport technology within the public broadband integrated service digital network [1] have chosen ATM.

Most of the produced ATM switches are considered as a type of LAN's interconnection switches. For example the Fore Runner ATM backbone switches (ASX – 200 BX and ASX - 1600) are designed specifically to meet the unique needs of LAN backbone networks. Other example is the power hub LAN switches (power hub 7000). It is considered as multi protocol LAN switch to increase the LAN bandwidth without replacing the existing network infrastructure. Therefore, they are defined as data oriented ATM switch providers [2]. On the other hand, most of the proposed ATM switches can support only a limited number of ATM lines, i.e. a switch fabric size of about 400 high speed ports is currently considered a large switch.

To realize a public broadband ATM network, it is expected to require an ATM switch of size not less than 10,000 high speed ports. The fundamental theory of building a switching system places certain limitations on the size of the switch module (about 64 ports) [3]. Thus many switch modules must be organized in a certain way and interconnected to construct the required large scale switch. The switch modules are organized in horizontal and vertical stages such that the required system capacity can be realized, While the number of the horizontal stages depending on the switch required capacity, but the number of the vertical stages increases proportional to the number of horizontal stages.

The switch modules can be organized in two vertical stages such that a system capacity ranging from one to nine hundreds of lines can be realized. To construct a large scale switch in the range of thousands of lines, at least three vertical stages are required. Very large switches use more than three stages.

Three stage switches to provide a five stage-switching network replace the center stage switches.

While providing the required large scale ATM switch with switch modules arrangement many architecture parameters must be considered [4,5].
These parameters can be summarized as follows;

1. Reducing the hardware complexity in terms of the number of stages and building elements.
2. Modularity considerations that determine the Expandability of he switching network.
3. Minimizing the required number of interconnection lines.
4. Maintaining the cell output sequence.
5. Maintaining a high throughput, a low cell loss probability and reducing the cell delay.

In the literature, most of switch interconnection architecture is either two-stage or three-stage to provide the required connectivity and to satisfy the previous objectives [6-10]. In [6], Lee proposed a two-stage nonblocking switch (fig.1). Since several packets at the first-stage switch may address to the same second-stage switch, more than one line, say "r" lines, are provided for each path between the two stages. The latter method is termed "grouping" [7]. Even this architecture is considered a multistage configuration but according to [3] the system capacity is limited.

One of the classical particular architectures, which based on the multistage configuration, is the Clos Network [8]. As shown in fig.2 the clos network is a three-stage switch in which the inlets and outlets are partitioned into subgroups of n inlets and n outlets each. The first stage switch modules are considered as the input switch modules. The third stage represents the output switch modules. Each one output of the input switch modules is connected to one of the center stage switches. Each one output of the center stage switches is connected to one of the output switch modules. One problem with this architecture is the need of a centralized controller for providing routing paths. Liew and lu in [9] proposed a three-stage Dilated-Banyan switch where first-stage and second-stage switches are grouped into partitions as shown in fig.3. One unique path is provided for each pair of input and output switches. This system and the next one will be discussed in some details in section V.

A growable ATM architecture is proposed in [10]. Its main drawback is that, for expanding the size of the switch, the number of the intermediate modules (middle stage) and respectively the total number of modules for the switch grow very fast which has a significant influence upon the fabric's total cost. In this paper, an interconnection architecture for large scale ATM multicast switch (LSMA) is presented. The important feature of the LSMA is the fact that, the system growth is independent of the number of vertical stages, i.e. the system always contains three stages only. In section II, LSMA interconnection architecture is introduced. The hardware description of the modules and the routing procedure for the traffic inside the switch is presented in section III and IV respectively. Finally, for evaluating the system, a comparison between the number of the middle stage switch modules and the total number of modules for different fabric sizes is calculated.

