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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
17 July 2003 (17.07.2003)

PCT

(10) International Publication Number
WO 03/058896 A1

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

(21) International Application Number: PCT/US02/41566

(22) International Filing Date:
26 December 2002 (26.12.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
10/037,669 3 January 2002 (03.01.2002) US

(71) Applicant: **INTEL CORPORATION** [US/US]; 2200 Mission College Boulevard, Santa Clara, CA 95052 (US).

(72) Inventors: **FEUERSTRAETER, Mark, T.**; 6122 Oakbridge Drive, Granite Bay, CA 95746 (US). **BOOTH, Bradley, J.**; 3851 Sendero Drive, Austin, TX 78735 (US).

(74) Agents: **MALLIE, Michael, J.** et al.; BLAKELY, SOKOLOFF, TAYLOR & ZAFMAN LLP, 12400 Wilshire Boulevard, 7th Floor, Los Angeles, CA 90025 (US).

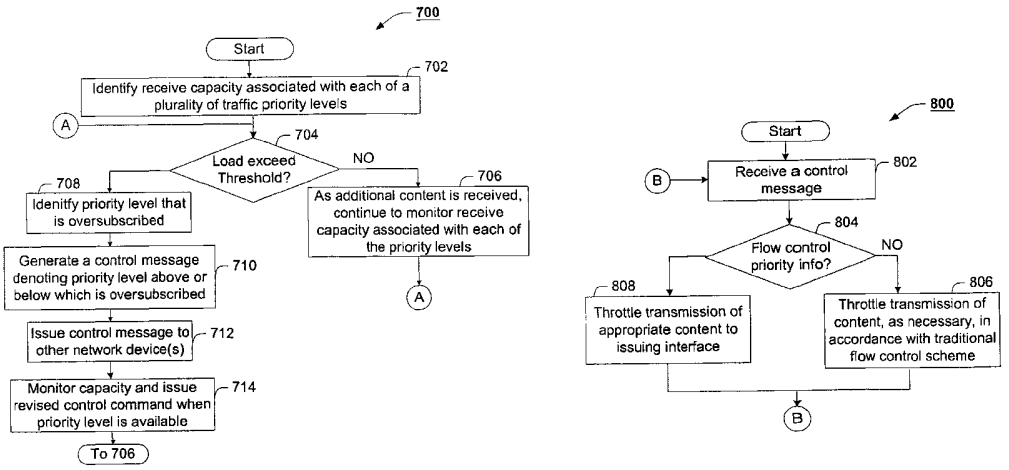
(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, 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, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

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*

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: A METHOD AND APPARATUS FOR PRIORITY BASED FLOW CONTROL IN AN ETHERNET ARCHITECTURE



(57) Abstract: A method and apparatus for priority-based flow control in an Ethernet architecture is generally described. In accordance with one aspect of the invention, a method is presented comprising identifying a receive capability associated with one or more priority levels of Ethernet traffic for a network device, and generating a control message including a flow control priority level, the flow control priority level denoting the identified priority level above or below which the network device has the ability to receive Ethernet traffic.

WO 03/058896 A1

A METHOD AND APPARATUS FOR PRIORITY BASED FLOW CONTROL IN AN ETHERNET ARCHITECTURE

TECHNICAL FIELD

[0001] The present invention generally relates to the field of data networks and, more particularly, to a method and apparatus for priority based flow control within an Ethernet network architecture.

BACKGROUND

[0002] As computer technology has evolved, so too has the use of networks to communicatively couple computer systems together enabling the exchange information. One example of just such a network topology is the *Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specification*, more commonly known as the Ethernet standard topology, developed and promoted under the auspices of the Institute of Electrical and Electronic Engineers as standard 802.3 (2000). One of the provisions of the Ethernet standard (802.3x) facilitates point-to-point flow control between two network devices.

[0003] Simplistically, flow control is a mechanism implemented within a media access controller (MAC) of an Ethernet network interface of a networked device to manage the flow of data between two full-duplex Ethernet devices (i.e., workstation, hub, switch, router, etc.). In accordance with the Ethernet flow control mechanism of 802.3x, a device (or port) that has reached a receive buffer threshold, a condition colloquially referred to as

oversubscribed, sends a control message instructing the transmitting device to pause transmission of Ethernet traffic until further notice (or for a set period of time). Once the receive buffer of the previously oversubscribed device is able to accept additional content, another control message may be sent to resume transmission of the Ethernet traffic. In alternate implementations, a timer is used wherein if a receiving device needs to continue an "off" state, another control message must be issued to maintain this state once the timer has expired. In this regard, the Ethernet flow control provisions are colloquially referred to as an XON/XOFF (transmit on/transmit off) schema, wherein the entire Ethernet communication link between two networked devices is either enabled or disabled.

[0004] One of the limitations often associated with the conventional Ethernet flow control mechanism of 802.3x is that the mechanism does not discriminate between type(s) of traffic, or priorities of traffic. That is, there is no provision within the rudimentary XON/XOFF Ethernet network interface to distinguish between different priority levels of Ethernet traffic, e.g., produced by applications requiring different class of service, quality of service, etc.. As a result, there is no provision to prioritize time-sensitive network content (e.g., multimedia content) over non-time sensitive content (e.g., email content).

Until recently, the conventional Ethernet flow control provisions of IEEE 802.3x were adequate to support the traffic of a typical data network. Today, however, with the increasing popularity of applications requiring uninterrupted flow of content from network resources, the lack of priority-based flow control in the Ethernet architecture is becoming an unbearable limitation. Accordingly, a new flow_control mechanism, unencumbered by the inherent limitations commonly associated with conventional Ethernet flow control is urgently required. Just such a solution is provided in the discussion to follow.

