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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
10 July 2003 (10.07.2003)

PCT

(10) International Publication Number
WO 03/056766 A1

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

(21) International Application Number: PCT/EP01/15371

(22) International Filing Date:
28 December 2001 (28.12.2001)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant *(for all designated States except US)*: **NOKIA CORPORATION** [FI/FI]; Keilalahdentie 4, FIN-02150 Espoo (FI).

(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: **LAKKAKORPI, Jani** [FI/FI]; Hakolahdentie 3 A 3, FIN-00200 Helsinki (FI).

(74) Agent: **UNGERER, Olaf**; Eisenführ, Speiser & Partner, Association No. 15, Arnulfstr. 25, 80335 Munich (DE).

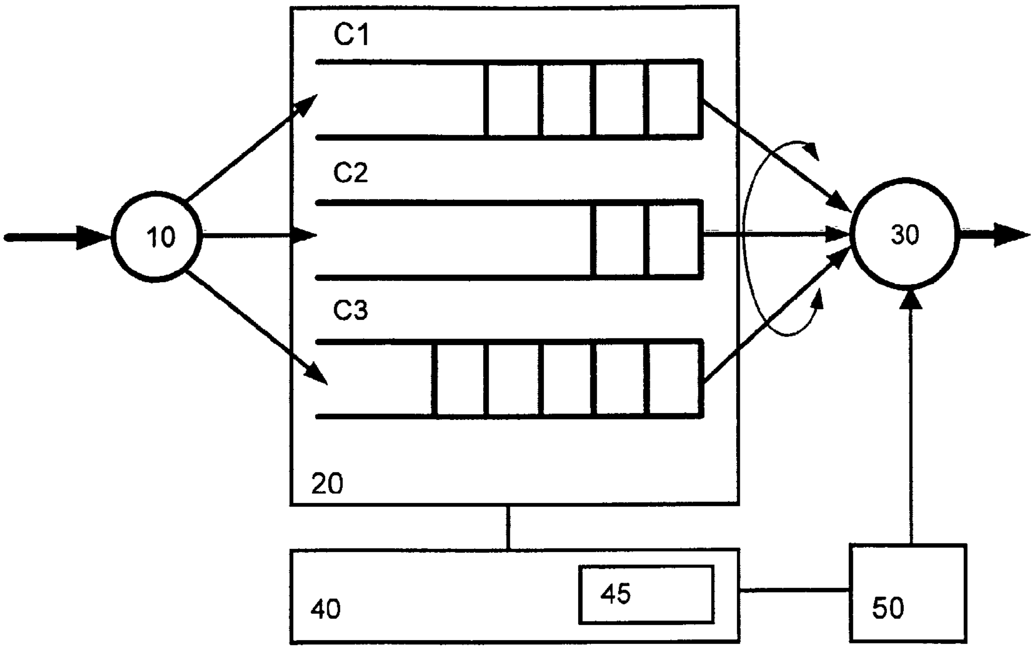
(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PH, PL, PT, RO,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



(57) Abstract: The present invention relates to a method for scheduling data packets in a network element of a packet data network, such as an IP network, wherein queue weights and sizes are adjusted at the same time so that the maximum queuing delay is as predictable as possible. Respective sizes of at least two data packet queues are adjusted at a predetermined or triggered timing based on at least one predetermined parameter indicating a change in the traffic mix routed through the network element or within a set of network elements. Thereby, more predictable maximum delays can be achieved.

Packet Scheduling Method and Apparatus

FIELD OF THE INVENTION

The present invention relates to a method and apparatus for scheduling data packets in a network element of a packet data network, e.g. a router in an IP (Internet Protocol) network.

BACKGROUND OF THE INVENTION

Traditional packet data networks, e.g. IP networks, can provide all customers with Best Effort (BE) services only. The whole traffic competes equally for network resources. With development of new applications of Internet, such as voice, video and Web services, the desire of manageable and/or predictable QoS (Quality of Service) becomes stronger.