II. INTERCONNECTION ARCHITECTURE FOR LSMA:

The LSMA permits the modular growth of its size from hundreds of ports to very large switch size. Adding additional identical horizontal partitions to the existed one and interconnecting them with each other can do this. As a result, the expansion of the switch size is implemented in a smooth way with simple modification of the already existing internal connections. Fig.4 indicates the LSMA with duplicating the switch size by using two identical partitions and interconnecting them. Each partition contains three stages of switch modules. The first stage contains a group of input switch modules with dimensions n x m, where m > n. The second stage contains a group of interconnection modules of dimensions m x L, where m ≥ L. The number of the second stage switch modules has to be equal to the number of the first stage switch modules. The third stage contains a group of output switch modules of dimensions m` x n` where m` > n`. Each first switch module in the first stage is connected to the first switch module of the second stage in both partitions. The same technique is repeated similarly for all other switch modules of the first stage. For example, the fourth switch module of the first stage is connected to the fourth switch module of the second stage of both partitions. Therefore the number of the first stage switch modules must be equal to the number of the second stage switch modules. Each second stage switch module is connected to all the third stage switch modules within the same partition. The interconnection between the LSMA switch modules is done by using the channel grouping concept [7]. By using this concept, a group of channels are used to interconnect any two-switch modules such that the blocking probability can be reduced. The determination of a certain channel within any channel group between the first and the second stage is done by using the traffic distribution and contenation resolution network existing at the outputs of the group translators (in the first stage). The determination of a certain channel within any channel group between the second and the third stage can be done by adding another network at the inputs of each second stage switch modules. The input switch modules are used to add the appropriate routing tag to the incoming cells. It also generates the requested number of copies for each multicast cell. Finally, the input switch modules distribute the generated copies over the second stage switch modules of both partitions. This distribution is done according to the destination address. For example, if the requested destination are connected to one of the output switch modules existing in the first partition, then the incoming cell is directed to the second stage switch module of the first partition. The middle stage switch modules are used to gather the incoming cells from different partitions, which are specified, to this partition. Then each middle stage switch module distributes the gathered cells to the output switch modules within the same partition. Each output switch module routes the incoming cells to a specific ATM output port according to their destination address.

In general, to expand the size of the LSMA system k times, it must contain k identical partitions, which are interconnected by the same previous technique. The number of switch modules within each partition is constant and not affected by the system expansion. The interconnection modules are used for interconnecting the system partitions with each other and their numbers within the same partition are constant and not affected by the system growth. The number of interconnection lines between the second and the third stage within the same partition are constant and not affected by the expansion process. The number of interconnection lines between the first and the second stage are increased with the increase of the system growth. The maximum allowable number of partition are affected by two main parameters, the first parameter is the maximum number of outputs of the first stage switch module. The second parameter is the number of links within a channel group between the first and the second stage.

III. HARDWARE DESCRIPTION OF THE LSMA SWITCH MODULES

In this section, the hardware description of the three types of switch modules, which are used in the three stages, are presented. The hardware modification of the switch modules, which is required to increase the maximum capacity of the system, is also proposed.

III. 1- The Input Switch Module

The proposed multicast ATM switch is shown in fig.5 [11]. It is composed of two main systems, the control system and the switching system. The switching system is divided into three subsystems a concentration network, a multicast network and a point to point switch The concentration network places the cells on consecutive outputs so as to ensure nonblocking operation of the subsequent multicast network. The multicast network is a Banyan network with switch elements that are capable of cell replication in addition to cell routing [12].An expansion multicast network is applied to reduce the traffic overflow within the switching system. This overflow occurs when the required number of the incoming cells exceeds the multicast network output. Finally the point to point switch is used to route each incoming cell to its proper destination.

A three level hierarchical control system is applied in the proposed switch module. The three levels are represented by a group of port controllers, a module controller per switch module and a set of group translators. The port controller provides the interface between the physical media used to transport ATM cells and the switching system. The adding of the appropriate routing tag for each incoming cell is also one of the main functions of the port controller. The essential part of the control system is the module controller. Its two main functions are reserving the multicast network outputs and updating the multicast translation tables, respectively.