BRIEF DESCRIPTION OF THE DRAWINGS

[0005] The present invention is illustrated by way of example, and not by way of limitation, in the figures of the accompanying drawings in which like reference numerals refer to similar elements and in which:

Fig. 1 is a block diagram of an example data network incorporating an innovative flow control mechanism in accordance with the teachings of the present invention;

Fig. 2 is a block diagram of an example network interface enhanced with an innovative flow control agent, in accordance with one example implementation of the present invention;

Fig. 3 is a graphical illustration of an example receive buffer(s) suitable for use in accordance with the priority based flow control mechanism of the present invention;

Fig. 4 is a graphical illustration of an example transmit buffer suitable for use in accordance with the priority based flow control mechanism of the present invention;

Fig. 5 is a graphical illustration of a flow control management data structure used, for example, in a switch to manage certain aspects of the point-to-point communication link

with one or more communicatively coupled network elements, in accordance with one aspect of the present invention;

Fig. 6 is a graphical illustration of an example Ethernet control message including a flow control priority field, in accordance with one aspect of the present invention;

Fig. 7 is a flow chart of an example method of implementing flow control, in accordance with one aspect of the present invention;

Fig. 8 is a flow chart of an example method of throttling a subset of Ethernet traffic, in accordance with one aspect of the present invention; and

Fig. 9 is a block diagram of an example storage medium comprising a plurality of executable instructions which, when executed, cause an accessing machine to implement one or more aspects of the enhanced network interface, in accordance with an alternate embodiment of the present invention.

DETAILED DESCRIPTION

[0006] The present invention is generally directed to a method and apparatus for priority based flow control in an Ethernet architecture. In this regard, an enhanced network interface (ENI) is introduced incorporating a flow control agent. According to one example implementation, the flow control agent is integrated within a media access controller (MAC) of the ENI to facilitate priority-based flow control.

[0007] In accordance with one aspect of the present invention, the flow control agent monitors the receive capability of the ENI, and selectively generates a control message denoting a priority level above or below which the ENI is able to receive Ethernet traffic. The control message is sent to communicatively coupled device(s) to effectively throttle

transmission of at least a subset of Ethernet traffic intended for the ENI from devices receiving the control message. In this regard, the flow control agent effectively discriminates between priorities of Ethernet traffic, and selectively throttles a mere subset of such traffic, as necessary, in accordance with the identified receive capability of the a host network interface (e.g., ENI).

[0008] In accordance with another aspect of the present invention, the enhanced network interface receives content having different priority levels into a transmit buffer. The media access controller (MAC) of the ENI identifies the destination of received content, an associated flow control priority, if any, associated with the identified destination and selectively forwards only a permissible subset of the content destined for a given network device based, at least in part, on the associated flow control priority level. In this regard, the enhanced network interface facilitates multiple priorities of data to span an Ethernet link, dynamically invoking flow control on only a subset of such priorities. Those skilled in the art will appreciate that having the ability to invoke flow control on a subset of the Ethernet traffic (e.g., only lower priority traffic) effectively enhances the perceived performance of the Ethernet network by only allowing high priority data to traverse the link during periods of congestion.

[0009] Reference throughout this specification to "one embodiment" or "an embodiment" means that a particular feature, structure or characteristic described in connection with the embodiment is included in at least one embodiment of the present invention. Thus, appearances of the phrases "in one embodiment" or "in an embodiment" in various places throughout this specification are not necessarily all referring to the same embodiment.

Furthermore, the particular features, structures or characteristics may be combined in any suitable manner in one or more embodiments.

Example Point-to-Point Ethernet Communication Link

[0010] Those skilled in the art will appreciate that communication within an Ethernet network topology is managed on a per-communication link basis. Thus, when discussing flow control within an Ethernet network topology, it is typically understood as flow control between two Ethernet enabled network devices, i.e., workstation(s), hub(s), switch(es), router(s), and the like. An example of just such a network topology is illustrated with reference to Fig. 1.

[0011] **Fig. 1** illustrates a block diagram of an example Ethernet communication link coupling two network devices, in accordance with the teachings of the present invention. More particularly, in accordance with a first embodiment 100, a computing device/network element 102 is coupled to another computing device/network element 104 through Ethernet communication link 106. In accordance with example implementation of the present invention, each of the computing device/network elements 102 and 104 are depicted comprising an enhanced network interface 120, introduced above. In accordance with the teachings of the present invention, to be developed more fully below, enhanced network interface (ENI) 120 integrated within each of the network device(s) 102, 104 effectively establishes a point-to-point Ethernet communication link between the two devices with a heretofore unavailable traffic prioritization (e.g., based on class of service, type of service, quality of service, etc.) capabilities.

[0012] As used herein, each of computing device/network element are intended to represent a wide variety of Ethernet network elements known in the art such as, for example, a desktop computing platform, a notebook computing platform, a handheld device (e.g., a personal digital assistant), a mobile communications device, or any of a number of network management devices (hub, switch, router, etc.) with Ethernet networking resources. But for the integration of the enhanced network interface (ENI) 120 described more fully below, each of the computing/network elements are intended to represent such conventional devices as they are currently known in the art – no special capability aside from the ENI 120 is required to practice the teachings of the present invention. Accordingly, the architectural details of the host device(s) 102, 104 need not be described further.

[0013] As introduced above, and developed more fully below, ENI 120 is endowed with advanced flow control capability to effectively establish and independently manage different priorities of Ethernet traffic within an Ethernet communication link 106. Accordingly to one aspect of the invention, ENI 120 monitors the receive capability of each of up to a plurality of buffers associated with different content priority levels, and generates a control message denoting a priority level above or below which the ENI has capacity to receive content. According to one implementation, the priority level denotes a level above which the receive buffers of ENI 120 can receive content. In such an implementation, one or more buffers at and/or below that associated with the designated priority level have reached a capacity threshold.

[0014] In accordance with another aspect of the invention, ENI 120 is responsive to receipt of control messages denoting a flow control priority level. In this regard, ENI 120 monitors content in a transmission buffer and forwards select content from the transmission buffer to

a remote network entity based, at least in part, on control messages received from the remote network element.

[0015] Although not particular denoted in Fig. 1, it is to be appreciated that enhanced network interface (ENI) 120 may well be compatible with legacy Ethernet interface(s) to implement the conventional flow control provisions enumerated in the aforementioned IEEE 802.3x, the disclosure of which is incorporated herein by reference for all purposes. That is, if both network elements of the point-to-point Ethernet communication link are endowed with the innovative ENI 120, priority based flow control may well be established over the communication link 106. Alternatively, if one of the network elements is endowed with ENI 120, while the other element of the point-to-point Ethernet communication link is populated with a conventional network interface, ENI 120 will participate in the conventional flow control mechanism defined in 802.3x.

