PCT

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(54) Title: MULTIPOINT-TO-POINT ARBITRATION IN A NETWORK SWITCH

(57) Abstract

An apparatus and a method are disclosed for arbitrating between streams of ATM cells, or sources for a connection, on multiple input port processors vying for an opportunity to be transmitted as a fixed bandwidth, or allocated, connection on a single output port through a network switch. The network switch maintains a plurality of input port processors, at least one output port, and input and output buffers associated with the respective input and output ports. Streams of ATM cells enter the network switch as sources for a connection through the multiple input port processors and are buffered in the input buffers. The ATM cells are then routed from the input buffers to an output buffer in the output port. The network switch also provides a multipoint topology controller (MTC) and a bandwidth arbiter (BA) for performing the arbitration. The MTC looks up a Fan-In-Number (FIN) identifier for every cell time slot in a switch allocation table (SAT) for which a scheduling list has an entry, assuming that the input buffers, which are arranged into queues, that are listed on the scheduling list have ATM cells to send and the scheduling list has not exceeded its bandwidth allocation. The BA compares all the valid FIN identifies and, if there is a match, performs multipoint-to-point arbitration among the input port processors that have scheduling lists which reference that FIN. The BA reads the FIN state information and gives the bandwidth to the input port processor having the highest priority. The BA then updates the FIN state information with data that indicates which input port processor will have the highest priority on the next arbitration for this multipoint-to-point connection.

- 1 -

MULTIPOINT-TO-POINT ARBITRATION IN A NETWORK SWITCH

CROSS-REFERENCE TO RELATED APPLICATIONS

A claim of priority is made to provisional application 60/001,498, entitled COMMUNICATION METHOD AND APPARATUS, filed July 19, 1995.

FIELD OF THE INVENTION

The present invention is generally related to network switching and, more particularly, to an apparatus and a method for arbitrating between streams of data cells, or sources for a connection, on multiple input port processors vying for an opportunity to be transmitted as a fixed bandwidth, or allocated, connection on a single output port through a network switch.

BACKGROUND OF THE INVENTION

Telecommunications networks such as asynchronous transfer mode (ATM) networks are used for the transfer of audio, video, and other data. ATM networks deliver data by routing data units such as ATM cells from a source to a destination through switches. Switches typically include multiple input/output (I/O) ports through which ATM cells are received and transmitted. The appropriate output port to which a received ATM cell is to be routed to and thereafter transmitted from is determined based upon an ATM cell header.

Oftentimes, the cell headers of ATM cells received on several different input port processors specify a single particular output port. The occurrence of such an event does not usually pose a problem unless the bandwidth capability of the connection through the output port is less than the sum of the bandwidth rates of the connections on the input port processors for an extended period of time. Such a problem is resolved by either discarding ATM cells which

- 2 -

cannot be included within the allocated output bandwidth or asserting flow control back at the connection source.

A further problem exists, however, in that the discarding of ATM cells has heretofore not been evenly distributed among the several different input port processors. That is, ATM cells have been disproportionately discarded among the connections from the several different input port processors. Such can result in a corresponding disproportionate diminishment of data quality causing, for example, undesirable interruptions in audio and video data transmissions or other more serious damage to other types of data transmissions. Accordingly, it would be desirable to devise a scheme whereby the allocated output bandwidth of a connection at an output port is evenly distributed among sources for a connection from the several different input port processors in a multipoint-to-point switching scenario.

SUMMARY OF THE INVENTION

An apparatus and a method are disclosed for arbitrating between streams of data cells, or sources for a connection, on multiple input port processors vying for an opportunity to be transmitted as a fixed bandwidth, or allocated, connection on a single output port through a network switch. The network switch comprises a plurality of input port processors, at least one output port, and input and output buffers associated with the respective input and output ports. Streams of data cells enter the switch as sources for a connection through multiple input port processors and are buffered in the input buffers. The data cells are then routed from the input buffers to the output buffer in the output port.

The network switch also comprises a multipoint topology controller (MTC) and a bandwidth arbiter (BA) for performing the arbitration. The MTC looks up a Fan-In-Number (FIN) identifier for every cell time slot in a switch allocation table (SAT) for which a scheduling list has an entry,

- 3 -

assuming that the input buffers, which are arranged into queues, that are listed on the scheduling list have data cells to send and the scheduling list has not exceeded its bandwidth allocation. The BA compares all the valid FIN identifiers and, if there is a match, performs multipoint-to-point arbitration among the input port processors that have scheduling lists which reference that FIN. The BA reads the FIN state information and gives the bandwidth to the input port processor having the highest priority. The BA then updates the FIN state information with data that indicates which input port processor will have the highest priority on the next arbitration for this multipoint-to-point connection.