Congestion management features allow to control congestion by determining the order in which packets are sent out at an interface, and the order in which packets are dropped – if needed, based on priorities assigned to those packets. Congestion management entails the creation of queues, assignment of packets to those queues based on a classification of the packet, and scheduling of the packets in a queue for transmission. There are numerous types of queuing mechanisms, each of which allows to specify creation of a different number of queues, affording greater or lesser degrees of differentiation of traffic, and to specify the order in which the traffic is sent.

During periods with light traffic, that is, when no congestion exists, packets are sent out the interface as soon as they arrive. During periods of transmit congestion at the outgoing interface, packets arrive faster than the interface can send them. If congestion management features are used, packets accumulating at an interface are either queued until the interface is free to send them, or dropped if the congestion is heavy, and the packet is marked as low priority packet. Packets are then scheduled for transmission according to their assigned priority and the queuing mechanism configured for the interface. A respective router of the packet data network determines the order of packet transmission by controlling which packets are placed in which queue and how queues are serviced with respect to each other.

Queuing types for congestion management QoS control are e.g. FIFO (First-In-First-Out), Weighted Fair Queuing (WFQ) and Priority Queuing (PQ). With FIFO, transmission of packets out the interface occurs in the order the packets arrive. WFQ offers dynamic, fair queuing that divides bandwidth across traffic queues based on weights. And, with PQ, packets belonging to one priority class of traffic are sent before all lower priority traffic to ensure timely delivery of those packets.

Heterogeneous networks include many different protocols used by applications, giving rise to the need to prioritize traffic in order to satisfy time-critical applications while still addressing the needs of less time-dependent applications, such as file transfer. Different types of traffic sharing a data path through the network can interact with one another in ways that affect their application performance. If a network is designed to support different traffic types that share a single data path between routers, congestion management techniques should be applied to ensure fairness of treatment across various traffic types.

For situations in which it is desirable to provide consistent response time to heavy and light network users alike without adding excessive bandwidths, the solution is WFQ. WFQ is a flow-based queuing algorithm which does two things simultaneously. It schedules interactive traffic to the front of the queue to reduce response time, and it fairly shares the remaining bandwidth between high bandwidth flows, wherein the bandwidth indicates the number of bits per second which can be output from the router interface.

WFQ ensures that queues do not starve for bandwidth, and that traffic gets predictable service. Low-volume traffic streams which make up the majority of traffic receive preferential service, so that their entire offered loads are transmitted in a timely fashion. High-volume traffic streams share the remaining capacity or bandwidth proportionally between them. WFQ is designed to minimize configuration effort and adapts automatically to changing network traffic conditions in that it uses whatever bandwidth is available to forward traffic from lower priority flows if no traffic from higher priority flows is present. This is different from Time Division Multiplexing (TDM) which simply carves up the bandwidth and lets it go unused if no traffic is present for a particular traffic type.

Further details of WFQ can be gathered from Hui Zhang, "Service Disciplines for Guaranteed Performance Service in Packet-Switching Networks", in Proceedings of the IEEE, Volume 83, No. 10, October 1995, and from "Weighted Fair Queuing (WFQ)", Cisco Systems, Inc., http://www.cisco.com/warp/public/732/Tech/wfq/.

Assured Forwarding (AF) is an IETF standard in the field of Differentiated Services. Routers implementing AF have to allocate certain resources (buffer space and bandwidth) to different traffic aggregates. Each of the four AF classes has three drop precedences: in the event of congestion, packets with low drop precedence (within a class) are dropped first. Assured Forwarding can basically be implemented with any weight-based scheduling mechanism e.g., with Cisco's Class-Based Weighted Fair Queueing (CB-WFQ). The mutual relationships of different AF classes are open, but one reasonable approach is to use them as delay classes. This approach, however, demands automatic weight adjustments. If weight for a particular AF class stays the same while the amount of traffic in this class increases, delay in this AF class will also increase (assuming that the output link is congested).