Example Network Interface with Priority Based Flow Control

[0016] **Fig. 2** illustrates a block diagram of an example enhanced network interface (ENI) incorporating the teachings of the present invention. In accordance with the illustrated example embodiment of Fig. 2, ENI 120 is depicted comprising a host system interface 202, one or more input/output buffer(s) 204, a media access controller 206, an encoder/decoder block 208, an attachment unit interface 210, and one or more physical media interface(s) 212, each coupled as depicted. In accordance with the teachings of the present invention, ENI 120 is depicted comprising an innovative flow control agent 214 to facilitate priority based flow control with remote network interface(s) similarly endowed with priority-based flow control features. In the illustrated example implementation of Fig. 2, flow control

agent 214 is integrated within one (or more) media access controller(s) 206. In accordance with such an implementation, flow control agent 214 may well be embodied in software or firmware content executed by MAC to implement the priority-based flow control features described herein. Those skilled in the art will appreciate, however, that flow control agent 214 may well be practiced as an independent element of the network interface as, for example, an application specific integrated circuit (ASIC), a dedicated microcontroller, and the like. In this regard, alternate implementations of the flow control agent 214 are anticipated within the spirit and scope of the claimed invention.

[0017] As used herein, host system interface 202 provides a communication link with a host processing system. In certain implementations, host system interface 202 provides a communication interface with, for example, applications executing on a host computing system that beneficially utilize the network communication resources of ENI 120. As will be discussed more fully below, in accordance with such an example implementation (e.g., as a network interface card (NIC) of a computing system) content from such host applications is received through host system interface 202 and buffered in one or more transmit buffer(s) of I/O buffer(s) 204.

[0018] As used herein, I/O buffer(s) 204 are intended to represent any of a wide variety of memory systems known in the art. According to one implementation, I/O buffer(s) 204 include transmit data structure(s), or queues, and receive data structure(s). According to one example implementation, a number of receive queues are maintained and managed by the MAC 206, wherein each queue is associated with a particular content priority level. Thus, as will be discussed more fully below, flow control agent 214 of MAC 206 monitors the receive capacity of such receive queues in making priority based flow control decisions. It

will be appreciated by those skilled in the art that I/O buffer(s) 204 may well be comprised of any of a number of many different types of physical memory/storage devices.

[0019] As used herein, encoder/decoder 208, AUI 210 and PMI 212 are each intended to represent such elements of a typical network interface. That is, but for their relationship with MAC 206 in general, and flow control agent 214 in particular, such elements 208-212 are intended to represent any of a wide variety of such device(s) known in the art and, as such, need not be further described herein.

[0020] As introduced above, media access controller (MAC) 206 controls the flow of information within the network interface. In accordance with the illustrated example implementation of Fig. 2, MAC 206 is depicted comprising an innovative flow control agent 214. As will be discussed more fully below, flow control agent 214 selectively implements priority-based flow control within a point-to-point Ethernet communication link between two network devices. In this regard, flow control agent 214 detects a priority level associated with content, and selectively forwards only a subset of content to a coupled device in accordance with a priority level identified in a control message received from the coupled device. As used herein, the priority level can be dictated by an application generating the content (e.g., if received from a host application), or embedded within an administrative section (header/footer) of the content. According to one implementation, flow control agent 214 checks administrative information of content received, e.g., from coupled network element for one or more of a class-of-service (CoS), type-of-service (ToS), and/or a quality-of-service (QoS) indication denoting a priority level, and stores such received content in a buffer commensurate with its identified priority level.

[0021] In addition, flow control agent 214 monitors the capacity of receive buffers of the network interface, and selectively generates and issues control message(s) to throttle transmission of content from a coupled device. As will be developed more fully below, flow control agent 214 monitors the receive buffers of I/O buffers to identify the receive capability of each of the buffers associated with one or more content priority levels. Once a buffer has reached a particular threshold, flow control agent 214 generates a control message denoting the priority level above which content can still be received. Issuance of the control message with the priority level information causes an appropriately endowed receiving network interface to throttle traffic having a priority level below that which has been denoted in the control message. According to one implementation, the network device receiving the control message pauses transmission of traffic having a priority level below that which is denoted in the control message until a subsequent control message is received modifying/eliminating the hold for the particular priority level. As indicated above, in alternate implementations, a timer may well be used wherein communication associated with an otherwise suspended priority level may resume after a particular period of time (either pre-determined, or denoted within the received control message). In this regard, flow control agent 214 effectively implements a priority-based flow control mechanism, improving the perceived capacity and capability of the Ethernet network.

Example Data Structure(s)

[0022] Turning to **Fig. 3** a graphical representation of at least a subset of an example memory system 300 is presented. In accordance with the illustrated example of Fig. 3, memory system 300 is depicted comprising a plurality of receive buffer queues 302, 304 and

306. According to one example implementation of the present invention, flow control agent 214 effectively uses each of the buffers 302, 304, 306 as receive buffers to receive content from a remote network device. In accordance with the teachings of the present invention, each of the buffers is associated with content of a disparate priority level, e.g., priority level 1 content, priority level 2 content, through priority level N content (where N denotes the number of priority levels supported by the interface or the communication protocol). As introduced above, each of the buffer queues 302, 304 and 306 may well be comprised of a number of memory devices. Alternatively, each of the different buffers may merely occupy disparate space of a single memory device. Moreover, although illustrated as roughly the same size, those skilled in the art will appreciate that in certain implementations it may be advantageous to prioritize one level of traffic over another. One way in which this may be implemented is to provide the higher priority level with more memory and a higher threshold, while lower priority levels are allocated less memory and/or a lower buffer threshold.

[0023] As will be developed more fully below, each of the buffer queues 302, 304 and 306 are marked with a threshold line 308. As used herein, flow control agent 214 monitors the content (e.g., 310, 312, etc.) within each of the buffers 302, 304, 306, and when such content reaches and/or exceeds the threshold 308, flow control agent 214 generates a control message denoting the priority level associated with the buffer which has reached the threshold 308. A receiving network device implementing the teachings of the present invention will then suspend transmission of content having a priority level at or below the priority level denoted in the received control message.

[0024] **Fig. 4** provides a graphical illustration of a transmission queue 400 suitable for use in accordance with the teachings of the present invention. In accordance with the illustrated example implementation of Fig. 4, transmission queue 400 is depicted comprising a plurality of entries 402, wherein each entry is associated with a datagram (or a stream) of content. According to one example implementation, each of the entries 402 represents a plurality of content having a similar priority level. It is to be appreciated that the number of transmit and receive queues need not be the same. In one example implementation, flow control agent 214 may well implement a number of transmit queues to facilitate precise flow control as enabled by a receiving network element, while implementing two receive buffers, one for high and another for low priority traffic. That is, the number of transmit queues and receive queues are independent of one another.

