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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
8 September 2017 (08.09.2017)

WIPO | PCT

(10) International Publication Number
WO 2017/149364 A1

(51) International Patent Classification:
H04L 12/709 (2013.01) *H04L 12/939* (2013.01)

(21) International Application Number:
PCT/IB2016/052308

(22) International Filing Date:
22 April 2016 (22.04.2016)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/301,505 29 February 2016 (29.02.2016) US

(71) Applicant: **TELEFONAKTIEBOLAGET LM ERICSSON (PUBL)** [SE/SE]; SE-164 83, Stockholm (SE).

(72) Inventors: **LU, Juan**; 986 La Mesa Ter #C, Sunnyvale, California 94086 (US). **WANG, Sunny**; 20401 Chalet Lane, Saratoga, California 95070 (US).

(74) Agents: **DE VOS, Daniel M.** et al.; Nicholson De Vos Webster & Elliott LLP, 99 Almaden Boulevard, Suite 710, San Jose, CA 95113 (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, BN, BR, BW, BY, BZ, CA, CH, CL, 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, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, 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, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, 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, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

(54) Title: COORDINATED TRAFFIC REROUTE IN AN INTER-CHASSIS REDUNDANCY SYSTEM

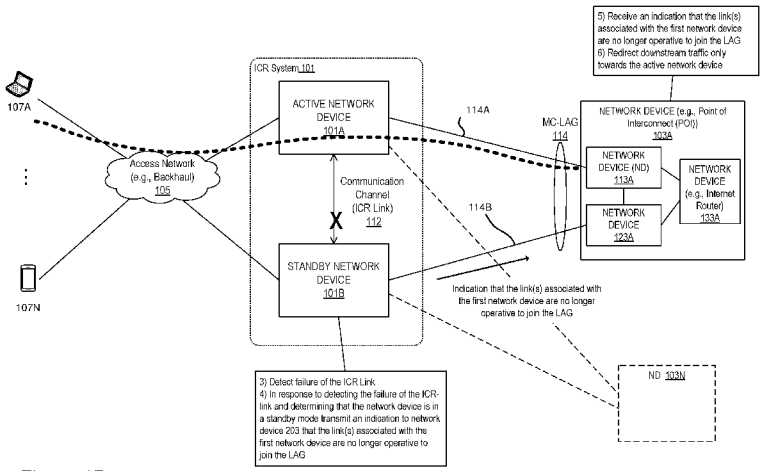


Figure 1B

(57) Abstract: A method and apparatus for enabling traffic reroute in an inter-chassis redundancy (ICR) system are described. A first network device (101B) of the ICR system (101) is coupled with a second network device (101A) of the ICR system (101), and each one of the first and the second network devices is coupled with a third network device (103A) through a multi-chassis link aggregation group (MC-LAG) (114). The first network device (101B) monitors (302) an ICR link (112) coupling the first network device (101B) to the second network device (101A) of the ICR system (101); detects (304) a failure of the ICR link; and in response to detecting the failure of the ICR link, and determining (306) that the first network device (101B) is in a standby mode, transmits (308) to the third network device (103A) an indication that one or more links associated with the first network device (101B) are no longer operative to join the MC-LAG (114).

WO 2017/149364 A1

COORDINATED TRAFFIC REROUTE IN AN INTER-CHASSIS REDUNDANCY SYSTEM

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of U.S. Provisional Application No. 62/301,505, filed on February, 29, 2016, which is hereby incorporated by reference.

TECHNICAL FIELD

[0002] Embodiments of the invention relate to the field of packet networks; and more specifically, to traffic reroute in an ICR system.

BACKGROUND

[0003] In communication networks it is generally desirable to prevent service outages and/or loss of network traffic. By way of example, such service outages and/or loss of network traffic may occur when a network element fails, loses power, is taken offline, is rebooted, a communication link to the network element breaks, etc. In order to help prevent such service outages and/or loss of network traffic, the communication networks may utilize inter-chassis redundancy (ICR). ICR is a high availability (HA) solution that increases the availability of network devices, and may optionally be used to provide geographical redundancy. ICR is commonly implemented through a mated pair of an active network device and a standby network device. The active network device handles current sessions using session state that is built up over runtime. The session data is synchronized or replicated from the active network element to the standby network element. The standby network element begins to handle the sessions when an ICR switchover event occurs.

[0004] The active and standby network devices are coupled with a communication channel referred to as an Inter-Chassis Redundancy (ICR) link. The ICR link enables the two network devices to exchange control messages as well as traffic (i.e., data packets). Under some scenarios, a network administrator may desire to use the resources of the active and standby network devices to forward downstream traffic towards an end user's device. The downstream traffic can be distributed on either one of the network devices of the ICR system. The standby network device can then forward the traffic through the ICR link to the active network device to be processed by active applications prior to being forwarded towards an end user's device. With multiple Next-Hop network devices connections (where multiple next-hop network devices are coupled to each one of the standby and active network devices), this flexibility allows load

balancing of traffic on the two network devices of the ICR system, such that each device can process a portion of the traffic before forwarding it to end user's devices. However, when the ICR link is down, traffic can no longer be transmitted from the standby network device to the active network device resulting in the standby network device dropping data packets.