From the above descriptive summary it is apparent how the present invention apparatus and method overcome the shortcomings of the above-mentioned prior art.

Accordingly, the primary object of the present invention is to provide an apparatus and a method for arbitrating between sources for a connection on multiple input port processors vying for an opportunity to be transmitted as a fixed bandwidth, or allocated, connection on a single output port through a network switch.

The above-stated primary object, as well as other objects, features, and disadvantages, of the present invention will become readily apparent from the following detailed description which is to be read in conjunction with the appended drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

In order to facilitate a fuller understanding of the present invention, reference is now made to the appended drawings. These drawings should not be construed as limiting the present invention, but are intended to be exemplary only.

Fig. 1 is a block diagram of a network switch according to the present invention.

- 4 -

Fig. 2 is illustrates the operation of several switch allocation tables in the network switch according to the present invention.

Fig. 3 is a block diagram illustrating a multipoint-to-point switching scenario.

DETAILED DESCRIPTION OF THE PRESENT INVENTION

Referring to Fig. 1, there is shown a network switch 1 comprising a Data Crossbar 10, a Bandwidth Arbiter (BA) 12, a plurality of input port processors 14, a plurality of output ports 16, and a plurality of Multipoint Topology Controllers (MTC) 18. The Data Crossbar 10, which may be an N x N crosspoint switch, is used for data cell transport and, in this particular embodiment, yields N x 670 Mbps throughput. The BA 12 controls switch interconnections, dynamically schedules momentarily unused bandwidth, and resolves multipoint-to-point bandwidth contention. Each input port processor 14 schedules the transmission of data cells to the Data Crossbar 10 from multiple connections. Each output port 16 receives data cells from the Data Crossbar 10 and organizes those data cells onto output links.

In order to traverse the switch 1, a data cell 22 first enters the switch 1 on a link 24 to an input port processor 14 and is buffered in a queue 26 of input buffers. The data cell 22 is then transmitted from the queue 26 of input buffers through the Data Crossbar 10 to a queue 28 of output buffers in an output port 16. From the queue 28 of output buffers, the data cell 22 is transmitted onto a link 30 outside of the switch 1, for example, to another switch. To facilitate traversal of the switch 1, each input port processor 14 includes a cell buffer RAM 32 and each output port 16 includes a cell buffer RAM 34. The cell buffer RAM's 32 and 34 are organized into the respective input and output queues 26 and 28. All data cells 22 in a connection pass must through a unique input queue 26 and a unique output queue 28 for the life of the connection. The queues 26 and

- 5 -

28 thus preserve cell ordering. This strategy also allows quality of service ("QoS") guarantees on a per connection basis.

Three communication paths are used to facilitate traversal of the switch 1 via probe and feedback messages: a Probe Crossbar 42, an XOFF Crossbar 44, and an XON Crossbar 46. The Probe Crossbar 42, which in this particular embodiment is an N x N crosspoint switch, is used to transmit a multiqueue number from an MTC 18 to an output port 16. Each input port processor 14 includes a plurality of scheduling lists 47, each of which is a circular list containing input queue numbers for a particular connection. Each multiqueue number is derived from information provided to the MTC 18 from a scheduling list 47 in an input port processor 14. A multiqueue number identifies one or more output queues 28 to which a data cell may be transmitted when making a connection. An output port 16 uses the multiqueue number to direct a request message probe to the appropriate output queue or queues 28 and thereby determine if there are enough output buffers available in the output queue or queues 28 for the data cell.

The XOFF Crossbar 44, which in this particular embodiment is an N x N crosspoint switch, is used to communicate "DO NOT SEND" type feedback messages from an output port 16 to an input port processor 14. An XOFF feedback message is a two bit message, wherein the first bit of the message is an ACCEPT/REJECT bit and the second bit of the message is an XOFF/NO-OP bit. When the first bit of the XOFF feedback message, namely the ACCEPT/REJECT bit, is set to "0", the output port 16 has "accepted" a request message probe from an input port processor 14 to transfer a data cell to a particular output queue 28. Thus, a data cell may then be transferred from an input queue 26 through the Data Crossbar 10 to that output queue 28.

