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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
12 September 2002 (12.09.2002)

PCT

(10) International Publication Number
WO 02/071703 A1

(51) International Patent Classification[7]: H04L 12/56

(21) International Application Number: PCT/US02/04601

(22) International Filing Date: 14 February 2002 (14.02.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/800,678 6 March 2001 (06.03.2001) US

(71) Applicant: **PLURIS, INC.** [US/US]; 10455 Bandley Drive, Cupertino, CA 95014 (US).

(72) Inventors: **MANSHARAMANI, Deepak**; 1738 Commodore Drive, San Jose, CA 95133 (US). **BASTURK, Erol**; 10246 Will Court, Cupertino, CA 95014 (US).

(74) Agent: **BOYS, Donald, R.**; P.O. Box 187, Aromas, CA 95004 (US).

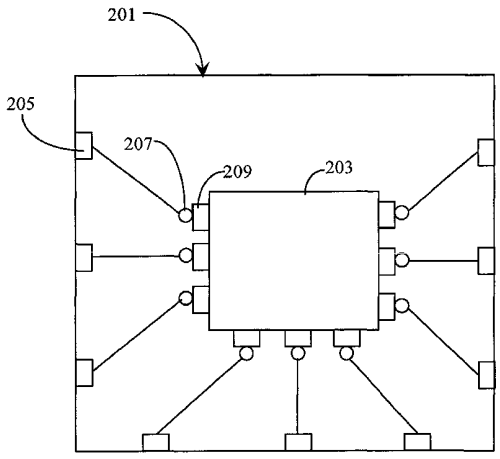
(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

**Published:**
— *with international search report*

*For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.*

(54) Title: AN IMPROVED SYSTEM FOR FABRIC PACKET CONTROL




(57) Abstract: A method for managing data traffic in nodes in a fabric network (Fig. 2), each node having internally-coupled ports, follows the steps of establishing a managed queuing system (209) comprising one or more queues associated with each port, for managing incoming data traffic; and accepting or discarding data directed to a queue according to the quantity of data in the queue relative to queue capacity. In one preferred embodiment the managed system accepts all data directed to a queue less than full, and discards all data directed to a queue that is full. In some alternative embodiments the queue manager (209) monitors quantity of data in a queue relative to queue capacity, and begins to discard data at a predetermined rate when the quantity of queued data reaches the threshold. In other cases the queue manager increases the rate of discarding as the quantity of queued data increases above the preset threshold, discarding all data traffic when the queue is full.

WO 02/071703 A1

# An Improved System for Fabric Packet Control

## Field of the Invention

The present invention is in the field of routing packets through alternative paths between nodes in a routing fabric, and pertains in particular to methods by which back-ups in a fabric may be avoided.

## Background of the Invention

With the advent and continued development of the well-known Internet network, and of similar data-packet networks, much attention has been paid to computing machines for receiving, processing, and forwarding data packets. Such machines, known as routers in the art, typically have multiple interfaces for receiving and sending packets, and circuitry at each interface, including typically a processor, for handling and processing packets. The circuitry at the interfaces is implemented on modules known as line cards in the art. In some routers the line cards are interconnected through what is known as the internal fabric, which comprises interconnected fabric cards handling transmissions through the fabric. Fabric interconnection has not always been a part of routers in the art, and is a fairly recent innovation and addition for packet routers.

Fig. 1, labeled prior art, illustrates a number of interconnected fabric nodes, labeled in this example A through J, each node of which may be fairly considered to comprise a fabric card in a switching fabric in a router. It will be apparent to the skilled artisan that Fig. 1 is an exemplary and partial representation of nodes and interconnections in a switching fabric, and that there are typically many more nodes and interconnections than those shown.

One purpose of Fig. 1 in this context is to illustrate that there are a wide variety of alternative paths that data may take within a switching fabric. For example,

transmission from node E to node J may proceed either via path E-F-H-G-J, or alternatively via E-F-D-G-J. The skilled artisan will also recognize that the nodes and interconnections shown are but a tiny fraction of the nodes and interconnections that might be extant in a practical system.