[0005] Conventional ICR systems, do not support load balancing of downstream traffic over both the active and the standby network devices as only the active network device has an active L2 connections. In conventional ICR systems the ports of the standby network device which are part of a multi-chassis Link Aggregation Group (MC-LAG) coupling the standby network device to the next-hop network device are shut down such that traffic is only forwarded from the next-hop network device to the active network device. In these conventional systems, when a switchover is performed from the active network device to the standby network device (i.e., the standby and active NDs switch roles), the MC-LAG ports of the new standby network device are administratively shut down. This administrative shut down of the MC-LAG ports on the new standby network device is a heavy handed forceful approach that may result to large traffic loss when redirecting the traffic.

SUMMARY

[0006] According to one embodiment, a method in a first network device of an inter-chassis redundancy (ICR) system coupled with a second network device of the ICR system, of enabling traffic reroute in the ICR system, where each one of the first and the second network devices is coupled with a third network device through a multi-chassis link aggregation group (MC-LAG), is described. The method includes monitoring an ICR link coupling the first network device to the second network device of the ICR system; detecting a failure of the ICR link; and in response to detecting the failure of the ICR link, and determining that the first network device is in a standby mode, transmitting to the third network device an indication that one or more links associated with the first network device are no longer operative to join the MC-LAG.

[0007] According to one embodiment, a first network device of an inter-chassis redundancy (ICR) system to be coupled with a second network device of the ICR system, for enabling traffic reroute in the ICR system, where each one of the first and the second network devices is to be coupled with a third network device through a multi-chassis link aggregation group (MC-LAG), is described. The first network device includes one or more processors and a non-transitory computer readable storage medium, said non-transitory computer readable storage medium containing instructions, which when executed by the one or more processors, causes the first network device to monitor an ICR link coupling the first network device to the second network device of the ICR system. The first network device is further to detect a failure of the ICR link;

and in response to detecting the failure of the ICR link, and determining that the first network device is in a standby mode, the first network device is to transmit to the third network device an indication that one or more links associated with the first network device are no longer operative to join the MC-LAG.

[0008] In one embodiment, a non-transitory computer readable storage medium that provide instructions, which when executed by a processor of a first network device of an inter-chassis redundancy (ICR) system to be coupled with a second network device of the ICR system, for enabling traffic reroute in the ICR system, where each one of the first and the second network devices is to be coupled with a third network device through a multi-chassis link aggregation group (MC-LAG), cause said processor to perform operations including: monitoring an ICR link coupling the first network device to the second network device of the ICR system; detecting a failure of the ICR link; and in response to detecting the failure of the ICR link, and to determining that the first network device is in a standby mode transmitting to the third network device an indication that one or more links associated with the first network device are no longer operative to join the MC-LAG.

BRIEF DESCRIPTION OF THE DRAWINGS

[0009] The invention may best be understood by referring to the following description and accompanying drawings that are used to illustrate embodiments of the invention. In the drawings:

[0010] Figures 1A and 1B are block diagrams illustrating an exemplary network including an ICR system according to some embodiments.

[0011] Figure 1C illustrates a flow diagram of exemplary operations performed in a network device of an ICR system to enable traffic reroute when failure of an ICR Link coupling a first ND and a second ND of the ICR system in accordance with some embodiments.

[0012] Figure 1D illustrates a transactions diagram of detailed exemplary operations performed in a network device of an ICR system to enable traffic reroute when failure of an ICR link occurs in accordance with some embodiments.

[0013] Figure 2A illustrate a block diagram of an exemplary network including the ICR system for enabling traffic reroute when an ICR Link resumes operation in accordance with some embodiments.

[0014] Figure 2B illustrates a transactions diagram of detailed exemplary operations performed in a network device of an ICR system to enable traffic reroute in accordance with some embodiments.

[0015] Figure 2C illustrates a flow diagram of exemplary operations performed in a network device of an ICR system to enable traffic reroute when an ICR Link resumes operation in accordance with some embodiments.

[0016] Figure 3A illustrates connectivity between network devices (NDs) within an exemplary network, as well as three exemplary implementations of the NDs, according to some embodiments of the invention.

[0017] Figure 3B illustrates an exemplary way to implement a special-purpose network device according to some embodiments of the invention.

[0018] Figure 3C illustrates various exemplary ways in which virtual network elements (VNEs) may be coupled according to some embodiments of the invention.

[0019] Figure 3D illustrates a network with a single network element (NE) on each of the NDs, and within this straight forward approach contrasts a traditional distributed approach (commonly used by traditional routers) with a centralized approach for maintaining reachability and forwarding information (also called network control), according to some embodiments of the invention.

[0020] Figure 3E illustrates the simple case of where each of the NDs implements a single NE, but a centralized control plane has abstracted multiple of the NEs in different NDs into (to represent) a single NE in one of the virtual network(s), according to some embodiments of the invention.

