

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
30 July 2009 (30.07.2009)

PCT

(10) International Publication Number
WO 2009/094339 A2

(51) International Patent Classification:
G06F 13/42 (2006.01) **H04L 12/56** (2006.01)
G06F 13/14 (2006.01)

(21) International Application Number:
PCT/US2009/031461

(22) International Filing Date: 20 January 2009 (20.01.2009)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
12/009,970 23 January 2008 (23.01.2008) US

(71) Applicant (for all designated States except US): **INTEL CORPORATION** [US/US]; 2200 Mission College Boulevard, Santa Clara, California 95052 (US).

(72) Inventor; and

(75) Inventor/Applicant (for US only): **WAGH, Mahesh** [IN/US]; 13527 NW Hogan St., Portland, Oregon 97229 (US).

(74) Agent: **TROP, Timothy N.**; Trop, Pruner & Hu, P.C., 1616 S. Voss Rd., Ste. 750, Houston, Texas 77057-2631 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— without international search report and to be republished upon receipt of that report

(54) Title: ENABLING FLEXIBILITY OF PACKET LENGTH IN A COMMUNICATION PROTOCOL

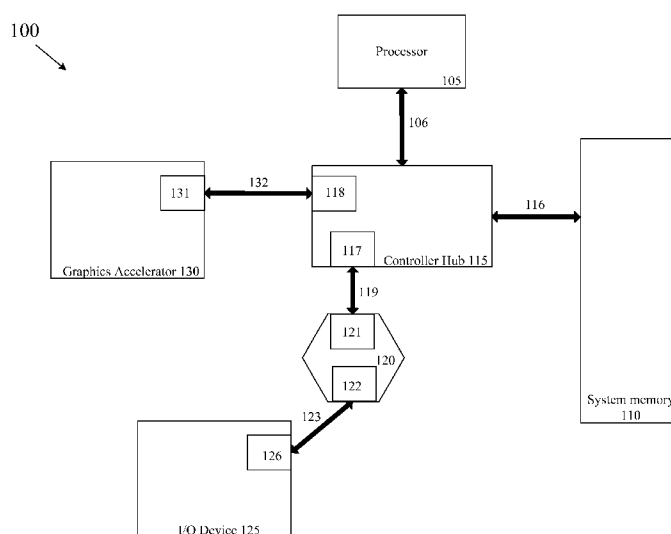


FIG. 1

(57) Abstract: In one embodiment, the present invention includes a method for transmitting a packet from a transmitter to a receiver along an interconnect and terminating transmission of the packet at a packet disconnect boundary, which has a length less than a payload length of the packet. After such termination, another packet such as a higher priority packet can be transmitted, or a remainder of the original packet may be transmitted as a separate packet. Other embodiments are described and claimed.

WO 2009/094339 A2

ENABLING FLEXIBILITY OF PACKET LENGTH IN A COMMUNICATION PROTOCOL

Background

[0001] Computer systems include a number of components and elements, which are typically coupled via a bus or interconnect. Previously, input/output (IO) devices were coupled together through a conventional multi-drop parallel bus architecture referred to as Peripheral Component Interconnect (PCI). More recently, a new generation of an IO bus referred to as PCI-Express (PCIe) has been used to facilitate faster interconnection between devices having a serial physical layer communication protocol.

[0002] A PCIe architecture includes a layered protocol to communicate between devices. As an example, a physical layer, link layer, and transaction layer form a PCIe protocol stack. The PCIe link is built around dedicated unidirectional pairs of serial point-to-point connections referred to as a lane. A link between devices includes some number of lanes, such as one, two, sixteen, thirty-two, and so-on. The current PCIe specification, PCI ExpressTM Specification Base Specification version 2.0 (published January 17, 2007), is available at <http://www.pcisig.com/specifications/pciexpress/>.

[0003] A PCIe protocol utilizes credit-based flow control mechanisms to transfer packets across a link. The receiver advertises credits equal to the amount of storage buffers available at the receiver. The transmitter is not permitted to issue transactions that can consume more credits than what the receiver has advertised. The payload length supplied in the transmitter request header is required to match exactly to the amount of data payload supplied and be less than or equal to credits available at the receiver in order to service the transaction. This can unnecessarily limit flexibility in data communication.