The module controller is also responsible for applying two control schemes. The first one is traffic overflow control on the multicast network outputs. The second scheme is the control of the maximum simultaneous multicast connections within the switch module. The main function of the group translators are to assign a new routing tag to each copy of the multicast cell so that it can be routed to its final destination. One of the major functions of the hierarchical control system is the identification and classification of the incoming cell. This classification must be done to determine the appropriate processing task, which must be implemented on each type. The classification is done according to the ATM forum recommendation.

As indicated in fig.6, the inputs of the first stage are divided into channel groups. The number of these channel groups has to be equal to the maximum allowable number of partitions, which can be added to the system. For example if the maximum number of partitions are k, then the number of channel groups must be from G1 to Gk. Each incoming cell must be directed to a certain channel group according to the location of the required destination. For example, if the required destination is included in the fourth partition, then the cell must be routed to the channel group number 4 (G4).

If the maximum available capacity of the system is duplicated, then the maximum number of partitions, which can be connected to the system, has to be duplicated. For example, if the system contains k identical partitions, to duplicate the maximum number of partitions, to 2k, then the number of channel groups within the first stage must be duplicated (from G1 to G2K). As indicated in fig.7, to duplicate the number of the first stage switch module outputs, an additional switchboard must be added to its point to point switch. The first stage switch module can be designed in a way such that future expansion can be implemented up to four times as indicated in fig.8. This can be done by using a group of binary trees with one input to four outputs each. By using these trees, up to four switchboards can be added to the point to point switch. The additional switchboard is added and easily connected to the binary trees (one output of each binary tree is connected to one input of the new switchboard). The interconnection is done without changing the already existing connections. As indicated in the figure, the output B, C, D of the binary trees are spare for future expansion.

III. 2- The Middle Stage Switch Module

As mentioned before, the middle stage switch modules are used to gather the incoming cells from different partitions, then distributes them over the output switch modules within the same partition. Therefore, the middle stage switch module is a point to point switch, which used as an interconnection stage between the input and the output switch modules. The dimensions of this switch module are determined according to the maximum number of partitions within the system and the number of switch modules within the partition.

As indicated in fig.9, the inputs of the middle stage switch module are divided into channel groups G1 to Gk. where k is the maximum number of partitions within the system. The outputs of the switch module are divided into channel groups from G1 to Gh`, where h` is the number of output switch modules within the partition. The number of inputs of the middle stage switch module are dependent on two parameters, the first parameter is the number of partitions within the system. The second parameter is the number of channels within the channel group between the first and the second stage. The number of outputs of the middle stage switch module is dependent on two another parameters, the first one is the number of output switch modules within the partition. The second parameter is the number of channels within the channel group between the second and the third stage.

To duplicate the system capacity, the number of inputs of the second stage switch module must be duplicated. This can be done by adding a new switchboard to the existing one. As indicated in fig.10, the two switchboards are interconnected by using a group of multiplexers. The number of the required multiplexes is equal to the number of outputs of the middle stage switch module. As a result, the number of outputs of the second stage switch module are constant and not affected by the system expansion. For example, if the dimensions of the middle stage are m × L, when the system capacity is duplicated, the new dimensions must be 2m x L.

III. 3- The Third Stage Switch Module (The Output Switch Module).

While the dimensions of the first and the second stage switch modules are determined according to the maximum capacity of the system, the dimensions of the output switch module are constant and not affected by the variation of the system size. The number of inputs of the third stage switch module are determined according to the number of middle stage switch modules. The number of links within the channel group between the second and the third stage are also taken into account while determining the number of inputs of the output switch module.

The inputs of the switch module are divided into channel groups G1 to Gh, where h is the number of the second stage switch modules. The number of outputs of the third stage switch module are determined according to the number of ATM output ports which must be served within one switch module. Therefore, a concentration process must be implemented to concentrate the number of outputs to the required number of ATM output ports.