[0021] Figure 3F illustrates a case where multiple VNEs are implemented on different NDs and are coupled to each other, and where a centralized control plane has abstracted these multiple VNEs such that they appear as a single VNE within one of the virtual networks, according to some embodiments of the invention.

[0022] Figure 4 illustrates a general purpose control plane device with centralized control plane (CCP) software 650), according to some embodiments of the invention.

DETAILED DESCRIPTION

[0023] The following description describes methods and apparatus for enabling traffic reroute in an ICR system. In the following description, numerous specific details such as logic implementations, opcodes, means to specify operands, resource partitioning/sharing/duplication implementations, types and interrelationships of system components, and logic partitioning/integration choices are set forth in order to provide a more thorough understanding of the present invention. It will be appreciated, however, by one skilled in the art that the invention may be practiced without such specific details. In other instances, control structures, gate level circuits and full software instruction sequences have not been shown in detail in order

not to obscure the invention. Those of ordinary skill in the art, with the included descriptions, will be able to implement appropriate functionality without undue experimentation.

[0024] References in the specification to "one embodiment," "an embodiment," "an example embodiment," etc., indicate that the embodiment described may include a particular feature, structure, or characteristic, but every embodiment may not necessarily include the particular feature, structure, or characteristic. Moreover, such phrases are not necessarily referring to the same embodiment. Further, when a particular feature, structure, or characteristic is described in connection with an embodiment, it is submitted that it is within the knowledge of one skilled in the art to affect such feature, structure, or characteristic in connection with other embodiments whether or not explicitly described.

[0025] Bracketed text and blocks with dashed borders (e.g., large dashes, small dashes, dot-dash, and dots) may be used herein to illustrate optional operations that add additional features to embodiments of the invention. However, such notation should not be taken to mean that these are the only options or optional operations, and/or that blocks with solid borders are not optional in certain embodiments of the invention.

[0026] In the following description and claims, the terms "coupled" and "connected," along with their derivatives, may be used. It should be understood that these terms are not intended as synonyms for each other. "Coupled" is used to indicate that two or more elements, which may or may not be in direct physical or electrical contact with each other, co-operate or interact with each other. "Connected" is used to indicate the establishment of communication between two or more elements that are coupled with each other.

[0027] An electronic device stores and transmits (internally and/or with other electronic devices over a network) code (which is composed of software instructions and which is sometimes referred to as computer program code or a computer program) and/or data using machine-readable media (also called computer-readable media), such as machine-readable storage media (e.g., magnetic disks, optical disks, read only memory (ROM), flash memory devices, phase change memory) and machine-readable transmission media (also called a carrier) (e.g., electrical, optical, radio, acoustical or other form of propagated signals – such as carrier waves, infrared signals). Thus, an electronic device (e.g., a computer) includes hardware and software, such as a set of one or more processors coupled to one or more machine-readable storage media to store code for execution on the set of processors and/or to store data. For instance, an electronic device may include non-volatile memory containing the code since the non-volatile memory can persist code/data even when the electronic device is turned off (when power is removed), and while the electronic device is turned on that part of the code that is to be executed by the processor(s) of that electronic device is typically copied from the slower non-

volatile memory into volatile memory (e.g., dynamic random access memory (DRAM), static random access memory (SRAM)) of that electronic device. Typical electronic devices also include a set or one or more physical network interface(s) to establish network connections (to transmit and/or receive code and/or data using propagating signals) with other electronic devices. One or more parts of an embodiment of the invention may be implemented using different combinations of software, firmware, and/or hardware.

[0028] A network device (ND) is an electronic device that communicatively interconnects other electronic devices on the network (e.g., other network devices, end-user devices). Some network devices are "multiple services network devices" that provide support for multiple networking functions (e.g., routing, bridging, switching, Layer 2 aggregation, session border control, Quality of Service, and/or subscriber management), and/or provide support for multiple application services (e.g., data, voice, and video).

[0029] A method and apparatus for enabling traffic reroute in an Inter-Chassis Redundancy (ICR) System is described. In one embodiment of the invention, a first network device of an inter-chassis redundancy (ICR) system coupled with a second network device of the ICR system, of enabling traffic reroute in the ICR system is described. Each one of the first and the second devices is coupled with a third network device through a Multi-Chassis Link Aggregation Group (MC-LAG). The first network device is operative to monitor an inter-chassis redundancy (ICR) link coupling the first network device to the second network device of the ICR system; and to detect a failure of the ICR link. In response to detecting the failure of the ICR link, and to determining that the first network device is in a standby mode, the first network device is operative to transmit to the third network device an indication that one or more links associated with the first network device are no longer operative to join the MC-LAG. The third network device is operative to load balance downstream traffic to the first network device and the second network device, and transmitting the indication that one or more links associated with the first network device are no longer operative to join the MC-LAG causes the third network device to redirect downstream traffic only towards the second network device without loss or interruption of traffic.