Brief Description of the Drawings

[0004] FIG. 1 is a block diagram of a system in accordance with one embodiment of the present invention.

[0005] FIG. 2A is an example of transmission of data packets in accordance with one embodiment of the present invention.

[0006] FIG. 2B is an example of three packets that are part of a data stream in accordance with one embodiment of the present invention.

[0007] FIG. 3 is a flow diagram of a method in accordance with one embodiment of the present invention.

Detailed Description

[0008] Referring to FIG. 1, a block diagram of a system in accordance with one embodiment of the present invention including devices coupled to a controller hub via a serial link is illustrated. System 100 includes a processor 105 and a system memory 110 coupled to a controller hub 115. Processor 105 includes any processing element, such as a microprocessor, a host processor, an embedded processor, a co-processor, or other processor. Processor 105 is coupled to controller hub 115 through a front-side bus (FSB) 106. In one embodiment, FSB 106 is a serial point-to-point (PtP) interconnect as described below.

[0009] System memory 110 includes any memory device, such as random access memory (RAM), non-volatile (NV) memory, or other memory accessible by devices in system 100. System memory 110 is coupled to controller hub 115 through a memory interface 116.

[0010] In one embodiment, controller hub 115 is a root hub or root controller in a PCIe interconnection hierarchy. Examples of controller hub 115 include a chipset, a memory controller hub (MCH), a northbridge, an input/output controller hub (ICH) a southbridge, and a root controller/hub. Here, controller hub 115 is coupled to a switch/bridge 120 through a serial link 119. Input/output modules 117 and 121, which may also be referred to as interfaces/ports 117 and 121, include/implement a layered protocol stack to provide communication between controller hub 115 and switch 120. In one embodiment, multiple devices are capable of being coupled to switch 120. Embodiments may also apply where a processor and chipset are integrated into the same package. Further, this technique does not require a new revision of PCIe protocol, and can be implemented by PCIe 2.0, 1.1 compatible devices, among others. Thus embodiments can be used anywhere a PCIe interconnect is used for communication.

[0011] Switch 120 routes packets/messages from a device 125 upstream, i.e., up a hierarchy towards controller hub 115 and downstream, i.e., down a hierarchy away from controller hub 115 to device 125. IO modules 122 and 126 implement a layered protocol stack to communicate between switch 120 and device 125. Device 125 includes any internal or external device or component to be coupled to an electronic system, such as an IO device, a network interface controller (NIC), an add-in card, an audio processor, a network processor, a hard-drive, a storage device, a monitor, a printer, a mouse, a keyboard, a router, a portable storage device, a Firewire device, a Universal Serial Bus (USB) device, a scanner, and other input/output devices.

[0012] A graphics accelerator 130 is also coupled to controller hub 115 through serial link 132. In one embodiment, graphics accelerator 130 is coupled to an MCH, which is coupled to an ICH. Switch 120, and accordingly IO device 125, is then coupled to the ICH. IO modules 131 and 118 are also to implement a layered protocol stack to communicate between graphics accelerator 130 and controller hub 115.

[0013] In one embodiment, a PCIe protocol stack may include a transaction layer, a link layer, and a physical layer. In one embodiment, the transaction layer is to provide an interface between a device's processing core and the interconnect architecture, such as a data link layer and a physical layer. In this regard, a primary responsibility of the transaction layer is the assembly and disassembly of packets (i.e., transaction layer packets, or TLPs). PCIe implements split transactions, i.e., transactions with request and response separated by time, allowing a link to carry other traffic while the target device gathers data for the response.

[0014] In addition PCIe utilizes credit-based flow control. In this scheme, a device advertises an initial amount of credit for each of the receive buffers in the transaction layer. An external device at the opposite end of the link, such as controller hub 115 in FIG. 1, counts the number of credits consumed by each TLP. A transaction may be transmitted if the transaction does not exceed a credit limit. Upon receiving a response an amount of credit is restored. An advantage of a credit scheme is that the latency of credit return does not affect performance, provided that the credit limit is not encountered.

[0015] The link layer acts as an intermediate stage between the transaction layer and the physical layer. In one embodiment, a responsibility of the data link layer is to provide a reliable mechanism for exchanging TLPs between two components by a link. One side of the link layer accepts TLPs assembled by the transaction layer, applies identifiers, calculates and applies an error detection code, i.e., cyclic recovery codes (CRC), and submits the modified TLPs to the physical layer for transmission across a physical link to an external device.