The process of concentration is done by using a group of multiplexers. The number of the required multiplexes is equal to the number of ATM output ports within the switch modules. For example, if the dimensions of the third stage switch module are m` x n`, then n` multiplexers are required as indicated in fig.11.

IV. THE ROUTING THROUGH THE LSMA

In LSPM system, the destination address is divided into three parts. The first part is used to route the cell through the first stage to the appropriate partition. The second part is used to route the cell through the second stage to the appropriate output switch module on which the required destination is connected. The third part of the destination address is used to route the cell through the output switch module to the final destination.

The three parts of the destination address are added to each incoming cell through the first stage. This can be done by using the three level hierarchical control system.

The three parts of the destination address are:
Gk: The channel group number between the first and the second stage. Where k is the number of partitions within the system.
Gh`. The channel group number between the second and the third stage. Where h` is the number of output switch modules within the partition.
Tn : The output port number at the output switch module.
Where n is the number of output ports within the output switch module.

The steps of the cell header modification through the stages of the LSMA system are indicated in fig. 12. At the first stage, the hierarchical control system adds the OR parameter which is used to route the cell through the multicast network. The MCN and IR, which are used to determine the destination address, are also added by the control system. At the output of the multicast network, the OR parameter is replaced by the output address value (OA) of the multicast network on which the cell appears. At the group translator, the IR and OA are used to calculate a copy index (CI) for each copy. The CI and MCN are used to index the multicast translation tables (MTT) such that the appropriate destination address can be added to each copy. At the output of the group translator, each cell header contains the three parts of the destination address,

For example, if the second input port in the first switch module existing in the first partition, sends a copy to each of the following destinations:

Partition	Output module	output port
1	2	3
2	1	4
3	3	2

By using the appropriate multicast routing tag, three copies are generated through the multicast network. At the output of the multicast network, the destination address must be added to each copy. This can be done by indexing the multicast translation table.

In this example if the MCN = 30 and the calculated copy indexes are 0, 1, 2 then the destination address are determined as in fig.13. The process of routing through the LSMA system is indicated in fig.14.

V. EVALUATION Of LSMA

One of the major factors affecting the system cost and complexity of the system hardware is the number of the required switch modules. Therefore, in order to evaluate the proposed system the number of the switch modules required for constructing the system is calculated. Then a comparison between the number of switch modules of the proposed system, and other systems is presented. The middle stage switch modules have great effect on the total number on modules because they are used to interconnect the horizontal partitions, i.e. to construct the large-scale switch. Therefore, one way of optimizing the system design is to reduce the number of the required middle stage switch modules. In the following section the number of the middle stage switch modules for LSMA is calculated and compared with the systems in [9,10] in order to evaluate the proposed architecture.

The switch fabric size, the number of switch modules within the system, the required number of middle stage switch modules within the partition and the total number of middle stage switch modules within the system is calculated as following

Number of first stage switch modules in the partition = Number of second stage switch modules in the partition and is given by :

$$h = \frac{m`}{r} \tag{1}$$

Number of the third stage switch modules in the partition is:

$$h` = \frac{L}{r} \tag{2}$$

The maximum number of partitions within the system is calculated by:

$$K = \frac{m}{r} \tag{3}$$

The number of switch modules within the partition $= 2 * \frac{m`}{r} + \frac{L}{r}$ (4)

For the purpose of comparison with other systems, the following equations must be defined.

The number of middle stage switch modules within the partition is:

$$S = \frac{m`}{r} \tag{5}$$

The number of switch modules within the system is:
Given by:

$$Y = K\left[2\frac{m`}{r} + \frac{L}{r}\right] \tag{6}$$

The switch fabric size is given by:

$$N = n * h * K = n * \frac{m`}{r} * \frac{m}{r} \tag{7}$$

Next, the number of switch modules, and the number of middle stage switch modules of the Dilated Banyan system and the growable system must be defined.