Especially for real time traffic (such as streaming video), it is essential to keep the delays in different output queues as predictable as possible. Bearing this in mind, it is not sufficient to adaptively change only queue weights. If the queue size remains constant while the weight is changed, also the maximum queuing delay changes. Thus, an IP router with multiple output queues per interface needs a maximum size and weight for each queue. Setting of these queue sizes and weights can be quite difficult if the traffic mix is unknown and not stable.

Further details of Differentiated Services, Assured Forwarding and different queueing mechanisms can be gathered e.g. from Kalevi Kilkki, "Differentiated Services for the Internet", Macmillan Technical Publishing, ISBN 1-57870-132-5, 1999.

SUMMARY OF THE INVENTION

It is therefore an object of the present invention to provide a packet scheduling method and apparatus, by means of which predictability of queuing delays can be improved.

This object is achieved by a method of scheduling data packets in a network element of a packet data network, said method comprising the steps of:

assigning respective weights to at least two data packet queues, said weights determining a transmit order of queued data packets of said at least two data packet queues; and
adjusting the respective sizes of said at least two data packet queues at a predetermined or triggered timing based on at least one predetermined traffic parameter indicating a change in the traffic mix routed through said network element or within a set of network elements.

Additionally, the above object is achieved by a network element for scheduling data packets in a packet data network, said network element comprising:
weight control means for assigning respective weights to at least two data packet queues, said weights determining a transmit order for queued data packets of said at least two data packet queues; and
size adjusting means for adjusting the respective sizes of said at least two data packet queues at a predetermined or triggered timing based on at least one predetermined traffic parameter indicating a change in the traffic mix routed through said network element or within a set of network elements.

Accordingly, in addition to weights, queue sizes are also set adaptively at the same time. Thereby, the maximum queuing delay in every queue can be kept as predictable as possible by binding the weight and size for each output queue together. Thus, an adaptation to changes in the traffic mix is achieved.

The at least one predetermined parameter may comprise at least one of
- weight of the respective one of said at least two data packet queues
- output link bandwidth of said network element, and
- desired per-hop maximum delay.

Furthermore, the respective sizes may be adjusted every predetermined number of seconds or the adjustment procedure may be triggered by some event (e.g. dramatic change in traffic mix).

Preferably, predetermined minimum weights can be used for said at least two data packet queues. The respective weights may be converted into byte limits which can be taken from each of said at least two data packet queues in its turn.

The size adjusting means may be arranged to adjust the respective size of said at least two data packet queues based on at least one of the weight of the respective one of said at least two data packet queues, the output link bandwidth of said network element, and the desired per-hop maximum delay.

Additionally, timer means may be provided for setting said predetermined intervals. Some events may trigger the adjustment procedure as well.

BRIEF DESCRIPTION OF THE DRAWINGS

In the following, the present invention will be described in greater detail based on a preferred embodiment with reference to the accompanying drawing figures, in which:

Fig. 1 shows a schematic block diagram of a packet scheduling architecture according to the preferred embodiment; and

Fig. 2 shows a schematic flow diagram of a scheduling method according to the preferred embodiment.

DESCRIPTION OF THE PREFERRED EMBODIMENT

The preferred embodiment will now be described based on a packet scheduling architecture for output queues in an IP router.

According to Fig. 1, the scheduling architecture according to the preferred embodiment is based on a scheme which provides bandwidth allocation to all network traffic. To achieve this, a classifier 10 is provided to classify traffic into different classes, i.e. to select packets based the content of packet headers, e.g. DiffServ Code Point (DSCP). However, any other type of classification based on predetermined characteristics of the received traffic can be applied.

The classifier 10 places packets of various conversations in queues C1 to C3 for transmission. The order of removal from the queues C1 to C3 is determined by weights allocated to them. The queues C1 to C3 are arranged in a configurable queuing buffer resource architecture 20.

