

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
6 February 2003 (06.02.2003)

PCT

(10) International Publication Number
WO 03/010909 A1

(51) International Patent Classification⁷: **H04J 3/14**

(US). **CARTIER, Jean-Francois**; F- 4700 Orvault (FR).
PHILIPPE, Levillain; 4044 Cottonwood Grove Trail,
Agoura Hills, CA 91301-5312 (US).

(21) International Application Number: PCT/US02/23055

(22) International Filing Date: 23 July 2002 (23.07.2002)

(74) **Agent: CHANG, Josephine, E.**; Christie, Parker & Hale,
LLP, 350 West Colorado Boulevard,, Post Office Box 7068,
Pasadena, CA 91109-7068 (US).

(25) Filing Language: English

(26) Publication Language: English

(84) **Designated States (regional)**: European patent (AT, BE,
BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT,
LU, MC, NL, PT, SE, SK, TR).

(30) Priority Data:

60/307,269 23 July 2001 (23.07.2001) US
10/068,710 6 February 2002 (06.02.2002) US

Published:

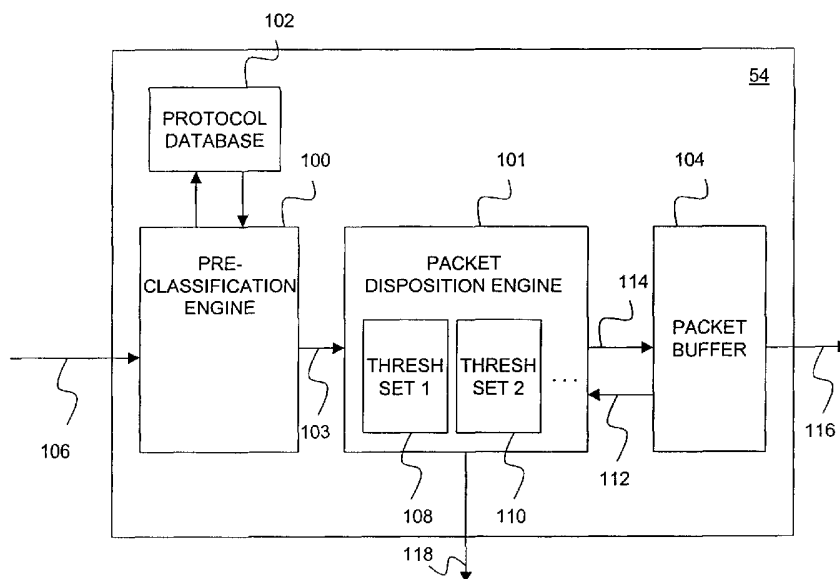
- with international search report
- before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments

(71) **Applicant: ALCATEL INTERNETWORKING, INC.**
[US/US]; 26801 West Agoura Road, Calabasas, CA 91301
(US).

(72) **Inventors: NARSINH, Anees**; 1393 Avenida De Cortez,
Pacific Palisades, CA 90272 (US). **HITZELBERGER,**
James, A.; 3182 Sierra Drive, Westlake Village, CA 91362

For two-letter codes and other abbreviations, refer to the "Guid-
ance Notes on Codes and Abbreviations" appearing at the begin-
ning of each regular issue of the PCT Gazette.

(54) **Title: SWITCHING NODE WITH CLASSIFICATION-DEPENDENT MAC BUFFER CONTROL**



(57) **Abstract:** A system and method for traffic congestion control at the media access controller (MAC) level. The media access controller (54) receives an inbound packet and preclassifies the packet for determining its priority. A utilization level of a queue associated with the determined priority is analyzed for determining its level of congestion. The packet is admitted or not based on the congestion information. Admitted packets are forwarded to the packet switching controller which engages in further classification and congestion control for determining whether the admitted packet is to be forwarded to a destination address.

WO 03/010909 A1

SWITCHING NODE WITH CLASSIFICATION-DEPENDENT
MAC BUFFER CONTROL

FIELD OF THE INVENTION

This invention relates generally to controlling traffic passing through a data communication switch, and more particularly, to preclassifying traffic at a media access controller for traffic congestion control.

BACKGROUND OF THE INVENTION

Because data connections tend to be used intermittently, oversubscription of a switching controller of a data communication switch is generally an effective strategy for achieving more economical network connectivity. Oversubscription is achieved by allocating bandwidth on one or more ports coupled to the switching controller in excess of the actual bandwidth of the switching controller.