[0016] In one embodiment, the physical layer physically transmits a packet to an external device. The physical layer includes a transmit section to prepare outgoing information for transmission and a receiver section to identify and prepare received information before passing it to the link layer. The physical layer includes a transmitter and a receiver. The transmitter is supplied with symbols, which the transmitter serializes and transmits onto to an external device. The receiver is supplied with serialized symbols from an external device and transforms the received signals into a bit-stream. The bit-stream is de-serialized and supplied to a logical sub-block. In one embodiment, an 8b/10b transmission code is employed, where

ten-bit symbols are transmitted/received. In addition, in one example, the receiver also provides a symbol clock recovered from the incoming serial stream. As stated above, although the transaction layer, link layer, and physical layer are discussed in reference to a specific embodiment of a PCIe protocol stack, a layered protocol stack is not so limited. In fact, any layered protocol may be included/implemented.

[0017] In accordance with the PCIe protocol, the following rules apply to a transmitter that intends to initiate a transaction with data: (1) a transmitter must issue requests with a data length equal or less than a configured maximum payload size; (2) a transmitter must guarantee that the length of the request is less than or equal to the credits available at the receiver to service the request; (3) a transmitter must guarantee that once a packet transmission begins, it can provide/stream data as indicated in the request length field, in other words the transmitter cannot stall a packet transfer, thus the transmitter must buffer data or guarantee that it can stream data at the packet rate; and (4) any high priority packet must wait for the current packet with data to finish before it can be issued. However, some implementations may terminate a packet with an 'End Bad' termination symbol, in which case any data that was transmitted is dropped by the receiver.

[0018] To improve flexibility in this protocol, embodiments may define an allowable packet disconnect boundary. In this way, the transmitter is permitted to terminate a request at the allowable packet disconnect boundary. While the scope of the present invention is not limited in this regard, the allowable packet disconnect boundary may be 64 bytes (B) or multiple of 64B or other configurable quantity. Some embodiments may also include flow control and length protocol enhancements to enable such variable length transmissions. The transmitter can terminate a packet as per PCIe packet termination rules, with the CRC and normal packet end symbol inserted to terminate the packet.

[0019] In some embodiments, length rules can be enhanced such that the length field indicates the payload size the transmitter intends to transfer, however, the packet may be terminated at an allowable packet disconnect boundary. For example, a transmitter may initiate a packet transfer with a length indicating a transfer size of 256B but terminate the packet after transmitting the first 128B of payload data. If the transmission is successful, i.e., does not incur any transmission errors, then the receiver consumes the first 128B of packet that was successfully transferred, rather than dropping the data as per an "End Bad" process. The transmitter may transfer the remaining payload as a separate newly issued transaction; the receiver is not required to keep a history of the previously received transaction. The

transmitter may initiate a new transaction to complete the original intended payload with a length equal to 128B and provide the remaining payload, or the transmitter may start a new packet with a different length starting from the previously disconnected packet boundary, or may initiate a new high priority packet.

[0020] In some embodiments, the flow control rules may be enhanced such that the transmitter is permitted to initiate a transaction as long as it has credits up to the allowable disconnect boundary and it can guarantee data to be streamed. If the receiver releases credits up to the next allowable disconnect boundary, the transmitter is permitted to continue its transmission beyond the current allowable disconnect boundary up to the next allowable disconnect boundary, provided sufficient credits are available and data can be provided without stalls.

[0021] Thus embodiments provide protocol enhancements to permit a transmitter to specify a payload length that may exceed the actual amount of data supplied in the payload. The amount of data supplied in the payload has to guarantee that it does not exceed the amount of credits allocated by the receiver. The packet can be terminated/suspended at a specified disconnect boundary. With this scheme a large packet, for example a 4 kilobyte (KB) packet (length = 4KB), is permitted to be terminated at a disconnect boundary in order to interleave a high priority packet.