Conversely, when the first bit of the XOFF feedback message is set to "1", the output port 16 has "rejected" a

- 6 -

request message probe from an input port processor 14 to transfer a data cell to a particular output queue 28. In this case, a data cell may not be transferred from an input queue 26 through the Data Crossbar 10 to that output queue

5   28, usually because insufficient buffer space is available to receive a data cell. However, further request message probes may still be able to be sent from the input port processor 14 depending upon the state of the XOFF/NO-OP bit in the XOFF message as described below.

10   When the second bit of the XOFF feedback message, namely the XOFF/NO-OP bit, is set to "1", the scheduling list 47 in the input port processor 14 that provided information to the MTC 18 in order to derive the multiqueue number for the request message probe is placed in an XOFF state. This means

15   that the scheduling list 47 may no longer be used to initiate the sending of request message probes. The scheduling list 47 remains in an XOFF state until receiving an XON message from the output port 16, as described below. An input port processor 14 responds to an asserted XOFF feedback message

20   by modifying XOFF state bits in a descriptor of the scheduling list 47. The XOFF state bits prevent the input port processor 14 from attempting to send a request message probe from the input port processor 14 to the output port 16 until notified by the output port 16 that output buffers are

25   available for a corresponding connection.

When the second bit of the XOFF feedback message, namely the XOFF/NO-OP bit, is set to "0", a no operation message is indicated, meaning that no adverse action has been taken with respect to the scheduling list 47 in the input port processor

30   14 that provided information to the MTC 18 in order to derive the multiqueue number for the request message probe. In other words, the scheduling list 47 may still be used to provide request message probes.

The XON Crossbar 46, which in this particular embodiment

35   is an N x N crosspoint switch, is used to communicate "ENABLE SEND" type messages from an output port 16 to an input port

- 7 -

processor 16. More particularly, the XON Crossbar 46 communicates an XON feedback message from an output port 16 to an input port processor 14. When an XOFF feedback message has been asserted by an output port 16 in response to a request probe message from an input port processor 14, the output port 16 sets a state bit in a queue descriptor of a corresponding output queue 28. When the number of data cells in that output queue 28 drops below an XON threshold, an XON message is sent from that output port 16 to the input port processor 14. The XON message enables the scheduling list 47 in the input port processor 14 to be used in the sending of request probe messages, and hence data cells.

The Probe & XOFF communication paths operate in a pipelined fashion. First, an input port processor 14 selects a scheduling list 47, and information associated with that scheduling list 47 is used to determine the output port 16, or the output queue 28, to which a data cell will be transmitted. More particularly, a multiqueue number, which is derived from information provided to an MTC 18 from a scheduling list 47 in an input port processor 14, is transmitted from the MTC 18 to one or more output ports 16 using the Probe Crossbar 42. Each output port 16 then tests for buffer availability and asserts a "DO NOT SEND" type feedback message through the XOFF Crossbar 44 if output buffering is not available for that connection. If output buffering is available for that connection, the input port processor 14 transmits a data cell to one or more output queues 28 through the Data Crossbar 10.

Referring now to Figs. 1 and 2, each input port processor 14 within the switch 1 includes a Switch Allocation Table (SAT) 20 for mapping bandwidth allocation. SAT's 20 are the basic mechanism behind the scheduling of data cells. Each SAT 20 includes a plurality of sequentially ordered cell time slots 50 and a pointer 52 which is alway directed to one of the cell time slots 50. All of the pointers 52 in the switch 1 are synchronized such that at any given point in

- 8 -

time each of the pointers 52 is directed to the same cell
time slot 50 in the respective SAT 20 with which the pointer
52 is associated, e.g., the first cell time slot.   In
operation, the pointers 52 are advanced in lock-step, with

5   each cell time slot 50 being active for 32 clock cycles at
50 MHz.   When a pointer 52 is directed toward a cell time
slot 50, an input port processor 14 uses the corresponding
entry 51 in the cell time slot 50 to obtain a data cell for
launching into the Data Crossbar 10.