Oversubscription tends to increase the volume of traffic supported by a switching controller, leaving the switching controller more prone to congestion. Existing congestion control mechanisms are typically implemented at the switching controller level of the data communication switch. Generally, one or more media access controllers (MAC) resident in the switch receive incoming packets and provide them to the switching controller at wire speed as long as the switching controller is capable of accepting the packets. The packets are transmitted to the switching controller without regard to the type of priority associated with the incoming packets.

If the switching controller stops accepting packets, the MACs attempt to store the packets in one or more temporary buffers, again without any regard to the priority of the packets. Unfortunately, when the temporary buffers are full, the MACs generally discard further incoming packets until space becomes available again. The incoming packets are dropped even if they are associated with a high priority.

There is a need, therefore, for a data switch that provides oversubscription traffic management at the MAC level in addition to the traffic management at the switching controller level. The traffic management at the MAC level should, whenever possible, allow high priority packets to pass to the switching controller and drop low priority packets.

SUMMARY OF THE INVENTION

The present invention is directed to an oversubscription traffic management at an access controller level.

According to one embodiment of the invention, a data communication node forwarding inbound packets includes an access controller and a switching controller. The access controller receives an inbound packet, classifies the packet, and determines whether the packet is to be admitted for forwarding to the switching controller or not based on congestion status data determined from the classification information. If the packet is forwarded, the switching controller receives the packet for further classifying the packet and determines whether the packet is to be further forwarded or not based on additional congestion status data determined from additional classification information.

In another embodiment of the invention, an access controller in a data communication node includes an input receiving an inbound packet, a classification engine coupled to the input classifying the inbound packet, a buffer storing inbound packets, and a disposition engine coupled to the classification engine and the buffer. The disposition engine receives the classification information and determines whether the inbound packet is to be admitted or not based on the classification information and a buffer utilization level. The disposition engine forwards the inbound packet to the switching controller, if the packet is admitted, for determining whether the admitted packet is to be further forwarded.

In a further embodiment of the invention, a method is provided for packet traffic management in a data communication node that includes an access controller and a switching controller. The method includes, at the access controller, receiving an inbound packet, classifying the inbound packet, and obtaining congestion status data from the classification information, admitting the inbound packet or not based on the congestion status data, and forwarding the inbound packet to the switching controller if the packet is admitted. The method further includes, at the switching controller, determining whether the admitted packet is to be further forwarded.

In another embodiment of the invention, a congestion control method is provided in a switching node that includes a media access controller having at least one input coupled to a switching controller having at least one output. The method selectively forwards packets received on the at least one input to the at least one output, and includes receiving a packet on an input and forwarding the packet on an output only if a first congestion check is passed at the media access controller and a second congestion check is passed at the switching

1 controller. The first congestion check is based on first packet classification information and a first buffer utilization level.

5 It should be appreciated, therefore, that the present invention allows packets to be preclassified at an access controller level for use in determining whether the packet is to be admitted and forwarded to the switching controller. The preclassification and congestion avoidance mechanism at the access controller allow packets of higher priority to be admitted over packets of lower priority. Thus, the problem of the prior art of indiscriminately dropping packets at the access controller when the access controller buffer is full is avoided.

10 BRIEF DESCRIPTION OF THE DRAWINGS

These and other features, aspects and advantages of the present invention will be more fully understood when considered with respect to the following detailed description, appended claims, and accompanying drawings where:

15 FIG. 1 is a schematic block diagram of a packet switching node according to one embodiment of the invention;

FIG. 2 is a schematic block diagram of a switching interface according to one embodiment of the present invention;

20 FIG. 3 is schematic block diagram of an access controller according to one embodiment of the present invention; and

FIG. 4 is a flow diagram of a process for traffic congestion control at an access controller level according to one embodiment of the invention.

25 DETAILED DESCRIPTION OF THE SPECIFIC EMBODIMENTS

FIG. 1 is a schematic block diagram of a packet switching node 10 according to one embodiment of the invention. The packet switching node may also be referred to as a switch, a data communication node, or a data communication switch. The packet switching node 10 includes switching interfaces 14, 16 and 18 interconnected to respective groups of local area networks (LANs) 30, 32, 34, and interconnected to one another over data paths 20, 22, 24 via switching backplane 12. The switching backplane 12 preferably includes a switching fabric in a manner that is conventional in the art. The switching interfaces may also be coupled to one another over control paths 26 and 28.