For the dilated Banyan system: It has to be noticed that,, Liew assumed an identical switched are used ;i.e. r=r`,m=m`and L=L.`

The maximum number of partitions within the system is:

$$K = \frac{m}{r} \tag{8}$$

The number of middle stage switch modules within the partition is:

$$S = \frac{m}{r} \tag{9}$$

The number of middle stage switch modules within the system is:

$$X = K * \frac{m}{r} = K^{2} \tag{10}$$

The number of first stage switch modules within the partition = the number of third stage switch modules within the partition ; is given by :

$$h = \frac{L}{r} \tag{11}$$

The number of switch modules within the system is:

$$Y = K\left[2\frac{L}{r} + \frac{m}{r}\right] \tag{12}$$

The switch fabric size is given by:

$$N = n * h * K = n * \frac{m}{r} * \frac{L}{r} \tag{13}$$

For the growable ATM switching fabric:

The number of switch modules within the partition is equal to 3h.
The number of interconnection modules within the partition is S= h.
The number of interconnection modules within the system =hk (K-1).

Thus the number of switching elements within the system is given by:

$$Y = 3hK + K(K-1)h \tag{14}$$

Finally the switch fabric size is given by:

$$N = n * h * K \tag{15}$$

Table 1 shows a comparison between the proposed system and the other two discussed systems. The number of the required middle stage switch modules within the partition and within the system is calculated for different system capacities. As indicated in the table, for the proposed system, the number of middle stage switch modules within the partition are constant for different system capacities. For the dilated Banyan system, they are equal to the number of partitions within the system. Thus their value increase with the increase of the system capacity. Finally it is shown that, for the proposed system, the number of the required middle stage switch modules within the system are less than the two other systems for the some switch fabric size.

Table 2, shows the overall number of switch modules within the system for the proposed system and the two other systems for the switch fabric size of 16384 x 16384.
As shown is the table, for the same switch size, the proposed system uses the smallest overall number of switch modules.

CONCLUSION

In this paper, an architecture for scalable multicast ATM switch is introduced. The switch capacity can be extended to 20,000 port with slide change in the interconnection channel groups. On the other hand, the number of the switch modules and respectively, the cost and the complexity of the switch are controlled.

Table 1 Comparison of the Number of Middle Stage Switch Modules

				The proposed system		Dilated Banyan		Growable	
N	n	h	K	S	X	S	X	S	X
8192	64	8	16	8	128	16	256	8	1920
12288	64	8	24	8	192	24	576	8	4416
16384	64	8	32	8	256	32	1024	8	7936
24576	64	8	48	8	384	48	2304	8	18048

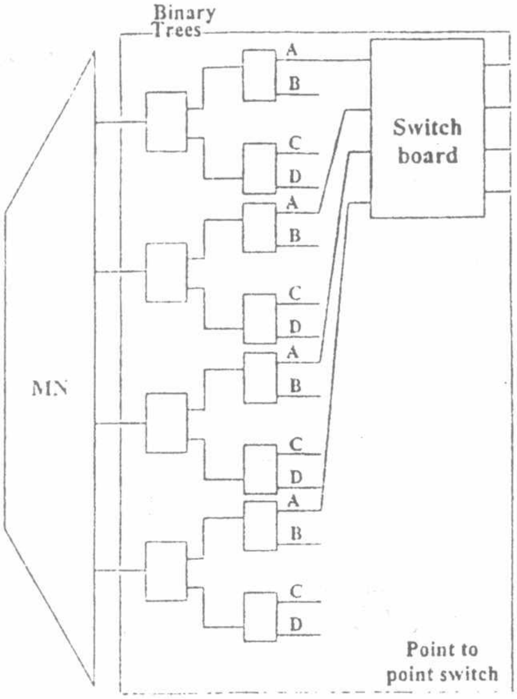
Table 2 Comparison of the Number of Switch Modules

