US009270489B1

(12) **United States Patent**
Wells et al.

(10) **Patent No.: US 9,270,489 B1**
(45) **Date of Patent: Feb. 23, 2016**

(54) **EXPLICIT CONGESTION NOTIFICATION IN MIXED FABRIC NETWORKS**

(75) Inventors: **Philip Wells**, Madison, WI (US); **Michael Marty**, Madison, WI (US)

(73) Assignee: **Google Inc.**, Mountain View, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 458 days.

(21) Appl. No.: **13/423,902**

(22) Filed: **Mar. 19, 2012**

(51) **Int. Cl.**
H04L 12/00 (2006.01)
H04L 12/54 (2013.01)

(52) **U.S. Cl.**
CPC ***H04L 12/56*** (2013.01)

(58) **Field of Classification Search**
CPC H04L 12/56; H04L 12/28; H04L 12/26; H04L 12/863; H04W 28/10; H04W 28/02; G06F 15/16
USPC 370/466, 236, 235, 230.1, 392, 401, 370/359, 368, 241, 252, 229, 395.21, 335, 370/310, 474, 389; 709/224, 212, 228
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

8,213,427	B1 *	7/2012	Eiriksson et al.	370/391
8,457,138	B1 *	6/2013	Boling	370/401
8,660,141	B2	2/2014	Green	
2002/0198927	A1	12/2002	Craddock et al.	
2010/0220740	A1 *	9/2010	Hufferd	370/401
2011/0075563	A1 *	3/2011	Leung et al.	370/236
2011/0222402	A1 *	9/2011	Gai et al.	370/230.1
2012/0147750	A1 *	6/2012	Pelletier et al.	370/235

FOREIGN PATENT DOCUMENTS

GB	2 441 950 B	3/2008
WO	WO-2012/094380 A2	7/2012

OTHER PUBLICATIONS

Allman, M., et al. TCP Congestion Control, Network Working Group, Standards Track, Purdue University, Sep. 2009 (18 pages).

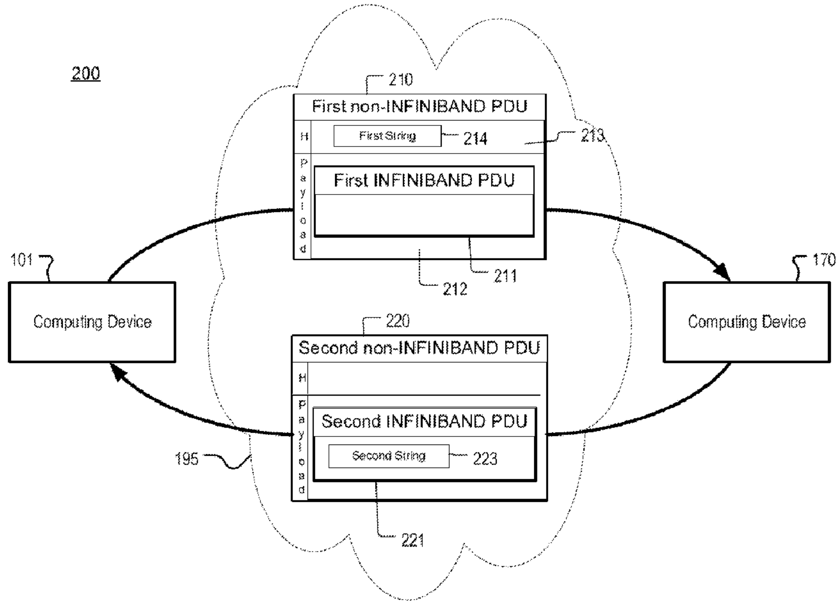
(Continued)

Primary Examiner — Jamal Javaid
Assistant Examiner — Wali Butt
(74) *Attorney, Agent, or Firm* — Edward A. Gordon; Foley & Lardner LLP

(57) **ABSTRACT**

Systems and methods are provided for congestion notification in mixed-fabric InfiniBand networks. In one aspect, a system and apparatus is provided wherein a receiving endpoint receives, from a sending endpoint, InfiniBand messages over a mixed-fabric network. The mixed-fabric network may include an InfiniBand transport layer and a non-InfiniBand messaging fabric. The non-InfiniBand massaging fabric may be any type of non-InfiniBand data-link-layer and network-layer network (e.g. Ethernet, IP). For example, the receiving endpoint may receive a non-InfiniBand protocol data unit (PDU) that contains at least a part of a first InfiniBand PDU as payload. The receiving endpoint may extract signaling from the received non-InfiniBand PDU that indicates whether congestion has been detected in the non-InfiniBand layers of the mixed-fabric network. The receiving endpoint may include the same or similar signaling in the header portion of a second InfiniBand PDU, and transmit the second InfiniBand PDU to the sending endpoint.

15 Claims, 6 Drawing Sheets



(56) **References Cited**

OTHER PUBLICATIONS

InfiniBand(TM) Architecture Release 1.2.1, vol. 1—General Specification. Annex A10: Congestion Control, Final Release, Nov. 2007 (48 pages).

Ramakrishnan, K., et al. The Addition of Explicit Congestion Notification (ECN) to IP, Networking Group, Standard Track, The Internet Society, EMC, Sep. 2001.

* cited by examiner

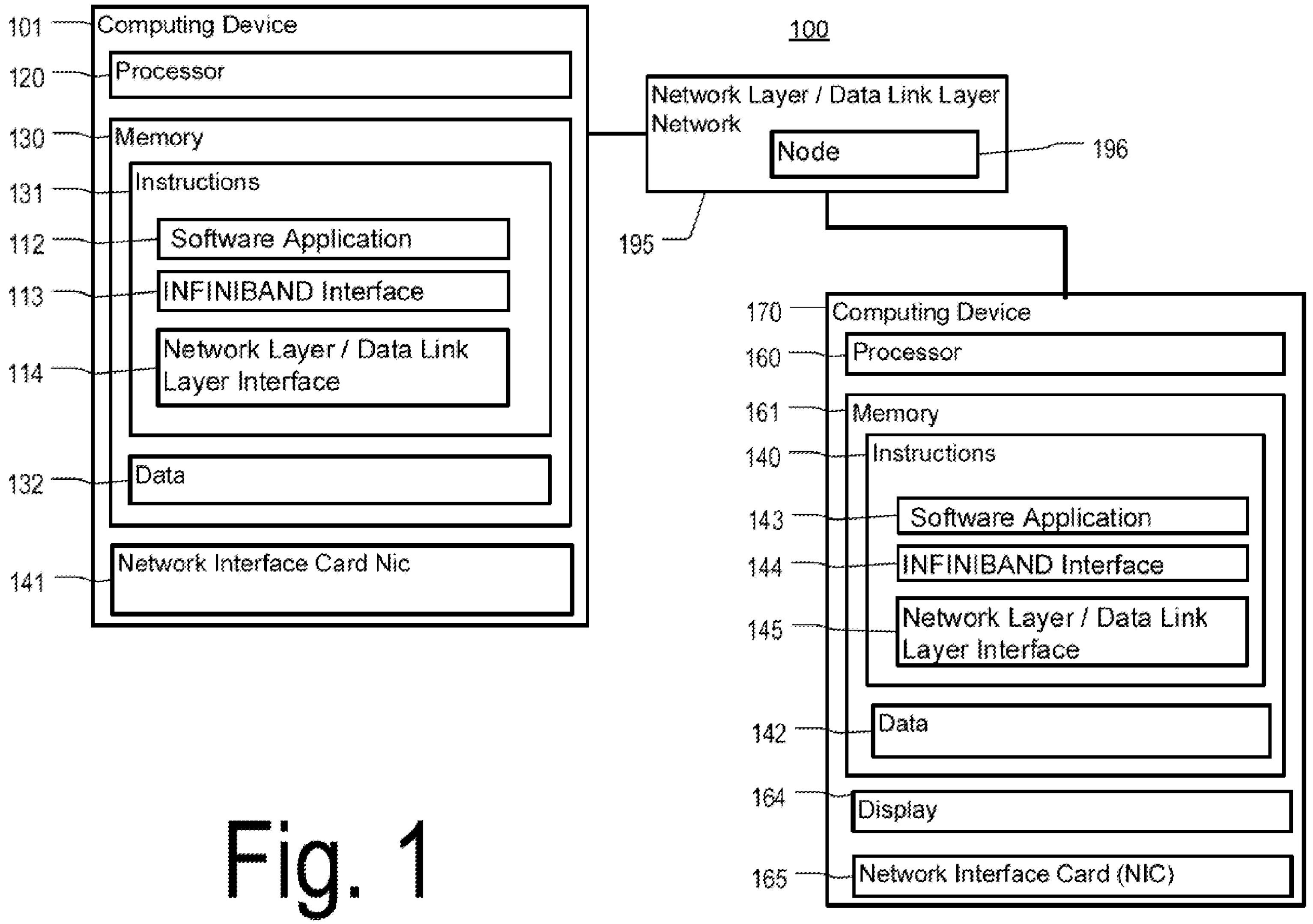

100
101
Computing Device
120
Processor
130
Memory
131
Instructions
112
Software Application
113
INFINIBAND Interface
114
Network Layer / Data Link Layer Interface
132
Data
141
Network Interface Card Nic
Network Layer / Data Link Layer Network
Node
196
195
170
Computing Device
160
Processor
161
Memory
140
Instructions
143
Software Application
144
INFINIBAND Interface
145
Network Layer / Data Link Layer Interface
142
Data
164
Display
165
Network Interface Card (NIC)

