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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
25 September 2003 (25.09.2003)

PCT

(10) International Publication Number
WO 03/079612 A1

(51) International Patent Classification[7]: H04L 12/28, 12/56

(21) International Application Number: PCT/US02/39784

(22) International Filing Date:
11 December 2002 (11.12.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
10/099,734 15 March 2002 (15.03.2002) US

(71) Applicant: **ADAPTEC, INC.** [US/US]; 691 S. Milpitas Blvd., Milpitas, CA 95035 (US).

(72) Inventor: **MUKUND, Shridhar**; 22232 Woodbury Lane, San Jose, CA 95121 (US).

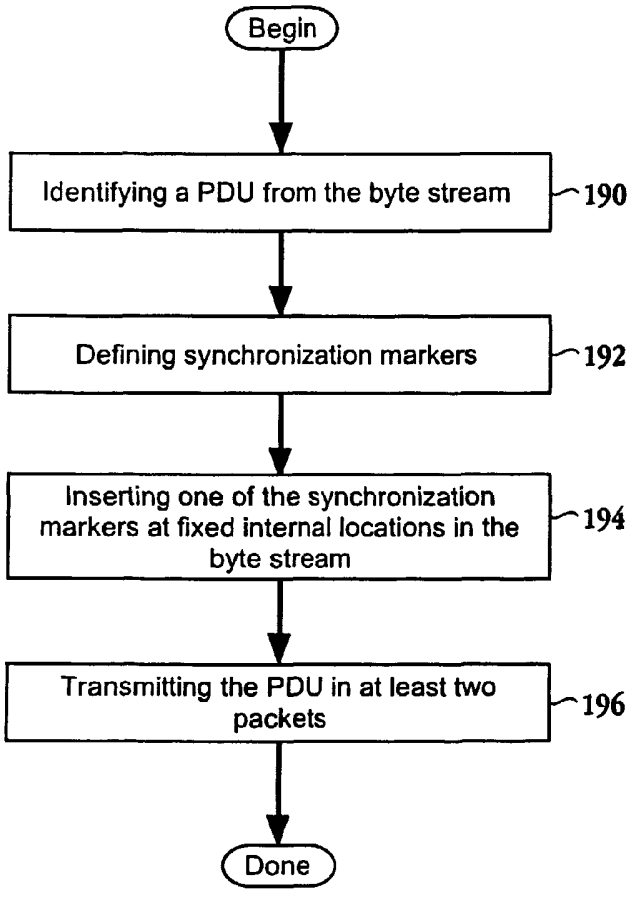
(74) Agent: **GENCARELLA, Michael, L.**; Martine & Penilla, LLP, 710 Lakeway Drive, Suite 170, Sunnyvale, CA 94085 (US).

(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),

[Continued on next page]

(54) Title: METHOD AND APPARATUS FOR DIRECT DATA PLACEMENT OVER TCP/IP



(57) Abstract: Methods and an apparatus capable of placing data packets received out of order directly into the memory of a receiving device is provided. In one embodiment, a method (Fig. 10) for receiving a byte stream at a host from a networked transmitting device is provided. The method initiates with a packet of a protocol data unit (PDU) being received (109). Next, a synchronization marker contained within the packet is located (192) .Then, PDU parameters contained in the synchronization marker are read to determine a position of the received packet within the PDU (194) Next, the packet is written directly into memory of the host (196). A method for processing a byte stream to be transmitted to a host over a network is also provided. A device for processing a data stream to be transmitted over a network and a system configured to receive a byte stream from a networked transmitting device are also provided.

WO 03/079612 A1

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).

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.

Published:

— *with international search report*

Method and Apparatus for Direct Data Placement Over TCP/IP

by Inventor

Shridhar Mukund

BACKGROUND OF THE INVENTION

1. Field of the Invention

This invention relates generally to networking and more particularly to a method and apparatus for allowing a receiver's network interface to place packets of data directly to its final destination irrespective of the order in which the packets of data are received.

2. Description of the Related Art

One aspect of the quest for faster connections over distributed networks, such as the Internet, is being hindered by the splitting of a byte stream into packets as it is transmitted over a network from an origin to a destination. Much of today's usage of the Internet and Internet Protocol (IP) networks is for buffer-to-buffer data transfers, often in the form of bulk data and inter-process communications. These end-to-end bulk data transfer operations taking place over IP networks are typically limited by data copying overhead. Additionally, gigabit speed buffer-to-buffer network transfers use significant memory bandwidth and central processing unit (CPU) time of the device receiving the data due to the interrupts. The structure of IP based networks currently requires a high price in overhead because of the necessity of copying the transferred data to a local memory of the network interface of the receiving unit in order to place the data in the receiving unit's memory buffer. More specifically, the receiving device such as a network card must copy data segments received out of order, reassemble the data in order when it has all been received and send the reassembled data to the host memory.

Figure 1 is a schematic diagram of a system configured to receive data from a distributed network. Network interface card (NIC) 102 is connected to a distributed network 100, such as the Internet. As data is sent over network 100, it is received by NIC 102 prior to being stored in memory 110 of host 108. The data sent over network 100 is typically transmitted as packets of a protocol data unit (PDU). Figure 2 is a schematic diagram of a PDU split into packets by a transmitting device for transmission over a network. PDU 112 includes a header 114, and data 116. As is well known in the art, header 114 of PDU 112 includes buffer

parameters, offset, and the length of the PDU. When PDU 112 is transmitted over a network, such as the Internet, the PDU is split into packets. As shown by Figure 2, PDU 112 is split into packets 120-1 through 120-5. Each of packets 120-1 through 120-5 has header 122-1 through 122-5, respectively, inserted at the beginning of each packet. For example, if packets 120-1 through 120-5 are Transmission Control Protocol (TCP) packets, headers 122-1 through 122-5 are TCP headers which are eventually stripped off before the data in the packets is presented to an upper level protocol (ULP). Since packets 120-1 through 120-5 are individual units, they may be transmitted through different routes from an origin to a destination over a network. Thus, when using any IP based network it should be expected that the packets of a transmitted PDU will arrive at a destination out-of-order, i.e., differently than the order of the packets in the original PDU.

Referring back to Figure 1, the out-of-order packets received by NIC 102 are sent to local memory 104 of NIC 102. The out of order packets of the PDU are re-assembled in memory 104 and sent to the proper location of memory 110 of host 108 through bus 106. One skilled in the art will appreciate that the header of a PDU provides location data for where the PDU should be placed in memory 110. Thus, if header 114 of PDU 112 of Figure 2 arrives first to NIC 102 of Figure 1, then the data contained in the header, i.e., buffer parameters, offset and length, allow for the data to be placed into memory 110 without having to reassemble the packets of the PDU.

However, if the PDU header is lost or even if it comes late, the non-PDU header packets must be stored in the local memory of the NIC until the packet containing the PDU header is received. The local storage of the PDU packets require more memory space and bandwidth being provided to accommodate the packets. Network interface accelerators have been adopted to offload and accelerate transport protocol processing. As these accelerators use large, high speed memories to buffer and reassemble out of order transport packets, cost is an issue in terms of price of the increased memory capacity and overhead. More particularly, overhead is dominated by the costs of processing and copying incoming data in order to place it at its ultimate destination.

Remote direct memory access (RDMA) has been used to transfer data from a local computer directly to a memory address space of a remote computer. However, current RDMA protocols do not address the issues of receiving data split into packets and transmitted over IP networks and subsequently received in a different order than how the data was transmitted. Attention is being focused on introducing alternate protocols to avoid the perceived inherent

problems with TCP, such as the splitting up of PDUs into packets and the out-of-order receipt of the packets. Since TCP has been tried and tested resulting in its wide use and acceptance, it would be beneficial to adopt a RDMA technique to be used with any IP based protocol such as TCP.