In conventional switching fabric at the time of the present patent application fabric nodes in such a structure are implemented on fabric cards or chips that do Flow Control. Such Flow Control is very well-known in the art, and comprises a process of monitoring ports for real or potential traffic overflow, and notifying an upstream port to stop or slow sending of further data. That is, if node G as shown in Fig. 1, becomes overloaded at a particular input port, for example, the port from D, the Flow Control at G will notify D to restrict or suspend traffic to G In this example, D may receive traffic from upstream neighbors that it cannot forward to G, and it may then have to notify these neighbors to suspend sending traffic to D. This example illustrates how Flow Control may cause traffic changes made by nodes as a result of an overflow condition at a downstream node to propagate further upstream affecting further nodes, and further stopping or diverting traffic. In Fig. 1 arrows between nodes are indicative of Flow Control indicators passed, and the skilled artisan will also understand that traffic may be in any direction, and that Flow Control indicators are therefore passed in both directions as well.

A serious problem with Flow Control as conventionally practiced is that the upstream notifications, inherent in flow control, propagate further upstream and hinder or stop traffic that there is no need to stop, partly because the interconnections of nodes may be quite complicated and the alternative paths quite numerous. Further, a node that has been informed of a downstream overload condition cannot select to stop or divert traffic just for that particular link, but only to stop or divert all traffic. These effects, because of the complexity and interconnection of nodes in a fabric, can result in complete stultification of parts of a system, or of an entire network.

There have been in the art several attempts to improve upon flow control, but all such solutions have only been partly successful, and still use upstream propagation of control indicators, which always still have a good chance of causing unwanted difficulty.

What is clearly needed is a way to deal with temporary overloads at fabric nodes without resorting to problematic upstream messaging without impacting traffic that does not need to use the overloaded link.

## Summary of the Invention

In a preferred embodiment of the present invention a method for managing data traffic at switching element in a fabric network, each node having two or more internally coupled ports is provided, comprising the steps of (a) establishing a managed queuing system comprising one or more queues associated with each port, for managing incoming data traffic; and (b) accepting or discarding data directed to a queue according to the quantity of data in the queue relative to queue capacity.

In some embodiments all data is discarded for a full queue. In some other embodiments the queue manager monitors quantity of queued data in relation to a preset threshold, and begins to discard data at a predetermined rate when the quantity of queued data reaches the threshold. In still other embodiments the queue manager increases the rate of discarding as quantity of queued data increases above the preset threshold, discarding all data traffic when the queue is full.

In another aspect of the invention a switching element for a fabric network is provided, comprising two or more internally-coupled ports, and a managed queuing system comprising one or more queues associated with each port, for managing incoming data traffic. The switching element is characterized in that the queue manager accepts or discards data directed to a queue according to the quantity of data in the queue relative to queue capacity.

In some embodiments all data is discarded for a full queue. In some other embodiments the queue manager monitors quantity of queued data against a preset threshold, and begins to randomly discard data when the quantity of queued data exceeds the threshold. In still other embodiments the queue manager increases the rate of discarding as the quantity of queued data increases above the preset threshold.

In still another aspect of the invention a data router having external connections to other data routers is provided, comprising an internal fabric network, and a plurality of switching elements in the internal fabric network, each having internally-coupled ports, and a managed queuing system comprising one or more queues associated with each port, for managing incoming data traffic. The router is characterized in that the queue manager accepts or discards data directed to a queue according to the quantity of data in the queue relative to queue capacity.

In some embodiments all data is discarded for a full queue. In some other embodiments the queue manager monitors quantity of queued data against a preset threshold, and begins to randomly discard data when the quantity of queued data exceeds the threshold. In still other embodiments the queue manager increases the rate of discarding as the quantity of queued data increases above the preset threshold.

In various embodiments of the invention taught below in enabling detail, for the first time a system is provided for routers that accomplished the purposes of flow control without requiring upstream notification of problems, which can often result in extensive and unnecessary cessation or diversion of traffic.

## Brief Descriptions of the Drawing Figures

Fig. 1 is a prior art diagram illustrating fabric node interconnections and upstream propagation of flow control indicators.