35 The switching interfaces 14, 16, 18 preferably forward packets to and from their respective groups of LANs 30, 32, 34 in accordance with one or more operative

1 communication protocols, such as, for example, media access control (MAC) address based bridging, and Internet Protocol (IP) routing. The switching node 10 is shown for illustrative purposes only. In practice, packet switching nodes may include more or less than three switching interfaces.

5 FIG. 2 is a schematic block diagram of a switching interface 50 according to one embodiment of the present invention. The switching interface 50 may be similar, for example, to the switching interfaces 14, 16, 18 of FIG. 1. The switching interface 50 includes an access controller 54 coupled between LANs and a packet switching controller 10 52. The access controller 54, which may, for example, include a media access controller (MAC), preferably receives inbound packets off LANs, performs physical and MAC layer operations on the inbound packets, and transmits the inbound packets to the packet switching controller 52 for flow-dependent processing. According to one embodiment of the invention, the access controller 54 performs access control operations including 15 preclassification of inbound packets for determining whether the inbound packets are to be admitted based on the preclassification information and a detected congestion level.

The packet switching controller 52 preferably receives the admitted packets forwarded by the access controller 54, classifies the packets, and queues them for 20 downstream congestion control. If the admitted packets are to be forwarded to their destination address based on the congestion control mechanism at the switching controller, the packet switching controller modifies the packets in accordance with flow information and transmits the modified packets on a switching backplane, such as the switching 25 backplane 12 of FIG. 1.

The packet switching controller 52 preferably also receives packets modified by other packet switching controllers via the switching backplane and transmits them to the access controller 54 for forwarding on LANs. The packet switching controller 52 may also subject 30 selected ones of the packets to egress processing prior to transmitting them to the access controller 54 for forwarding on LANs.

FIG. 3 is a more detailed block diagram of the access controller 54 according to one embodiment of the present invention. The access controller 54 preferably includes a packet preclassification engine 100, packet disposition engine 101, protocol database 102, and 35 packet buffer 104. Although the packet preclassification engine 100 and packet disposition engine 101 are illustrated as separate engines, a person skilled in the art should recognize that they may be combined into a single engine or distributed over multiple engines.

1 It is also understood, of course, that FIG. 3 illustrates a block diagram of an access
controller without obfuscating inventive aspects of the present invention with additional
elements and/or components which may be required for the access controller. These
additional elements and/or components, which are not shown in FIG. 3, are well known to
5 those skilled in the art.

The packet preclassification engine 100 is preferably coupled to the protocol database
102 for preclassifying inbound packets 106 based on information contained in the protocol
database. The packet preclassification engine 100 is preferably implemented in an
10 application-specific integrated circuit (ASIC). The protocol database 102 is preferably a
form of content addressable memory (CAM) storing Layer 3 protocol identifiers and their
associated priorities. The packet preclassification engine 100 assigns a priority to the
inbound packets 106 based on the protocol information and/or other header data such as, for
example, 802.1P/Q tag status, Layer 2 encapsulation type, ToS (type of service) values, other
15 connection information, embedded priority information, and/or the like.

The preclassification information 103 is transmitted to the packet disposition engine
101 for determining whether the inbound packets 106 are to be admitted based on the
preclassification information and one or more thresholds 108, 110 set for the packet buffer
20 104. The packet disposition engine 101 is preferably implemented in an ASIC.

The packet buffer 104 preferably includes one or more queues for storing the inbound
packets 106 that have been admitted by the packet disposition engine 101. According to one
embodiment of the invention, the packet buffer 104 includes queues having different
25 priorities for storing packets 106 based on their priority. The packets stored in the queues are
dequeued and forwarded to the switching controller 52, preferably based on a priority-based
dequeuing which is commonly referred to as class-based dequeuing. A person skilled in the
art should recognize, however, that other algorithms may also be utilized for dequeuing the
packets, such as, for example, a paycheck round robin algorithm or a combination of strict
30 priority and weighted round robin.