[0022] Embodiments may further enhance this protocol enhancement to: (1) optimize receiver buffers, streaming data from memory; (2) improve link utilization based on credit returns; and (3) optimize transmitter store and forward mechanisms, thus enabling streaming data over a PCIe interconnect. Accordingly, a transmitter can initiate the transfer of a large packet once it has stored data equal to the disconnect boundary or can guarantee data to be streamed without stalls at least up to the disconnect boundary and has sufficient receiver credits allocated. Upon approaching the disconnect boundary, the transmitter may continue with data payload if it has data stored up to the next disconnect boundary, or it can guarantee data to be streamed without stalls until the next disconnect boundary and has enough header credits allocated. If data stored is less than the disconnect boundary or data streaming up to the disconnect boundary cannot be guaranteed or lack of receiver credits, the packet is terminated by transmitter.

[0023] Embodiments thus enable implementations to optimize receiver buffers, improve link utilization and optimize transmitter buffers. In contrast, conventional PCIe mechanisms are more rigid and require the requestor to delay the transmission of requests until a payload

worth of credits is received and does not permit 'on-the-fly' optimization. Using embodiments of the present invention, data from a memory controller can be streamed over a PCIe interconnect. Completions (e.g., memory reads issued by an IO device) can indicate the completion payload equal to the maximum payload size and optimize for completion payloads. This enables completions to be returned at least equal to 64B and up to the maximum payload size, thus improving the link utilization for read completions.

[0024] Referring now to FIG. 2A, shown is an example of transmission of data packets 200 in accordance with one embodiment of the present invention. As shown in FIG. 2A, a transmitter may transmit packet 205, originally intended to have a payload length of 256 bytes, as indicated in the header portion of packet 205, with an early termination. More specifically, packet 205 is terminated after a first allowable disconnect boundary of 64 bytes is reached. Thus as shown in FIG. 2A, packet 205 includes a start portion, a header length portion, a payload portion (which corresponds to a length of the allowable disconnect boundary) followed by an end-to-end CRC (ECRC), which in turn is followed by a termination sequence, which includes a link CRC portion and an end portion. Note that two CRC protocols may be used, a link CRC, which is used for all packets, and an optional ECRC that is used to protect packets end-end. The distinction is as follows: the link CRC is applied by the link to protect the packet transmission from transmitter to receiver across the link, and the optional ECRC (if supported) is applied by the transaction layer to protect the packet, and is only inspected by the final destination, which could be multiple hops away. The inspection is also carried out by the final destination's transaction layer. After transmission of this terminated packet 205, another packet 210 is transmitted to include the remaining data portion of the terminated packet 205. More specifically, packet 210 is identified as having a payload length of 192 bytes (i.e., corresponding to the remainder of the packet length of terminated packet 205). Thus as shown in FIG. 2A, packet 210 transmits 192B of payload, followed by an ECRC portion and an end portion.

[0025] By providing a transmitter with the ability to terminate a packet, embodiments may enable interleaving of high priority packets after a terminated packet, which can then be followed by the remainder of the disconnected packet. As shown in FIG. 2B, three packets are part of a data stream 250, namely terminated packet 255, high priority packet 260, and remainder packet 265. After termination of terminated packet 255, high priority packet 260, in the embodiment shown in FIG. 2B, includes only a start portion, header portion and ECRC and end portion (i.e., a high priority packet without data). After transmission of this high

priority packet, the transmitter may then send remainder packet 265, which provides the remainder portion of the disconnected packet. Thus as shown in FIG. 2B, remainder portion 265 transmits the remaining 192B of the data payload. While shown with these particular examples in the embodiments of FIGS. 2A and 2B, the scope of the present invention is not limited in this regard.

[0026] Referring now to FIG. 3, shown is a flow diagram of a method in accordance with one embodiment of the present invention. As shown in FIG. 3, method 300 may be used to transmit a packet from a transmitter to a receiver. Specifically, as shown in FIG. 3, method 300 may begin by transmitting a first packet from the transmitter to the receiver (block 310). Such a packet may include a header to indicate a packet length which may be a multiple number of data portions, each corresponding to a length of an allowable packet disconnect boundary, in one embodiment. For example, the header length may be for a relatively long payload, e.g., of 4 KB, where the communication protocol of the interconnect that couples the transmitter and receiver may have an allowable packet disconnect boundary of 64 bytes.