	Fabric Size	No. of Switches

Growable
n = 64, h = 8
K = 16
	n * h * K
	64 * 8 * 16
	16384
		8320

Dilated Banyan
n = 64, h = 8
k = 16
	n * h * K
	64 * 8 * 16.
		1536

The proposed system
n = 64, h = 8
k = 16
	n * h * K
	16384
		768



B, C, D, Spare for system expansion

Fig. 8 The first stage switch module with future expansion facility up to four times

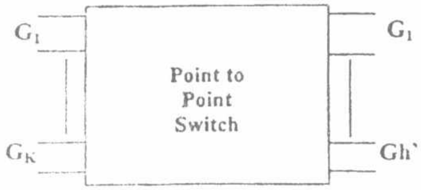


Fig. 9 The middle stage switch module

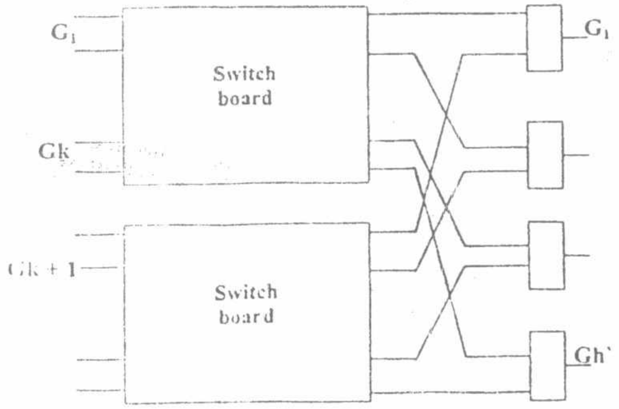


Fig. 10 The middle stage switch module with basic system size duplication

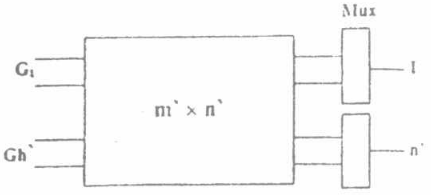


Fig. 11 The third stage switch module

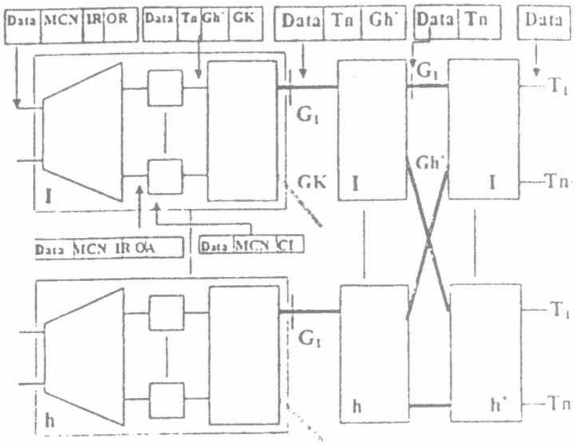
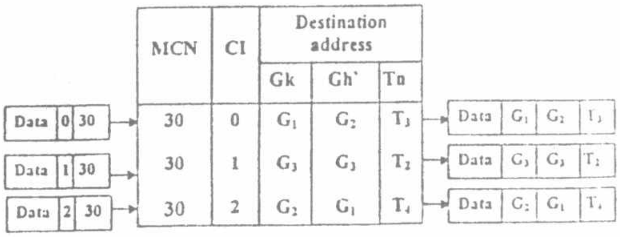


Fig. 12 The cell header modification during the LSMA stage



	MCN	CI	Destination address			
			Gk	Gh`	Tn	
Data 0 30	30	0	G_1	G_2	T_3	Data G_1 G_2 T_3
Data 1 30	30	1	G_3	G_3	T_2	Data G_3 G_3 T_2
Data 2 30	30	2	G_2	G_1	T_4	Data G_2 G_1 T_4