[0030] Figures 1A and 1B are block diagrams illustrating an exemplary network including an ICR system 101 according to some embodiments. Referring first to Figure 1A, network 100 includes, but is not limited to, one or more end users' devices 107A-N. End users' devices are electronic devices enabling one or more users to transmit and receive data packets through a network. Examples of suitable subscriber end stations include, but are not limited to, servers, workstations, laptops, netbooks, palm tops, mobile phones, smartphones, multimedia phones, tablets, phablets, Voice Over Internet Protocol (VOIP) phones, user equipment, terminals,

portable media players, GPS units, gaming systems, set-top boxes, and combinations thereof. End users' devices 107A-N access content/services provided over the Internet and/or content/services provided on virtual private networks (VPNs) overlaid on (e.g., tunneled through) the Internet. The content and/or services are typically provided by one or more provider end stations (e.g., server end stations) belonging to a service or content provider. Examples of such content and/or services include, but are not limited to, public webpages (e.g., free content, store fronts, search services), private webpages (e.g., username/password accessed webpages providing email services), and/or corporate networks over VPNs, etc.

[0031] As illustrated, end users' devices 107A-N are communicatively coupled (e.g., through a wired and/or wireless customer premise equipment) to access network devices (not illustrated) of access network 105 (e.g., a backhaul network). Access network devices can be communicatively coupled to provider edge network devices (e.g., network devices 101A and 101B) of the ICR system 101. The provider edge network devices may be communicatively coupled through one or more core network devices 103A to one or more provider end stations (not illustrated). In some cases, the provider edge network devices of provider edge network 101 may host on the order of thousands to millions of wire line type and/or wireless end users' devices, although the scope of the invention is not limited to any known number.

[0032] In one embodiment, network devices 101A and 101B form inter-chassis redundancy (ICR) system/cluster 101. In an ICR system, there are typically two ICR devices. There may, however, be more than two ICR devices in an ICR system. During normal operation, one ICR device is configured to be in an active state (herein referred to as active network device (ND)) while the other is configured to be in a standby state (herein referred to as standby ND). The active ND is responsible for handling network traffic with a plurality of other network devices (e.g., end users' devices 107A-N), including, for example, allocating Internet Protocol (IP) addresses to such devices. During an ICR switchover (herein referred to simply as a "switchover"), the active and standby network devices switch roles (e.g., the active network device becomes the standby network device, and the standby network device becomes the active network device). Figures 1A-1B illustrate that network devices 101A and 101B are configured to serve as the active and standby device of ICR system 101, respectively. In some embodiments, Virtual Router Redundancy Protocol (VRRP) can be implemented on each one of the network devices of the ICR system 101 and define a VRRP group. In other embodiments, another redundancy protocol can be used without departing from the scope of the present embodiments.

[0033] Each one of ND 101A and ND 101B is connected using a Multi-Chassis Link Aggregation Group (MC-LAG) 114 to the same network device 103A of a core network. In

some embodiments, network device 103A can be a redundant (virtual) next-hop network device (e.g., a Multi-Chassis (MC) LAG master). Each one of ND 101A and ND 101B is connected using a Multi-Chassis Link Aggregation Group (MC-LAG) 114 to a third network device (which can be referred to as a point of interconnect (POI)) 103A. ND 103A may include one or more network devices and is operative to couple the PE with application servers. In some embodiments, ND 103A is used to shield and provide resiliency to the application servers. In order to provide better resiliency, Multi-Chassis Link Aggregation Group (MC-LAG) (e.g., MC-LAG 114 which includes links 114A and 114B) is configured to connect ND 103A to ND 101A and 101B. While the present embodiments will be described with reference to ND 103A including a redundant system (e.g., network device 113A and 123A) for providing increased availability; in other embodiments, ND 103A may include a single network device to which each of the network devices ND 101A and 101B are connected. The operations described herein will still apply in this scenario. Furthermore, each of the NDs 101A and 101B may be coupled to more than one points of interconnect devices (e.g., N number of POIs 103A to 103N).

[0034] A "LAG" comprises multiple links directly connecting two network devices, and a load distribution decision across these different link paths is performed at the network device forwarding plane. A "MC-LAG" refers to a LAG that directly connects one network device with two or more other network devices. In the illustrated example, MC-LAG 114 directly connects network device 103A with network devices 101A and 101B, via links 114A and 114B, respectively. While Figures 1A and 1B illustrate a single link 114A coupling network device 103A and active ND 101A, and a single link 114B coupling ND 103A and standby ND 101B, other embodiments may include more than one link coupling the respective NDs, where these links are part of the MC-LAGs.

[0035] The two network devices ND 101A and ND 101B are coupled to each other through a communication channel referred to herein as an inter-chassis redundancy link 112. The ICR link enables the two network devices to exchange control messages as well as traffic (i.e., data packets). In some embodiments, the block diagram of Figure 1 illustrates the distribution of downstream traffic over the two network devices of the ICR system 101. The downstream traffic (downstream data packets transmitted from the core network to the end users' devices 107A-N) is distributed on either one of the network devices of the ICR system 101. The standby ND 101B forwards the traffic through the ICR link to the active ND 101A to be forwarded towards a user's device (e.g., user's device 107A).