A scheduler 30 is provided to assign a weight to each flow, i.e. to each of the queues C1 to C3, which weight determines the transmit order for queued packets. The assigned weight may be determined by the required QoS, the desired flow throughput, and the like. Based on the assigned weights, the scheduler 30 supplies queued packets from the queues C1 to C3 to a transmit queue, from which they are output to an output link towards the IP network.

A weight setting unit 50 is arranged to control the scheduler 30 to adjust the respective weights *weight_i* of the queues C1 to C3 at predetermined intervals, i.e. every T seconds, based on the following procedure:

$\text{traffic}_i := a \cdot \text{traffic}_i + (1\text{-}a) \cdot \text{traffic}_{i,\,\text{last period}},$ where $0 < a < 1$
$\text{traffic}_{i,\,\text{last period}} := 0$

$$weight_i = F(traffic_i)/(\sum_{i=1}^{N} F(traffic_i)),$$

wherein traffic_i denotes the moving average of traffic characteristics (e.g. byte count, flow count etc.) at queue C_i within the measurement period T. It is noted that in the example shown in Fig. 1, N equals three, since three queues are provided in the queuing buffer resource architecture 20. The parameter a, i.e. the weight for the previous moving average value and the new moving average value, can be chosen freely. Furthermore, $\text{traffic}_{i,\,\text{last period}}$ denotes the traffic characteristics (e.g. number of bytes arrived) within the last measurement period, and F denotes any suitable predetermined functional relationship between the traffic characteristic and a desired weight. After the moving averages have been updated, the respective counters provided e.g. in the weight setting unit 50 are set to zero in order to start a new counting operation. The weight setting unit 50 may be arranged to use predetermined minimum weights for each queue.

The measurement period T may be set and controlled by a timer 45 which may be provided in a size setting unit 40, as indicated in Fig. 1, or alternatively in the weight setting unit 50 or in any other unit or as a separate unit for the IP router.

The sizes of queues C1 to C3, i.e. the maximum number of data packets in the queues, are set by the size setting unit 40 according to determined parameters indicating the traffic mix. In the preferred embodiment, these parameters are the assigned weights, the output link bandwidth and the desired per-hop maximum delays. However, other suitable parameters may be used for this purpose. The size setting may be performed based on the following equation:

$$\text{size}_i\text{: F(weight}_i\text{, OLB, delay}_i\text{),}$$

wherein weight_i denotes the weight assigned to the queue C_i, OLB denotes the output link bandwidth of the output link of the IP router, and delay_i denotes the desired per-hop maximum delay of the queue C_i. It is noted that the function F may be any suitable function defining a relationship between the allowed queue size and the traffic-specific parameters to thereby keep the delays in the different queues C1 to C3 as predictable as possible.

The scheduler 30 is arranged to convert the weights into bytes which can be dequeued (taken) from one of the queues C1 to C3 in its turn.

Fig. 2 shows a schematic flow diagram of the scheduling operation according to the preferred embodiment.

When the timer 45 has expired in step S200, the procedure proceeds to step S201 where the queue weights are adjusted by the weight setting unit 50 according to any changes in the traffic parameters, i.e. any changes in the traffic mix. Then, the queue sizes are adjusted in step S202 by the size setting unit 40, e.g. using the weight information determined in the weight setting unit 50. Thereafter, the timer 45 is rescheduled or reset to zero (step S203) in order to start a new measurement period or cycle for determining the moving average of bytes and/or other parameters required for the scheduling operation. Finally, the procedure loops back to the initial step S200 and applies the determined sizes and weights until the timer 45 expires again. It is noted that some other events (than expired timer) may be used as well to trigger the adjusment process.

In the following, a specific implementation example of the preferred embodiment is described. In this example, the sizes (in bytes) of the queues C1 to C3 are set according to the queue weights, output link bandwidth and desired per-hop maximum queuing delay, using the following equation:

$$\text{size}_i := (\text{weight}_i \cdot \text{OLB} \cdot \text{delay}_i)/8.$$

It is noted that the blocks indicated in the architecture of Fig. 1 may be implemented as software routines controlling a corresponding processor in the IP router, or as discrete hardware units.