Fig. 1

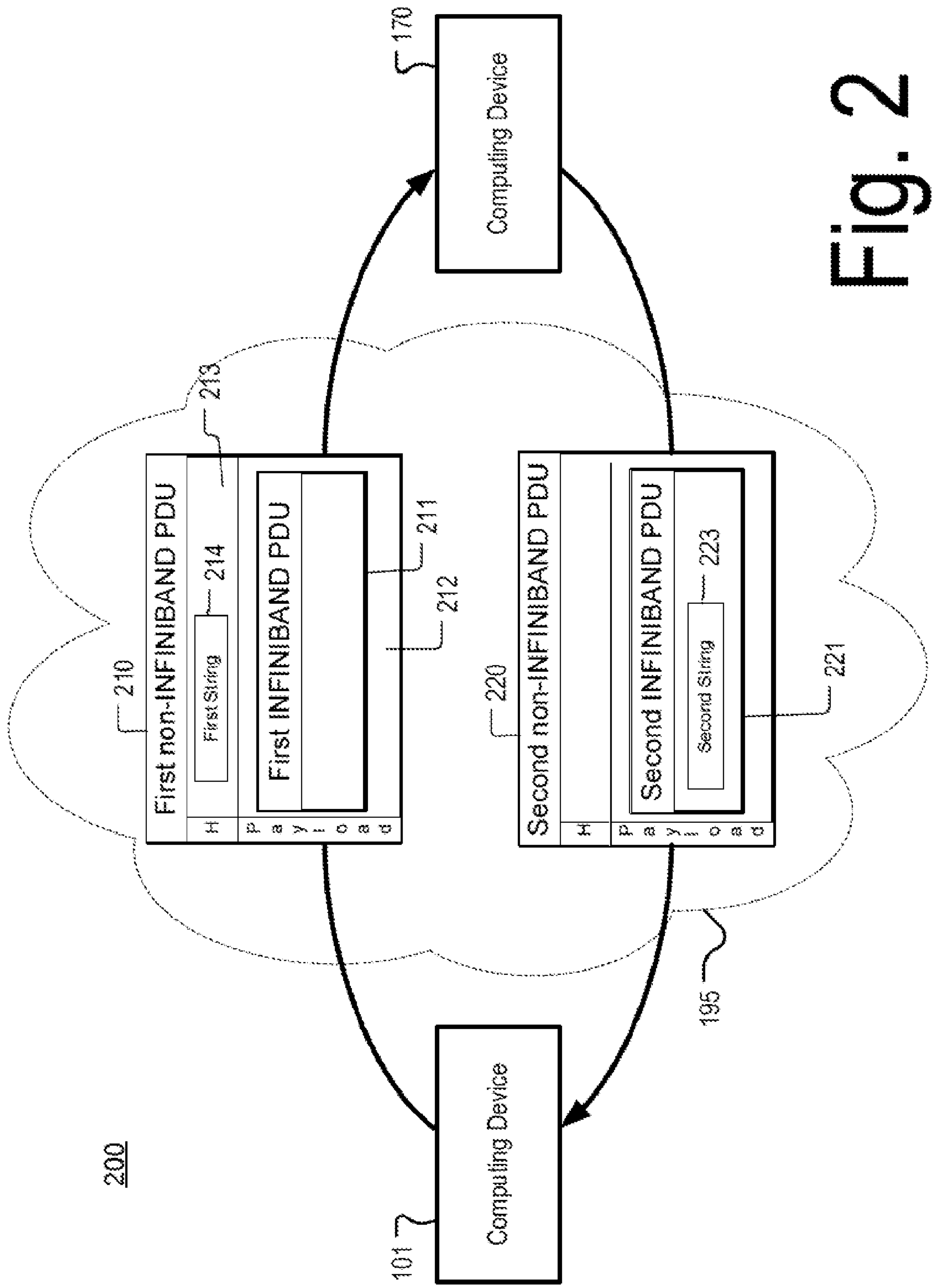

200
101
Computing Device
170
Computing Device
210
First non-INFINIBAND PDU
H
Payload
213
214
First String
First INFINIBAND PDU
211
212
220
Second non-INFINIBAND PDU
H
Payload
Second INFINIBAND PDU
223
Second String
221
195

Fig. 2

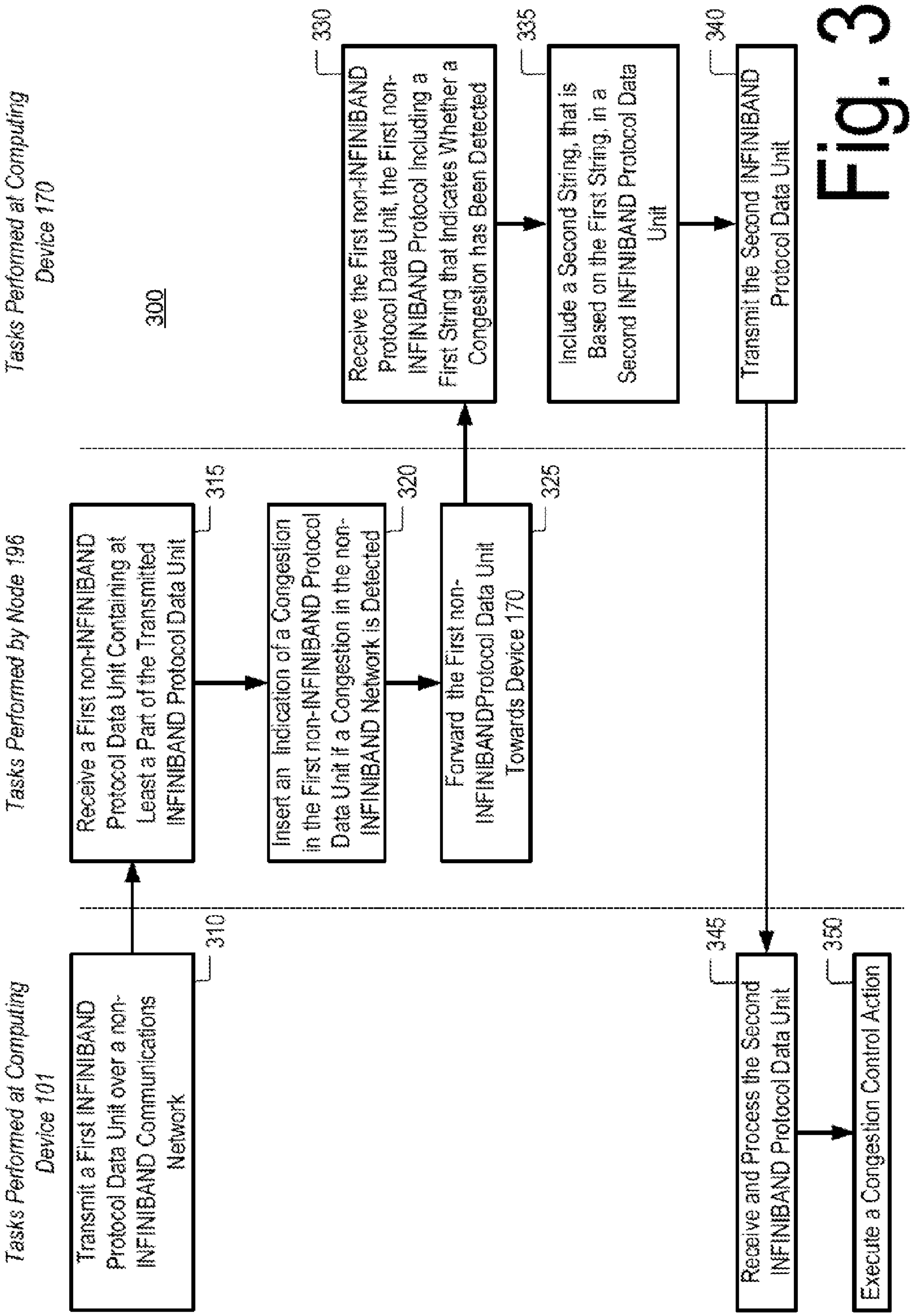

Tasks Performed at Computing Device 101
Tasks Performed by Node 196
Tasks Performed at Computing Device 170
300
Transmit a First INFINIBAND Protocol Data Unit over a non-INFINIBAND Communications Network
310
Receive a First non-INFINIBAND Protocol Data Unit Containing at Least a Part of the Transmitted INFINIBAND Protocol Data Unit
315
Insert an Indication of a Congestion in the First non-INFINIBAND Protocol Data Unit if a Congestion in the non-INFINIBAND Network is Detected
320
Forward the First non-INFINIBANDProtocol Data Unit Towards Device 170
325
Receive the First non-INFINIBAND Protocol Data Unit, the First non-INFINIBAND Protocol Including a First String that Indicates Whether a Congestion has Been Detected
330
Include a Second String, that is Based on the First String, in a Second INFINIBAND Protocol Data Unit
335
Transmit the Second INFINIBAND Protocol Data Unit
340
Receive and Process the Second INFINIBAND Protocol Data Unit
345
Execute a Congestion Control Action
350