10   A counter in each input port processor 14 is incremented
once for each cell time, and each pointer 52 returns to the
first cell time slot 50 after reaching the last cell time
slot 50.   Hence, given an SAT 20 with a depth of 8k, which
defines a frame, each pointer 52 scans its respective SAT 20

15   approximately every 6 msec, thereby providing a maximum delay
for a transmission opportunity of approximately 6 msec.   The
delay can be decreased by duplicating a given entry 51 at a
plurality of cell time slots 50 within a SAT 20.   The maximum
delay that an incoming data cell will experience corresponds

20   to the number of cell time slots 50 between the pointer 52
and the cell time slot 50 containing the entry 51 which
specifies the location of the data cell.   When multiple
entries 51 are made in order to decrease the maximum possible
number of separating cell time slots 50, the duplicate

25   entries 51 are preferably spaced equidistantly within the SAT
20.   Maximum delay for a transmission opportunity therefore
corresponds to the frequency and spacing of duplicate entries
51 within the SAT 20.

          The amount of bandwidth allocated to a particular
30   connection corresponds to the frequency at which a given
entry 51 appears in a SAT 20.   Each cell time slot 50
provides 64 kbps of bandwidth.   Since a pointer 52 cycles
through a SAT 20 at a constant rate, the total bandwidth
allocated to a particular connection is equal to the product

35   of 64 kbps times the number of occurrences of that entry 51.
For example, connection identifier "g (4,6)," which occurs

- 9 -

in five cell time slots 50, is allocated 320 kbps of bandwidth.

Significantly, unused cell time slots 60 correspond to unused bandwidth, which becomes available during the operation oí the switch 1. Such unused bandwidth may occur because that bandwidth, i.e., that cell time slot 50 in the SAT 20, has not been allocated to any connection. Such bandwidth is referred to as "unallocated bandwidth." Unused bandwidth may also occur when a SAT entry 51 is allocated to a connection, but the connection does not have a data cell enqueued for transmission across the Data Crossbar 10. Such bandwidth is referred to as "allocated-unused" bandwidth. Both types of unused bandwidth are collectively referred to as "dynamic" bandwidth, and some connections, such as connections assigned an Available Bit Rate ("ABR") QoS level utilize such dynamic bandwidth. The BA 12 operates to increase efficiency within the switch 1 by granting dynamic bandwidth to such connections.

If valid, the contents of each SAT entry 51 point to a scheduling list 47. The contents of each (non-empty) entry in a scheduling list 47 consists of an input queue number. Each input queue number points to a input queue descriptor which contains state information that is specific to a particular connection. Each input queue descriptor, in turn, points to the head and the tail of a corresponding input queue 26, which contains data cells for transmission through the switch Data Crossbar 10.

If a SAT entry 51 does not contain a pointer to a scheduling list, i.e. the SAT entry 51 is set to zero, then the corresponding cell time slot 50 in the SAT 20 has not been allocated and that cell time slot 50 is available for dynamic bandwidth. Also, if a SAT entry 51 does contain a pointer to a scheduling list 47 but no queue number is listed in scheduling list 47, then there are no data cells presently available for transmission and the corresponding cell time slot 50 is also available for dynamic bandwidth.

- 10 -

Fig. 3 illustrates a multipoint-to-point switching scenario, i.e., transmission from multiple input queues 26 to a single output queue 28. As previously described, each output queue 28 has a threshold, and an XON message is sent from an output port 16 to an input port processor 14 when an output queue 28 drains below that threshold. In multipoint-to-point operation, the XON threshold of an output queue 28 is dynamically set to reserve enough output buffering for multiple input queues 26 to transmit to the output queue 28. For example, if four input queues 26 are transmitting data cells, the threshold is set to four so that the output queue 28 will have sufficient output buffering to receive all four of the data cells contemporaneously in serial fashion.

Referring now to Figs. 1 and 3, in the case of multipoint-to-point connections, the XON Crossbar 46 is used to broadcast to every input port processor 14 in the switch 1, regardless of whether or not any particular input port processor 14 is transmitting to the output queue 28 asserting the broadcast. For the broadcast, the MTC 18 transmits a reverse broadcast channel number on behalf of the output port 16. The receiving MTC 18 then performs a reverse broadcast channel to scheduling list number look-up to determine which scheduling list 47 to enable. Any input port processor 14 without an input queue 26 transmitting to that particular output queue 28 is unaffected by the broadcast XON message since the reverse broadcast channel number look-up entry will be marked invalid.