[0036] At operation 1, the network device 103A is operative to distribute downstream data packets destined to a user's device (e.g., 107A) over the active ND 101A and the standby ND 101B. The downstream data packets are forwarded through the link 114A and 114B respectively

coupling ND 103A with active ND 101A and standby ND 101B. The downstream data packets follow the paths 115A, and 115B to reach the user's device 107A. At operation 2, the standby ND 101B monitors the ICR Link 112 coupling the standby network device with the active ND 101A.

[0037] In some embodiments, a link aggregation protocol can be used to manage and operate the MC-LAG 114. For example, Link Aggregation Control Protocol (LACP) which is part of the IEEE specification 802.3ad (which included herein by reference) allows to bundle several physical ports to form a single logical channel. Per the LACP protocol, when an actor link joins the LAG and is ready to carry traffic, an "IN-SYNC" bit is set in a LACP protocol data unit (PDU) packet. The IN-SYNC bit is an indication that the actor link has joined the LAG and is now operative to receive traffic. The LACP PDU is then transmitted to the network device 103. Upon receipt of this packet, the network device 103A can transmit downstream data packets to the standby ND 101B. While the embodiments are described with respect to the LACP protocol, one of ordinary skill in the art will recognize that other link aggregation protocol may be used without departing from the spirit of the present invention.

[0038] Referring now to Figure 1B, when the ICR link 112 fails, a reroute of the downstream traffic towards the active ND 101A is performed. Thus at operation 3, the standby detects the failure of the ICR Link. At operation 4, in response to detecting the failure of the ICR-link and determining that the network device ND 101B is in a standby mode, ND 101B transmits an indication to network device 103A that the link associated with ND 101B and which are part of the MC-LAG are no longer operative to join the MC-LAG. At operation 5, ND 103A receives the indication that the link 114B associated with the ND 101B and which is part of the MC-LAG 114 is no longer operative to join the MC-LAG. ND 103A redirects downstream traffic, at operation 5, only towards the active network device 101A.

[0039] For example, in the embodiments where LACP is used, upon detection of the failure, the standby ND 101B resets the appropriate LAG's LACP PDU packets from "IN-SYNC" bit to "OUT-SYNC" bit. ND 103A receives the LACP PDU packet with the "OUT-SYNC" bits set and an MC-LAG manager of ND 103A will know that its MC-LAG partner has lost the ability to carry traffic. ND 103A stops sending traffic to the link for which the LACP PDU packet was received (where the PDU packet included the "OUT-SYNC" bit), and switches the traffic to links connected to the active ND 101A.

[0040] The operations in the flow diagrams will be described with reference to the exemplary embodiments of the other figures. However, it should be understood that the operations of the flow diagrams can be performed by embodiments of the invention other than those discussed with reference to the other figures, and the embodiments of the invention discussed with

reference to these other figures can perform operations different than those discussed with reference to the flow diagrams.

[0041] Figure 1C illustrates a flow diagram of exemplary operations 300 performed in a network device of an ICR system to enable traffic reroute when failure of an ICR Link coupling a first ND and a second ND of the ICR system in accordance with some embodiments. In some embodiments the operations 300 of Figure 3A are performed at network device 101A when the network device is in a standby mode with respect to the ICR system. In some embodiments both ND 101A and 101B are configured to be operative to perform the operations 300, however these operations are only performed when the state of the ND is in a standby mode. Thus, even though the operations are described with respect to ND 103A, when the two network devices switch roles within the ICR system (i.e., ND 101A becomes standby and ND 101B becomes active), the operations 300 will then be performed by the standby network device.

[0042] At operation 302, ND 101A monitors the ICR Link coupling ND 101B with ND 101A of the ICR system 101. Flow then moves to operation 304, at which ND 101B detects a failure of the ICR link, and determines at operation 306 that it is in a standby mode. Upon determining that a failure of the ICR link has occurred and that the ND is in a standby mode, flow then moves to operation 308, at ND 101B transmits to the third network device (ND 103A) an indication that one or more links associated with the ND 101B are no longer operative to join the LAG. In some embodiments, the indication that ND 101B is no longer operative to receive traffic is transmitted according to the LACP protocol. LACP PDU packets with an "OUT-SYNC" bit set are transmitted to ND 103A, informing the network device that these links are no longer operative to receive traffic.

[0043] The receipt of the indication that one or more links associated with the ND 101B are no longer operative to join the MC-LAG cause ND 103A to reroute downstream traffic towards remaining active links of the MC-LAG (which are the links associated with the active ND 101A). Thus upon detection of a fault in the ICR Link the traffic is rerouted towards the active network device without any interruption or loss of traffic.

[0044] Figure 1D illustrates a transactions diagram of detailed exemplary operations performed in a network device of an ICR system to enable traffic reroute when failure of an ICR link occurs in accordance with some embodiments. While Figure 1D will be described with reference to standby network device 101B ND, the components illustrated in Figure 1D are also present in other network devices of the ICR system 101 (e.g., ND 101A). ND 101B includes a reroute manager 210, which includes an ICR manager 212 and a LAG manager 214. The ICR manager 212 is operative to manage the ICR system and enabled ND 101B to communicate with ND 101A through the ICR link 112. The LAG manager 214 is operative to enable

communication links (including ports of ND 101B) to join or leave the MC-LAG 114. The LAG manager 214 is operative to communicate through the link 114A with a MC-LAG manager of ND 103A. The Reroute manager 210 may be implemented as described in further details with reference to the Reroute Manager 521 or 551 of Figure 3A below.