In general terms, the access controller 54 receives the inbound packets 106 via an
input, such as an inbound cable. The inbound packets 106 may include, but are not limited
to, Ethernet frames, ATM cells, TCP/IP and/or UDP/IP packets, and may consist of Layer 2
35 (Data Link/MAC Layer), Layer 3 (Network Layer), Layer 4 (Transport Layer), or Layer 5
(ATM Adaptation Layer) data units. All or portions of the inbound packet are transmitted to
the packet preclassification engine 100 for preclassification. In this regard, the packet

1 preclassification engine 100 preferably examines an inbound packet's header data and preclassifies the packet for determining its priority. According to one embodiment of the invention, the preclassification engine 100 accesses the protocol database 102 and determines the packet's priority based on its protocol information. According to another embodiment of
5 the invention, the packet's priority is determined based on the packet's information, encapsulation type, ToS values, other connection information, embedded priority information, and/or the like.

The preclassification information and/or all or portions of the inbound packet is
10 transmitted to the packet disposition engine 101. The packet disposition engine determines whether the packet is to be admitted or dropped based on the preclassification information and detected congestion level of the packet buffer 104. The packet disposition engine 101 preferably invokes a weighted random early discard (WRED) algorithm for determining whether the preclassified packet is to be dropped or admitted. The WRED algorithm is a
15 derivative of the RED algorithm, both of which are well known to those skilled in the art. The RED algorithm is described in detail in S. Floyd et. al, "Random Early Detection Gateways for Congestion Avoidance," *IEEE/ACM Transactions on Networking*, 1(4):397-413, August 1993, the content of which is incorporated herein by reference. A person skilled in
20 the art should recognize that the packet disposition engine 101 may utilize other algorithms for determining whether a packet is to be admitted, such as, for example, strict priority, weighted round robin, or other traffic shaping methods which are well known to those skilled in the art.

25 Preferably, the packet disposition engine 101 maintains at least two thresholds for each priority queue in the packet buffer 104, a minimum threshold and a maximum threshold. If a RED or WRED congestion control algorithm is utilized, the packet disposition engine further maintains a discard probability for each priority. According to one embodiment of the invention, the thresholds and discard probabilities set for the queues vary based on their
30 priorities.

The packet disposition engine 101 preferably receives periodic updates 112 about the level of utilization of the packet buffer 104 for comparing against the minimum and maximum thresholds. According to one embodiment of the invention, the thresholds are
35 recomputed based on the periodic updates 112.

According to both the RED and WRED algorithms, if the number of packets contained in a queue of the packet buffer 104 is less than a minimum threshold, the packet

1 disposition engine 101 admits the inbound packet 106 destined for the queue and adds it to
the queue. If the number of packets contained in the queue is more than the maximum
threshold, the packet disposition engine 101 discards the packet. If the queue contains
5 packets in between the minimum and maximum thresholds, the packet disposition engine 101
preferably discards the inbound packet according to a pre-determined discard probability
associated with the queue.

Packets that are not dropped by the packet disposition engine 101 are admitted into
the node and passed 114 to the packet buffer 104 for storage. Packets that are stored are
10 preferably held in a common buffer where the utilization for each priority is monitored.
When ready to be forwarded to the packet switching controller, the packet buffer 104
dequeues the packets, preferably according to a class based dequeuing, where packets in the
higher priority queues are dequeued before packets in the lower priority queues. This allows
15 higher priority queues to be emptied before lower priority queues, causing higher priority
packets destined for the high priority queues to be admitted more often than lower priority
packets. The dequeued packets are forwarded as outgoing packets 116 to the packet
switching controller 52.

20 The packet switching controller 52 receives the admitted packets and engages in
further classification of the packets. The admitted packets may be classified for determining
their priority, and recommended to be dropped or forwarded to their destination address
based on the classification information and congestion at the switching controller level.

FIG. 4 is a flow diagram of a process for traffic congestion control at an access
25 controller level according to one embodiment of the invention. The process starts, and in step
200, the packet preclassification engine 100 receives an inbound packet. In step 202, the
packet preclassification engine 100 preclassifies the packet and determines a priority
associated with the packet. In step 203, the packet preclassification engine 203 assigns the
30 determined priority to the packet.

In step 204, the packet disposition engine 101 receives the priority information and
compares the utilization level of the associated queue in the packet buffer 104 against
minimum and maximum thresholds set for the queue. If a determination is made, in step 206,
that the queue utilization level is less than the set minimum threshold, the packet is admitted
35 in step 208 for forwarding to the packet switching controller 52.

If, on the other hand, a determination is made, in step 210, that the queue utilization
level is greater than the set maximum threshold, the packet is discarded in step 212.