Fig. 3

Fig. 4

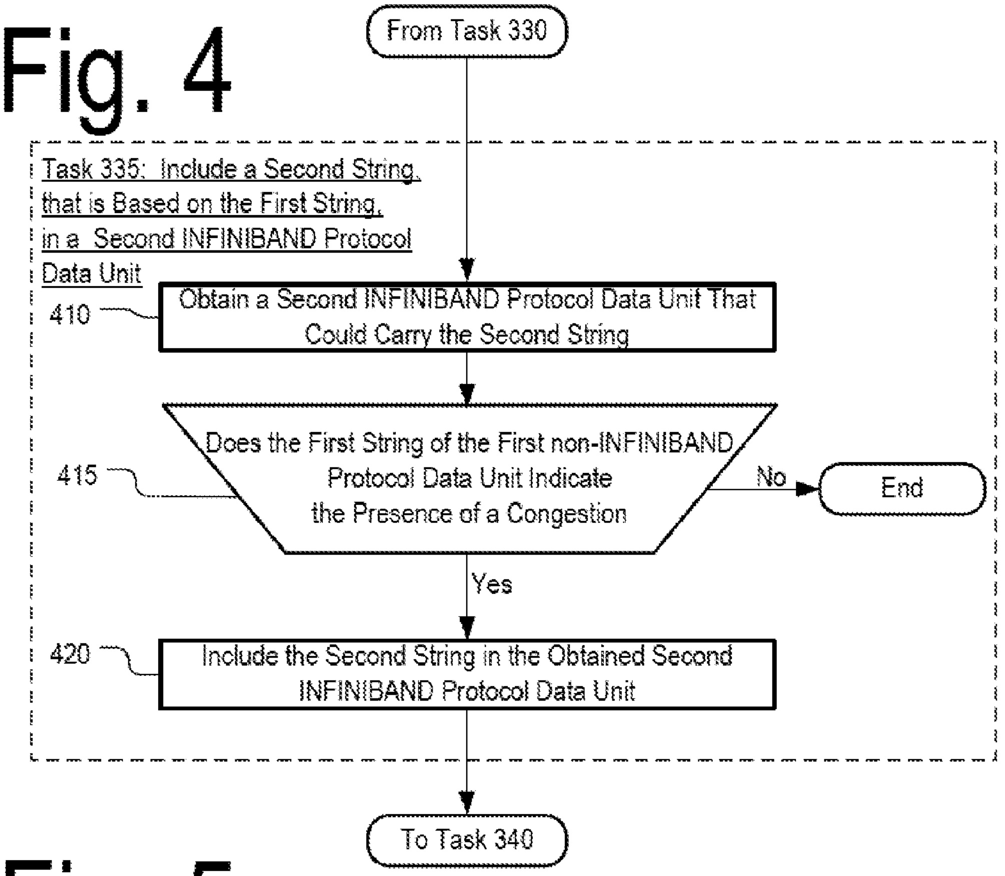

From Task 330
Task 335: Include a Second String, that is Based on the First String, in a Second INFINIBAND Protocol Data Unit
410
Obtain a Second INFINIBAND Protocol Data Unit That Could Carry the Second String
415
Does the First String of the First non-INFINIBAND Protocol Data Unit Indicate the Presence of a Congestion
No
End
Yes
420
Include the Second String in the Obtained Second INFINIBAND Protocol Data Unit
To Task 340

Fig. 5

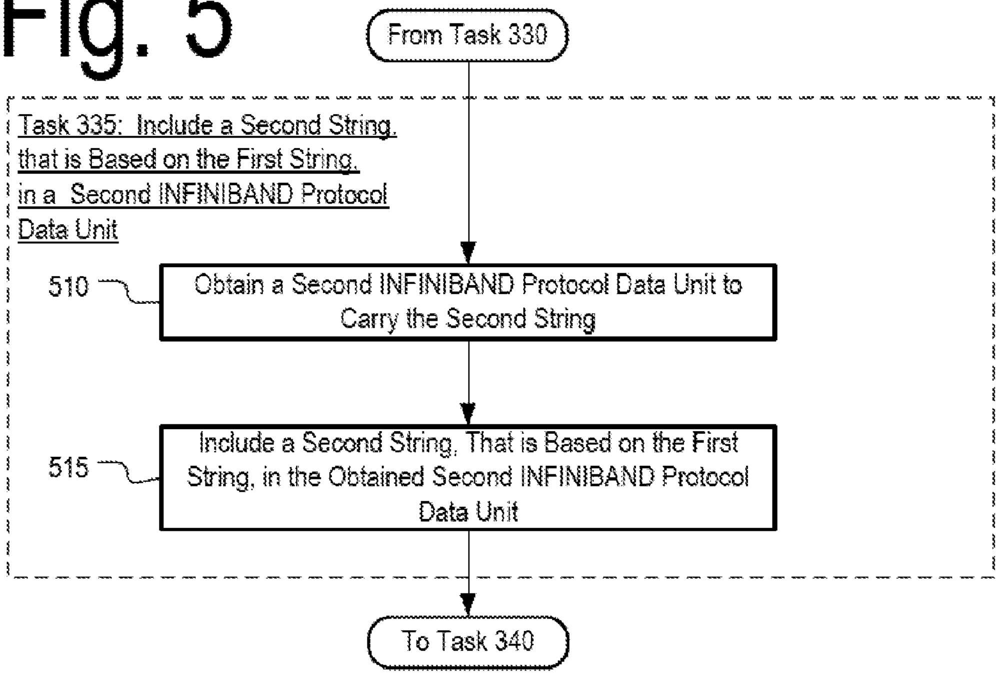

From Task 330
Task 335: Include a Second String, that is Based on the First String, in a Second INFINIBAND Protocol Data Unit
510
Obtain a Second INFINIBAND Protocol Data Unit to Carry the Second String
515
Include a Second String, That is Based on the First String, in the Obtained Second INFINIBAND Protocol Data Unit
To Task 340

Fig. 6

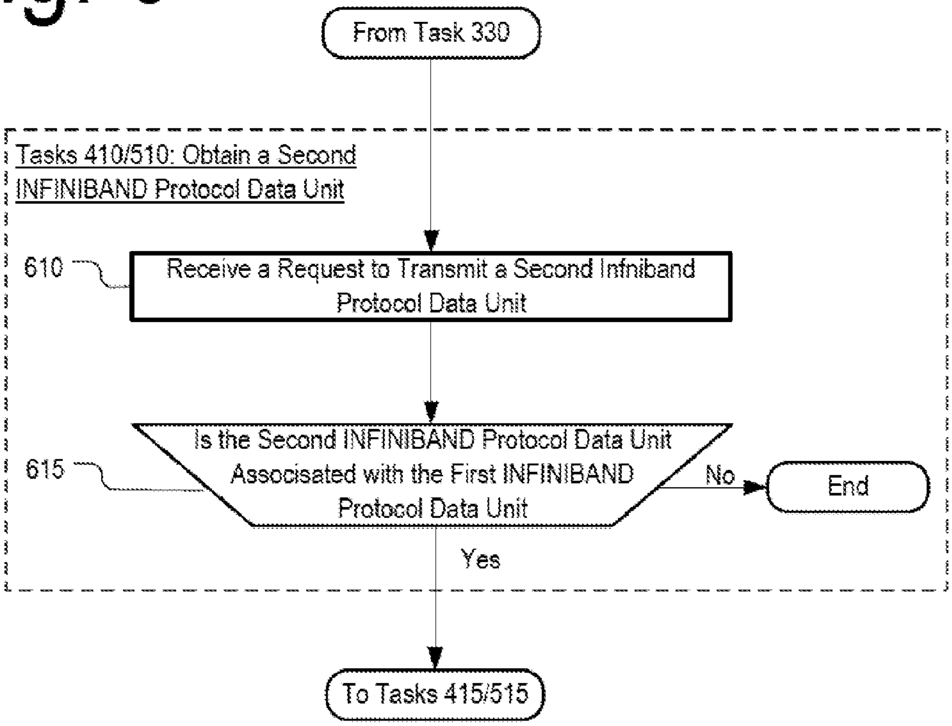

From Task 330
Tasks 410/510: Obtain a Second INFINIBAND Protocol Data Unit
610
Receive a Request to Transmit a Second Infniband Protocol Data Unit
615
Is the Second INFINIBAND Protocol Data Unit Associsated with the First INFINIBAND Protocol Data Unit
No
End
Yes
To Tasks 415/515

Fig. 7

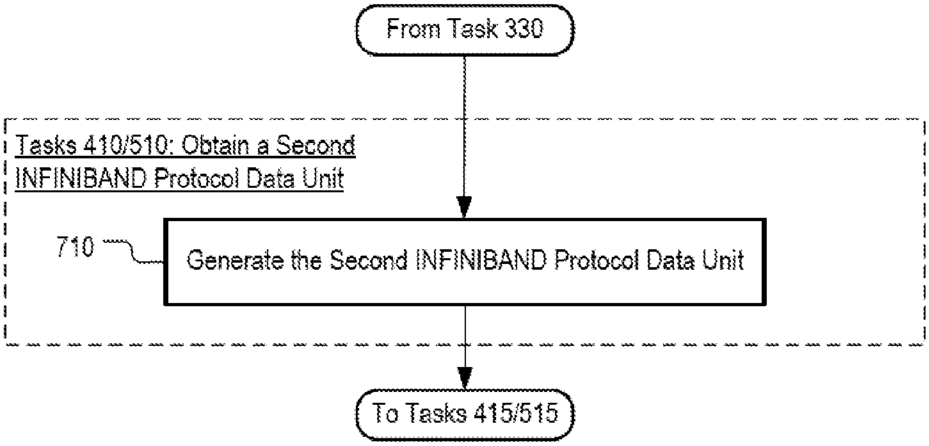

From Task 330
Tasks 410/510: Obtain a Second INFINIBAND Protocol Data Unit
710
Generate the Second INFINIBAND Protocol Data Unit
To Tasks 415/515