A Fan-In-Number, or FIN, is a mechanism to associate scheduling lists 47 from different input port processors 14 in a multipoint-to-point connection, and have them compete against each other for bandwidth using an arbitration mechanism. This arbitration mechanism allows switch control software to load SAT entries 51 on multiple input port processors 14 with scheduling lists 47 that are transmitting to the same output port processor 16. Without this arbitration mechanism, such would be flagged as a programming

- 11 -

error, i.e. it is not legal for multiple input port
processors 14 to send data cells to a single output port
processor 16. A FIN is required for every multipoint-to-
point connection. The BA 12 uses the FIN to look-up state
information for each multipoint-to-point connection.
Basically, the FIN state information indicates which input
port processor 14 was awarded bandwidth for the last
arbitration for a particular multipoint-to-point connection.
Based on this state information, the BA 12 either grants
bandwidth to an input port processor 14, and hence a
scheduling list 47, requesting bandwidth or denies the
bandwidth because the bandwidth limit for a connection has
been exceeded. The FIN state information also includes which
scheduling list 47 received the bandwidth from the last
arbitration. The BA 12 then picks the scheduling list 47
with the next highest priority as the winner of the
multipoint-to-point arbitration. This is effectively round-
robin sharing of the multipoint-to-point bandwidth.

        If all of the input queues 26 for a connection competing
in a multipoint-to-point arbitration are on a single input
port processor 14, FIN's do not have to be used. In such a
case, the input port processor 14 will perform the
multipoint-to-point arbitration by sharing the bandwidth
among the input queues 26 listed in a scheduling list 47.
This case occurs when there are sources to a connection on
multiple input links 24 feeding into an input port processor
14.

        To ensure that all scheduling lists 47 competing in a
multipoint-to-point arbitration all get a turn, the BA 12
grants the bandwidth to the scheduling lists 47 in a round-
robin fashion. To perform this, the BA 12 keeps track of the
last input port processor 14 associated with the scheduling
list 47 that received bandwidth for each FIN identifier. The
next time scheduling lists 47 compete for bandwidth having
that FIN identifier, the BA 12 gives the bandwidth to the
input port processor 14 having the next highest priority.

- 12 -

A FIN identifier is associated with each scheduling list 47 and is stored in the MTC 18. If a scheduling list 47 is not part of a multipoint-to-point connection, its FIN is set to invalid. To perform the arbitration the following sequence is followed:

1.   The MTC 18 looks up a FIN identifier for every cell time slot 50 in a SAT 20 for which a scheduling list 47 has an entry, assuming that the input queues 26 listed on that scheduling list 47 have data cells to send and the scheduling list 47 has not exceeded its bandwidth allocation.

2.   The BA 12 compares all of the valid FIN identifiers and, if there is a match, performs multipoint-to-point arbitration among the input port processors 14 that have scheduling lists 47 which reference that FIN.

3.   The BA 12 reads the FIN state information, and gives the bandwidth to the scheduling list 47 having the next highest priority.

4.   The BA 12 updates the FIN state information with data that indicates which input port processor 14 will have the highest priority on the next arbitration for this multipoint-to-point connection.

Of course, the above-described arbitration may be simultaneously conducted for numerous multipoint-to-point allocated connections competing for different output ports in the network switch 1.

Multipoint-to-point arbitration and flow control are coupled in the network switch 1 to provide fairness guarantees on the allocated portion of a multipoint-to-point connection. Typically, when a group of scheduling lists 47 are disabled due to flow control and then enabled at the same time a starvation condition may occur if one or a set of scheduling lists 47 consistently get chosen so that data cells from input queues 26 pointed to by those scheduling lists 47 get to be transmitted before data cells from input queues 26 pointed to by any other particular scheduling list 47 in the group get to be transmitted. To implement

- 13 -

fairness, each input port processor 14 communicates XOFF state information to the BA 12 in conjunction with the FIN state information.   The BA 12 uses this XOFF state information during multipoint-to-point arbitration to prevent starvation for the allocated portion of the connection by implementing round-robin allotment.

Table 1 shows how the XOFF history information is utilized by the BA 12.

- 14 -

TABLE 1

SAT XOFF bit	History bit	Comment
0	0	No special action
0	1	The scheduling list will be granted the bandwidth. If several scheduling lists with the same FIN have their history bit set, one arbitrary one is picked. Eventually all scheduling lists which have a set history bit will be granted the bandwidth. If a scheduling list had its history bit set, the FIN state information will not be updated.
1	x	If one of the scheduling lists are XOFFed, none of the scheduling lists with the same FIN are granted the bandwidth.