In view of the foregoing, there is a need to allow a byte stream transmitted as packets of data to be placed directly into the memory of a receiving device without having to reassemble the byte stream in its original sequential order prior to placing the byte stream in memory.

SUMMARY OF THE INVENTION

Broadly speaking, the present invention fills these needs by providing a method and apparatus capable of placing packets that are received out-of-order directly into the memory of the receiving device. It should be appreciated that the present invention can be implemented in numerous ways, including as a process, an apparatus, a system, or a device. Several inventive embodiments of the present invention are described below.

In one embodiment, a method for processing a byte stream to be transmitted from a storage device to a requesting host over a network is provided. The method initiates with a protocol data unit (PDU) from the byte stream being identified. The PDU includes a header and a tail. Then, synchronization markers containing parameters of the PDU are defined. Next, one of the synchronization markers is inserted at each of the fixed locations of the PDU. The fixed locations are equally spaced apart. Then, the PDU is transmitted in at least two packets, and at least one of the two packets contains one of the inserted synchronization markers.

In another embodiment, a method for processing a byte stream to be transmitted to a host over a network is provided. The method initiates with a protocol data unit (PDU) from the byte stream being identified, where the PDU has a header and a tail. Then, synchronization markers containing parameters of the PDU are defined. Next, one of the synchronization markers is inserted at fixed locations of the PDU. The fixed locations are equally spaced apart. Then, the PDU is transmitted in two or more packets, and each of the two or more packets contain one of the inserted synchronization markers.

In yet another embodiment, a method for receiving a byte stream at a host from a networked transmitting device is provided. The method initiates with a packet of a protocol data unit (PDU) being received. Next, a synchronization marker contained within the packet is located. Then, PDU parameters contained in the synchronization marker are read to determine

a position of the received packet within the PDU. Next, the packet is written directly into memory of the host.

In still another embodiment, a method for receiving a byte stream at a host from a networked transmitting device is provided. The method initiates with (a) a packet of a protocol data unit (PDU) being received. Then, (b) a synchronization marker contained within the packet is located. Next, (c) PDU parameters contained in the synchronization marker are read to determine a position of the received packet within the PDU. Then, (d) the packet is written directly into memory of the host. Next, operations (a) – (d) above are repeated for each successive packet received for the PDU.

In another embodiment, a method for receiving a byte stream at a host having a network card for connecting to a networked transmitting device is provided. The method initiates with a packet of a protocol data unit (PDU) being received. Then, it is determined whether the packet contains a synchronization marker. If the synchronization marker is present, the method includes reading PDU parameters contained in the synchronization marker to determine a position of the received packet within the PDU and writing the packet directly into memory of the host. If the synchronization marker is not present, the method includes holding the packet in memory of the network card, assembling the packet with another received packet that does contain the synchronization marker, and writing the packet without the synchronization marker and the packet with the synchronization marker into memory of the host.

In yet another embodiment, a device for processing a byte stream to be transmitted over a network is provided. The device includes a network interface. The network interface includes circuitry for identifying a protocol data unit (PDU) from the byte stream and circuitry for defining synchronization markers containing parameters of the PDU. Circuitry for inserting one of the synchronization markers at fixed locations of the PDU is included, wherein the fixed locations are equally spaced apart. Circuitry for transmitting the PDU in at least two packets is included, wherein at least one of the two packets contains one of the inserted synchronization markers.

In still another embodiment, a computer system configured to receive a byte stream from a networked transmitting device is provided. The computer system includes a central processing unit (CPU) and a memory configured to receive the byte stream from a network card. The network card includes circuitry for receiving a packet of a protocol data unit (PDU) of the byte stream and circuitry for locating a synchronization marker contained within the

packet. Circuitry for reading PDU parameters contained in the synchronization marker to determine a position of the received packet within the PDU is included. Circuitry for writing the packet directly into memory is also included.

Other aspects and advantages of the invention will become apparent from the following detailed description, taken in conjunction with the accompanying drawings, illustrating by way of example the principles of the invention.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention will be readily understood by the following detailed description in conjunction with the accompanying drawings, and like reference numerals designate like structural elements.

Figure 1 is a schematic diagram of a system configured to receive data from a distributed network.

Figure 2 is a schematic diagram of a PDU split into packets by a transmitting device for transmission over a network.

Figure 3 is a schematic diagram of a protocol data unit (PDU) of a byte stream in accordance with one embodiment of the invention.

Figure 4 is a schematic diagram of a PDU of a byte stream that includes synchronization markers in accordance with one embodiment of the invention.

Figure 5 is a block diagram listing PDU parameters contained by a synchronization marker inserted into a PDU in accordance with one embodiment of the invention.

Figure 6 is a schematic diagram of a PDU of a byte stream that includes synchronization markers for each packet of the PDU in accordance with one embodiment of the invention.

Figure 7 is a schematic diagram of a packet of a PDU where the synchronization marker included in the packet provides PDU parameters for a receiving device so that the packet can be directly placed into memory in accordance with one embodiment of the invention.

Figure 8 is a schematic diagram of a packet of a PDU where the synchronization markers are not included in every packet of the PDU in accordance with one embodiment of the invention.

Figure 9 illustrates a schematic of a system configured to transmit packets of data over a network and receive the packets of data directly into memory of the receiving device in accordance with one embodiment of the invention.

Figure 10 is a flowchart diagram of the method operations performed in processing a byte stream to be transmitted from a storage device to a requesting host over a network in accordance with one embodiment of the invention.

Figure 11 is a flowchart diagram of the method operations for receiving a byte stream at a host from a networked transmitting device in accordance with one embodiment of the invention.

Figure 12 is a flowchart diagram of the method operations for receiving a byte stream for a host where each packet of the byte stream does not contain a synchronization marker in accordance with one embodiment of the invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

An invention is described for an apparatus and method for transmitting and receiving Transmission Control Protocol (TCP) packets of a data stream where the transmitting device conditions the TCP packets to allow the TCP packets to be placed directly into memory of the receiving device. It will be obvious, however, to one skilled in the art, that the present invention may be practiced without some or all of these specific details. In other instances, well known process operations have not been described in detail in order not to unnecessarily obscure the present invention.

The embodiments of the present invention described below provide a method and apparatus extending direct data placement (DDP) and remote direct memory access (RDMA) to TCP/ Internet Protocol (IP) based protocols. TCP/IP based protocols include any protocols or applications that can sit on top of the TCP, such as Internet small computer system interface (ISCSI), and other similar upper layer protocols. As is well known, the data packets of the IP based protocols are not necessarily transmitted over the same route through the IP network. Thus, the packets will not arrive at the destination in the order that the packets were sent. By including synchronization markers in the transmitted data stream, a receiving device is able to determine the location in memory for any received packet of the data stream from the information contained within the synchronization markers. Thus, the increased memory requirements of the network interface of the receiving device are avoided. It should be appreciated that while the embodiments disclosed herein are described in terms of protocol data units (PDU) of upper layer protocols (ULP) for illustrative purposes, they are not meant to be limiting as the method and apparatus can function with any packet based protocol or TCP/IP based networking protocol.