Fig. 8

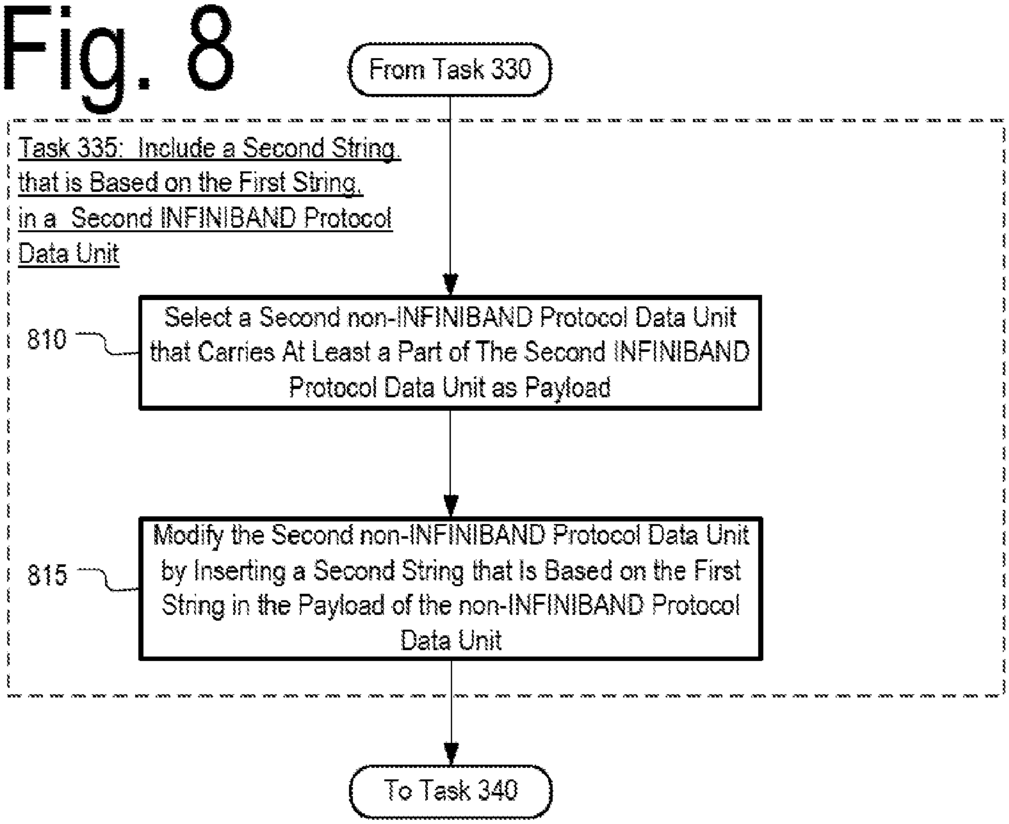
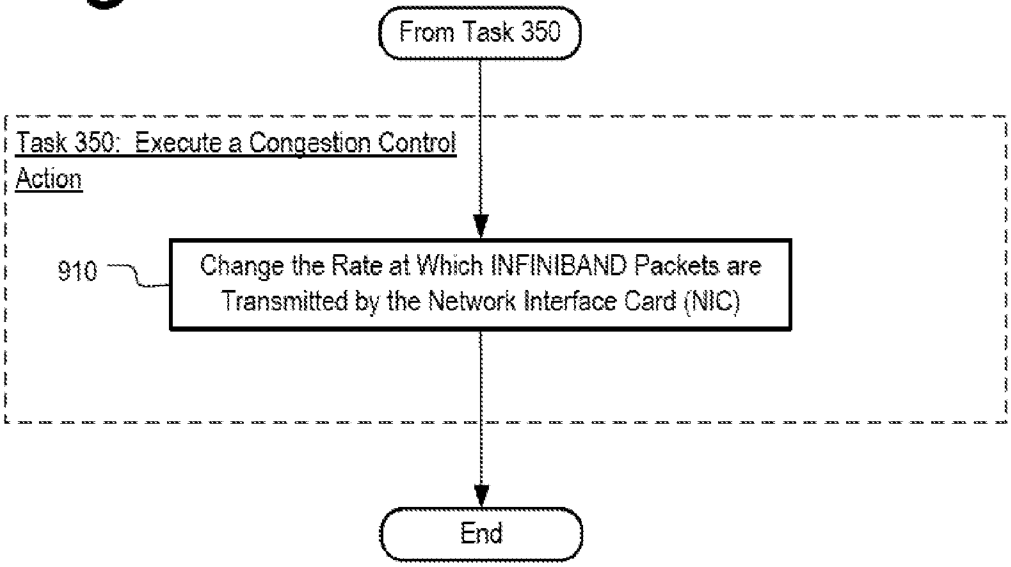

From Task 330
Task 335: Include a Second String, that is Based on the First String, in a Second INFINIBAND Protocol Data Unit
810
Select a Second non-INFINIBAND Protocol Data Unit that Carries At Least a Part of The Second INFINIBAND Protocol Data Unit as Payload
815
Modify the Second non-INFINIBAND Protocol Data Unit by Inserting a Second String that Is Based on the First String in the Payload of the non-INFINIBAND Protocol Data Unit
To Task 340

Fig. 9


From Task 350
Task 350: Execute a Congestion Control Action
910
Change the Rate at Which INFINIBAND Packets are Transmitted by the Network Interface Card (NIC)
End

EXPLICIT CONGESTION NOTIFICATION IN MIXED FABRIC NETWORKS

BACKGROUND

The InfiniBand Architecture (IBA) is an industry standard for inter-computer communication specified by the InfiniBand Trade Association (IBTA). It emerged in 1999 as the joining of two competing proposals known as Next Generation I/O and Future Generation I/O. The InfiniBand Architecture provides high levels of reliability, performance, and scalability.

The InfiniBand Architecture spans several layers of the OSI model. It includes a physical layer, data-link layer, network layer, transport layer, and upper layers. The physical InfiniBand layer defines both the electrical and mechanical aspects of the InfiniBand Architecture including, cables, receptacles, and hot swap characteristics. The link layer encompasses packet layout, point-to-point link operations and switching within subnets and the network layer handles the routing of packets between subnets.

The transport layer may be responsible for in-order packet delivery, partitioning, channel multiplexing, data segmentation and reassembly. In performing its functions, the transport layer may provide several services such as Reliable Connection (RC), Unreliable Connection (UC), Reliable Datagram (RD), Unreliable Datagram (UD), and Raw. In pure InfiniBand networks, the transport services may operate over an InfiniBand data-link fabric. However, when InfiniBand data-link fabric is not available, mixed-fabric networks may be formed where InfiniBand transport services operate atop the network and/or data link layers of non-InfiniBand networks. In that regard, the InfiniBand transport services may utilize the resources of ubiquitous network infrastructure like that of Ethernet and/or Internet Protocol (IP) networks.

SUMMARY

In one aspect, a system is provided that includes a processor coupled to a receiver and a transmitter. The receiver is configured to receive, over a communications network, a protocol data unit of a first protocol that includes (i) a first header portion and (ii) a payload that includes at least part of a first protocol data unit of a second protocol. The first header portion is used in a first explicit congestion notification standard to indicate congestion. Furthermore, the first header portion contains a first string that indicates whether congestion has been detected in the communications network. The processor is configured to prepare a second string based on the first string, insert the second string in a second protocol data unit of the second protocol, and provide to the transmitter the second protocol data unit of the second protocol for transmission to a sender of the first protocol data unit of the second protocol.

In one aspect, a system is provided that includes a receiver and a transmitter. The receiver is configured to receive, over a communications network, a protocol data unit of a second type having (i) a payload including protocol at least part of a first protocol data unit of a first type, and (ii) a header portion including a first string that indicates whether congestion has been detected in the communications network. The system also includes means for preparing a second string based on the first string, inserting the second string in a second protocol data unit of the first type, and providing to the transmitter, the second protocol data unit of the first type for transmission to the sender of the first protocol data unit of the first type.

In yet another aspect, a method is provided wherein a non-InfiniBand protocol data unit that comprises at least a part of a first InfiniBand protocol data unit is received over a non-InfiniBand communications network. The non-InfiniBand protocol data unit includes an indication of a congestion of the non-InfiniBand network. The indication of congestion is transferred, by a processor, to a second InfiniBand protocol data unit that is associated with the first InfiniBand protocol data unit. After the transfer is completed, the second InfiniBand protocol data unit is transmitted to a sender of the first InfiniBand protocol data unit over the non-InfiniBand communications network.

In yet another aspect, a method is provided wherein an incoming non-InfiniBand protocol data unit that comprises a payload portion including at least a part of a first InfiniBand protocol data unit is received. The non-InfiniBand protocol data unit includes a first string that indicates whether congestion is detected in a non-InfiniBand communications network and is part of a header of the non-InfiniBand protocol data unit. After the incoming non-InfiniBand protocol data unit is received, an outgoing non-InfiniBand protocol data unit is identified that comprises a payload portion including at least a part of a second InfiniBand protocol data unit. The payload portion of the outgoing non-InfiniBand protocol data unit is modified, by a processor, so as to insert a second string in a header portion of the second InfiniBand protocol data unit. The second string is based on the first string and is one or more characters long. Following the insertion of the second string, the outgoing non-InfiniBand protocol data unit is transmitted.

In yet another aspect, a network interface card is provided. The network interface card includes programmable logic that is configured to receive an indication of a congestion, wherein the indication of the congestion is associated with a first InfiniBand protocol data unit that has been previously transmitted by the network interface card over a non-InfiniBand network, change, in response to receiving the indication of the congestion, the transmission rate of non-InfiniBand protocol data units that comprise InfiniBand protocol data units as a payload. The change of the transmission rate of non-InfiniBand protocol data units that comprise InfiniBand protocol data units as payload does not affect the transmission rate of non-InfiniBand protocol data units having payload that includes non-InfiniBand messages.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. **1** depicts a schematic diagram of a system.