The SAT XOFF bit indicates the XOFF state of a scheduling list 47, wherein 0 = NO-OP and 1 = XOFF. The history bit indicates the ACCEPT/REJECT status of the scheduling list 47, i.e., whether the previous request message probe resulted in an ACCEPT or REJECT feedback message, wherein 0 = ACCEPT, 1 = REJECT. If the scheduling list 47 is in the XOFF state, no priority is given. Further, if the scheduling list 47 is in the XON state and the previous request message probe resulted in an ACCEPT feedback message, no priority is given. However, if the scheduling list 47 is in the XON state and the previous request message probe resulted in a REJECT feedback message, priority is given to the scheduling list 47. If multiple scheduling lists 47 have priority, the bandwidth is randomly allocated to the scheduling list having the highest priority. It will thus be appreciated that previously rejected scheduling lists 47 are given priority over other scheduling lists 47 based on the history bit, and fairness is thereby guaranteed.

It will be understood that various changes and modifications to the above described method and apparatus may

- 15 -

be made without departing from the inventive concepts disclosed herein. Accordingly, the present invention is not to be viewed as limited to the embodiments described herein.

- 16 -

CLAIMS

What is claimed is:

5   1.   A method for arbitrating between connections on multiple
input ports vying for an opportunity to be transmitted as an
allocated connection to a single output port through a
network switch, said method comprising the steps of:
          looking up an identifier for a slot in a plurality of
10   switch allocation tables, wherein each slot has at least one
of a plurality of scheduling lists as an entry, wherein each
scheduling list is associated with a connection on an input
port, and wherein each connection is seeking an opportunity
to be transmitted as an allocated connection to an output
15   port;
          comparing every identifier; and
          if there is a match between identifiers, performing an
arbitration among the scheduling lists corresponding to the
matching identifiers, wherein the scheduling list having the
20   highest priority is awarded the opportunity to be transmitted
as an allocated connection to the output port, and wherein
the highest priority is based upon which scheduling list was
the least most recent scheduling list to be awarded the
opportunity to be transmitted as an allocated connection to
25   the output port.

2.   The method as defined in claim 1, further comprising
updating a priority list to reflect which scheduling list was
awarded the opportunity to be transmitted as an allocated
30   connection to the output port.

3.   An apparatus for arbitrating between connections on
multiple input ports vying for an opportunity to be
transmitted as an allocated connection to a single output
35   port through a network switch, said apparatus comprising:

- 17 -

a controller for looking up an identifier for a slot in a plurality of switch allocation tables, wherein each slot has at least one of a plurality of scheduling lists as an entry, wherein each scheduling list is associated with a connection on an input port, and wherein each connection is seeking an opportunity to be transmitted as an allocated connection to an output port;