Figure 3 is a schematic diagram of a protocol data unit (PDU) of a byte stream in accordance with one embodiment of the invention. Here, the origin of a TCP connection is represented by vertical line 124. A byte stream 126 includes PDU 128. PDU 128 is defined between header 130 and tail 132. One skilled in the art will appreciate that header 130 includes information on where PDU 128 should be placed in the memory of a receiving device. For example, header 130 includes region identifiers (RID), a buffer offset and a length of the PDU. As is well known, the RID, also referred to as buffer parameter index (BPI), is an opaque identifier of a memory region on a node, wherein the memory region is a contiguous range of bytes in a particular address space that has an associated length. As used herein, a node is a computer attached to one or more links of a network. Tail 132 includes data, such as a RID, the buffer offset and a forward distance, which are described in more detail below with respect to Figure 5.

Figure 4 is a schematic diagram of a PDU of a byte stream that includes synchronization markers in accordance with one embodiment of the invention. Here, byte stream 126 includes PDU 128 defined between header 130 and tail 132. Synchronization markers 136-1 through 136-4 are distributed throughout byte stream 126. A sequence number is associated with each byte of byte stream 126. Scale 134 represents the sequential sequence numbers assigned to each byte of byte stream 126. The sequence numbers assist in the placement of PDU 128 into memory of a receiving device as will be explained in more detail below. In one embodiment of the invention, synchronization markers 136-1 through 136-4 are evenly distributed over byte stream 126. That is, the distance between each synchronization marker 136-1 through 136-4 is the same, which is another way of saying the frequency of the synchronization markers is uniform throughout the byte stream. In another embodiment, the number of markers is greater than the number of packets making up the transmitted PDU, but less than two times the number of packets making up the PDU. One skilled in the art will appreciate that TCP packets are less than or equal to 1.5 Kilobytes (KB) in length. Thus, a distance of approximately 1.5KB between synchronization markers will essentially ensure that each PDU contains a synchronization marker. However, as will be explained in more detail below, it is not required that each packet of the PDU contain a synchronization marker.

Still referring to Figure 4, synchronization markers 136-1 through 136-4 are inserted into byte stream 126 by a transmitting device. One skilled in the art will appreciate that the transmitting device can be any device configured to transmit data through a network interface over a distributed network, such as a networked personal or portable computer, a storage

device, a personal digital assistant (PDA), etc. The transmitting device includes logic for inserting the synchronization markers into byte stream 126 at a defined frequency. Each of the synchronization markers include parameters corresponding to the PDU within which the synchronization marker is contained. Thus, synchronization markers 136-1, 136-2 and 136-3 include parameters corresponding to PDU 128. Synchronization marker 136-4 contains parameters for a PDU following PDU 128. The PDU parameters are described in more detail below with reference to Figure 5. PDU 128 is split into packets for transmission over a network. Preferably, each packet contains a synchronization marker, however, it is not required that each packet contain a synchronization marker. Through the information from the PDU parameters of each synchronization marker, a receiving device is enabled to place each of the packets directly into memory of the receiving device. As will be explained further below, the sequence numbers associated with each byte of byte stream 126 enables a packet not containing a synchronization marker to be joined with another packet containing a synchronization marker. Accordingly, the packet without a synchronization marker is placed into memory along with the packet containing the synchronization marker. While byte stream 126 illustrates one PDU 128, it should be appreciated that byte stream 126 can contain any number of PDUs.

Figure 5 is a block diagram listing PDU parameters contained by a synchronization marker inserted into a PDU in accordance with one embodiment of the invention. Each synchronization marker includes a region identifier (RID), a buffer offset, a forward distance and a backward distance in accordance with one embodiment of the invention. A RID includes a pointer to the buffer parameters. One skilled in the art will appreciate that the RID will include data as to where a memory region for storing the PDU is located inside the memory of the receiving device. The offset defines the location of the PDU within the area of memory defined for the location of data. The forward distance represents the distance, measured in the number of bytes, from the synchronization marker to the beginning of the PDU. Likewise, the backward distance represents the distance, measured in the number of bytes, from the synchronization marker to the end of the PDU. The manner in which the backward distance and the forward distance are used to place a packet of a PDU directly into memory of a receiving device is explained in more detail in reference to Figure 7.

Figure 6 is a schematic diagram of a PDU of a byte stream that includes synchronization markers for each packet of the PDU in accordance with one embodiment of the invention. Byte stream 126 includes PDU 128 having header 130 and tail 132. PDU 128 has been split into 3 IP packets, 138-1, 138-2 and 138-3. One skilled in the art will appreciate that for TCP packets

a TCP header is added to the beginning of each packet as it is split from the PDU. The TCP header is eventually stripped off prior to the packet being stored at its ultimate destination. Each of the packets, 138-1, 138-2 and 138-3, contain a synchronization marker. Specifically, packet 138-1 includes synchronization marker 136-1, packet 138-2 includes synchronization marker 136-2 and packet 138-3 includes synchronization marker 136-3. Thus, as PDU 128 is transmitted from a transmitting device over a network, packets 138-1, 138-2 and 138-3 individually travel over the network to a receiving device. It should be appreciated that in a distributed network, such as the Internet, the packets can take different paths to the receiving destination. Thus, it is unlikely that the packets will be received by the receiving device in the order they were transmitted from the transmitting device.

Still referring to Figure 6, synchronization markers 136-1 through 136-3 contain PDU parameters for PDU 128. Additionally, each byte of byte stream 126 is assigned a sequence number as represented by scale 134. Thus, the position of each packet 138-1 through 138-3 of PDU 128 relative to header 130 and the tail 132 can be determined through the information provided by synchronization markers 136-1 through 136-3. That is, the memory location in a receiving device of each packet, 138-1 through 138-3, can be individually determined for each packet, thereby allowing each of the packets to be placed directly into the memory of the receiving device without the need for a header having been received.

Figure 7 is a schematic diagram of a packet of a PDU where the synchronization marker included in the packet provides PDU parameters for a receiving device so that the packet can be directly placed into memory in accordance with one embodiment of the invention. PDU packet 138-2 containing synchronization marker 136-2 is transmitted from a transmitting device to a receiving device. As mentioned above, the transmitting device inserts synchronization marker 136-2. Where packet 138-2 is a TCP packet, a TCP header is attached to the beginning of each packet. Thus, packet 138-2 would include a TCP header in this embodiment. Synchronization marker 136-2 includes PDU parameters such as a RID, a buffer offset, forward distance 142 and backward distance 140. Forward distance 142 is the distance from synchronization marker 136-2 to the beginning of PDU 128. The beginning of PDU 128 is represented by point 144 on the byte stream, which is just after header 130. Backward distance 140 is the distance from synchronization marker 136-2 to the end of PDU 128. The end of PDU 128 is a point on the byte stream, just before tail 132. As explained below, when packet 138-2 is received, the receiving device can directly place the packet into memory through the PDU parameters provided by synchronization marker 136-2.

Still referring to Figure 7, the PDU parameters contained in synchronization marker 136-2 provide the receiving device with the necessary information to directly place packet 138-2 into memory. In the example where packet 138-2 is a TCP packet sent over a distributed network such as the Internet, the receiving device can be a personal computer connected to the network through a network interface card (NIC). If the NIC receives packet 138-2 first, the PDU parameters within synchronization marker 136-2 allow the NIC to transfer the packet directly to the memory of the personal computer without holding the packet in the memory of the NIC. That is, forward distance 142 provides the logical position of header 130 and in combination with the offset, the beginning 144 point of PDU 128 can be determined, as each byte has been assigned a sequential number represented by scale 134. Likewise, backward distance 140 provides the logical position of tail 132. It will be apparent to one skilled in the art that the offset provides a logical position for beginning point 144 of PDU 128 in the memory region. Forward distance 142 provides the distance from beginning point 144 of PDU 128 to synchronization marker 136-2. Since each byte of PDU 128 is assigned a sequential number represented by scale 134, which correlates to a length in bytes, the distance from beginning point 144 of PDU 128 to beginning point 146 of packet 138-2 can be determined. This provides a logical position for beginning 146 of packet 138-2 in the memory region. Likewise, end point 148 of packet 138-2 can be determined from the sequential numbering of the bytes to provide a logical position for the end of the packet in the memory region. Thus, the logical position in the memory region of the receiving device is known, enabling the direct placement of packet 138-2 into memory of the receiving device.