FIG. **2** depicts an example of the operation of the system of FIG. **1**.

FIG. **3** depicts a flowchart of a process in accordance with aspects of the disclosure.

FIG. **4** depicts a flowchart of subtasks associated with FIG. **3**.

FIG. **5** depicts an example of a subtask associated with FIG. **3**.

FIG. **6** depicts a flowchart of subtasks associated with FIG. **4**.

FIG. **7** depicts an example of a subtask associated with FIG. **4**.

FIG. **8** depicts an example of a subtask associated with FIG. **3**.

FIG. **9** depicts an example of a subtask associated with FIG. **3**.

DETAILED DESCRIPTION

In one aspect, a system and apparatus are provided wherein a receiving endpoint receives, from a sending endpoint, a first protocol data unit (PDU) from a first type. The first PDU from the first type may be an InfiniBand transport PDU or any other transport-layer, application layer, session layer, or presentation layer PDU. The first protocol data unit of the first type is received over a network of a second type. The network of the second type may be an Internet Protocol (IP) network or any type data-link-layer and/or network-layer network. For example, the receiving endpoint may receive a protocol data unit (PDU) of the second type that contains at least a part of the first PDU of the first type as payload. The receiving endpoint may extract signaling from the received PDU of the second type that indicates whether congestion has been detected in the network of the second type. The receiving endpoint may include the same or similar signaling in the header portion of a second PDU of the first type, and transmit the second PDU of the first type to the sending endpoint. In that regard, signaling from the network of the second type is communicated to the sending endpoint inside protocol data units of the first type.

In another aspect, a system and apparatus are provided wherein a receiving endpoint receives, from a sending endpoint, InfiniBand messages over a mixed-fabric network. The mixed-fabric network may include an InfiniBand transport layer and a non-InfiniBand messaging fabric. The non-InfiniBand massaging fabric may be any type of non-InfiniBand data-link-layer and/or network-layer network (e.g. Ethernet, IP). For example, the receiving endpoint may receive a non-InfiniBand protocol data unit (PDU) that contains at least a part of a first InfiniBand PDU as payload. The receiving endpoint may extract signaling from the received non-InfiniBand PDU that indicates whether congestion has been detected in the non-InfiniBand layers of the mixed-fabric network. The receiving endpoint may include the same or similar signaling in the header portion of a second InfiniBand PDU, and transmit the second InfiniBand PDU to the sending endpoint. In that regard, signaling from the non-InfiniBand network layer is transferred onto the secondary InfiniBand network layer (of the mixed fabric network) that relies on the non-InfiniBand layer for transport.

In yet another aspect, a network interface card (NIC) is provided that is capable of transmitting InfiniBand messages over a non-InfiniBand fabric. The non-InfiniBand fabric may be any type of data-link-layer or network-layer network (e.g. Ethernet, IP). When an indication of congestion in the non-InfiniBand fabric is detected, the transmission rate of network-layer (or data-link layer) PDUs that carry InfiniBand messages as payload is changed, while the transmission rate of network-layer (or data-link layer) PDUs that carry other types of payload remains unaffected. In that way, rate-based flow control for InfiniBand PDUs is implemented that does not interfere with the flow control mechanisms of networks that use the non-InfiniBand fabric (e.g. TCP networks).

As shown in FIG. **1**, an exemplary system **100** may include computers **101** and **170** interconnected by network **195**. Computing device **101** may contain a processor **120**, memory **130** and other components typically present in general purpose computers.

Memory **130** of computing device **101** stores information accessible by processor **120**, including instructions **131** that may be executed by the processor **120**. Memory also includes data **132** that may be retrieved, manipulated or stored by the processor. The memory may be of any type capable of storing information accessible by the processor, such as a hard-drive, memory card, ROM, RAM, DVD, CD-ROM, write-capable, and read-only memories. The processor **120** may be any well-known processor, such as commercially available processors. Alternatively, the processor may be a dedicated controller such as an ASIC.

The instructions **131** may be any set of instructions to be executed directly (such as machine code) or indirectly (such as scripts) by the processor. In that regard, the terms "instructions," "steps" and "programs" may be used interchangeably herein. The instructions may be stored in object code format for direct processing by the processor, or in any other computer language including scripts or collections of independent source code modules that are interpreted on demand or compiled in advance. Functions, methods and routines of the instructions are explained in more detail below.

Data **132** may be retrieved, stored or modified by processor **120** in accordance with the instructions **131**. For instance, although the system and method is not limited by any particular data structure, the data may be stored in computer registers, in a relational database as a table having a plurality of different fields and records, or XML documents. The data may also be formatted in any computer-readable format such as, but not limited to, binary values, ASCII or Unicode. Moreover, the data may comprise any information sufficient to identify the relevant information, such as numbers, descriptive text, proprietary codes, pointers, references to data stored in other memories (including other network locations) or information that is used by a function to calculate the relevant data.

Although FIG. **1** functionally illustrates the processor and memory as being within the same block, it will be understood by those of ordinary skill in the art that the processor and memory may actually comprise multiple processors and memories that may or may not be stored within the same physical housing. For example, the processor may actually comprise a collection of processors which may or may not operate in parallel. Furthermore, the processor may comprise a supplemental component such as a co-processor, FPGA, or another type of special-purpose circuitry for supplementing the functions of the processor. In some aspects, the supplemental component may be located in a separate packaging from the processor. In other aspects, the supplemental component may be mounted on the same system board as the processor. Moreover, some of the instructions and data may be stored on removable CD-ROM and others within a read-only computer chip. Some or all of the instructions and data, that are used by the processor, may be stored in a location physically remote from, yet still accessible by, the processor.

The computing device **101** may be at one node of the network **195** and capable of directly and indirectly communicating with other nodes of the network. For example, computing device **101** may comprise a web server that is capable of communicating with computing device **170** via network **195** such that computing device **101** uses network **195** to transmit and display information on a display screen of computing device **170**. Computing device **101** may also comprise a plurality of computers, e.g., a load balanced server farm, that exchange information with different nodes of a network for the purpose of receiving, processing and transmitting data to the client devices. In this instance, the client devices will typically still be at different nodes of the network than any of the computers comprising computing device **101**.

Each computing device **170** may be configured similarly to the computing device **101**, with a processor **160**, memory **161**, instructions **140**, and data **142**. Each computing device **170** may be a personal computer, intended for use by a person, having all the internal components normally found in a personal computer such as a central processing unit (CPU), display device **163** (for example, a monitor having a screen, a projector, a touch-screen, a small LCD screen, a television, or another device such as an electrical device that is operable to display information processed by the processor), CD-ROM, hard drive, user input (for example, a mouse, keyboard, touch screen or microphone), speakers, modem and/or network interface device (telephone, cable or otherwise) and all of the components used for connecting these elements to one another. Moreover, computers in accordance with the systems and methods described herein may comprise any device capable of processing instructions and transmitting data to and from humans and other computers including general purpose computers, PDAs, network computers lacking local storage capability, set-top boxes for televisions, and other networked devices.

Although the computing device **170** may comprise a full-sized personal computer, the system and method may also be used in connection with mobile devices capable of wirelessly exchanging data with a server over a network such as the Internet. By way of example only, computing device **170** may be a wireless-enabled PDA, hand-held or in-car navigation device, tablet PC, netbook, or a cellular phone capable of obtaining information via the Internet. The user may input information, for example, using a small keyboard, a keypad, or a touch screen. Furthermore, computing device **170** may be a communications switch or another type of communications network node.

The computing device **101** and computing device **170** are capable of direct and indirect communication, such as over network **195**. Although only a few computers are depicted in FIG. **1**, it should be appreciated that a typical system can include a large number of connected computers, with each different computer being at a different node of the network **195**. The network, and intervening nodes, may comprise various configurations and protocols. Such communication may be facilitated by any device capable of transmitting data to and from other computers, such as modems (e.g., dial-up, cable or fiber-optic) and wireless interfaces.

Furthermore, although certain advantages are obtained when information is transmitted or received as noted above, other aspects of the system and method are not limited to any particular manner of transmission of information. For example, in some aspects, information may be sent via a medium such as a disk, tape or CD-ROM. Yet further, although some functions are indicated as taking place on a single client device having a single processor, various aspects of the system and method may be implemented by a plurality of computers, for example, communicating information over network **195**.

InfiniBand interface **113** of computer **101** may include logic for executing at least a part of a process **300** which is illustrated in FIG. **3**. Software application **112** may use InfiniBand interface **113** to establish an InfiniBand communications channel with software application **143** and exchange InfiniBand protocol data units (PDUs) over it. The InfiniBand protocol data units (PDUs) may be exchanged, for example, by using SEND/RECEIVE or RDMA WRITE/READ routines. InfiniBand interface **113** may use interface **114** to send and transmit the InfiniBand protocol data units (PDUs) over network **195**. The interface **114** may be a part of a communications protocol stack that is available on computing device **101** and it may comprise processor executable instructions for receiving and transmitting layer-3 or layer-2 protocol data units (PDUs) over network **195**. For example, interface **114** may be an Internet Protocol (IP) interface. Although in this example, InfiniBand interface **113** and interface **114** are depicted as comprising instructions stored in memory **130**, they can also comprise instructions stored in memory of network interface card (NIC) **141** that are executable by the network card's own processor (not shown). Moreover, InfiniBand interface **113** and interface **114** can be implemented, at least in part, in hardware as digital logic circuit(s) that are part of a processor, or network interface card (NIC) **141**, or another computer component.