Fig. 2 is a diagram of a fabric card in an embodiment of the present invention.

Fig. 3 is a diagram of a fabric network of fabric cards in an embodiment of the present invention.

## Description of the Preferred Embodiments

Fig. 2 is a plan view of a fabric card 201 in an embodiment of the present invention. In this embodiment there are nine (9) ports on each card, rather than four

as indicated in the prior art diagram of Fig. 1. This is not meant to imply that the prior art is limited to four ports per node, as Fig. 1 was exemplary only.

In the fabric card of this embodiment, as shown in Fig. 2, there are nine queue managers 209, one for each external port 205, with each queue manager isolated from its connected external port by an optical interface 207. The inter-node communication in this embodiment is by optical links. Queue managers 209 interface with crossbar 203, which connects each of the nine ports with the other eight ports internally in this embodiment, although these internal connections are not shown in the interest of simplicity.

Fig. 3 is a diagram illustrating a fabric having interconnected fabric cards according to the embodiment described above with reference to Fig. 2. In this diagram one card 319 is shown connected to nine neighbor cards 301, 303, 305, 307, 309, 311, 313, 315, and 317. Each of the neighbor cards is illustrated as having eight additional ports for interconnecting to further neighbors in addition to the one port connecting the near neighbor with card 319. It will be clear to the skilled artisan from this diagram that interconnection complexity escalates at a very great rate as ports and cards (nodes) proliferate.

Referring now back to Fig. 2, each port on each card in this example passes through a queue management gate 209 as indicated in Fig. 2. Each queue manager comprises a set of virtual output queues (VOQ), with individual VOQs associated with individual ones of the available outputs on a card. This VOQ queuing system manages incoming flows based on the outputs to which incoming packets are directed. Data traffic coming in on any one port, for example, is directed to a first-in-first-out (FIFO) queue associated with an output port, and the queue manager is enabled to discard all traffic when the queue to which data is directed is full. There are, in this scheme, no Flow Control indications generated and propagated upstream as is done in the prior art.

In this unique arrangement the size of each queue is set to provide adequate flow under ordinary, and to some extent extraordinary, load conditions without data loss, but under extreme conditions, when a queue is full, data is simply discarded until the situation corrects, which the inventors have found to be less conducive to data loss

than the problems associated with conventional Flow Control, which uses the previously described upstream-propagated Flow Control indicators.

In an alternative embodiment of the present invention each queue manager on a card has an ability to begin to drop packets at a pre-determined rate at some threshold in queue capacity short of a full queue. In certain further embodiments the queue manager may accelerate the rate of packet dropping as a queue continues to fill above the first threshold. In these embodiments the incidence of dropping packets is minimized and managed, and spread over more traffic than would be the case if dropping of packets were to begin only at a full queue, wherein all packets would be dropped until the queue were to begin to empty.

A distinct advantage of the queue management scheme of the present invention is that the intelligence required is considerably lessened, and there is no addition to the traffic load by generating Flow Control indicators.

It will be apparent to the person with ordinary skill in the art that the embodiments of the invention described in this specification are exemplary, and may vary in a number of ways without departing form the spirit and scope of the present invention. For example, there may be more or fewer than nine ports and queue managers per fabric card, the system may be implemented on a chip or a set of chips, and the size of each queue may vary. There are many other alterations within the spirit and scope of the invention as well, and the scope of the invention is limited only by the claims which follow.

What is claimed is:

1. A method for managing data traffic at switching element in a fabric network, each node having two or more internally coupled ports, comprising the steps of:

(a) establishing a managed queuing system comprising one or more queues associated with each port, for managing incoming data traffic; and

(b) accepting or discarding data directed to a queue according to the quantity of data in the queue relative to queue capacity.

2. The method of claim 1 wherein, in step (b), all data is discarded for a full queue.

3. The method of claim 1 wherein the queue manager monitors quantity of queued data in relation to a preset threshold, and begins to discard data at a predetermined rate when the quantity of queued data reaches the threshold.