[0025] As will be developed more fully below, if a flow control agent 214 receives a control message denoting a priority level of four (4), transmission of content in entries associated with a priority level at or below priority level four (4) 404 will be suspended, while transmission of content in entries 406 associated with priority levels above four (4), i.e., priority level five (5) and above, can continue.

[0026] **Fig. 5** is a graphical illustration of an example management data structure 500, suitable for use in accordance with the teachings of the present invention. More particularly, Fig. 5 graphically illustrates an example management data structure 500 used by a network interface implementing priority-based flow control with multiple devices, i.e., over multiple point-to-point Ethernet communication links. As shown, the example management data structure 500 includes at least a destination identifier field 502, a flow control priority field 504 and a number of entries 506 associated with the number of active point-to-point

communication links. Those skilled in the art will appreciate that such a data structure 500 may well be implemented in a network element (hub, switch, router, etc.) where a single flow control agent 214 supports multiple communication ports (i.e., with up to a commensurate number of active point-to-point communication links). It will be appreciated that in the case where each port of such a network device has a dedicated flow control agent 214, e.g., within a dedicated MAC 206, such a data structure may not be required.

[0027] As used herein, the destination identifier field is used to uniquely identify the communication link and/or the remote network element coupled through the communication link. In this regard, it may be a hard-coded port value, dynamically updated with a network element identifier, and the like. The flow control priority field 502 is used to store a current flow control priority level associated with the particular communication link. As shown, certain of the entries 506 denote a flow-control priority level. In accordance with one example implementation of the present invention, flow control agent 214 is compatible with legacy network interface that rely on the conventional Xon/Xoff flow control scheme, as depicted.

[0028] With reference to **Fig. 6**, an example control message data structure is graphically illustrated. In accordance with the illustrated example of Fig. 6, control message datagram 600 is depicted comprising an administrative section 602 including at least a subset of datagram source information field 604, destination information field 606, a type/length field 608 and a flow control priority level field 610. As used herein, the administrative info 602 may well be embodied in a header, a footer, or some other subset of the datagram.

[0029] The source information field 604 denotes the source of the datagram, while the destination info denotes the ultimate destination of the datagram. The type/length field 608

denotes that the datagram is of the control message type, and provides an indication as to the size of the datagram. In accordance with the teachings of the present invention, the priority level field 610 denotes the flow control priority provisions implemented on a per-link basis. According to certain implementations, control messages 600 are effective on a per-link basis and, in this regard, the flow control priority level 610 merely denotes the flow control provisions of the link. Alternate implementations are envisioned wherein the priority level information is used across multiple point-to-point links communicatively coupling a content source 102 with a content destination 104. Those skilled in the art will appreciate that in accordance with one or more implementations of the priority-based flow control of the present invention, control message 600 may well be used to enable communication for a given priority level(s), disable communication for a given priority level(s), and/or refresh a disable of communication for a given priority level(s) (i.e., when a timer mechanism is used to automatically enable a previously disabled priority level).

Example Operation and Implementation

[0030] Having introduced the operating environment and architectural elements of the present invention, above, attention is now directed to Figs. 7 and 8, wherein an example implementation of the priority-based flow control mechanism is presented in greater detail. For ease of illustration, and not limitation, the methods of Figs. 7 and 8 will be developed with continued references to Figs. 1-6, as appropriate. Nonetheless, it is to be appreciate that the teachings of Figs. 7 and 8 may well be implemented in alternate network architectures/configurations without deviating from the spirit and scope of the present invention.

[0031] **Fig. 7** is a flow chart of an example method of implementing flow control, in accordance with one aspect of the present invention. In accordance with the illustrated example implementation of Fig. 7, the method 700 begins with block 702 by identifying a receive capacity/capability of a network interface. In accordance with the teachings of the present invention, flow control agent 214 scans each of a plurality of receive buffers 302, 304, 306 to determine whether the content in the buffer(s) has reached/exceeded a threshold 308, block 704. As introduced above, each of the buffers 302, 304, 306 is associated with a disparate content priority level.

[0032] If, in block 704, that the buffer load has not exceeded a threshold 308, network interface allows continue content to be received from the remote network device, as flow control agent 214 continues to monitor the receive capacity associated with each of the priority levels until a threshold in one of the buffers is reached.

[0033] If the threshold of a buffer is reached the process continues with block 708 wherein the flow control agent 214 identifies the priority level that is oversubscribed. In this regard, flow control agent 214 identifies which of the receive buffer(s) 302, 304, 306 has reached the threshold 308, and generates a control message 600 denoting the priority level 610 above or below which is oversubscribed, block 710. In accordance with one implementation, the priority level denotes the level at and below which the flow control agent 214 is suspending transmission. In alternate implementations, the flow control priority level denoted in field 610 of a control message 600 may well denote a content priority level above which is oversubscribed, suspending transmission of content at and above the level denoted in the control message 600.

[0034] In block 712 flow control agent 214 issues the control message 600 to other network device(s). More particularly, flow control agent 214 transmits the generated control message 600 denoting a flow control priority level 610 to a network element coupled through the point-to-point Ethernet communication link. According to one example implementation, once the control message 600 has been issued, flow control agent 214 monitors the capacity of buffer(s) associated with a priority level denoted in the generated control message 600 and issues a revised control command when the buffer associated with the priority level becomes available (i.e., falls below the threshold 308), block 714. The process then continues with block 706.

[0035] Turning to **Fig. 8,** a flow chart of an example method of throttling a subset of Ethernet traffic is presented, in accordance with one aspect of the present invention. In accordance with the illustrated example implementation of Fig. 8, the process begins with block 802 wherein the network interface receives a control message 600 from a remote network element coupled through a point-to-point Ethernet communication link (e.g., 106). In accordance with one example implementation, the control message 600 is routed to a media access controller 206 of the receiving network interface. In accordance with the teachings of the present invention, a flow control agent 214 implemented within the network interface determines whether the control message 600 includes flow control priority information, block 804.

[0036] If flow control agent 214 does not identify a priority level field 610 in the received control message 600, the process continues with block 806 wherein the flow control agent 214 throttles transmission of all content within the transmission buffer 400 on the point-to-

point link, as necessary, in accordance with the conventional Ethernet flow control mechanism of 802.3x.