[0027] During transmission of the first packet, the transmitter, and more specifically a transaction layer of the transmitter or an internal fabric of the device may choose to terminate the transmission, for example to enable interleaving of a higher priority packet. Thus as shown in FIG. 3 at block 320 the first packet may be terminated at an allowable packet disconnect boundary. In terminating the first packet, the transmitter may send an error coding portion, e.g., an ECRC, as well as an end signal to indicate termination of the first packet. Then, the transmitter may transmit a second packet from transmitter to receiver (block 330). In one embodiment, the second packet may correspond to a higher priority packet which may or may not include payload data.

[0028] Referring still to FIG. 3, after the second packet is transmitted, a third packet may be transmitted from transmitter to receiver. In various embodiments, this third packet may include a remainder payload of the first packet (block 340). For example, assume that the first packet was terminated at an allowable packet disconnect boundary after transmission of a 2 KB portion of the payload of the first packet. In this case, the third packet may include a header portion indicating a length of 2K, and the corresponding data payload may provide the remaining 2K of the payload data originally intended for the first packet. While shown with this particular implementation in the embodiment of FIG. 3, the scope of the present invention is not limited in this regard.

[0029] Embodiments may reduce the silicon foot print (gate count) for chipset and endpoint designs, and improves the link bandwidth utilization, thus providing additional data throughput that improves maximum achievable platform bandwidth and application performance.

[0030] Embodiments may be implemented in code and may be stored on a storage medium having stored thereon instructions which can be used to program a system to perform the instructions. The storage medium may include, but is not limited to, any type of disk including floppy disks, optical disks, compact disk read-only memories (CD-ROMs), compact disk rewritables (CD-RWs), and magneto-optical disks, semiconductor devices such as read-only memories (ROMs), random access memories (RAMs) such as dynamic random access memories (DRAMs), static random access memories (SRAMs), erasable programmable read-only memories (EPROMs), flash memories, electrically erasable programmable read-only memories (EEPROMs), magnetic or optical cards, or any other type of media suitable for storing electronic instructions.

[0031] While the present invention has been described with respect to a limited number of embodiments, those skilled in the art will appreciate numerous modifications and variations therefrom. It is intended that the appended claims cover all such modifications and variations as fall within the true spirit and scope of this present invention.

Claims

What is claimed is:

1. A method comprising:
transmitting a first packet from a transmitter to a receiver along an interconnect, the first packet including a header indicating a payload length of a first size; and
terminating transmission of the first packet at a packet disconnect boundary, the packet disconnect boundary having a length less than the payload length, wherein the packet transmission is terminated after transmission of a first portion of a payload corresponding to the payload length of the first size.
2. The method of claim 1, further comprising transmitting a second packet from the transmitter to the receiver after terminating the first packet, the second packet having a higher priority than the first packet.
3. The method of claim 2, further comprising transmitting a third packet after transmitting the second packet, the third packet including a remainder of the payload of the first packet.
4. The method of claim 3, further comprising consuming the first portion of the packet in the receiver and not maintaining a history of the first packet in the receiver.
5. The method of claim 4, further comprising transmitting the first packet when the transmitter has credits sufficient to reach the packet disconnect boundary, wherein transmitting the first packet comprises streaming data of the first portion of the payload from the transmitter to the receiver.
6. The method of claim 1, wherein the packet disconnect boundary corresponds to an allowable packet disconnect boundary of a communication protocol of the interconnect, wherein the allowable packet disconnect boundary is configurable.
7. The method of claim 1, wherein terminating the first packet includes transmitting an error code portion and a packet end symbol after the packet disconnect boundary.
8. An apparatus comprising:
a first component to communicate with a second component over a point-to-point (PtP) interconnect, wherein the first component is to transmit a first packet along the PtP interconnect, the first packet including a header indicating a payload length of a first size, and terminate transmission of the first packet at a packet disconnect boundary, the packet disconnect boundary having a length less than the payload length, wherein the first packet

transmission is terminated after transmission of a first portion of a payload corresponding to the payload length of the first size.

9. The apparatus of claim 8, wherein the first component is to transmit a second packet after termination of the first packet, the second packet having a higher priority than the first packet.

10. The apparatus of claim 9, wherein the first component is to transmit a third packet after transmission of the second packet, the third packet including a remainder of the payload of the first packet.

11. The apparatus of claim 9, wherein the second component is to consume the first portion of the packet and not maintain a history of the first packet.

12. The apparatus of claim 9, wherein the first component is to transmit an error code portion and a packet end symbol after the packet disconnect boundary and before the second packet.