4. The method of claim 3 wherein the queue manager increases the rate of discarding as quantity of queued data increases above the preset threshold, discarding all data traffic when the queue is full.

5. A switching element for a fabric network, comprising:

two or more internally-coupled ports; and

a managed queuing system comprising one or more queues associated with each port, for managing incoming data traffic;

characterized in that the queue manager accepts or discards data directed to a queue according to the quantity of data in the queue relative to queue capacity.

6. The switching element of claim 5 wherein all data is discarded for a full queue.

7. The switching element of claim 5 wherein the queue manager monitors quantity of queued data against a preset threshold, and begins to randomly discard data when the quantity of queued data exceeds the threshold.

8. The switching element of claim 7 wherein the queue manager increases the rate of discarding as the quantity of queued data increases above the preset threshold.

9. A data router having external connections to other data routers, comprising:

an internal fabric network; and

a plurality of switching elements in the internal fabric network, each having internally-coupled ports, and a managed queuing system comprising one or more queues associated with each port, for managing incoming data traffic;

characterized in that the queue manager accepts or discards data directed to a queue according to the quantity of data in the queue relative to queue capacity.

10. The data router of claim 9 wherein all data is discarded for a full queue.

11. The data router of claim 9 wherein the queue manager monitors quantity of queued data against a preset threshold, and begins to randomly discard data when the quantity of queued data exceeds the threshold.

12. The data router of claim 11 wherein the queue manager increases the rate of discarding as the quantity of queued data increases above the preset threshold.