[0037] If, however, flow control agent 214 identifies a priority level field 610 in the received control message 600, flow control agent 214 throttles transmission of an appropriate subset of the total content within the transmission buffer 400, block 808, in accordance with the priority-based flow control features of the present invention. As described above, if flow control agent 214 receives a control message 600 including a priority level field 610 populated with an indication that priority level(s) four (4) and below are oversubscribed, flow control agent 214 merely authorizes the continued transmission by MAC 206 of content 406 of the transmission buffer 400, representing that content having a priority level above that which is oversubscribed.

Alternate Embodiment(s)

[0038] **Fig. 9** is a block diagram of an example storage medium comprising a plurality of executable instructions which, when executed, cause an accessing machine to implement one or more aspects of the innovative priority-based flow control mechanism of the present invention. In this regard, storage medium 900 includes content for implementing an enhanced network interface 120 with priority based flow control features. scalable network interface 200 of the present invention, in accordance with an alternate embodiment of the present invention.

[0039] In the description above, for the purposes of explanation, numerous specific details are set forth in order to provide a thorough understanding of the present invention. It will be

apparent, however, to one skilled in the art that the present invention may be practiced without some of these specific details. In other instances, well-known structures and devices are shown in block diagram form.

[0040] The present invention includes various steps. The steps of the present invention may be performed by hardware components, such as those shown in Figs. 1-5, or may be embodied in machine-executable instructions, which may be used to cause a general-purpose or special-purpose processor or logic circuits programmed with the instructions to perform the steps. Alternatively, the steps may be performed by a combination of hardware and software. Moreover, although the invention has been described in the context of a network interface device, those skilled in the art will appreciate that such functionality may well be embodied in any of number of alternate embodiments such as, for example, integrated within a computing device, and is readily adapted to wireless Ethernet implementations as well as the wired environment described herein.

[0041] The present invention may be provided as a computer program product which may include a machine-readable medium having stored thereon instructions which may be used to program a computer (or other electronic devices) to perform a process according to the present invention. The machine-readable medium may include, but is not limited to, floppy diskettes, optical disks, CD-ROMs, and magneto-optical disks, ROMs, RAMs, EPROMs, EEPROMs, magnet or optical cards, flash memory, or other type of media / machine-readable medium suitable for storing electronic instructions. Moreover, the present invention may also be downloaded as a computer program product, wherein the program may be transferred from a remote computer to a requesting computer by way of data signals

embodied in a carrier wave or other propagation medium via a communication link (e.g., a modem or network connection).

Many of the methods are described in their most basic form but steps can be added to or deleted from any of the methods and information can be added or subtracted from any of the described messages without departing from the basic scope of the present invention. For example, in one embodiment a single receive buffer is employed utilizing different threshold set points, which would effectively distinguish the different levels of priority flow control. As an example, if 50% of the single que is full, send a flow control message with value 4 (half way point, i.e. high and low traffic), at 75% full, send a message with value 2, etc. It will be apparent to those skilled in the art that many further modifications and adaptations can be made. The particular embodiments are not provided to limit the invention but to illustrate it. The scope of the present invention is not to be determined by the specific examples provided above but only by the claims below.

CLAIMS:

What is claimed is:

1. A method comprising:

identifying a receive capability associated with one or more priority levels of Ethernet traffic for a network device; and

generating a control message including a flow control priority level, the flow control priority level denoting the identified priority level above or below which the network device has the ability to receive Ethernet traffic.

2. A method according to claim 1, further comprising:

transmitting the generated control message to a communicatively coupled network device, whereupon receipt of the generated control message the communicatively coupled network device acts in accordance with the received control message to suspend a subset of Ethernet traffic.

3. A method according to claim 1, wherein identifying comprises:

determining available buffer capacity for each of a plurality of buffers associated with a commensurate plurality of Ethernet priority levels.

4. A method according to claim 3, wherein the available buffer capacity associated with a particular Ethernet priority level denotes the ability of the buffer to receive additional Ethernet traffic of that priority level.

5. A method according to claim 3, wherein the buffer for each priority level is comprised of one or more memory device(s).

6. A method according to claim 3, wherein the buffers associated with each of the priority levels are virtual buffers implemented within a common physical buffer.

7. A method according to claim 3, wherein the generated control message includes an indication of the priority level above which a receive buffer has available capacity to receive Ethernet traffic of an associated priority level.

8. A method according to claim 7, wherein a receiving network device initiates a pause in transmission of Ethernet traffic having a priority level below that indicated in the received control message.

9. A method according to claim 1, wherein generating a control message comprises:

generating an Ethernet control packet including a priority field, the priority field denoting the flow control priority level.

10. A method according to claim 9, wherein the priority field is included in a header portion of the Ethernet control packet.

11. A method according to claim 1, further comprising:

receiving Ethernet traffic;

identifying a priority level associated with each packet of received Ethernet traffic; and

forwarding each received packet to a receive buffer based, at least in part, on the identified priority level associated with the Ethernet packet.

12. A method according to claim 11, further comprising:

monitoring the receive capability of buffers associated with each of the priority levels of Ethernet traffic; and

issuing control messages, as necessary, to throttle transmission of at least a subset of Ethernet traffic in accordance with the identified receive capability associated with the one or more priority levels.

13. A method according to claim 12, wherein throttling transmission of a subset of Ethernet traffic comprises temporarily suspending transmission of the subset of Ethernet traffic for a set period of time and/or until another control message is received denoting that transmission of the subset of Ethernet traffic may resume.

14. A method comprising:

receiving a control message denoting a flow control priority level from a network device; and

throttling transmission to the network device of a subset of Ethernet traffic having a priority level above or below that denoted in the received control message.

15. A method according to claim 14, wherein the flow control priority level denotes a priority level associated with a subset of Ethernet traffic above which the issuing network device has a receive capability.

16. A method according to claim 14, wherein the control message is an Ethernet control message.

17. A method according to claim 16, further comprising:

analyzing a header of the received Ethernet control message to identify a flow control priority level.

18. A method according to claim 14, wherein throttling transmission comprises:

suspending transmission of a subset of Ethernet traffic having a priority level below the flow control priority level denoted in the received control message until a subsequent control message is received denoting an ability of an issuing network device to receive the subset of Ethernet traffic.

19. A method according to claim 14, further comprising:

receiving content from a host network device for transmission to another network device communicatively coupled through an Ethernet network; and

assigning a priority level to the received content based, at least in part, on a source of such content.

20. A method according to claim 14, further comprising:

receiving content from one or more source applications executing on a host network device, the content tagged with a priority level associated with its source application; and

selectively transmitting received content to another network device communicatively coupled through an Ethernet network based, at least in part, on the priority level of the content and received control message(s) throttling transmission of a subset of such Ethernet traffic.