[0045] At operation 302, the ICR manager monitors the ICR Link coupling ND 101B with ND 101A of the ICR system 101. In some embodiments, the monitoring of the ICR link can be performed with according to Bidirectional Forwarding Detection (BFD) protocol. For example, the ICR manager 212 monitors the status of the ICR link interface by subscribing to BFD events on that interface. In one embodiment, a single BFD session is created on the interface and the BFD tracking object is created.

[0046] At operation 304, ND 101B detects failure of the ICR link, and determines at operation 306 that it is in a standby mode. Upon determining that a failure of the ICR link has occurred and that the ND is in a standby mode, the ICR manager 212 requests from LAG manager 214 (at operation 307) to stop forwarding traffic received from the links (e.g., link 114B) of the MC-LAG 114. For example, ICR manager 112A may transmit the request through an inter process communication messaging mechanism to the LAG manager 214 requesting the LAG manager to stop carrying traffic for these links.

[0047] Flow then moves to operation 308, at which LAG manager 214 upon receipt of the request from ICR manager 212 transmits to the third network device (ND 103A) an indication that one or more links associated with the ND 101B are no longer operative to join the LAG. In some embodiments, the LAG manager identifies at operation 314, active links from the MC-LAG which are connected with ND 103A. For example, the LAG manager 214 may manage multiple MC-LAGs coupling the ND 101B with one or more network devices. In these embodiments, LAG manager may need to identify which links are part of MC-LAG 114 and are in an active state receiving traffic from ND 103A. Upon identification of these links flow then moves to operation 316 at which LAG manager 214 sets the identified links to be in a temporary halted state, indicating that these links are no longer operative to receive traffic from ND 103A. Flow then moves to operation 318 at which LAG manager 214 send to its MC-LAG partners (i.e., ND 103A), LACP protocol data unit (PDU) packets with an "OUT-SYNC" bit set, informing ND 103A that these links are no longer operative to receive traffic.

[0048] Upon receipt of the LACP PDU packets with "OUT-SYNC" bit set, ND 103A stops sending downstream data packets and reroute the downstream traffic towards the remaining active links of the MC-LAG 114, which includes the link 114A. Thus upon detection of a fault in the ICR Link the traffic is rerouted towards the active network device without any interruption or loss of traffic.

[0049] Thus by tracking failure of the ICR Link, enabling the signaling of the failure to the LAG manager (i.e., the control element of the Link Aggregation Group), and transmitting an indication to the LAG partner network device, the present embodiments cause the master of the MC-LAG to stop load balancing packets over a standby and active ND of the ICR system and reroutes all traffic towards the active ND. Thus with this mechanism, the ICR system is provided with the flexibility of load balancing traffic over the two network devices of the system while enabling a quick and reliable reroute of the traffic when the ICR link fails.

[0050] Figure 2A illustrate a block diagram of an exemplary network including the ICR system 101 for enabling traffic reroute when an ICR Link resumes operation in accordance with some embodiments. The operations described below occur within the ICR system following a failure of the ICR Link and its recovery at a later time. Thus in some embodiments, the operations 7-9 are performed following the operations 1-6 of Figure 1A-B. At operation 7, ND 101A detects that ICR Link 112 has resumed normal operation (e.g., ND 101B may detect that the ICR Link has recovered from the failure and is now operational to forward traffic towards ND 101A through the use of BFD protocol). At operation 8, upon detection that ICR Link has resumed operation, ND 101B transmit to the ND 103A an indication that one or more links associated with ND 101A are operative to join the LAG. Upon receipt of the indication ND 103A starts at operation 9, to distribute downstream data packets over the active network device and the standby network device. Figure 2B illustrates a transactions diagram of detailed exemplary operations performed in a network device of an ICR system to enable traffic reroute in accordance with some embodiments. At operation 302, ND 101A continues to monitor the ICR Link 112. At operation 404, ND 101B detects that the ICR link has resumed normal operations; and determines that it is in a standby mode (operation 406). Upon determination that the ICR link has recovered from the failure and that the ND is in a standby mode, ND 101B transmits, at operation 408, a second indication that one or more links associated with ND 101B (which are part of the MC-Lag) are operative to join the MC-LAG.

[0051] Figure 2C illustrates a flow diagram of exemplary operations performed in a network device of an ICR system to enable traffic reroute when an ICR Link resumes operation in accordance with some embodiments. While Figure 2C will be described with reference to standby network device 101B ND, the components illustrated in Figure 2C are also present in other network devices of the ICR system 101 (e.g., ND 101A).

[0052] At operation 402, the ICR manager monitors the ICR Link coupling ND 101B with ND 101A of the ICR system 101. In some embodiments, the monitoring of the ICR link can be performed with according to Bidirectional Forwarding Detection (BFD) protocol. For example, the ICR manager 212 monitors the status of the ICR link interface by subscribing to BFD events

on that interface. In one embodiment, a single BFD session is created on the interface and the BFD tracking object is created.