Network interface card (NIC) **141** may include a transmitter and receiver (not shown). The transmitter and receiver may be arranged as separate devices or as a single transceiver. The transmitter may include hardware and/or software for transmitting protocol data units, such as PDU **210**, over network **195**. Similarly, the receiver may include hardware and/or software for receiving protocol data units, such as PDU **220**, over network **210**.

InfiniBand interface **144** of computer **170** may include logic for executing at least a part of the process **300**. Software application **143** may use InfiniBand interface **144** to establish an InfiniBand communications channel with software application **112** and exchange InfiniBand protocol data units (PDUs) over it. The InfiniBand protocol data units (PDUs) may be exchanged, for example, by using SEND/RECEIVE or RDMA WRITE/READ routines. InfiniBand interface **144** may use interface **145** to send and transmit the InfiniBand protocol data units (PDUs) over network **195**. The interface **145** may be a part of a communications protocol stack that is available on computing device **170** and it may comprise processor executable instructions for receiving and transmitting layer-3 or layer-2 protocol data units (PDUs) over network **195**. For example, interface **145** may be an Internet Protocol (IP) or Ethernet interface. Although in this example, InfiniBand interface **144** and interface **145** are depicted as comprising instructions stored in memory **161**, they can also comprise instructions stored in memory of network interface card (NIC) **165** that are executable by the network card's own processor (not shown). Moreover, InfiniBand interface **144** and interface **145** can be implemented, at least in part, in hardware as digital logic circuit(s) that are part of network interface card (NIC) **165** or another computer component.

Network interface card (NIC) **165** may include a transmitter and receiver (not shown). The transmitter and receiver may be arranged as separate devices or as a single transceiver. The transmitter may include hardware and/or software for transmitting protocol data units, such as PDU **220**, over network **195**. Similarly, the receiver may include hardware and/or software for receiving protocol data units, such as PDU **210**, over network **210**. In operation, the transmitter may provide incoming protocol data units such as PDU **220**, to the processor of the network interface card's **165** (not shown). The processor may process the incoming protocol data units and generate, at least in part, outgoing packets. After the outgoing packets are generated, they may be provided to the transmitter for transmission over the network **195**.

In one embodiment, network **195** is a non-InfiniBand communications network that is used as a messaging fabric for an InfiniBand transport layer in a mixed-fabric network. For instance, Network **195** may be a non-InfiniBand OSI layer-3 (network layer), OSI layer-2 (data link layer) network, such as an Internet Protocol (IP) network or an Ethernet network, or any non-InfiniBand network that falls at least partially within the spectrum of either OSI layer-2 or OSI layer-3. Furthermore, network **195** may include node **196**. Node **196** may be a switch or any other type of network node that is typically found in layer-3 or layer-2 networks (e.g., a bridge, a hub). Node **196** is capable of detecting congestions in network **195** and modifying outgoing protocol data units to carry an indication that congestion has been detected. In other words, node **196** implements an explicit congestion notification mechanism.

Explicit congestion notification is a mechanism for ensuring quality-of-service in communications networks. When network nodes operating below the transport layer become congested, they may indicate the congestion to nodes that operate at or above the transport layer. This gives the opportunity to nodes at the higher layer to ease the congestion by reducing the rate at which they transmit protocol data units.

The Internet Protocol (IP) incorporates the explicit congestion notification (ECN) standard specified in the Network Working Group Request for Comment: 3168 (RFC 3168). According to RFC 3168, the content of the CE codepoint of IP protocol data units is reflected (e.g., echoed) in the ECN-echo and Congestion Window Reduced (CWR) flags of Transport Control Protocol (TCP) protocol data units. The TCP protocol also incorporates the RFC 3168. The TCP protocol assigns the CWR and ECN-echo flags the function required by RFC 3168. Thus, the TCP and IP protocols incorporate (or implement) the same ECN specification (e.g., RFC 3168).

Unlike TCP, InfiniBand does not incorporate RFC 3168. InfiniBand uses a different explicit congestion notification standard. Different explicit congestion notification standards may use different bit fields and/or different message (or flag) formats to indicate congestion. Thus, when InfiniBand is used over an Internet Protocol fabric, it may lose the explicit congestion notification capabilities that are available when an InfiniBand messaging fabric is used.

FIG. **2** depicts a schematic diagram showing an example of the operation of system **200**. The system **200** may enable InfiniBand to retain its explicit congestion notification capabilities when used in mixed-fabric arrangements. As FIG. **2** shows, InfiniBand PDUs **211** and **221** are transmitted over a mixed fabric network. The mixed fabric network uses network **195** as its messaging fabric. When InfiniBand PDUs **211** and **221** are transmitted over network **195**, they are encapsulated in non-InfiniBand PDUs **210** and **220**, respectively. As suggested by the discussion of network **195**, non-InfiniBand PDUs **210** and **220** may be Internet Protocol (IP) PDUs or PDUs of another OSI layer-3 network (e.g., network layer network). Alternatively, PDUs **210** and **220** may be PDUs of an OSI layer-2 network (e.g., data-link layer network) or any network that falls at least partially within the spectrum of OSI layer-2 and OSI layer-3.

In the present example, device **101** transmits an InfiniBand protocol data unit (PDU) **211** to device **170** over network **195**. The InfiniBand PDU **211** is encapsulated in payload portion **212** of non-InfiniBand protocol data unit (PDU) **210**. PDU **210** includes a first string **214** which indicates whether congestion has been detected in network **195**. More precisely, the first string **214** is the value of a field in PDU **210** that has been designated by the protocol of network **195** to carry an indication of congestion. The first string **214** may be a part of the header **213** of PDU **210**, its payload portion **212** or its padding. For example, if PDU **210** is an Internet Protocol (IP) protocol data unit, the first string **214** may be value of the CE codepoint of the protocol data unit. Although the entire InfiniBand PDU **211** is depicted as being encapsulated in PDU **210**, the InfiniBand PDU **211** may be fragmented among multiple protocol data units (PDUs).

In response to receiving PDU **210**, device **170** may transmit an InfiniBand PDU **221** that includes a second string **223**. The second string **223** may be part of the header of PDU **221**, its payload portion or its padding. The second string **223** is based on the first string **214**. For instance, the second string **223** may be identical to the first string **214** or it may be a different string that is derived, at least in part, from the first string **214**. By including the second string **223** in InfiniBand PDU **221**, an indication of congestion that is present in a non-InfiniBand PDU (e.g., PDU **210**) is transferred to an InfiniBand PDU (e.g., PDU **221**). In one aspect, this transfer enables InfiniBand interface **113** (or another part of the InfiniBand layer of the mixed fabric network) to be notified of a congestion taking place and take anti-congestion measures in response.

FIG. **3** depicts a flowchart of example process **300**. At task **310**, InfiniBand PDU **211** is transmitted from device **101** to device **170** over network **195**. At task **315**, PDU **210** is received at node **196**. At task **320**, node **196** determines if network **195** is congested. Upon a positive determination, node **196** may insert an indication of the congestion in PDU **210**. In inserting the indication of the congestion, node **196** may set the value of a field of PDU **210** that is designated to carry an indication of congestion to a predetermined value. For example, it may set the CE codepoint of an Internet Protocol (IP) protocol data unit to "11". Alternatively, as another example, node **196** may insert a string such as "11000100" in the OPTIONS header field of an Internet Protocol (IP) protocol data unit (to indicate that the congestion has been detected at node **196**). In any event, as a result of task **320**, the value of the first string **214** in PDU **210** is set to indicate that congestion has been detected in network **195**.

At task **325**, PDU **210** is forwarded by node **196** towards device **170**. At task **330**, PDU **210** is received at device **170**. At task **335**, the second string **223** is included into InfiniBand PDU **221**. Different ways for including the second string **223** are described in the discussion with respect to FIGS. **4**, **5**, and **8**. At task **340**, the InfiniBand PDU **221** is transmitted. The InfiniBand PDU **221** may be transmitted to the sender of InfiniBand PDU **210** (e.g., device **101**) or to another InfiniBand-enabled device that is cable of taking a congestion control measure. Although in the example of FIG. **3**, tasks **325**-**340** are performed by device **170**, they can be performed by network interface card (NIC) **165** in a manner that is autonomous from processor **160**. Furthermore, tasks **325**-**340** may be performed by a node (e.g., node **196** in FIG. **1**) of network **195** that is located on a path between device **101** and device **170**.

At task **345**, the InfiniBand PDU **221** is received and processed. If the InfiniBand PDU **221** contains an indication of congestion, a congestion control action is executed (at task **350**). The congestion control action may be any congestion control measure, such as slow start or sliding windows. Moreover, executing the congestion control action may include performing the process described in the discussion with respect to FIG. **9**.

FIG. **4** depicts a flowchart of example subtasks associated with including, in a second InfiniBand protocol data unit, a second string that is based on the first string as shown by task **335** of FIG. **3**. In this example, task **335** includes the following subtasks: obtaining an InfiniBand PDU that could carry the second string **223** (**410**), determining whether PDU **210** contains an indication of congestion (**415**), and if it does, including an indication of the congestion in the obtained InfiniBand PDU (**420**).

At task **410**, an InfiniBand PDU is obtained that could carry the second string **223**. In this example, the obtained PDU is InfiniBand PDU **221**. The manner in which an InfiniBand PDU may be obtained is further described in the discussion with respect to FIGS. **6**-**7**.

At task **415**, the first string **214** is examined to determine whether it indicates the presence of congestion. If it does, task **420** is executed. Otherwise, the execution of process **300** may be stopped.