21. A network interface comprising:

a plurality of receive buffers, each associated with a particular priority level of Ethernet traffic; and

control logic, coupled to the receive buffers, to identify a receive capability of each of the receive buffers and selectively generate control message(s) including a flow control priority level denoting the identified priority level above or below which the network interface has the ability to receive Ethernet traffic.

22. A network interface according to claim 21, further comprising:

a transmit buffer, responsive to a host network device and the control logic, to receive content from one or more application(s) executing on the host network device for transmission to other network device(s) through an Ethernet network, the received content including an indication of priority level.

23. A network interface according to claim 22, wherein the indication of priority level in the received content is determined by its source application.

24. A network interface according to claim 22, wherein the control logic receives control message(s) from other network interface(s), wherein at least a subset of the control messages include a flow control priority level denoting an inability to receive Ethernet traffic having a priority level below that of the denoted flow control priority level.

25. A network interface according to claim 24, wherein the control logic suspends transmission of Ethernet traffic having a priority level below that of the denoted flow control priority level from the transmit buffer to the network device having issued the control message.

26. A network interface according to claim 21, wherein the control logic is a media access controller (MAC).

27. A network interface according to claim 26, the MAC including enhanced flow control capability to implement flow control on a mere subset of Ethernet traffic.

28. A machine accessible medium comprising content which, when executed by an accessing machine, causes the machine to implement a network interface with

enhanced Ethernet flow control capability to selectively throttle a mere subset of Ethernet traffic.

29. A machine accessible medium according to claim 29, wherein the network interface identifies a receive capability associated with one or more priority levels of Ethernet traffic for a network device, and selectively generates a control message including a flow control priority level, the flow control priority level denoting the identified priority level above or below which the network device has the ability to receive Ethernet traffic

30. A machine accessible medium according to claim 28, wherein the network interface receives content from a host network device, the received content denoting a priority level associated with its source application, and wherein the network interface throttles transmission of the received content to another network device communicatively coupled through an Ethernet network based, at least in part, on control messages received from the another network device denoting flow control priority information.