[0053] At operation 404, ND 101B detects that the ICR link has recovered from the failure which occurred previously, and determines at operation 406 that it is in a standby mode. Upon determining that the ICR link has resumed normal operations and recovered from failure, and that the ND is in a standby mode, the ICR manager 212 requests from LAG manager 214 (at operation 407) to resume forwarding traffic through the links (e.g., link 114B) of the MC-LAG 114. For example, ICR manager 112A may transmit the request through an inter process communication messaging mechanism to the LAG manager 214 requesting the LAG manager to resume carrying traffic for these links.

[0054] Flow then moves to operation 408, at which LAG manager 214 upon receipt of the request from ICR manager 212 transmits to the third network device (ND 103A) an indication that one or more links associated with the ND 101B are operative to join the LAG. In some embodiments, the LAG manager identifies at operation 414, active links from the MC-LAG which are connected with ND 103A. For example, the LAG manager 214 may manage multiple MC-LAGs coupling the ND 101B with one or more network devices. In these embodiments, LAG manager may need to identify which links are part of MC-LAG 114 and are in a temporary halted state. Upon identification of these links flow then moves to operation 416 at which LAG manager 214 sets the identified links to be in an active state, indicating that these links are now operative to receive traffic from ND 103A. Flow then moves to operation 418 at which LAG manager 214 sends to its MC-LAG partners (i.e., ND 103A), LACP protocol data unit (PDU) packets with an "IN-SYNC" bit set, informing ND 103A that these links are now operative to receive traffic.

[0055] Upon receipt of the LACP PDU packets with "IN-SYNC" bit set, ND 103A reroute the downstream traffic towards the two sets of active links of the MC-LAG 114, which includes the links 114A and 114B. Thus upon detection that the ICR Link is operational the traffic is rerouted towards the active network device as well as the standby network device without any interruption or loss of traffic.

[0056] Thus by tracking failure of the ICR Link, enabling the signaling of the failure to the LAG manager (i.e., the control element of the Link Aggregation Group), and transmitting an indication to the LAG partner network device, the present embodiments cause the master of the MC-LAG to stop load balancing packets over a standby and active ND of the ICR system and reroutes all traffic towards the active ND when a failure occurs while enabling a load balancing of the downstream traffic over both NDs of the ICR system when the link is operational. Thus with this mechanism, the ICR system is provided with the flexibility of load balancing traffic

over the two network devices of the system while enabling a dynamic and reliable reroute of the traffic when the ICR link fails.

[0057] Figure 3A illustrates connectivity between network devices (NDs) within an exemplary network, as well as three exemplary implementations of the NDs, according to some embodiments of the invention. Figure 3A shows NDs 500A-H, and their connectivity by way of lines between 500A-500B, 500B-500C, 500C-500D, 500D-500E, 500E-500F, 500F-500G, and 500A-500G, as well as between 500H and each of 500A, 500C, 500D, and 500G. These NDs are physical devices, and the connectivity between these NDs can be wireless or wired (often referred to as a link). An additional line extending from NDs 500A, 500E, and 500F illustrates that these NDs act as ingress and egress points for the network (and thus, these NDs are sometimes referred to as edge NDs; while the other NDs may be called core NDs).

[0058] Two of the exemplary ND implementations in Figure 3A are: 1) a special-purpose network device 502 that uses custom application–specific integrated–circuits (ASICs) and a special-purpose operating system (OS); and 2) a general purpose network device 504 that uses common off-the-shelf (COTS) processors and a standard OS.

[0059] The special-purpose network device 502 includes networking hardware 510 comprising compute resource(s) 512 (which typically include a set of one or more processors), forwarding resource(s) 514 (which typically include one or more ASICs and/or network processors), and physical network interfaces (NIs) 516 (sometimes called physical ports), as well as non-transitory machine readable storage media 518 having stored therein networking software 520. A physical NI is hardware in a ND through which a network connection (e.g., wirelessly through a wireless network interface controller (WNIC) or through plugging in a cable to a physical port connected to a network interface controller (NIC)) is made, such as those shown by the connectivity between NDs 500A-H. During operation, the networking software 520 may be executed by the networking hardware 510 to instantiate a set of one or more networking software instance(s) 522. Network software 520 further includes Reroute Manager 521, which when executed on hardware 510 enables the network device 502 to perform the operations described with reference to Figures 1A-2C. Each of the networking software instance(s) 522, and that part of the networking hardware 510 that executes that network software instance (be it hardware dedicated to that networking software instance and/or time slices of hardware temporally shared by that networking software instance with others of the networking software instance(s) 522), form a separate virtual network element 530A-R. Each of the virtual network element(s) (VNEs) 530A-R includes a control communication and configuration module 532A-R (sometimes referred to as a local control module or control communication module) and forwarding table(s) 534A-R, such that a given virtual network element (e.g., 530A) includes the

control communication and configuration module (e.g., 532A), a set of one or more forwarding table(s) (e.g., 534A), and that portion of the networking hardware 510 that executes the virtual network element (e.g., 530A).