The information contained within the synchronization marker of each packet and the assignment of sequential numbers to each byte in the PDU makes it possible to logically recreate the PDU so that the location in the memory region is known without having to receive the PDU header. Thus, it is not necessary to send the PDU to the memory of the NIC to await for the arrival of the header or remaining packets of the transmitted PDU. Therefore, the amount of memory required for the NIC is minimal and the overhead due to holding the packets in the memory of the NIC for reassembly is substantially eliminated. Furthermore, the insertion of the synchronization markers allows for remote direct memory access over TCP.

Figure 8 is a schematic diagram of a packet of a PDU where the synchronization markers are not included in every packet of the PDU in accordance with one embodiment of the invention. PDU 150 of data stream 152 is defined between header 168 and tail 170. PDU 150 consists of packets 154, 156 and 158. Packet 154 includes synchronization marker 160 and

packet 158 includes synchronization marker 162, while packet 156 does not include a synchronization marker. However, each byte of PDU 150 is assigned a sequential sequence number as represented by scale 134. As mentioned above, the sequence numbers correlate to a length in terms of bytes. Thus, a sequence number is known for beginning byte 164 of packet 156 and end byte 166 as well as the respective adjacent bytes of packets 154 and 158.

Still referring to Figure 8, as packets 154, 156 and 158 are transmitted over a network and subsequently received by a receiving device, it is possible that packet 156 will be received first. Since packet 156 does not include a synchronization marker, the PDU parameters are not available. Therefore, packet 156 will have to be stored in memory of the network card of the receiving device. However, if another packet containing a synchronization marker is received, the relative position of packet 156 to the other packet is known through the assignment of sequence numbers to each byte in PDU 150. For example, if packet 158 follows packet 156, then packet 156 can be assembled with packet 158 and placed into memory with packet 158. It should be appreciated that the assigned sequence numbers allow packets without synchronization markers to be combined with other packets of the PDU containing synchronization markers, so that the packets can be placed into memory together. Additionally, if a packet containing the packet header is received the position of the packet without the synchronization marker can likewise be determined. The assembly of the packets occurs at the network interface of the receiving device. Figures 4, 6 and 8 illustrate PDUs containing three packets for exemplary purposes only and are not meant to be limiting as a PDU can contain any number of packets. Additionally, more than one packet without a synchronization marker can exist within the PDU and the packets without a synchronization marker may be logically adjacent to each other.

In order to insert one synchronization marker in each of the packets of a PDU, it is preferable that the number of synchronization markers is greater than the number of packets in the PDU but no more than twice the number of packets in the PDU. For example, if there are 10 packets in the PDU, then in a preferred embodiment there would be between 10 and 20 synchronization markers evenly distributed throughout the PDU. As mentioned above, even if some packets do not contain synchronization markers, the assigned sequence numbers allow for the packet to be combined with a later received packet containing a synchronization marker.

Figure 9 illustrates a schematic of a system configured to transmit packets of data over a network and receive the packets of data directly into memory of the receiving device in

accordance with one embodiment of the invention. Transmitting device 172 transmits data over network 176 through network interface 174. Network interface 174 includes circuitry for identifying a byte stream of data consisting of PDUs and circuitry for defining and inserting synchronization markers in the byte stream at fixed locations. As mentioned above, the synchronization markers include parameters of the PDU and are inserted at equally spaced apart locations within the PDU. It will be apparent to one skilled in the art that network interface 174 is enabled to transmit packet based messages such as TCP packets. The transmitted packets are received by receiving device 182 through NIC 178. A synchronization marker for each packet is located by circuitry within NIC 178. By reading the PDU parameters contained within the synchronization markers, NIC 178 is able to write the transmitted packet directly into memory 184 of receiving device 182. Thus, the copying and holding of the packet in memory 180 of NIC 178 is not necessary for the packets containing synchronization markers.

However, if a packet without a synchronization marker should be received by NIC 178 of Figure 9, it will be copied to memory 180 of NIC 178. As described above, a packet without a synchronization marker is held in memory until a packet with a synchronization marker or a packet having the header of the PDU is received by NIC 178. Where a packet having a synchronization marker is received, the position of the packet without the synchronization marker relative to the packet containing the synchronization marker can be determined. Accordingly, the packet without the synchronization marker is assembled with the packet having the synchronization marker and both are written into memory 184. Likewise, if the packet containing the header information is received, the buffer parameters, offset and length of the PDU are contained by the header. Therefore, the location of the packet without the synchronization marker in memory may be determined from the header information. It should be appreciated that the memory of NIC 178 can remain relatively small since it will be infrequent where a packet does not contain a synchronization marker. As mentioned above, by setting the number of the synchronization markers for each PDU greater than the number of the packets for the PDU, the occurrence of a packet without a synchronization marker is minimized.

Figure 10 is a flowchart diagram of the method operations performed in processing a byte stream to be transmitted from a storage device to a requesting host over a network in accordance with one embodiment of the invention. The method initiates with operation 190 where a PDU from the byte stream is identified. The PDU has a header and a tail as described in reference to Figures 3, 4 and 6. The method then advances to operation 192 where

synchronization markers containing parameters of the PDU are defined. Here, the parameters of the PDU include the RID, the buffer offset, the forward distance and the backward distance as described with respect to Figure 5. The method then proceeds to operation 194 where one of the synchronization markers is inserted at periodic locations of the byte stream. The periodic locations are preferably equally spaced throughout the byte stream. In one embodiment, the number of periodic locations per PDU is greater than the number packets of the PDU and no more than twice the number of packets of the PDU. The method then advances to operation 196 where the PDU is transmitted in at least two packets and at least one of the two packets contains one of the inserted synchronization markers.

Figure 11 is a flowchart diagram of the method operations for receiving a byte stream at a host from a networked transmitting device in accordance with one embodiment of the invention. The method initiates with operation 198 where a packet of a PDU is received by a network interface of a receiving device. In one embodiment, the network interface is a network interface card (NIC). The method then advances to operation 200 where a synchronization marker contained in the packet is located. It will be apparent to one skilled in the art that the synchronization markers can be detected by observing the periodic sequence of the synchronization markers and looking for sequence numbers associated with the periodic sequence. The method then moves to operation 202 where the PDU parameters contained within the synchronization marker are read to determine the position of the received packet within the PDU. As mentioned above, the PDU parameters of RID, buffer offset, forward distance and backward distance allow for the logical recreation of the PDU in memory so that the position of each packet containing a synchronization marker in memory can be determined as described in reference to Figure 7. The method then proceeds to operation 204 where the packet is written directly into memory of the host. Since the position of the packet can be determined from the PDU parameters of the synchronization marker as described above, there is no need to copy the packet into local memory of the network interface. Thus, overhead is significantly reduces as compared to reassembling the packets of the PDU in local memory of the network interface and then writing the reassembled PDU to memory of the host.