FIG. 1

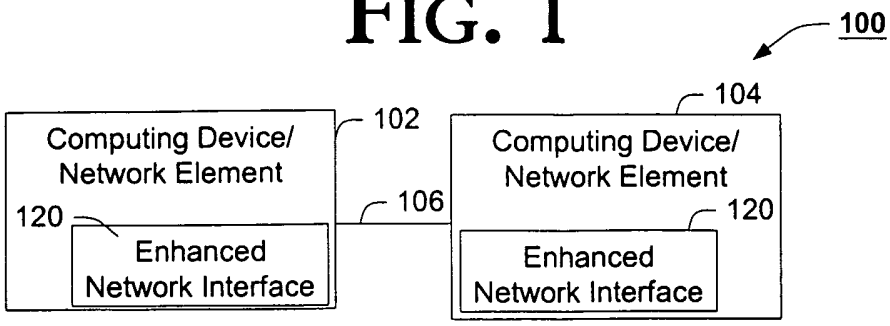


FIG. 2

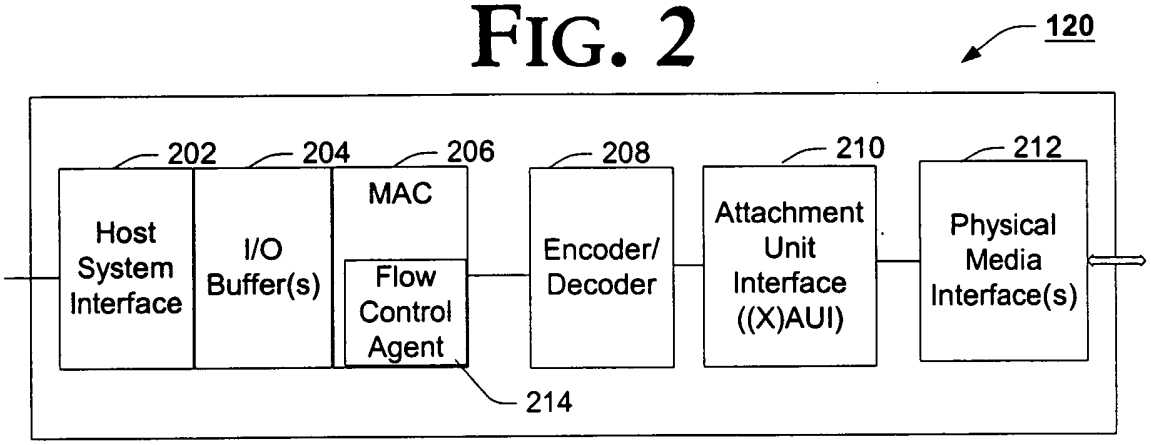


FIG. 3

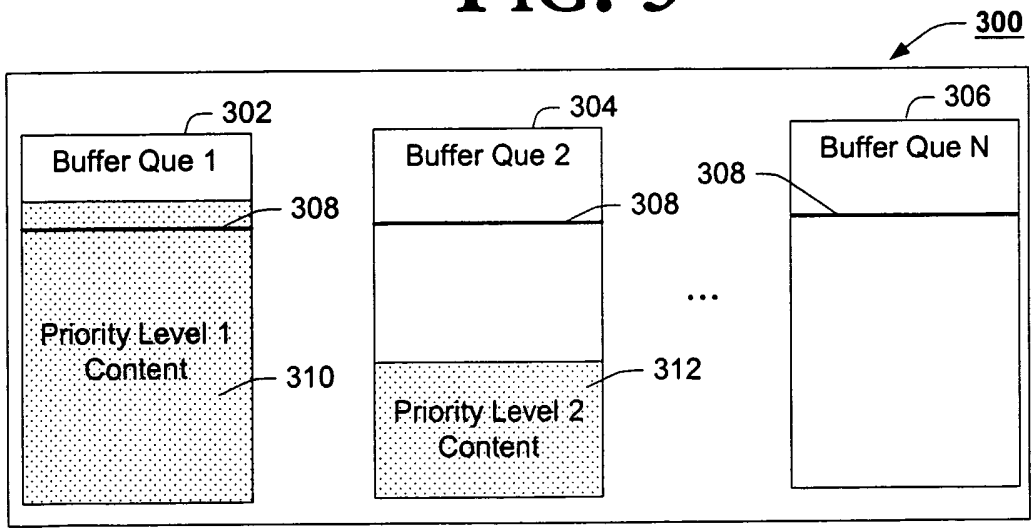


FIG. 4

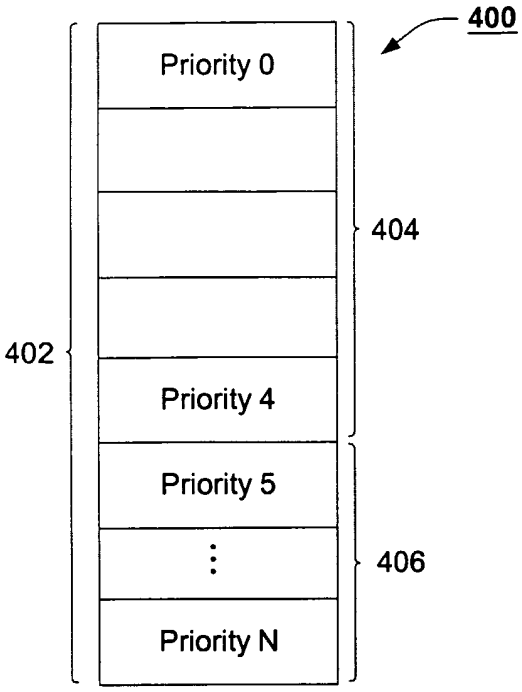


FIG. 5

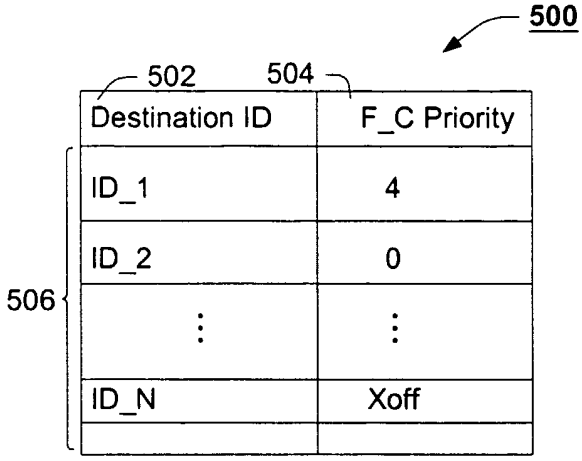


FIG. 6

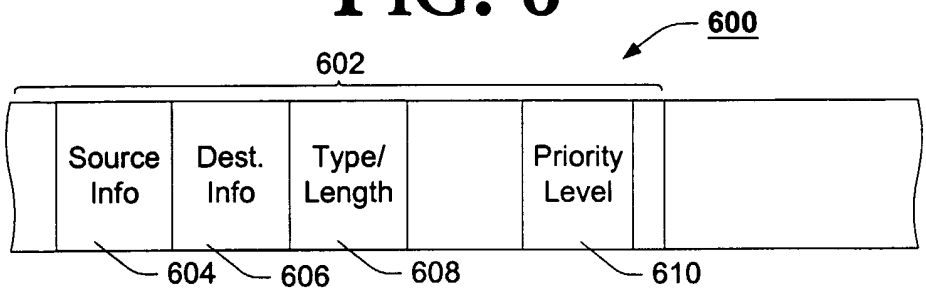


FIG. 9

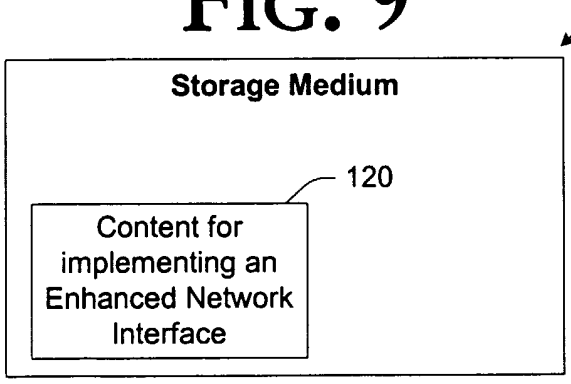


FIG. 7

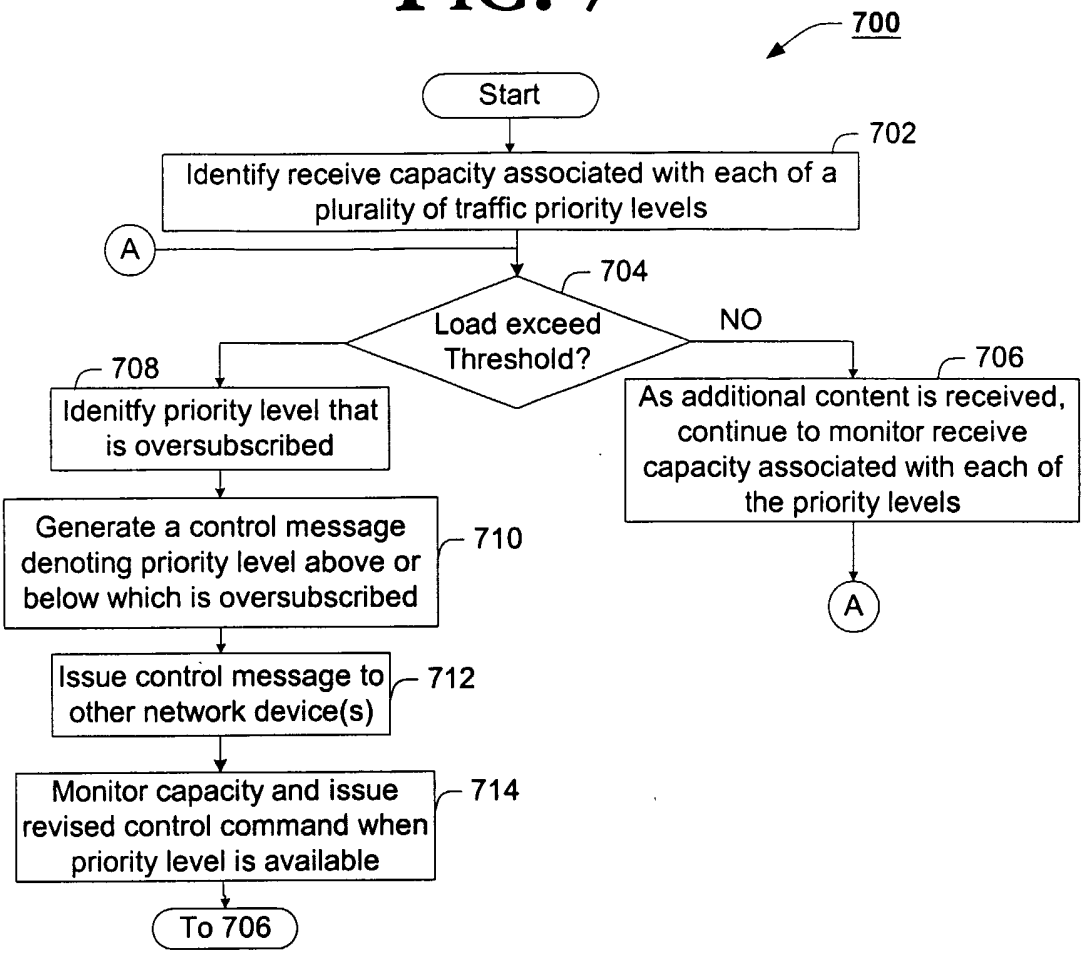
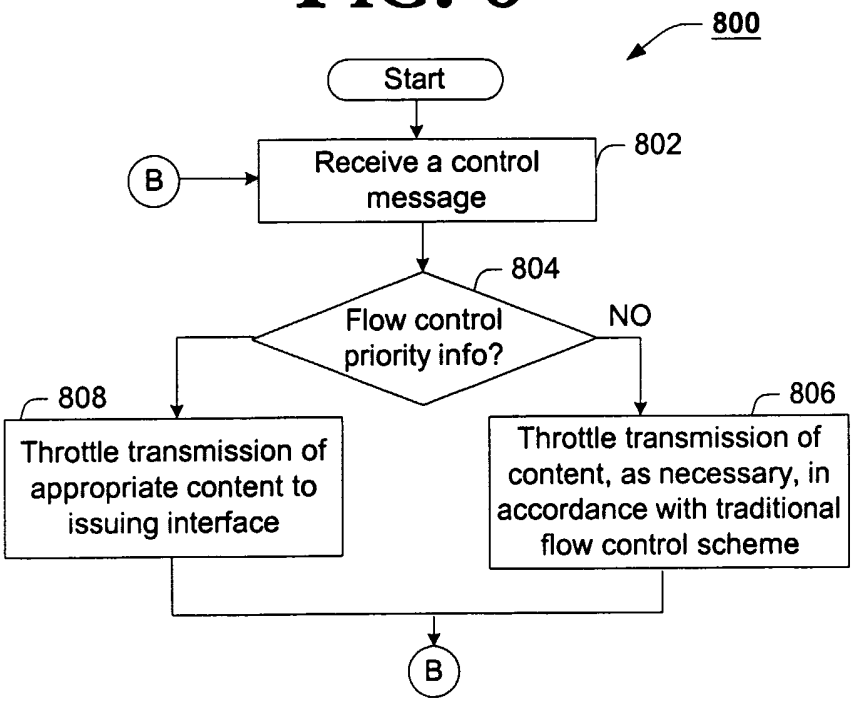

700
Start
702
Identify receive capacity associated with each of a plurality of traffic priority levels
A
704
Load exceed Threshold?
NO
708
Idenitfy priority level that is oversubscribed
706
As additional content is received, continue to monitor receive capacity associated with each of the priority levels
A
Generate a control message denoting priority level above or below which is oversubscribed
710
Issue control message to other network device(s)
712
Monitor capacity and issue revised control command when priority level is available
714
To 706

FIG. 8


800
Start
B
Receive a control message
802
804
Flow control priority info?
NO
808
Throttle transmission of appropriate content to issuing interface
806
Throttle transmission of content, as necessary, in accordance with traditional flow control scheme
B

INTERNATIONAL SEARCH REPORT

ıational Application No
PCT/US 02/41566

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 5 983 278 A (KATZ MARK ET AL) 9 November 1999 (1999-11-09) column 3, line 29 -column 4, line 46 column 8, line 60 -column 9, line 56 column 12, line 30 -column 14, line 63 figures 1,2,5 ---	1-29
Y	DE 101 23 821 A (IBM) 20 December 2001 (2001-12-20) column 2, line 36 -column 3, line 39 column 8, line 4 -column 9, line 47 figure 2 --- -/--	1-29