Fig. 13 Determination of the destination address

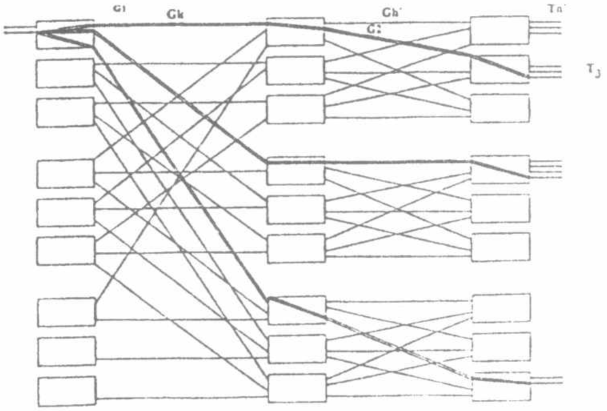


Fig. 14 The routing process through the LSMA system

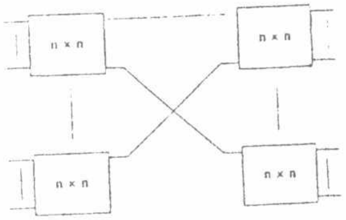


Fig. 1 The construction of large switch using two stage configuration

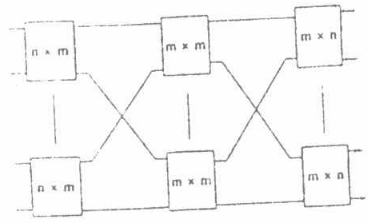


Fig. 2 Three stage switching network (clos network)

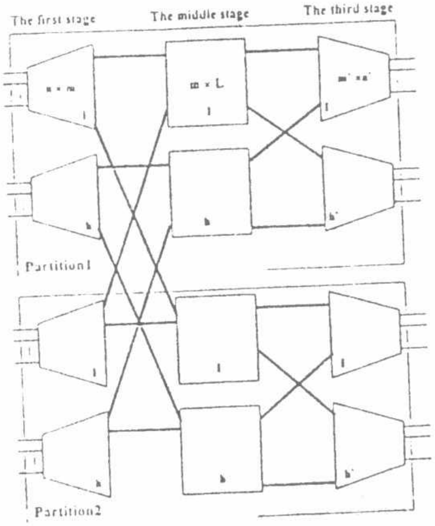
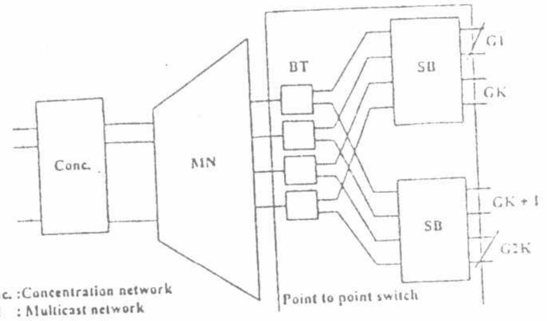
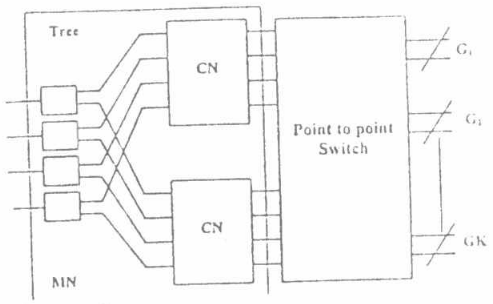


Fig. 4 duplication of LSMA switch size



Conc. :Concentration network
MN : Multicast network
SB : Switch board
BT : Binary tree