Figure 12 is a flowchart diagram of the method operations for receiving a byte stream for a host where each packet of the byte stream does not contain a synchronization marker in accordance with one embodiment of the invention. The method initiates with operation 206 where a packet of a PDU is received. In one embodiment, a network interface of a receiving device receives the packet from a distributed network such as the Internet. One skilled in the

art will appreciate that the PDU can be transmitted through a networking protocol such as TCP/IP. The method then advances to decision operation 208 where it is determined if the packet contains a synchronization marker. Detection methods well known in the art may be executed here. If a synchronization marker is present, then the method proceeds to operation 216 where the PDU parameters contained in the synchronization markers are read. The PDU parameters include the BPI which points to the region identifier (RID) and the offset. The RID and the offset provide a location within the memory region where the beginning of the PDU will be stored. Also included in the PDU parameters is a forward distance indicating the distance in bytes from the synchronization marker to the beginning of the PDU and a backward distance indicating the distance in bytes from the synchronization marker to the end of the PDU. The method then moves to operation 218 where a position of the received packet within the PDU is determined. For example, the position of the packet within the PDU may be determined as described with respect to Figure 7. The method then advances to operation 220 where the packet is directly written into the memory of the host. It should be appreciated that the memory of the host system is accessed by a remote system, i.e., transmitting device, without interrupting the CPU, thereby improving performance.

If a synchronization marker is not present in operation 208 of Figure 12, then the method proceeds to decision operation 209 where it is determined if a packet containing a synchronization marker has been received or if the packet containing the header has been received. If a packet containing a synchronization marker has been received or if a packet containing the PDU header has been received, then the method proceeds to operation 218 where a position of the received packet in the PDU is determined. As mentioned above, the position of the packet without the synchronization marker relative to the packet containing the synchronization marker can be determined through the sequence numbers associated with each byte of the byte stream. Likewise, if the packet containing the header information is received, the buffer parameters, offset and length of the PDU are contained by the header. The method will then proceed to operation 220 the packet without the synchronization marker is written into memory.

If a packet containing a synchronization marker has not been received and if a packet containing the PDU header has not been received, then the method proceeds to operation 210 where the packet is held in the local memory of the network card. That is, the packet is copied to the memory of the network card if a synchronization marker is not found by the detection method and both another packet containing a synchronization marker and a packet containing

the header information has not been received. It should be appreciated that since a synchronization marker or the header has not been received, there is no reference information available to place the packet into the memory of the host. The method then advances to operation 212 where the packet is reassembled with another received packet that contains a synchronization marker. As mentioned above, the relative positions of the packets following the packet without a synchronization marker are known through the associated sequence numbers. Thus, the distance in bytes between a packet with a synchronization marker and a packet without a synchronization marker can be determined. Accordingly, the packet without the synchronization marker can be assembled with the packet having a synchronization marker. It should be appreciated that the packet containing the synchronization marker and the packet containing the synchronization marker do not have to be logically adjacent to each other. Additionally, the packet not containing the synchronization marker can be assembled with a packet containing the header information as the necessary data to position the packet not containing the synchronization is contained within the PDU header. One skilled in the art will appreciate that the packet containing the header information need not contain a synchronization marker here. The method then moves to operation 214 where the assembled packet of operation 212 is written into memory of the receiving device. It should be appreciated that the packet having the synchronization marker or the packet containing the header information is not copied into the local memory of the network card.

As mentioned above, where a packet containing a PDU header arrives first to the receiving device, the data contained within the header i.e., the buffer parameters, a buffer offset and a length of the PDU, can be used to write each packet directly into the memory of the receiving host. As the beginning location in memory is known from the data in the header as well as a length of the PDU, a position for each subsequent packet received can be projected from the data contained on the header of each of the packets. Thus, when the packet containing the header arrives first, the data within the header combined with the data included with each subsequent packet allows for the direct placement of the packets within memory of the receiving device.

In summary, the embodiments of the present invention allow for RDMA with existing TCP based networking protocols. Thus, existing protocols, such as TCP/IP, that are universally accepted and in wide use, are capable of being adapted for placing packets of data directly into memory. The invention has been described herein in terms of several exemplary embodiments. Other embodiments of the invention will be apparent to those skilled in the art from

consideration of the specification and practice of the invention. The embodiments and preferred features described above should be considered exemplary, with the invention being defined by the appended claims.

With the above embodiments in mind, it should be understood that the invention may employ various computer-implemented operations involving data stored in computer systems. These operations are those requiring physical manipulation of physical quantities. Usually, though not necessarily, these quantities take the form of electrical or magnetic signals capable of being stored, transferred, combined, compared, and otherwise manipulated. Further, the manipulations performed are often referred to in terms, such as producing, identifying, determining, or comparing.

Any of the operations described herein that form part of the invention are useful machine operations. The invention also relates to a device or an apparatus for performing these operations. The apparatus may be specially constructed for the required purposes, or it may be a general purpose computer selectively activated or configured by a computer program stored in the computer. In particular, various general purpose machines may be used with computer programs written in accordance with the teachings herein, or it may be more convenient to construct a more specialized apparatus to perform the required operations.

The invention can also be embodied as computer readable code on a computer readable medium. The computer readable medium is any data storage device that can store data which can be thereafter be read by a computer system. Examples of the computer readable medium include hard drives, network attached storage (NAS), read-only memory, random-access memory, CD-ROMs, CD-Rs, CD-RWs, magnetic tapes, and other optical and non-optical data storage devices. The computer readable medium can also be distributed over a network coupled computer systems so that the computer readable code is stored and executed in a distributed fashion.

Although the foregoing invention has been described in some detail for purposes of clarity of understanding, it will be apparent that certain changes and modifications may be practiced within the scope of the appended claims. Accordingly, the present embodiments are to be considered as illustrative and not restrictive, and the invention is not to be limited to the details given herein, but may be modified within the scope and equivalents of the appended claims.

What is claimed is:

Claims

1. A method for processing a byte stream to be transmitted from a storage device to a requesting host over a network, comprising:

identifying a protocol data unit (PDU) from the byte stream, the PDU having a header and a tail;

defining synchronization markers containing parameters of the PDU;

inserting one of the synchronization markers at fixed locations of the PDU, the fixed locations being equally spaced apart; and

transmitting the PDU in at least two packets, and at least one of the two packets containing one of the inserted synchronization markers.

2. The method of claim 1, wherein the synchronization markers are used to locate a relative position of each of the at least two packets to the header and the tail of the PDU.

3. The method of claim 1, wherein each of the at least two packets is associated with Internet small computer system interface (iSCSI) protocol.

4. The method of claim 1, wherein the parameters of the PDU include a region identifier (RID), a buffer offset, a forward distance and a backward distance.

5. The method of claim 1, wherein each byte of the byte stream is assigned a sequence number.

6. The method of claim 1, wherein a number of synchronization markers is equal to or greater than a number of packets in the PDU.

7. The method of claim 1, further comprising:

receiving one of the at least two packets at the requesting host;

locating a synchronization marker contained within one of the at least two packets;

reading PDU parameters contained in the synchronization marker to determine a position of the received packet within the PDU; and

writing one of the at least two packets directly into memory of the requesting host.

8. The method of claim 7, wherein the one of the at least two packets is associated with Internet small computer system interface (iSCSI) protocol.

9. The method of claim 7, wherein the PDU parameters include a region identifier (RID), an offset, a forward distance and a backward distance.

10. The method of claim 7, wherein each byte of the byte stream is assigned a sequence number, the sequence number correlating to a length of the byte stream.

11. The method of claim 7, wherein the PDU includes a header and a tail, the header containing buffer parameters, a buffer offset, and a length of the PDU, the tail containing the RID, the buffer offset and a forward distance.