[X] Further documents are listed in the continuation of box C.

[X] 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	Date of mailing of the international search report
30 April 2003	09/05/2003

Name and mailing address of the ISA	Authorized officer
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	Kreppel, J

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

itional Application No
PCT/US 02/41566

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
A	NOUREDDINE W ET AL: "SELECTIVE BACK-PRESSURE IN SWITCHED ETHERNET LANS" 1999 IEEE GLOBAL TELECOMMUNICATIONS CONFERENCE. GLOBECOM'99. SEAMLESS INTERCONNECTION FOR UNIVERSAL SERVICES. RIO DE JANEIRO, BRAZIL, DEC. 5-9, 1999, IEEE GLOBAL TELECOMMUNICATIONS CONFERENCE, NEW YORK, NY: IEEE, US, vol. 2, 5 December 1999 (1999-12-05), pages 1256-1263, XP001016912 ISBN: 0-7803-5797-3 page 1260, right-hand column, paragraph 2 -page 1261, right-hand column ---	1-29
A	ROUSSOS J K ET AL: "CONGESTION CONTROL PROTOCOLS FOR INTERCONNECTED LANS SUPPORTING VOICE AND DATA TRAFFIC" COMPUTER COMMUNICATIONS, ELSEVIER SCIENCE PUBLISHERS BV, AMSTERDAM, NL, vol. 17, no. 1, 1994, pages 25-34, XP000415044 ISSN: 0140-3664 page 25, left-hand column -page 29, left-hand column, paragraph 2 ---	1-29
P,X	US 2002/087723 A1 (ERIMLI BAHADIR ET AL) 4 July 2002 (2002-07-04) paragraph '0041! - paragraph '0061! -----	1-29

INTERNATIONAL SEARCH REPORT

Information on patent family members

ıational Application No
PCT/US 02/41566

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5983278 A	09-11-1999	CA 2200135 A1	19-10-1997
		JP 10070555 A	10-03-1998
		US 6212582 B1	03-04-2001
DE 10123821 A	20-12-2001	DE 10123821 A1	20-12-2001
		US 2002031142 A1	14-03-2002
US 2002087723 A1	04-07-2002	WO 02054681 A1	11-07-2002