1 Otherwise, the queue utilization level is between the minimum and maximum thresholds, and
the packet preclassification engine, in step 216, determines if the packet is to be discarded
based on the discard probability set for the priority assigned to the packet. If the answer is
YES, the packet disposition engine 101 discards the packet in step 218. Otherwise, the
5 packet is admitted in step 220.

Although this invention has been described in certain specific embodiments, those
skilled in the art will have no difficulty devising variations which in no way depart from the
scope and spirit of the present invention. It is therefore to be understood that this invention
10 may be practiced otherwise than is specifically described. Thus, the present embodiments of
the invention should be considered in all respects as illustrative and not restrictive, the scope
of the invention to be indicated by the appended claims and their equivalents rather than the
foregoing description.

1 WHAT IS CLAIMED IS:

1. A data communication node forwarding inbound packets, the node comprising:

5 an access controller receiving an inbound packet and classifying the packet, the access controller determining whether the packet is to be admitted or not based on congestion status data determined from the classification information; and

10 a switching controller coupled to the access controller, the switching controller receiving the admitted inbound packet from the access controller for further classifying the packet, the switching controller determining whether the packet is to be forwarded or not based on additional congestion status data determined from additional classification information.

15 2. The data communication node of claim 1, wherein the access controller is a media access controller.

3. The data communication node of claim 1, wherein the classification information includes a priority associated with the inbound packet.

20 4. The data communication node of claim 3, wherein the access controller gives precedence in admitting packets associated with a first priority over packets associated with a second priority.

25 5. The data communication node of claim 1, wherein the access controller includes a buffer storing admitted inbound packets.

6. The data communication node of claim 5, wherein the congestion status data includes a buffer utilization level, the access controller admitting the inbound packet if the utilization level is lower than a predetermined threshold level.

30 7. The data communication node of claim 5, wherein the congestion status data includes a buffer utilization level, the access controller discarding the inbound packet if the utilization level is higher than a predetermined threshold level.

35 8. The data communication node of claim 1, wherein the access controller discards the inbound packet based on a discard probability that varies based on the classification information.

9. The data communication node of claim 1, wherein the access controller

1 includes:

an input receiving the inbound packet;

a classification engine coupled to the input classifying the inbound packet;

a buffer storing inbound packets; and

5 a disposition engine coupled to the classification engine and the buffer, the disposition engine receiving classification information from the classification engine and determining whether the inbound packet is to be admitted or not.

10 10. The data communication node of claim 9, wherein the disposition engine admits the inbound packet if a utilization level of the buffer is lower than a predetermined threshold level.

15 11. The data communication node of claim 9, wherein the disposition engine discards the inbound packet if a utilization level of the buffer is higher than a predetermined threshold level.

20 12. The data communication node of claim 9, wherein the disposition engine discards the inbound packet based on a discard probability that varies based on the classification information.

25 13. A method for packet traffic management in a data communication node including an access controller and a switching controller, the method comprising:

at the access controller:

receiving an inbound packet;

classifying the inbound packet;

determining congestion status data from the classification information;

admitting the inbound packet or not based on the congestion status data; and

forwarding the inbound packet to the switching controller if the packet is admitted; and

30 at the switching controller, determining whether the admitted packet is to be further forwarded.

35 14. The method of claim 13, wherein the access controller is a media access controller.

15. The method of claim 13, wherein the classification information includes a priority associated with the inbound packet.

1 16. The method of claim 15, wherein the admitting the inbound packet further comprises giving precedence to packets associated with a first priority over packets associated with a second priority.

5 17. The method of claim 15 further comprising storing the inbound packet in a packet buffer associated with the access controller if the packet is admitted.

 18. The method of claim 17, wherein the determining of the congestion status data comprises determining a utilization level of the packet buffer.

10

 19. The method of claim 18, wherein the admitting of the inbound packet comprises admitting the inbound packet if the utilization level of the packet buffer is lower than a predetermined threshold level.

15 20. The method of claim 18 further comprising discarding the inbound packet if the utilization level of the packet buffer is higher than a predetermined threshold level.

 21. The method of claim 18 further comprising discarding the inbound packet based on a discard probability that varies based on the classification information.