12. A device for processing a byte stream to be transmitted over a network, the device comprising:

a network interface, the network interface including;

circuitry for identifying a protocol data unit (PDU) from the byte stream;

circuitry for defining synchronization markers containing parameters of the PDU;

circuitry for inserting one of the synchronization markers at fixed locations of the PDU, the fixed locations being equally spaced apart; and

circuitry for transmitting the PDU in at least two packets, wherein at least one of the two packets contains one of the inserted synchronization markers.

13. The device of claim 12, wherein the network interface further includes:

circuitry for assigning a sequence number to each byte of the byte stream.

14. The device of claim 12, wherein the parameters of the PDU include a region identifier (RID), an offset, a forward distance and a backward distance.

15. The device of claim 12, wherein the at least two packets are associated with Internet small computer system interface (iSCSI) protocol.

16. The device of claim 12, wherein the number of fixed locations where the synchronization markers are inserted is greater than the number of packets of the PDU.

17. A computer system configured to receive a byte stream from a networked transmitting device, the computer system comprising:

a central processing unit (CPU);

memory configured to receive the byte stream; and

a network card in communication with the memory and a network, the network card including;

circuitry for receiving a packet of a protocol data unit (PDU) of the byte stream;

circuitry for locating a synchronization marker contained within the packet;

circuitry for reading PDU parameters contained in the synchronization marker to determine a position of the received packet within the PDU; and

circuitry for writing the packet directly into memory.

18. The system of claim 17, wherein the network card further includes:

circuitry for copying the packet to a local memory of the network card if a synchronization marker is not located; and

circuitry for combining the packet copied to local memory with a later received packet containing a synchronization marker.

19. The system of claim 18, wherein the later received packet is an adjacent packet to the packet copied to local memory.

20. The system of claim 17, wherein writing the packet directly to memory avoids interrupting the CPU.

21. The system of claim 17, wherein the network card further includes:

circuitry for reading an assigned sequence number associated with each byte of the byte stream.

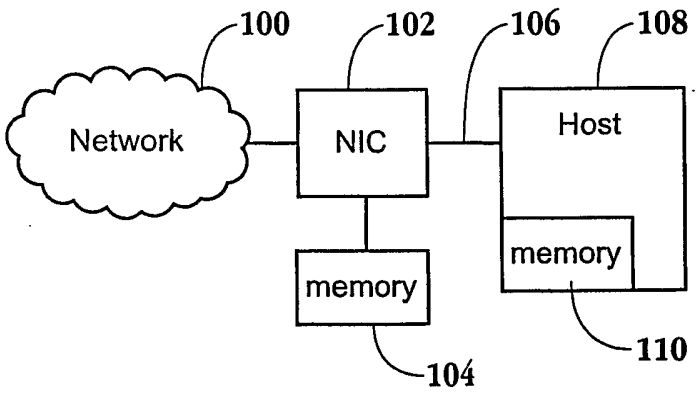


Fig. 1 (prior art)

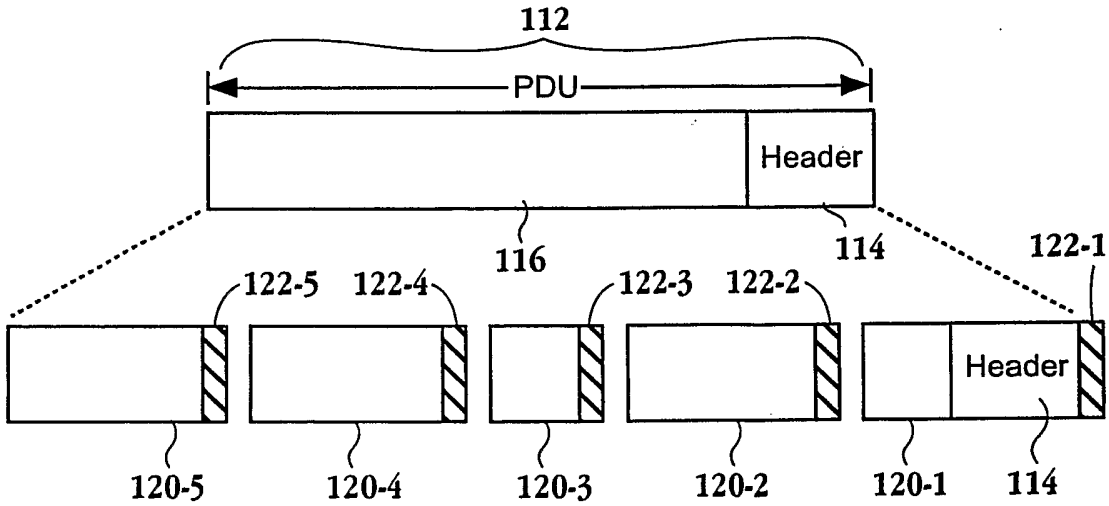


Fig. 2 (prior art)

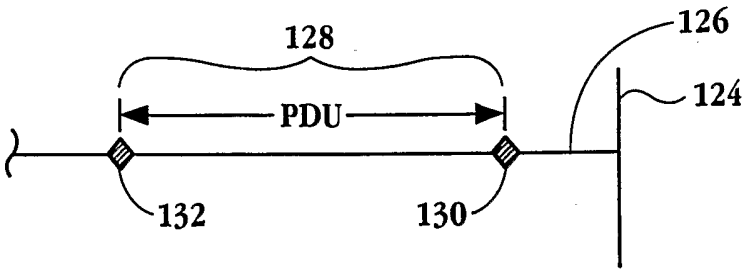

128
126
PDU
124
132
130

Fig. 3

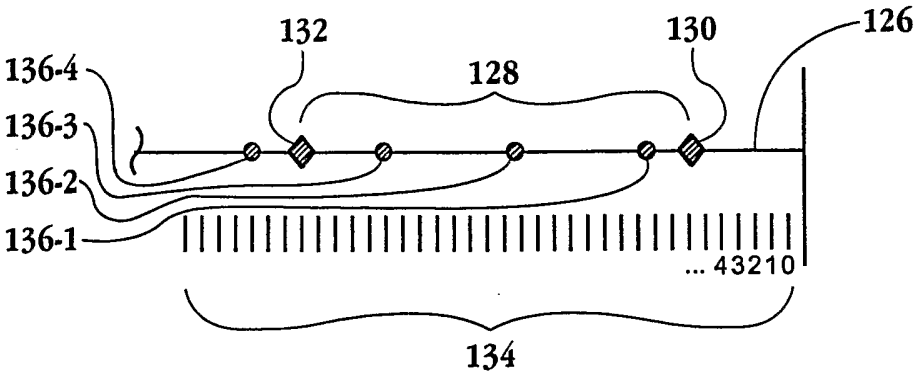

132
130
126
136-4
128
136-3
136-2
136-1
... 43210
134

Fig. 4

Synchronization Marker
Region Identifier (RID)
Buffer Offset
Forward Distance
Backward Distance