The proposed scheduling operation and architecture removes the need to manually update router queue sizes and provides an adaptive change of queue sizes and queue weights for output queues of routers or any other suitable network elements having a queuing function. Thereby, the scheduling can be adapted to changes in the traffic mix to achieve more predictable maximum delays.

It is noted, that the present invention is not restricted to the specific features of the above predetermined embodiment, but may vary within the scope of the attached claims. In particular, the determination of the queue size and the packet size is not restricted to the above implementation example. Any suitable weight-based scheduling scheme and way of determining suitable queue sizes based on a change in the traffic mix is intended to be covered by the present invention. Moreover, additional coefficients might be used for the different weighted queues C1 to C3 if it is intended that some queues are “faster” than others. If one or a number of priority queues have to be served before the weighted queues C1 to C3 can be served, rate limiters could be used for the priority queues so as to guarantee a minimum output link bandwidth for the weighted queues C1 to C3.

Claims

1. A method of scheduling data packets in a network element of a packet data network, said method comprising the steps of:
   a) assigning respective weights to at least two data packet queues (C1 to C3), said weights determining a transmit order for queued data packets of said at least two data packet queues; and
   b) adjusting the respective sizes of said at least two data packet queues (C1 to C3) at a predetermined or triggered timing based on at least one predetermined parameter indicating a change in the traffic mix routed through said network element or within a set of network elements.

2. A method according to claim 1, wherein said at least one predetermined parameter comprises at least one of
   - weight of the respective one of said at least two data packet queues (C1 to C3)
   - output link bandwidth of said network element, and
   - desired per-hop maximum delay.

3. A method according to claim 1 or 2, wherein said respective sizes are adjusted based on the following equation:

$$size_i = (weight_i \cdot OLB \cdot delay_i)/8,$$

wherein $size_i$ denotes the size of the i-th data packet queue in bytes, $weight_i$ denotes the weight of the i-th data packet queue, OLB denotes the output link bandwidth left for weighted queues of said network element, and $delay_i$ denotes the maximum per-hop delay of said i-th data packet queue.

4. A method according to any one of the preceding claims, wherein said respective sizes are adjusted every predetermined number of seconds or the adjustment procedure is triggered by some event.

5. A method according to any one of the preceding claims, wherein said respective weights are determined based on the following equation:

$$weight_i = F(traffic_i)/(\sum_{i=1}^{N} F(traffic_i)),$$

wherein $weight_i$ denotes the weight of the i-th data packet queue, $traffic_i$ denotes a moving average of traffic characteristics at said i-th data packet queue, F denotes a predetermined functional relationship, and N denotes the number of queues.

6. A method according to claim 5, wherein said moving average is obtained by applying respective weights (a, (1-a)) for previous information and new information.

7. A method according to any one of the preceding claims, wherein predetermined minimum weights are used for said at least two data packet queues (C1 to C3).

8. A network element for scheduling data packets in a packet data network, said network element comprising:
a) weight control means (30, 50) for assigning respective weights to at least two data packet queues (C1 to C3), said weights determining a transmit order for queued data packets of said at least two data packet queues (C1 to C3); and
b) size adjusting means (40) for adjusting the respective sizes of said at least two data packet queues (C1 to C3) at predetermined intervals based on at least one predetermined parameter indicating a change in the traffic mix routed through said network element or within a set of network elements.

9. A network element according to claim 8, wherein said size adjusting means (40) is arranged to adjust the respective sizes of said at least two data packet queues (C1 to C3) according to the following equation:

$$size_i: = (weight_i \cdot OLB \cdot delay_i)/8,$$

wherein $size_i$ denotes the size of the i-th data packet queue in bytes, $weight_i$ denotes the weight of the i-th data packet queue, OLB denotes the output link bandwidth of said network element, and delay_i denotes the maximum per-hop delay of said i-th data packet queue.

10. A network element according to claim 8 or 9, further comprising timer means (45) for setting said predetermined intervals.