20

 22. In a switching node comprising a media access controller having at least one input coupled to a switching controller having at least one output, a congestion control method for selectively forwarding packets received on the at least one input to the at least one output, comprising:

25 receiving a packet on an input; and

 forwarding the packet on an output only if a first congestion check is passed at the media access controller and a second congestion check is passed at the switching controller, wherein the first congestion check is based on first packet classification information and a first buffer utilization level.

30

 23. The method of claim 22 wherein the second congestion check is based on second packet classification information a second buffer utilization level.

35

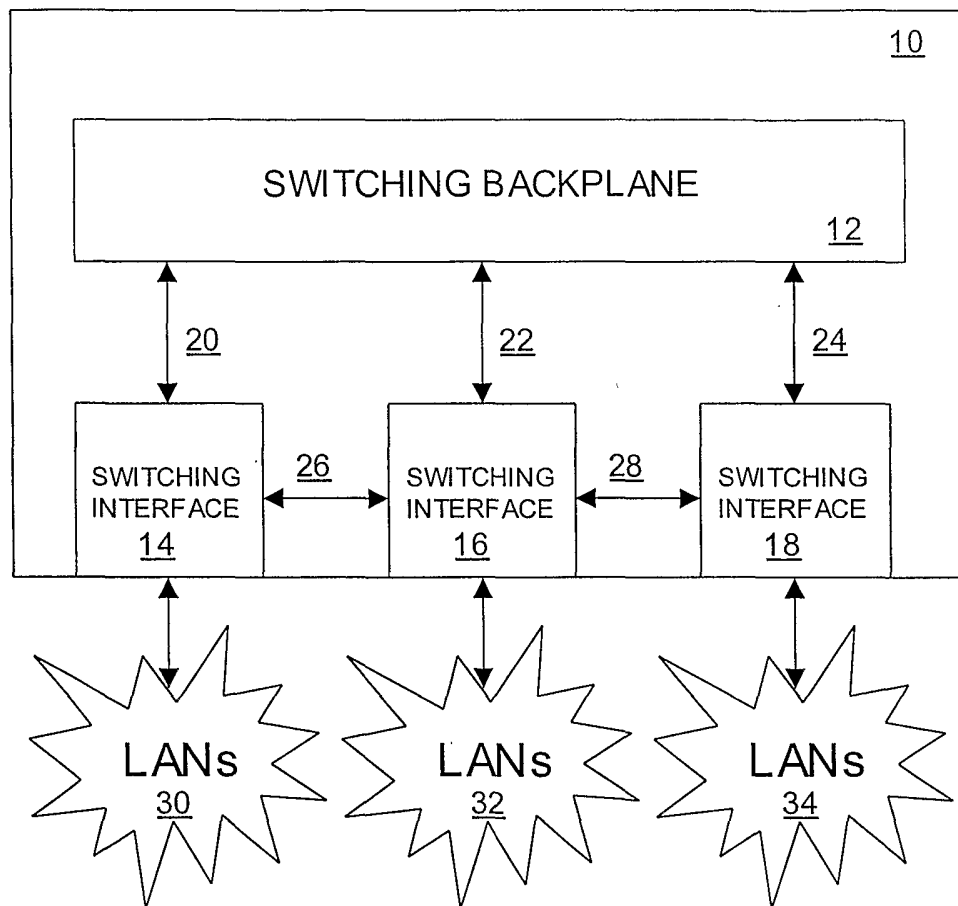


FIG. 1

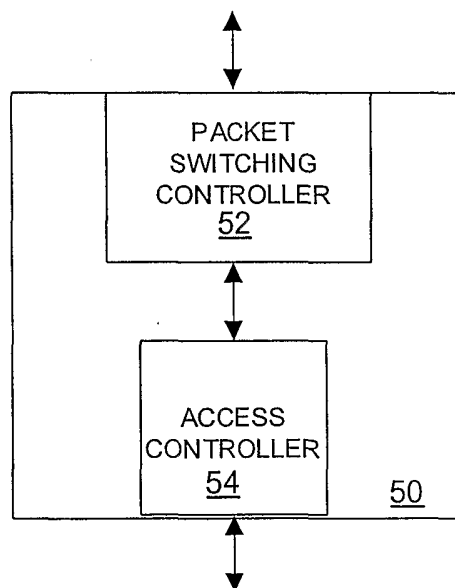


FIG. 2

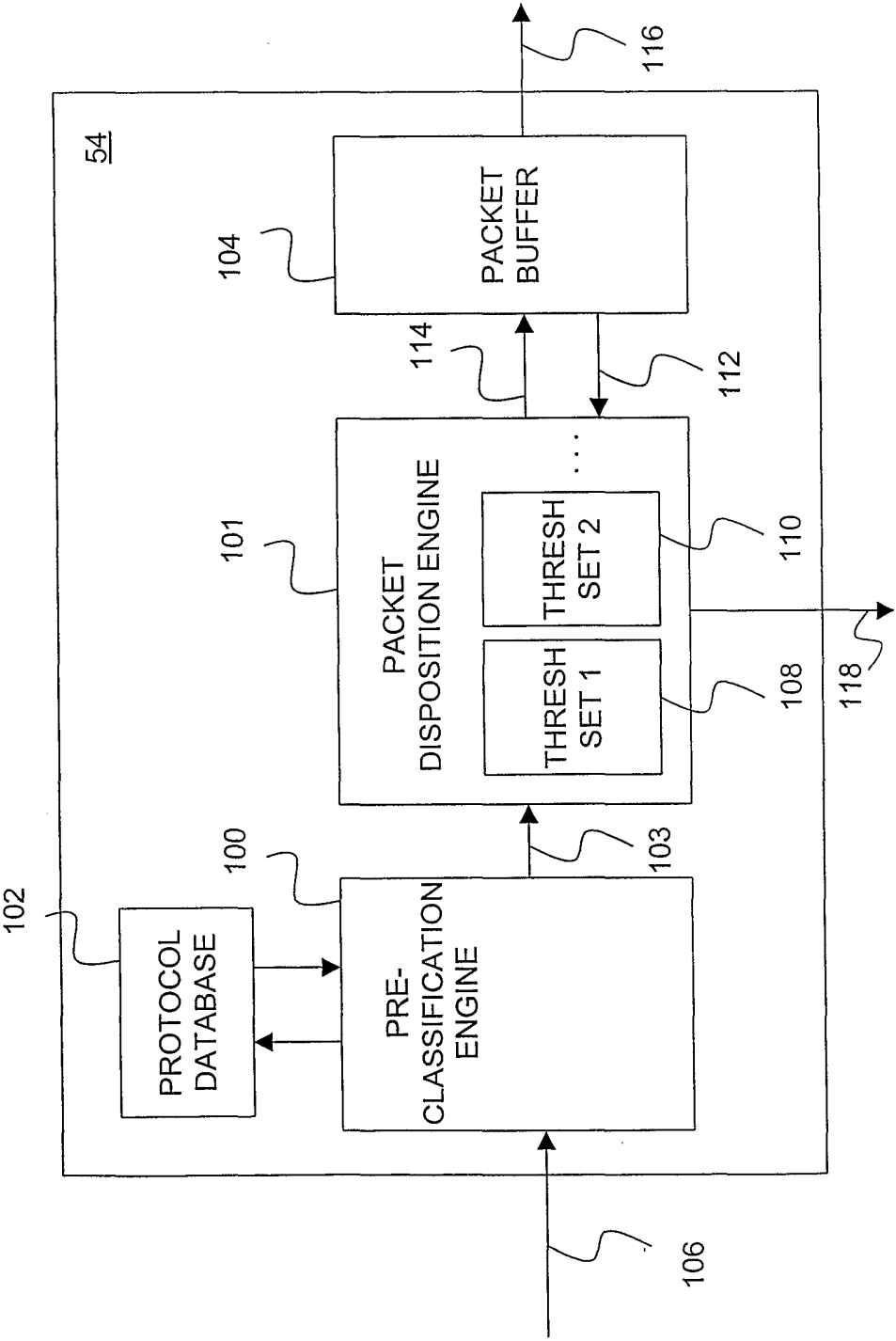
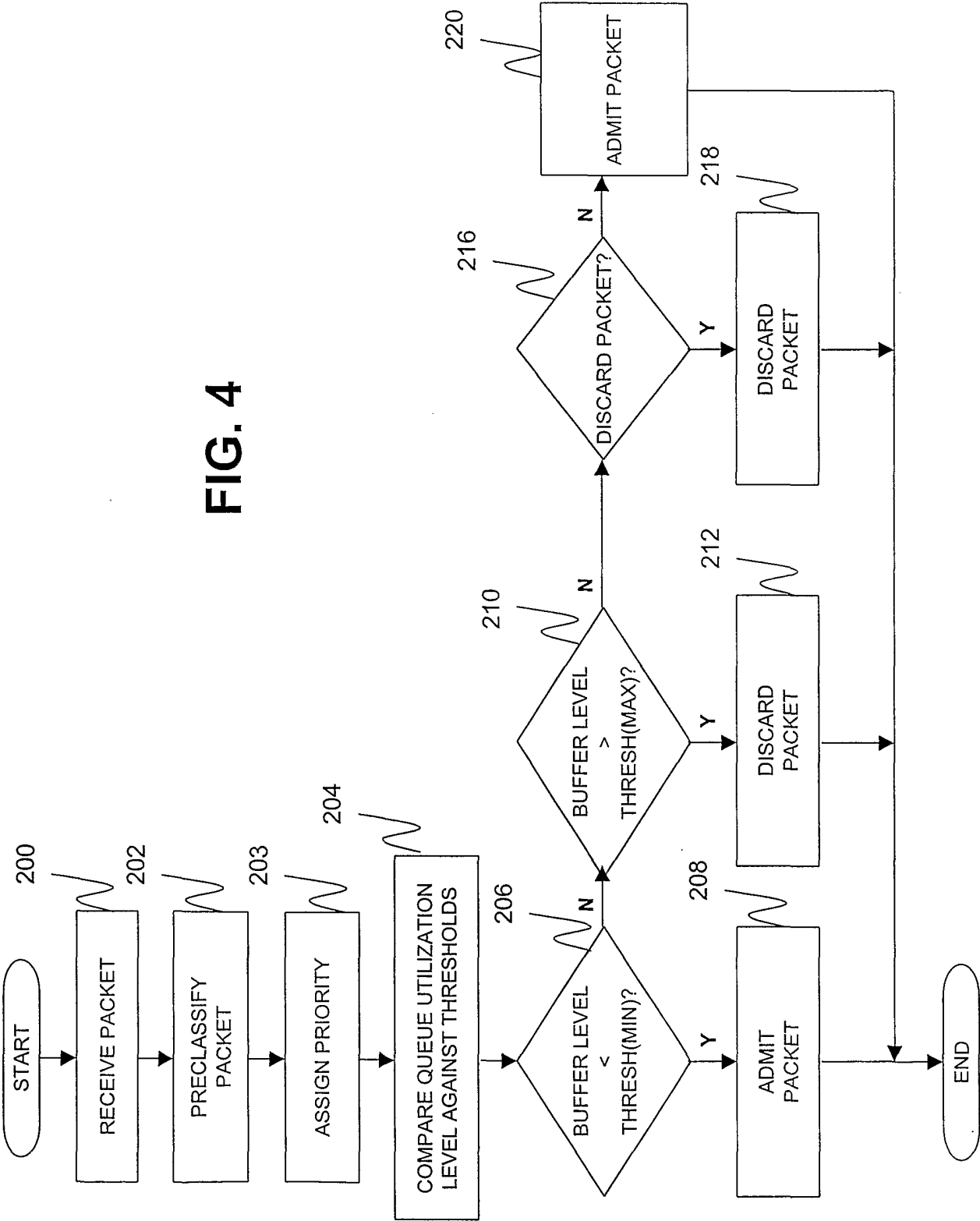


FIG. 3

FIG. 4



INTERNATIONAL SEARCH REPORT

International application No.

PCT/US02/23055

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) : H04J 3/14

US CL : 370/230

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 370/230, 370/232, 370/235, 370/235.1

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 6,047,326 (KILKKI) 4 April 2000 (04.04.2000), column 7, lines 28-42, column and Figures 6-7 and 10.	1-23
Y	US 6,252,849 B1 (ROM et al) 26 June 2001 (26.06.2001), column 4, lines 7-67, column 5, lines 1-36, and Figures 2-3.	1-23

☐

Further documents are listed in the continuation of Box C.

☐

See patent family annex.

* Special categories of cited documents:	
"A" document defining the general state of the art which is not considered to be of particular relevance	"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"E" earlier application or patent published on or after the international filing date	"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"O" document referring to an oral disclosure, use, exhibition or other means	"&" document member of the same patent family
"P" document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search

07 September 2002 (07.09.2002)

Date of mailing of the international search report

16 DEC 2002

Name and mailing address of the ISA/US

Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231

Facsimile No. (703)305-3230

Authorized officer

Scale
Huy Vu

Telephone No. (703) 305-3900