Fig. 5

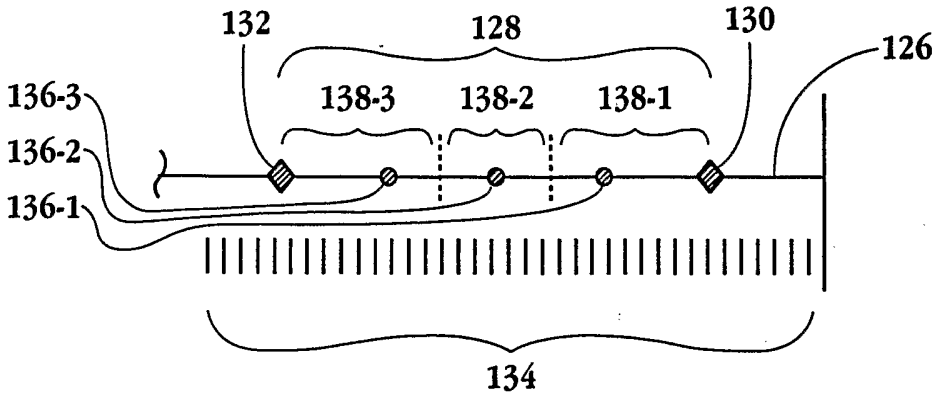


Fig. 6

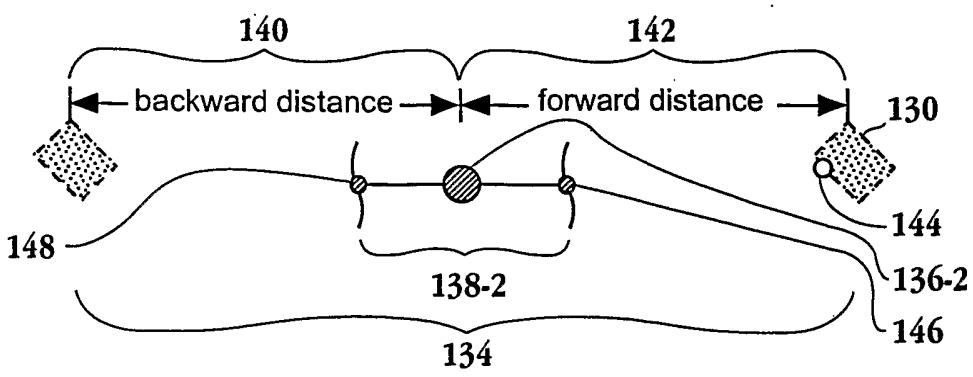


Fig. 7

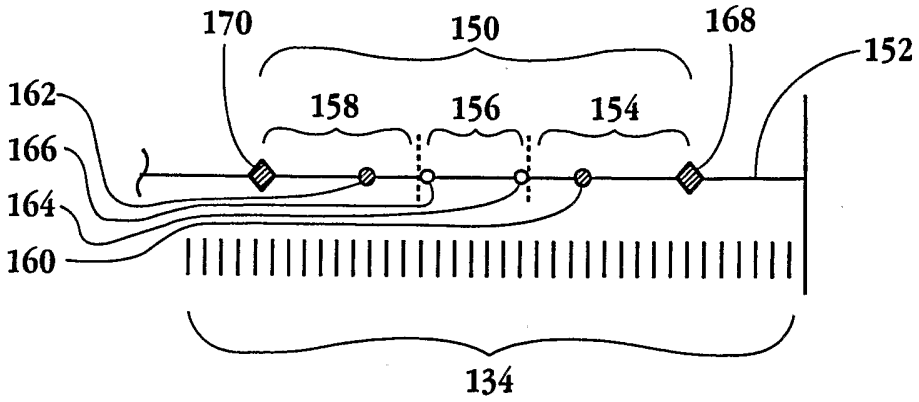

170
150
168
152
162
158
156
154
166
164
160
134

Fig. 8

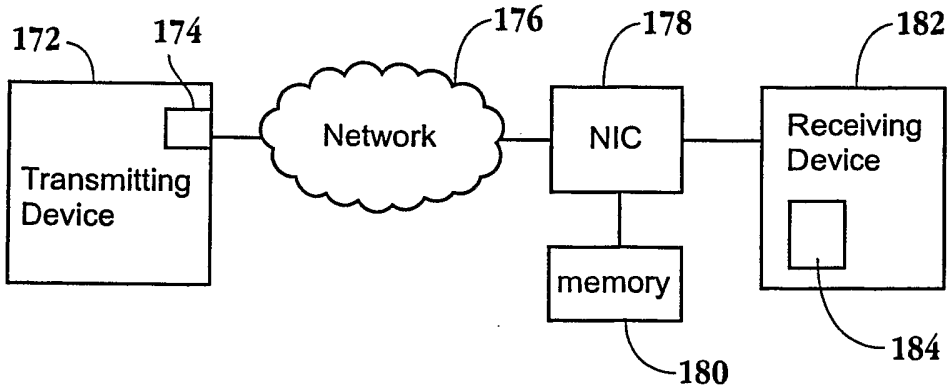

172
174
176
178
182
Network
NIC
Receiving Device
Transmitting Device
memory
184
180

Fig. 9

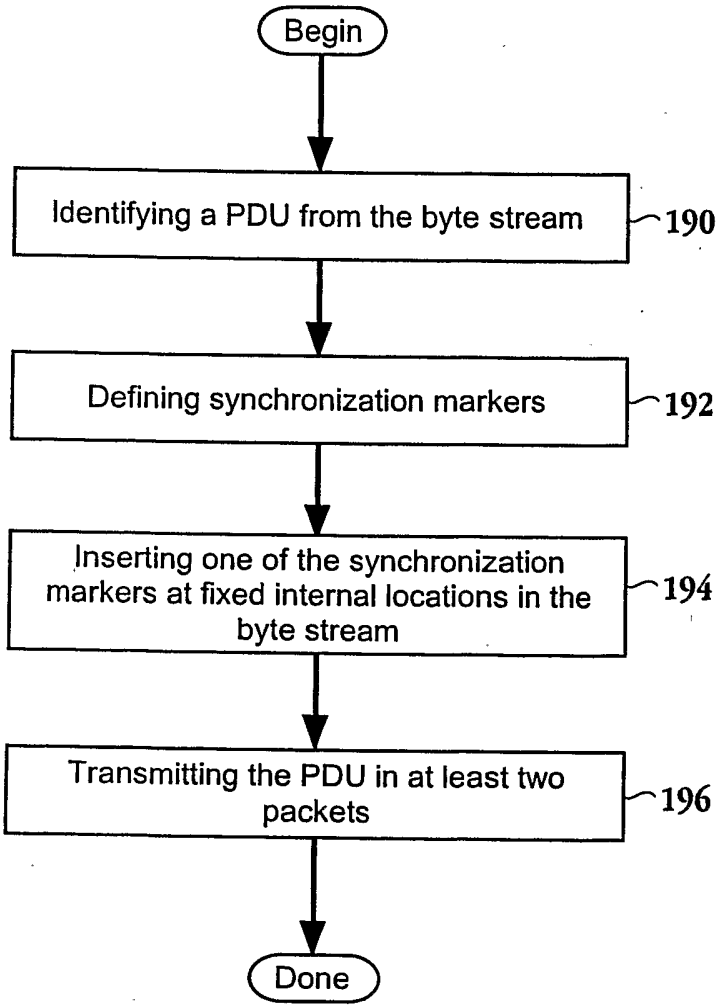

Begin
Identifying a PDU from the byte stream
190
Defining synchronization markers
192
Inserting one of the synchronization markers at fixed internal locations in the byte stream
194
Transmitting the PDU in at least two packets
196
Done