At task **420**, the value of the second string **223** is set to indicate the presence of congestion after which the second string **223** is included in InfiniBand PDU **221**. For example, the string may be written over bits of a specific section of InfiniBand PDU **221**. The section may be a specific header (e.g., Base Transport Header (BTH)), a part of the payload of InfiniBand PDU **221** (e.g., bits n, n+1, and n+3 from the start of the payload) or the padding of InfiniBand PDU **221**. In any event, as a result of executing the process of FIG. **4**, an indication of congestion is included in InfiniBand PDU **221** only when the first string **214** indicates that congestion has been detected in network **195**.

FIG. **5** depicts a flowchart of example subtasks associated with including, in a second InfiniBand protocol data unit, a second string that is based on the first string as shown by task **335** of FIG. **3**. In this example, task **335** includes the following subtasks: obtaining an InfiniBand PDU to carry the second string **223** (**510**) and including the second string **223** in InfiniBand PDU **221** (**515**).

At task **510**, an InfiniBand PDU is obtained to carry the second string **223**. In this example, the obtained PDU is InfiniBand PDU **221**. The manner in which an InfiniBand PDU may be obtained is further described in the discussion with respect to FIGS. **6**-**7**.

At task **515**, the second string **223** is generated based on the first string **214** and included in InfiniBand PDU **221**. For example, the string may be written over a specific section of InfiniBand PDU **221**. The section may be a specific header (e.g., Base Transport Header (BTH)), a part of the payload of InfiniBand PDU **221** (e.g., bits n, n+1, and n+3 from the start of the payload) or the padding of InfiniBand PDU **221**.

In one aspect, the second string **223** may be identical to the first string **214**. For example, if PDU **210** is an Internet Protocol (IP) protocol data unit, the value of a bit in the Base Transport Header (BTH) of the InfiniBand PDU **221** may be set to equal the value of a bit in the CE codepoint of PDU **210**. In another aspect, the second string **223** may be produced by translating the first string **214** from a first format to a second format that is recognized by InfiniBand interface **113** while retaining the same or similar meaning as the first string **214** (e.g., “there is congestion”, “there isn’t a congestion”). In any event, the second string **223** is generated and included in InfiniBand PDU **221** regardless of whether the first string **214** indicates that congestion has been detected in network **195**. In that regard, unlike the process of FIG. **4**, the process of FIG. **5** does not require detailed knowledge of the protocol used by PDU **210** in terms of being able to interpret the first string **214** and recognize whether it indicates the presence of congestion.

Moreover, in one aspect, in addition to the first string **214**, the value of the second string **223** may also be based on a string s_1 (not shown). String s_1 and the first string **214** are members of a set S. Each string in the set S is a part of a different layer-3 protocol data unit (PDU) from the set P and it indicates whether congestion has been detected in network **195**. The set P consists only of protocol data units (PDUs) among which InfiniBand PDU **211** has been fragmented (if at all) when transmitted. Thus, in one aspect, the value of the second string **223** is based on the value of sting **213** and another string found in another PDU that is used to carry a part of PDU **211** across network **195**. By way of example, the second string **223** may be set to equal one of: the sum of the first string **214** and string s_1, their logical sum, product, logical product, and so forth.

FIG. **6** depicts a flowchart of example subtasks associated with obtaining an InfiniBand PDU to carry the second string **223** as shown by tasks **410** and **510** of FIGS. **4**-**5**. In this example, the InfiniBand PDU is obtained by executing the following subtasks: receiving a request to transmit InfiniBand PDU **221** (**610**) and determining whether the InfiniBand PDU **221** is associated with InfiniBand PDU **211** (**615**).

At task **610**, a request to transmit the InfiniBand PDU **221** is received. The request may come in the form of a system call or a call to an API (such interfaces **144**, **145**). Furthermore, the request may be implicit in the sense that no specific instructions (or signaling) is provided along with PDU **221**, other than what might be in its header. In the example depicted in FIG. **3**, the request may be received at InfiniBand interface **144** of device **170**. The InfiniBand PDU **221** may be provided in a pre-packaged form with all necessary headers. Or alternatively, only the payload of the InfiniBand PDU **221** may be provided. In the latter case, the InfiniBand interface **144** may generate the InfiniBand PDU **221** by itself.

In an alternative example (not depicted in FIG. **3**), the request may be received at an intermediate node (e.g., node **196**) between device **170** and device **101**. The intermediate node may be part of the network **195** and it could lie on the path followed by PDU **210**, across network **195**, to its final destination. (e.g., device **170**). The intermediate node may intercept InfiniBand PDU **220** when it is asked to forward the PDU **220** to a next hop in its journey to device **101**. After the PDU **220** is intercepted, it may be further processed by the intermediate node (task **615**) to determine whether it or its payload (e.g., InfiniBand PDU **221**) is suitable to carry the second string **223**.

Upon a positive determination, the intermediate node may include the second string **223** into InfiniBand PDU **221**. The inclusion may be performed by manipulating the payload of PDU **220** in the manner described in the discussion with respect to FIG. **8**. Or alternatively, the inclusion may involve reconstituting InfiniBand PDU **221** from PDU **220**, and possibly other lower-level PDUs, modifying the reconstituted PDU **221** to include the second string **223**, generating PDU **220'** (not shown) which is identical to PDU **220** except for that it includes the second string **223**, and forwarding the PDU **220'** to the next hop in the path to device **101**. Once InfiniBand PDU **221** is reconstituted, the inclusion of the second string **223** into InfiniBand PDU **221** is performed in the manner discussed with respect to the example of FIG. **3**. Stated succinctly, in the alternative example, the second string **223** may be included in InfiniBand PDU **221** by a network element other than an endpoint (e.g., switch, bridge, hub).

Returning to the example of FIG. **3**, at task **615**, it is determined whether InfiniBand PDU **221** is suitable to carry the second string **223** by performing a test to determine whether InfiniBand PDU **221** and InfiniBand PDU **211** are associated with one another. If they are, InfiniBand PDU **221** is deemed suitable to carry the second string **223**. If however the test is failed, the execution of the process **300** is discontinued or task **610** is repeated.

The test may be based on a variety of criteria. For example, InfiniBand PDU **221** may be considered associated with the InfiniBand PDU **211** if it is directed to the sender of the InfiniBand PDU **211** (e.g., the final destination of InfiniBand PDU **221** is the originating node of InfiniBand PDU **211**). Similarly, in some aspects, the InfiniBand PDU **221** may be considered associated with the InfiniBand PDU **211** if the InfiniBand PDU **221** comprises a sequence number that is within a specific distance from the sequence number of the InfiniBand PDU **211** (e.g., the two InfiniBand PDUs are at most 5 sequence number units apart). Furthermore, to InfiniBand PDU **221** may be considered to be associated with a PDU **220** if a recipient and/or a sender of PDU **221** is identified in list of addresses (or other node identifiers) that is stored the memory of device **170** or another node. In one aspect, the list may identify one or more network nodes that are cable of taking a congestion control measure at the transport-layer level of the mixed fabric network which (network **195** and Interfaces **113** and **144** are part of.

FIG. **7** depicts a subtask associated with obtaining an InfiniBand PDU to carry the second string **223** as shown by tasks **410** and **510** of FIGS. **4-5**. In one example, tasks **410** and **510** include generating the InfiniBand PDU **221** (task **710**). The generated protocol data unit may be a response to the InfiniBand PDU **211** that does not carry a data payload (e.g., InfiniBand ACK protocol data unit). In the example depicted in FIG. **3**, the InfiniBand PDU **221** may be generated by device **170**. In another example (not depicted in FIG. **3**), the InfiniBand PDU **221** may be generated by an intermediate node. The intermediate node may be part of the network **195** and it could lie on the path followed by PDU **210**, across network **195**, to its final destination. (e.g., device **170**). In the latter case, InfiniBand PDU **221** may be generated before PDU **210** has reached its final destination (e.g., device **170**). In either case, InfiniBand PDU **221** is generated for the purpose of carrying the second string **223** and it may or may not carry other data.

FIG. **8** depicts a flowchart of the subtasks associated with including a second string that is based on the first string, in a second InfiniBand protocol data unit as shown by task **335** of FIG. **3**. In this example, task **335** includes the following subtasks: selecting PDU **220** that carries an InfiniBand protocol data unit in its payload (**810**), and modifying PDU **220** by inserting a string from the header of PDU **210** into the payload of PDU **220** so as to alter the content of the InfiniBand PDU that is stored in PDU **220**'s payload (**815**).

At task **810**, PDU **220** is selected to carry the second string **223**. In one aspect, PDU **220** may be selected based on carrying at least a part of an InfiniBand protocol data unit that is associated with InfiniBand PDU **211** (e.g., being directed to the sender of PDU **211**). In another aspect, the second layer-3 protocol data unit is selected based on being the first of a sequence of protocol data units among which an InfiniBand protocol data unit that is associated with InfiniBand PDU **211** is fragmented (e.g., selected based on containing the first N bits of InfiniBand PDU **221**).