13. The apparatus of claim 8, wherein the first component is to transmit the first packet when the first component has credits sufficient to reach the packet disconnect boundary, the first portion of the payload comprising streaming data.

14. The apparatus of claim 8, wherein the packet disconnect boundary corresponds to a configurable allowable packet disconnect boundary of a communication protocol of the PtP interconnect.

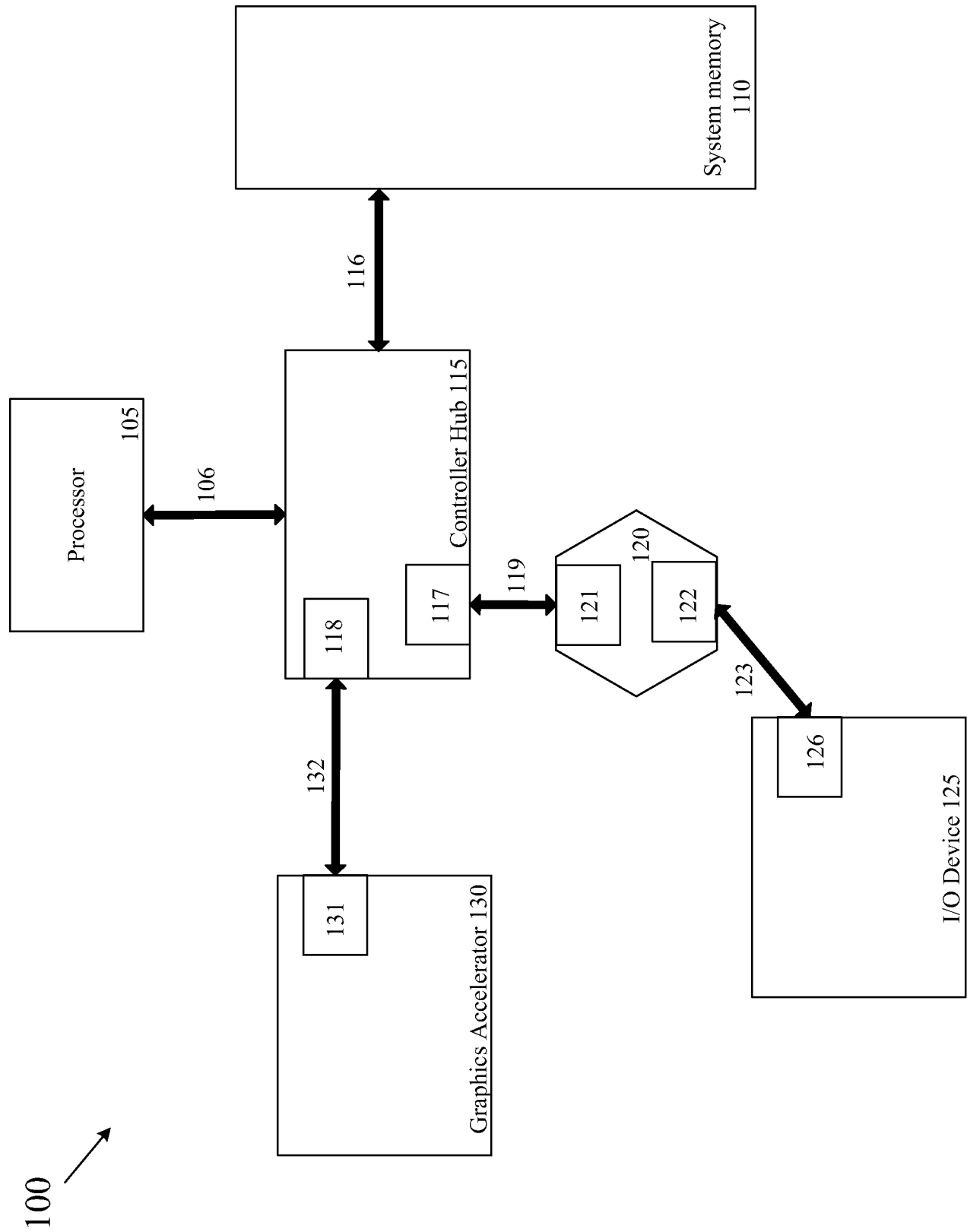


FIG. 1

200

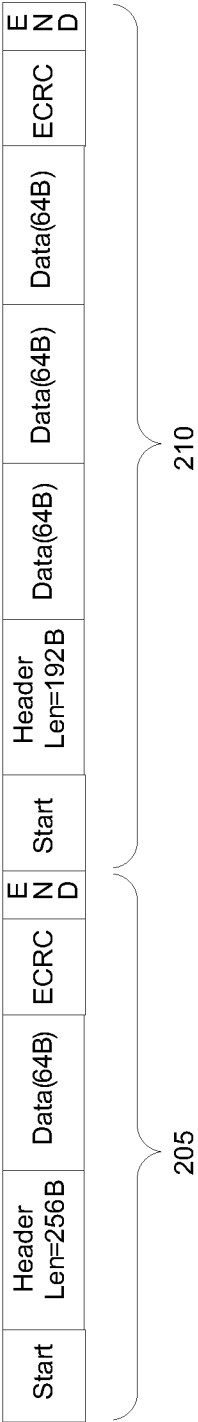


FIG. 2A

250

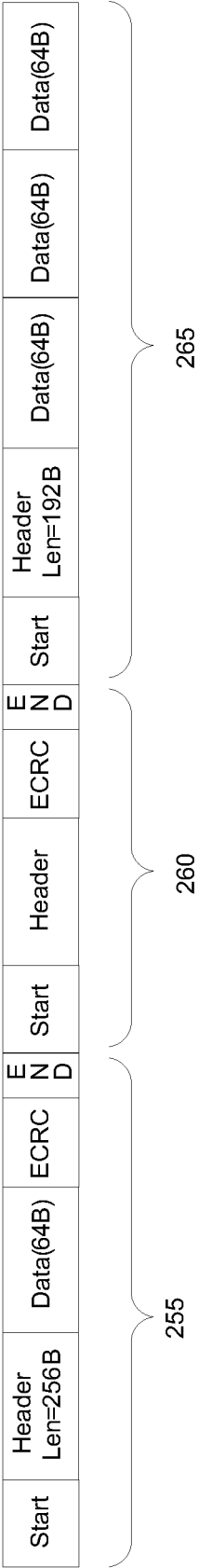


FIG. 2B

300

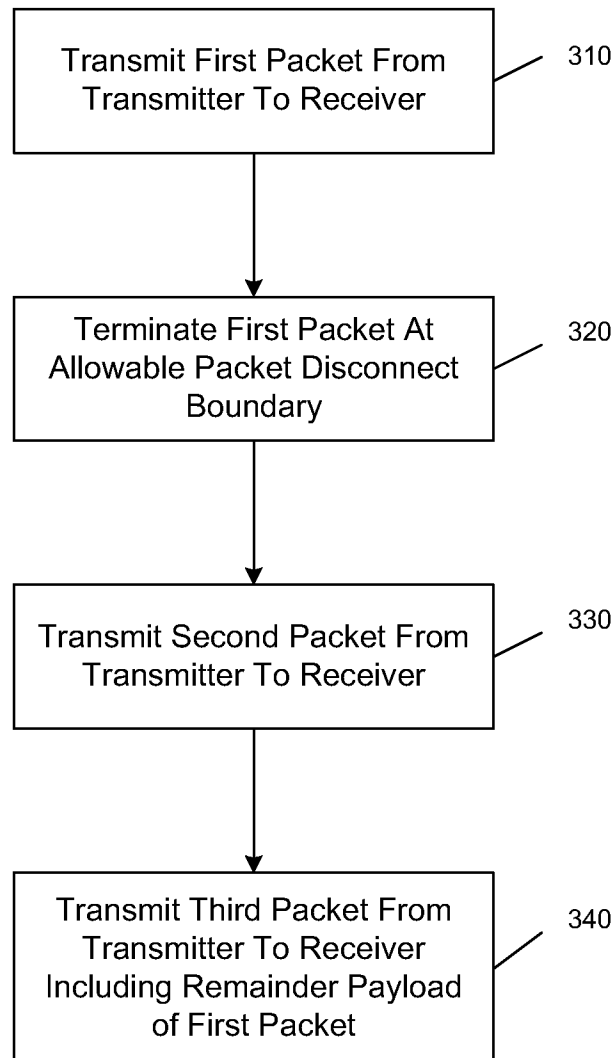


FIG. 3