Fig. 10

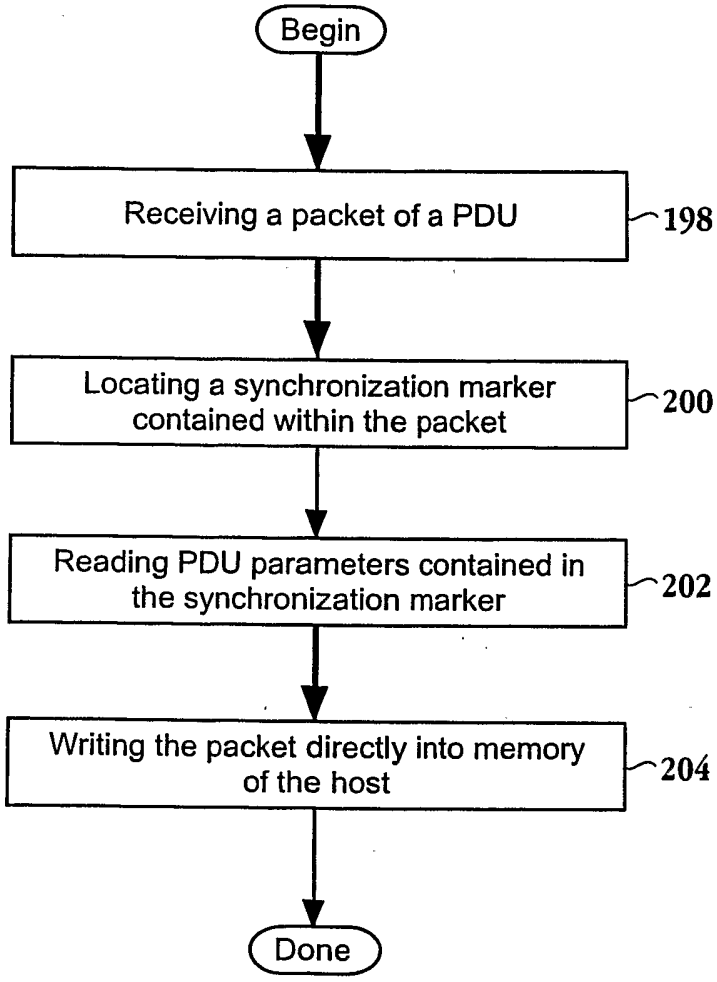

Begin
Receiving a packet of a PDU
198
Locating a synchronization marker contained within the packet
200
Reading PDU parameters contained in the synchronization marker
202
Writing the packet directly into memory of the host
204
Done

Fig. 11

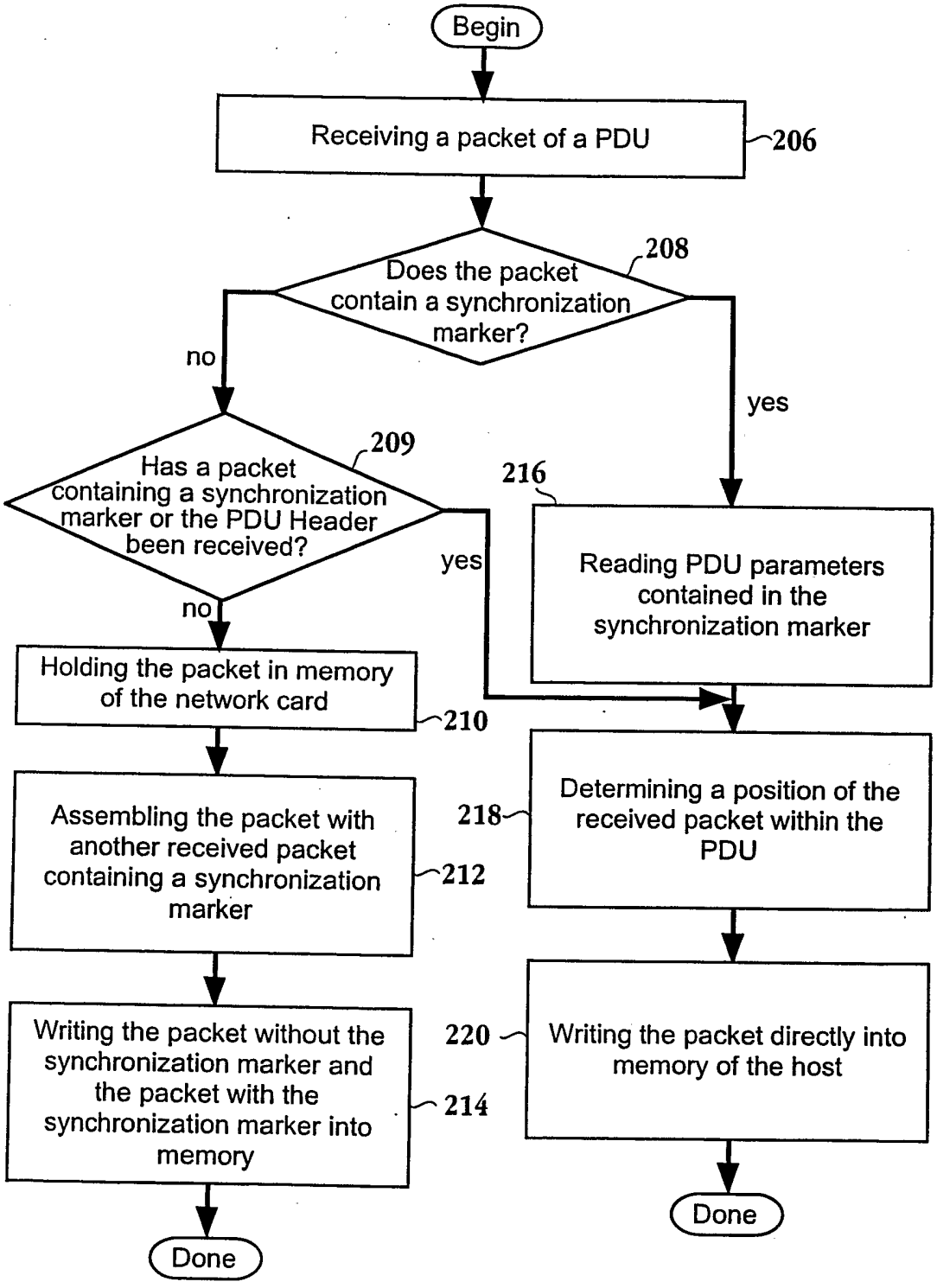

Begin
Receiving a packet of a PDU
206
Does the packet contain a synchronization marker?
208
no
yes
Has a packet containing a synchronization marker or the PDU Header been received?
209
yes
no
216
Reading PDU parameters contained in the synchronization marker
Holding the packet in memory of the network card
210
Assembling the packet with another received packet containing a synchronization marker
212
218
Determining a position of the received packet within the PDU
Writing the packet without the synchronization marker and the packet with the synchronization marker into memory
214
220
Writing the packet directly into memory of the host
Done
Done

Fig. 12

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US02/39784

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) :H04L 12/28, 12/56

US CL :Please See Extra Sheet.

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. : 370/235, 236.2, 389, 394, 400, 419, 464, 474; 709/201, 203, 213, 217, 218, 219

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

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y,P	US 6,493,343 B1 (GARCIA ET AL) 10 DECEMBER 2002, see the patent entirety.	1-3, 5-8, 10-13. 15-21
Y	ROMANOW ET AL, The Case for RDMA, INTERNET DRAFT, page 1-12, December 2000.	1-3, 5-8, 10-13, 15-21
A	US 6,223,270 B1 (CHESSON ET AL) 24 APRIL 2001, see the patent entirety.	1-21

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

*	Special categories of cited documents:	"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
"A"	document defining the general state of the art which is not considered to be of particular relevance	"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
"E"	earlier document published on or after the international filing date	"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
"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)	"&"	document member of the same patent family
"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		

Date of the actual completion of the international search	Date of mailing of the international search report
23 FEBRUARY 2003	27 MAR 2003
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer DUONG, FRANK Telephone No. (703) 306-5428

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

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US02/39784

A. CLASSIFICATION OF SUBJECT MATTER:
US CL :

370/235, 236.2, 389, 394, 400, 419, 464, 474; 709/201, 208, 213, 217, 218, 219