11. A network element according to any one of claims 8 to 10 wherein said network element is an IP router.

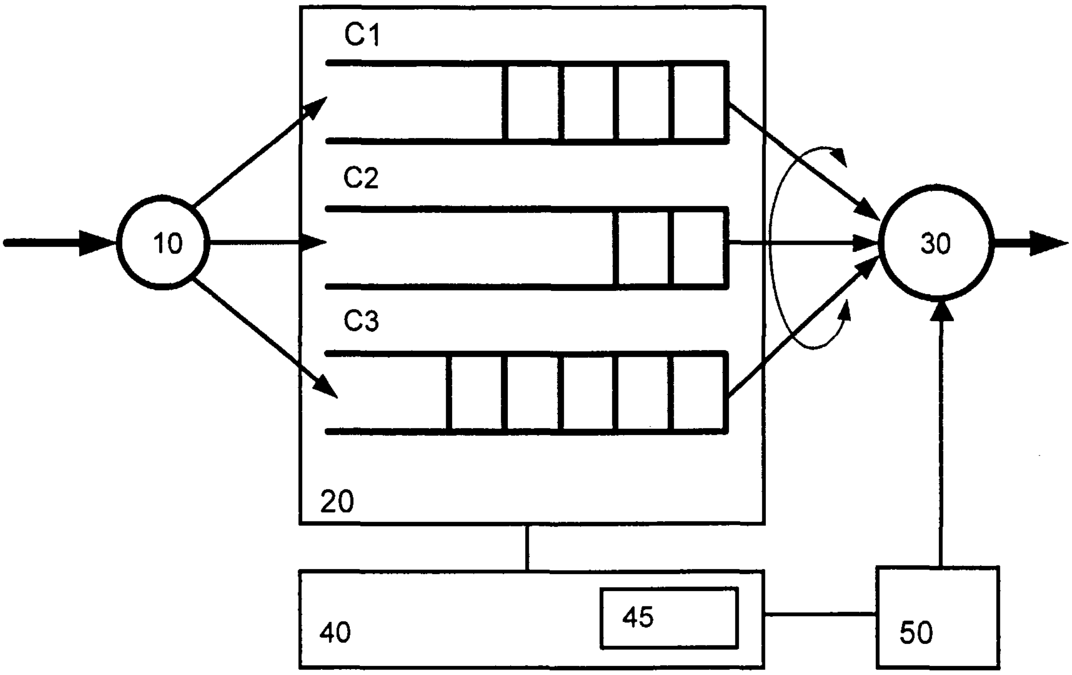


Fig. 1

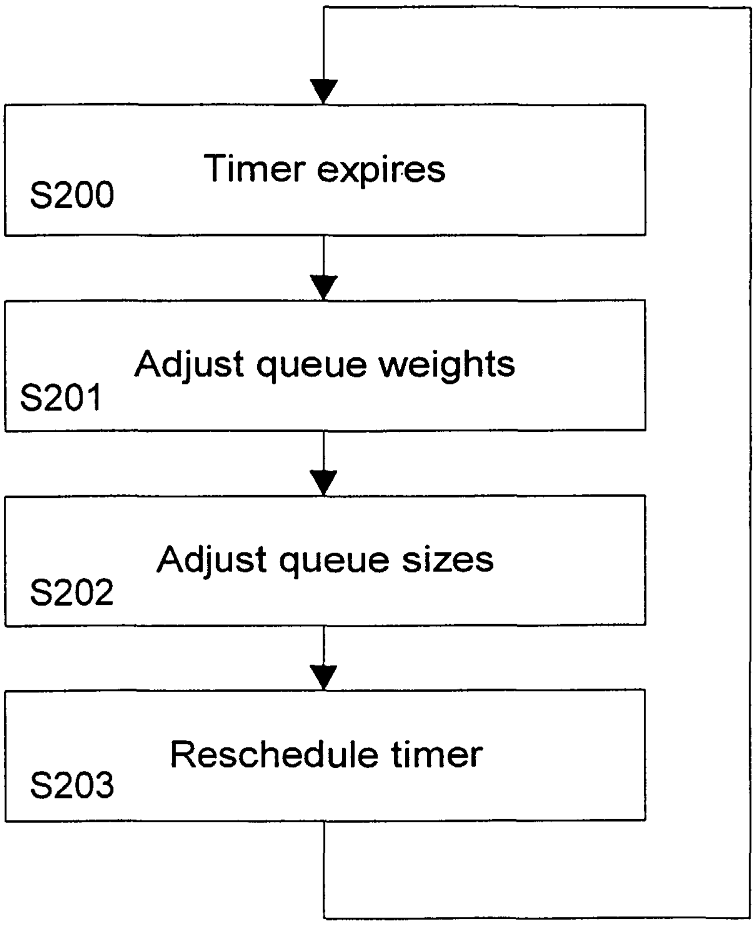

Timer expires
S200
Adjust queue weights
S201
Adjust queue sizes
S202
Reschedule timer
S203

Fig. 2

INTERNATIONAL SEARCH REPORT

PCT/EP 01/15371

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 H04L12/56

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)
IPC 7 H04L

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 practical, search terms used)

EPO-Internal, WPI Data, PAJ, INSPEC

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 01 69851 A (LIAO RAYMOND R F ;CAMPBELL ANDREW T (GB); UNIV COLUMBIA (US)) 20 September 2001 (2001-09-20) page 11, line 25 -page 12, line 14 page 12, line 25 - line 29 page 16, line 8 - line 10 page 18, line 12 - line 19 page 20, line 5 page 20, line 16 - line 18 ---	1-11
X	US 6 094 435 A (HOFFMAN DON ET AL) 25 July 2000 (2000-07-25) column 5, line 11 - line 16 column 5, line 51 - line 62 column 6, line 13 - line 15 column 18, line 58 -column 20, line 58 --- -/--	1,4,5,7,8,10,11

☒ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in 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 document but 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 23 August 2002	Date of mailing of the international search report 30/08/2002