At task **815**, PDU **220** is modified to include the second string **223**. As noted above, the second string **223** may be either an exact copy or a derivation of the first string **214**. Task **815** may be executed only when the first string **214** indicates the presence of congestion in network **195** or, alternatively, it may be executed regardless of whether the first string **214** indicates the presence of congestion. Once it is generated, the second string **223** may be written over bits in PDU **210**'s payload that correspond to a specific section of InfiniBand PDU **221**. The section may be a specific header (e.g., Backward Explicit Congestion Notification (BECN) bit or another bit in the Base Transport Header (BTH)), a part of the payload of InfiniBand PDU **221** or its padding. For example, the second string **223** may be inserted into the second layer-3 protocol data unit at an offset that corresponds to the location of the header field of InfiniBand PDU **221**. For example, if it is known that the header field starts at the N-th bit from the beginning of the second layer-3 protocol data unit (or its payload portion), the offset at which the string is stored in the second layer-3 protocol data unit will be N. Alternatively, the position of the header may be determined dynamically by processing the payload of PDU **220** to determine where in layer 3-PDU's payload (if at all) is the section of InfiniBand PDU **221** where the second string **223** should be stored. In any event, at task **815**, the second string **223** is included into InfiniBand PDU **221** without InfiniBand PDU **221** being reconstituted from PDU **220** or other lower-level PDUs.

FIG. **9** depicts a subtask associated with taking a congestion control measure as shown by task **350** in FIG. **2**. At task **910**, network interface card (NIC) **141** changes the rate at which PDUs that carry InfiniBand payload are transmitted. The change may affect the transmission rate of PDUs that carry InfiniBand payload, but not the transmission rate of PDUs that carry non-InfiniBand payload (e.g., TCP payload). This disparate treatment may be achieved by maintaining a separate queue for each type of PDU payload and reducing the burst rate at which protocol data units from one of the queues are transmitted. By selectively changing the transmission rate of PDUs that carry InfiniBand payload, the network interface card (NIC) **141** avoids interfering with congestion control measures that may be implemented by other transport-layer networks (e.g., TCP).

In one example, the congestion control measure may be a taken by network interface card (NIC) **141** autonomously from processor **120**. In that example, a controller of network interface card (NIC) **141** (not shown) may manage the queues of packets that carry InfiniBand and non-InfiniBand payload. Furthermore, the controller may manage the burst rates at which PDUs are transmitted from the queues. In that regard, process of FIG. **9** is performed at the NIC level thereby reducing the load placed of processor **120**.

FIGS. **3-9** are provided as an example. Although, in this example the second string **223** is included in InfiniBand PDU **221** by device **170** (which is the final destination of InfiniBand PDU **211**), in other examples the inclusion may be performed by an intermediate node that lies on the path between device **101** and device **170**. Furthermore, in some examples, the second string **223** may be inserted in PDU **221** by a network interface card (NIC). And still furthermore, at least some of the tasks associated with FIGS. **3-9** may be performed in a different order than represented, performed concurrently or altogether omitted.

Furthermore, in some examples, the second string **223** may be inserted in PDU **221** by a network interface card (NIC). The network interface card (NIC) may be configured to implement a mapping function for translating congestion indications from one standard for explicit congestion notification to another. For example, the NIC may be configured to generate values of fields used for explicit congestion notification in InfiniBand based on the value of the CE codepoints of IP packets. According to the mapping function, when the network interface card is part of a router, it may set the FECN bit of the InfiniBand protocol data unit to indicate the congestion. The mapping function may be implemented in software (e.g., by the network card's firmware), in hardware (e.g. FPGA, special purpose circuits), or both in software and hardware.

Although, in the above example PDUs **211** and **221** are InfiniBand type PDUs, they can be any other type of PDU that is cable of being transported as payload in PDUs **210** and **220**. For example, PDUs **221** and **211** may be transport-layer PDUs, application-layer PDUs, session-layer PDUs, and/or presentation-layer PDUs. Similarly, although, interfaces **113** and **144** are InfiniBand interfaces in the above example, in alternative examples they may be interfaces for communications in accordance with another type of protocol that is capable of using network **195** as a data-link-layer and/or network-layer messaging fabric. For instance, interfaces **113** and **144** may be transport-layer protocol interfaces, presentation-layer protocol interfaces, session-layer protocol interfaces, and application-layer protocol interfaces.

As these and other variations and combinations of the features discussed above can be utilized without departing from the subject matter as defined by the claims, the foregoing description of exemplary aspects should be taken by way of illustration rather than by way of limitation of the subject matter as defined by the claims. It will also be understood that the provision of the examples described herein (as well as clauses phrased as "such as," "e.g.", "including" and the like) should not be interpreted as limiting the claimed subject matter to the specific examples; rather, the examples are intended to illustrate only some of many possible aspects.

The invention claimed is:

1. A system comprising:

a transmitter;

a receiver configured to receive, over a communications network, a protocol data unit of a first protocol, the protocol data unit of the first protocol including:

(i) a first header portion containing a first string that indicates whether congestion has been detected in the communications network, the first header portion being used in a first explicit congestion notification standard to indicate congestion, and

(ii) a payload including at least part of a first protocol data unit of a second protocol; and

a processor coupled to the transmitter and the receiver, the processor being configured to:

prepare a second string based on the first string such that the second string is identical to the first string or derived from the first string,

insert the second string in a second header portion of a second protocol data unit of the second protocol, the second header portion not being used in the first explicit congestion notification standard to indicate congestion, and

provide to the transmitter the second protocol data unit of the second protocol for transmission to a sender of the first protocol data unit of the second protocol;

wherein the second header portion is used in a second explicit congestion notification standard to indicate congestion, the second explicit congestion notification standard being different from the first explicit congestion notification standard; and

a network interface card configured to maintain separate transmission queues for each type of protocol data unit payload, and in response to receiving an indication of congestion of the communications network for protocol data units of the first protocol, changing a transmission rate of protocol data units from at least one of the queues corresponding to protocol data unit payloads of the second protocol, the network interface card configured to translate notifications of congestion from the first explicit congestion notification standard to the second explicit congestion notification standard.

2. The system of claim **1**, wherein the first protocol implements the first explicit congestion notification standard and the second protocol implements the second explicit congestion notification standard.

3. The system of claim **1**, further comprising a memory coupled to the processor, the memory storing processor-executable instructions for transferring content of the first header portion into the second header portion.

4. The system of claim **1**, wherein the second string is inserted only when the first string indicates that congestion has been detected in the communications network.

5. The system of claim **1**, wherein:

the first protocol is the Internet Protocol (IP) and the first header portion is the Congestion Experienced (CE) codepoint; and

the second protocol is an InfiniBand protocol and the second header portion includes at least one of the Backward Explicit Congestion Notification (BECN) and the Forward Explicit Congestion Notification (FECN) bits.

6. The system of claim **1**, wherein the first protocol is the Internet Protocol (IP) and the second protocol is an InfiniBand protocol.

7. A method comprising:

receiving over a non-InfiniBand communications network a non-InfiniBand protocol data unit that comprises at least a part of a first InfiniBand protocol data unit, the non-InfiniBand protocol data unit including an indication of congestion of the non-InfiniBand communications network, the indication of congestion being a first string that is part of the non-InfiniBand protocol data unit;

transferring, by a processor, the indication of congestion from the non-InfiniBand protocol data unit to a second InfiniBand protocol data unit, wherein transferring comprises inserting a copy of a first string, or a string derived from the first string, into the header of the second InfiniBand protocol data unit, wherein the first string is part of a header of the non-InfiniBand protocol data unit;

transmitting the second InfiniBand protocol data unit to a sender of the first InfiniBand protocol data unit over the non-InfiniBand communications network;

maintaining separate transmission queues for each type of payload of the protocol data units;

in response to receiving the indication of congestion of the non-InfiniBand communications network for non-InfiniBand protocol data units, changing a transmission rate of protocol data units from at least one of the queues corresponding to payloads of an InfiniBand protocol;

translating the first string from a first format to a second format to produce a translated string; and

inserting the translated string into the second InfiniBand protocol data unit.

8. The method of claim **7**, wherein the non-InfiniBand communications network is an Internet Protocol network and the non-InfiniBand protocol data unit is an Internet Protocol (IP) protocol data unit.

9. The method of claim **8**, wherein the transferring of the indication of the congestion comprises copying a value of at least a part of the Congestion Experienced (CE) codepoint of the Internet Protocol (IP) protocol data unit into the Base Transport header of the second InfiniBand protocol data unit.

10. The method of claim **7**, wherein the transferring comprises inserting a copy of a string that is part of a header of the non-InfiniBand protocol data unit into the header of the second InfiniBand protocol data unit.

11. The method of claim **7** wherein the non-InfiniBand communications network is one of a transport-layer network and a data-link-layer network.

12. A method comprising:

receiving an incoming non-InfiniBand protocol data unit that comprises a payload portion including at least a part of a first InfiniBand protocol data unit, the non-InfiniBand protocol data unit also including a first string that indicates whether congestion is detected in a non-InfiniBand communications network, the first string being part of a header of the non-InfiniBand protocol data unit;

identifying an outgoing non-InfiniBand protocol data unit that comprises a payload portion including at least a part of a second InfiniBand protocol data unit;

when the first string indicates that congestion has been detected, modifying, by a processor, the payload portion of the outgoing non-InfiniBand protocol data unit so as to insert a second string in a header portion of the second InfiniBand protocol data unit, the second string being based on the first string such that the second string is identical to the first string or derived from the first string;

transmitting the outgoing non-InfiniBand protocol data unit;

maintaining separate transmission queues for each type of payload of the protocol data units;

in response to receiving an indication of congestion of the non-InfiniBand communications network for non-InfiniBand protocol data units, changing a transmission rate of protocol data units from at least one of the queues corresponding to payloads of an InfiniBand protocol; and

translating the first string from a first format to a second format to produce a translated string;

wherein the first string is one or more characters long.

13. The method of claim **12**, wherein the modifying is performed when the first string indicates that congestion has not been detected.

14. The method of claim **12**, wherein the first string is the same as the second string.

15. The method of claim **12**, wherein the non-InfiniBand communications network is one of a transport-layer network and a data-link-layer network.

* * * * *