Fig. 7 Size duplication of first stage switch module



MN : Multicast network
CN : Copy network
Gk : Channel group number K

Fig. 6. The first stage switch module of the LSMA system

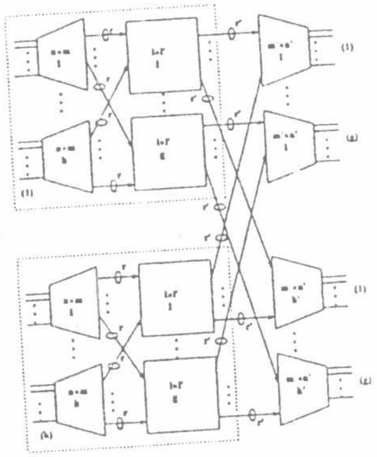


Fig. 3 Three – stage Dilated – Banyan Interconnection Architecture

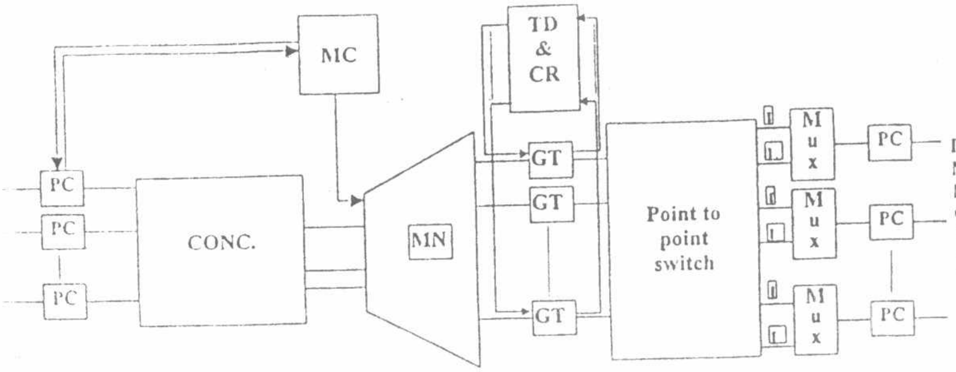


Fig. 5 Multicast ATM switch module

PC : Port Controller
MC : Module Controller
MN : Multicast Network
GT : Group Translator
MUX : Multiplxer
CONC : Concentration network
TD&CR : Traffic distribution and contenation resolution circuit

References

[1] **CCITT Rec. I.150,** "BISDN ATM functional characteristics," 1992

[2] **Nolle T.,** "Voice and ATM : Is anybody talking", Business comm. Review, PP. 42 – 47, June 95.

[3] **Tobagi F. A.,** "Fast packet switch architectures for broad band integrated services digital network", proc. IEEE, Vol 78, No. 1, PP. 133 – 167, Jan 90.

[4] **Wong P. C., Tung E. H.,** "A large scale switch interconnection architecture using overflow switches", IEEE-ICC, GENEVA, PP.708-714, May 23-26,1993.

[5] **Law K., Loan–Garcia E. A.,** "A large scalable ATM Multicast switch", IEEE JSIC, Vol. 15, No. S., PP. 844 – 854, June 97.

[6] **Lee T. T.,** "A modular architecture for very large packet switches", IEEE trans. Comm,. Vol., 35, PP. 1097 – 1106, July 90.

[7] **Pattvina A.,** "Multichannel bandwidth allocation in a broadband packet switch", IEEE J. select. Areas in comm., Vol. 6, PP. 1489 – 1499, Dec. 88.

[8] **Eng K. Y., Karol M. J., Yeh Y. S.,** "A growable packet (ATM) switch architecture Design principles and applications", Proc. of globecom 89, PP. 1159 – 1165.

[9] **Liew S. C., Lu K. W.,** "A- 3- stage interconnection structure for very large packet switches", Proc. of ICC 90, PP. 316.7.1 – 316.7.7.

[10] **Jajszezy K, Kabacinski W.,** "A growable ATM Switching fabric architecture", IEEE J. Select. Areas in comm., Vol. 43, No. 2 / 3 / 4, PP. 1155 – 1168, Feb / March / April 95.

[11] **Amin A. S., Ibrahim H. A.,** "A new hierarchical control system for multicast ATM switch", To be publicized.

[12] **Lee T. T.,** "Nonblocking copy networks for multicast packet switching" IEE J, select. Areas in comm., 6, no.9, PP. 1455 – 1467, Dec. 1988.