Name and mailing address of the ISA European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Authorized officer Tous Fajardo, J

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
A	US 6 317 416 B1 (AISSAOUI MUSTAPHA ET AL) 13 November 2001 (2001-11-13) page 2, line 37 - line 39 column 2, line 49 - line 52 ---	1-11
A	US 5 757 771 A (LI KWOK-LEUNG ET AL) 26 May 1998 (1998-05-26) column 3, line 3 - line 14 column 4, line 16 - line 29 column 5, line 15 - line 45 column 5, line 66 -column 6, line 12 column 6, line 42 - line 56 ---	1-11
A	WO 00 74432 A (NETWORK EQUIPMENT TECH) 7 December 2000 (2000-12-07) page 6, line 7 - line 16 -----	1-11

INTERNATIONAL SEARCH REPORT

Information on patent family members

PCT/EP 01/15371

Patent document cited in search report		Publication date	Patent family member(s)			Publication date
WO 0169851	A	20-09-2001	AU	4568201	A	24-09-2001
			WO	0169851	A2	20-09-2001
US 6094435	A	25-07-2000	EP	1005744	A1	07-06-2000
			JP	2002507366	T	05-03-2002
			WO	9900949	A1	07-01-1999
US 6317416	B1	13-11-2001	AU	7123596	A	30-04-1997
			DE	69618010	D1	24-01-2002
			DE	69618010	T2	22-08-2002
			EP	0872088	A1	21-10-1998
			CA	2234621	A1	17-04-1997
			WO	9714240	A1	17-04-1997
			US	2002044529	A1	18-04-2002
US 5757771	A	26-05-1998	NONE			
WO 0074432	A	07-12-2000	AU	5293400	A	18-12-2000
			AU	5589800	A	18-12-2000
			EP	1183900	A1	06-03-2002
			EP	1183833	A1	06-03-2002
			WO	0074432	A1	07-12-2000
			WO	0074321	A1	07-12-2000