an arbiter for comparing every identifier and, if there is a match between identifiers, performing an arbitration among the scheduling lists corresponding to the matching identifiers, wherein the scheduling list having the highest priority is awarded the opportunity to be transmitted as an allocated connection to the output port, and wherein the highest priority is based upon which scheduling list was the least most recent scheduling list to be awarded the opportunity to be transmitted as an allocated connection to the output port.
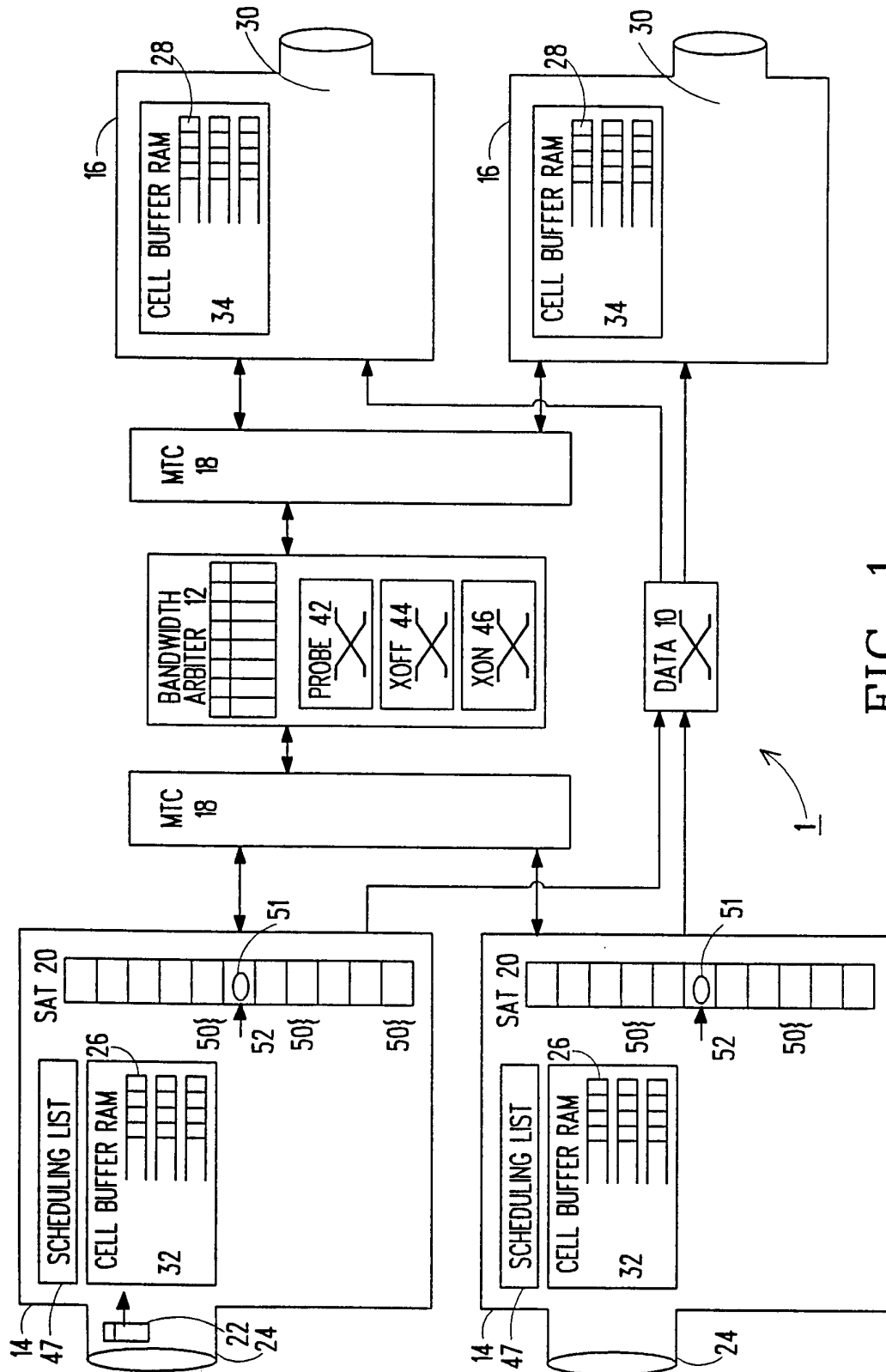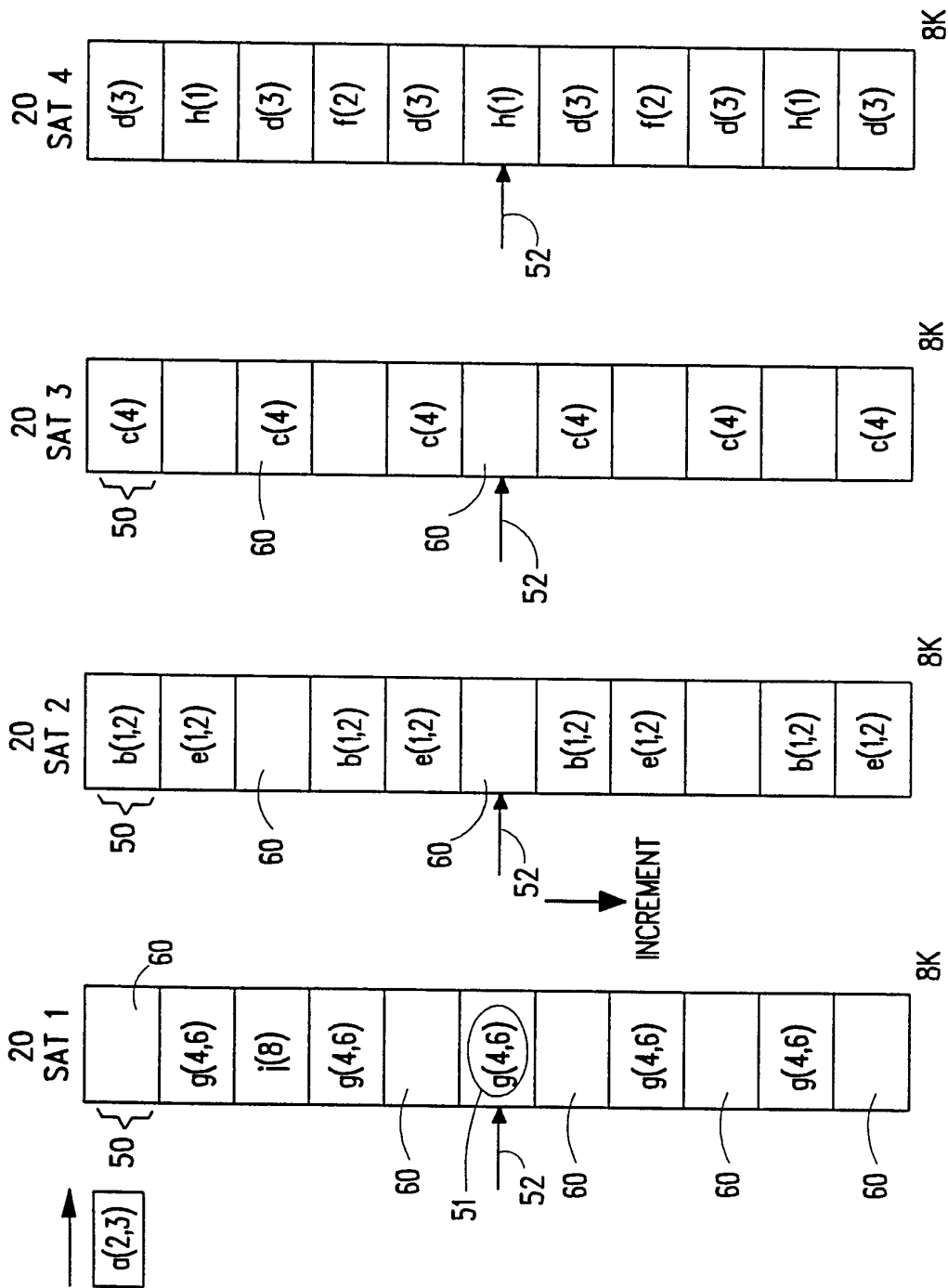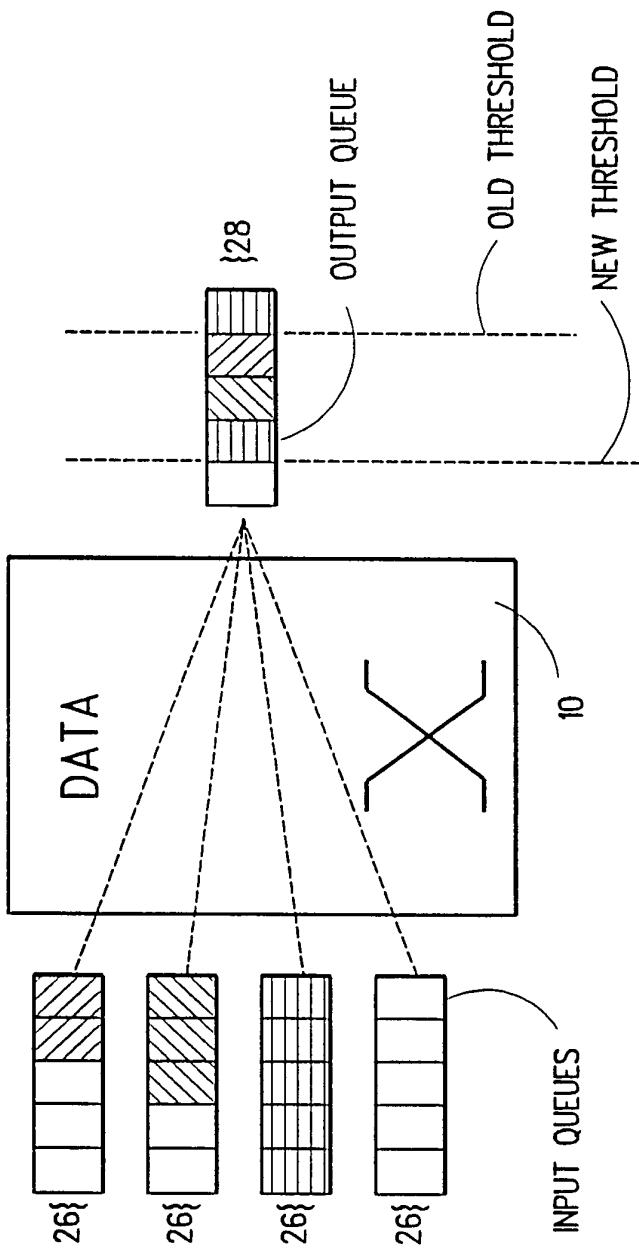
FIG. 1

FIG. 2

FIG. 3