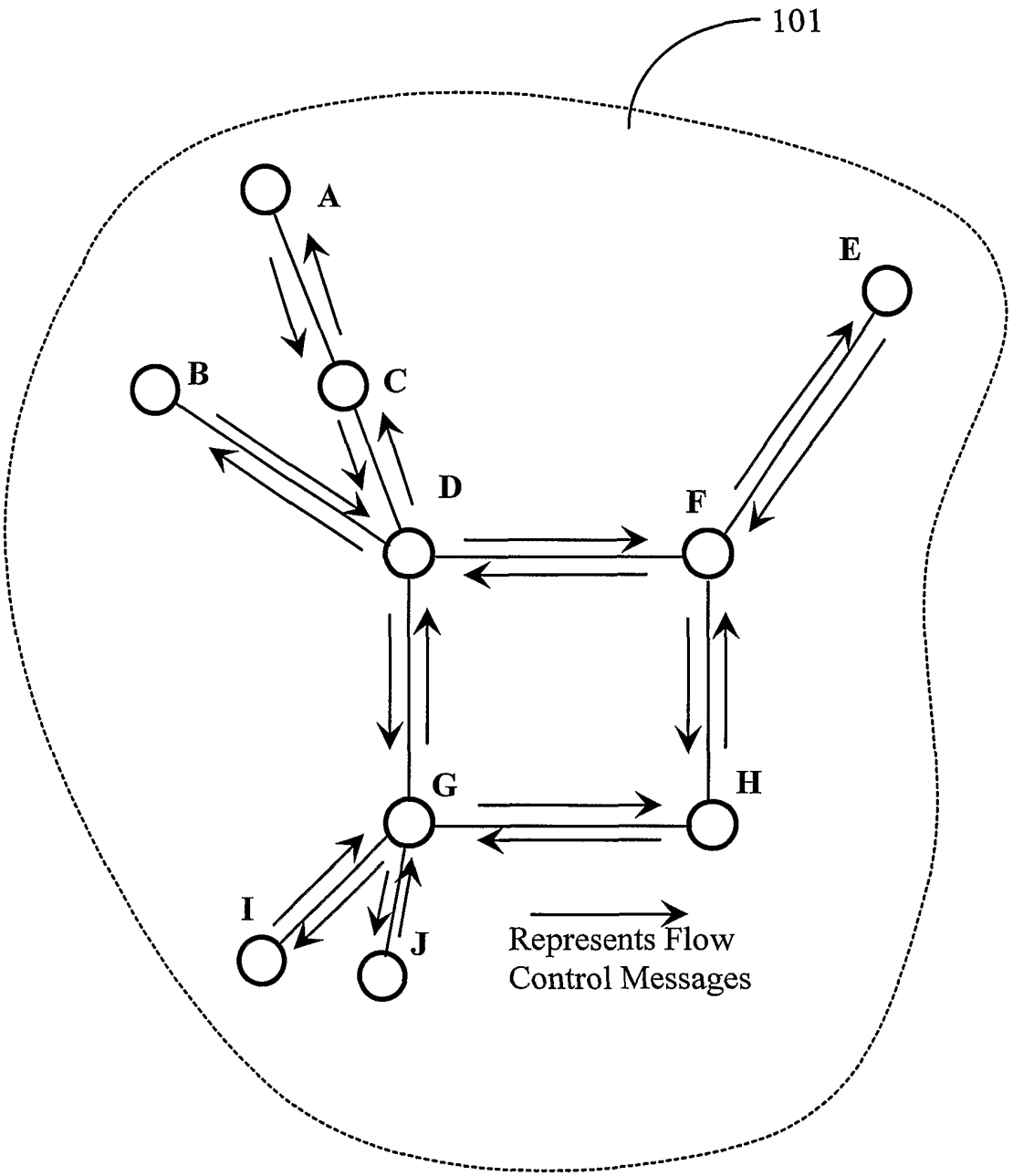



*Fig. 1 (Prior Art)*

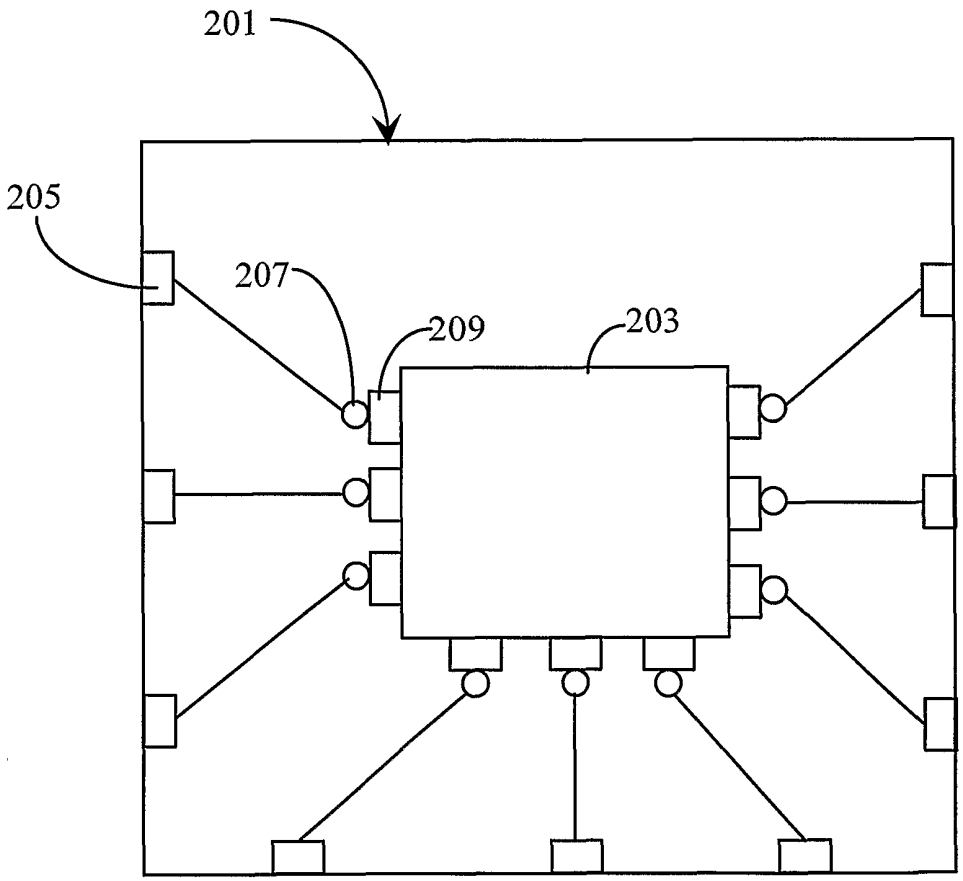



*Fig. 2*

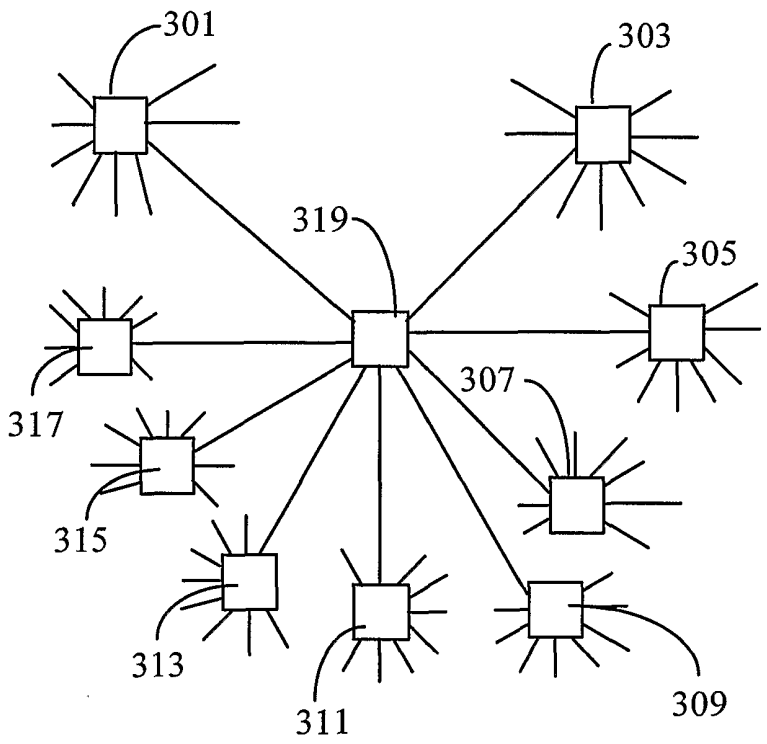



*Fig. 3*

# INTERNATIONAL SEARCH REPORT

International application No.
PCT/US02/04601

**A. CLASSIFICATION OF SUBJECT MATTER**

IPC(7) : H04L 12/56
US CL : 370/229,230.1,232-234,252,253,252,395.21,468,477

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. : 370/229,230.1,232-234,252,253,252,395.21,468,477

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)
Please See Continuation Sheet

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

| Category * | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| X | US 5,793,747 A (KLINE) 11 August 1998, figure 5, column 7, lines 31-59. | 1-12 |
| X | US 5,777,984 A (GUN) 07 July 1998, column 11, lines 6-61. | 1-12 |
| X, P | US 6,246,687 B1 (SIU) 12 June 2001, figure 2, column 3, lines 64-67 and column 4 , lines 1-4; column 6, line 9 to column 7, line 13; column 8, lines 41-63. | 1-12 |

☐ 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

"E" earlier application or patent published on or after the international filing date

"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)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"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

"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

"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

"&" document member of the same patent family

| Date of the actual completion of the international search | Date of mailing of the international search report |
|---|---|
| 03 May 2002 (03.05.2002) | 31 MAY 2002 |
| Name and mailing address of the ISA/US<br>Commissioner of Patents and Trademarks<br>Box PCT<br>Washington, D.C. 20231<br>Facsimile No. (703)305-3230 | Authorized officer<br>Phuongchau Ba Nguyen<br>Telephone No. 703-305-4700 |

Form PCT/ISA/210 (second sheet) (July 1998)

## INTERNATIONAL SEARCH REPORT

International application No.

PCT/US02/04601

**Continuation of B. FIELDS SEARCHED Item 3:**

((atm adj switch$3) and port?)
queue? and (((control$3 or manag$5) and (discard$3 and accept$3)) same (cell? or data)
((preset adj threshold) and (predetermin$3 adj rate))
((threshold) and (predetermin$3 adj rate)) and exceed$3 same threshold
overflow and discard$3