[0060] The special-purpose network device 502 is often physically and/or logically considered to include: 1) a ND control plane 524 (sometimes referred to as a control plane) comprising the compute resource(s) 512 that execute the control communication and configuration module(s) 532A-R; and 2) a ND forwarding plane 526 (sometimes referred to as a forwarding plane, a data plane, or a media plane) comprising the forwarding resource(s) 514 that utilize the forwarding table(s) 534A-R and the physical NIs 516. By way of example, where the ND is a router (or is implementing routing functionality), the ND control plane 524 (the compute resource(s) 512 executing the control communication and configuration module(s) 532A-R) is typically responsible for participating in controlling how data (e.g., packets) is to be routed (e.g., the next hop for the data and the outgoing physical NI for that data) and storing that routing information in the forwarding table(s) 534A-R, and the ND forwarding plane 526 is responsible for receiving that data on the physical NIs 516 and forwarding that data out the appropriate ones of the physical NIs 516 based on the forwarding table(s) 534A-R.

[0061] Figure 3B illustrates an exemplary way to implement the special-purpose network device 502 according to some embodiments of the invention. Figure 3B shows a special-purpose network device including cards 538 (typically hot pluggable). While in some embodiments the cards 538 are of two types (one or more that operate as the ND forwarding plane 526 (sometimes called line cards), and one or more that operate to implement the ND control plane 524 (sometimes called control cards)), alternative embodiments may combine functionality onto a single card and/or include additional card types (e.g., one additional type of card is called a service card, resource card, or multi-application card). A service card can provide specialized processing (e.g., Layer 4 to Layer 7 services (e.g., firewall, Internet Protocol Security (IPsec), Secure Sockets Layer (SSL) / Transport Layer Security (TLS), Intrusion Detection System (IDS), peer-to-peer (P2P), Voice over IP (VoIP) Session Border Controller, Mobile Wireless Gateways (Gateway General Packet Radio Service (GPRS) Support Node (GGSN), Evolved Packet Core (EPC) Gateway)). By way of example, a service card may be used to terminate IPsec tunnels and execute the attendant authentication and encryption algorithms. These cards are coupled together through one or more interconnect mechanisms illustrated as backplane 536 (e.g., a first full mesh coupling the line cards and a second full mesh coupling all of the cards).

[0062] Returning to Figure 3A, the general purpose network device 504 includes hardware 540 comprising a set of one or more processor(s) 542 (which are often COTS processors) and

network interface controller(s) 544 (NICs; also known as network interface cards) (which include physical NIs 546), as well as non-transitory machine readable storage media 548 having stored therein software 550. During operation, the processor(s) 542 execute the software 550 to instantiate one or more sets of one or more applications 564A-R. Network software 550 further includes Reroute Manager 551, which when executed on hardware 540 enables the network device 504 to perform the operations described with reference to Figures 1A-2C. While one embodiment does not implement virtualization, alternative embodiments may use different forms of virtualization. For example, in one such alternative embodiment the virtualization layer 554 represents the kernel of an operating system (or a shim executing on a base operating system) that allows for the creation of multiple instances 562A-R called software containers that may each be used to execute one (or more) of the sets of applications 564A-R; where the multiple software containers (also called virtualization engines, virtual private servers, or jails) are user spaces (typically a virtual memory space) that are separate from each other and separate from the kernel space in which the operating system is run; and where the set of applications running in a given user space, unless explicitly allowed, cannot access the memory of the other processes. In another such alternative embodiment the virtualization layer 554 represents a hypervisor (sometimes referred to as a virtual machine monitor (VMM)) or a hypervisor executing on top of a host operating system, and each of the sets of applications 564A-R is run on top of a guest operating system within an instance 562A-R called a virtual machine (which may in some cases be considered a tightly isolated form of software container) that is run on top of the hypervisor - the guest operating system and application may not know they are running on a virtual machine as opposed to running on a "bare metal" host electronic device, or through para-virtualization the operating system and/or application may be aware of the presence of virtualization for optimization purposes. In yet other alternative embodiments, one, some or all of the applications are implemented as unikernel(s), which can be generated by compiling directly with an application only a limited set of libraries (e.g., from a library operating system (LibOS) including drivers/libraries of OS services) that provide the particular OS services needed by the application. As a unikernel can be implemented to run directly on hardware 540, directly on a hypervisor (in which case the unikernel is sometimes described as running within a LibOS virtual machine), or in a software container, embodiments can be implemented fully with unikernels running directly on a hypervisor represented by virtualization layer 554, unikernels running within software containers represented by instances 562A-R, or as a combination of unikernels and the above-described techniques (e.g., unikernels and virtual machines both run directly on a hypervisor, unikernels and sets of applications that are run in different software containers).

[0063] The instantiation of the one or more sets of one or more applications 564A-R, as well as virtualization if implemented, are collectively referred to as software instance(s) 552. Each set of applications 564A-R, corresponding virtualization construct (e.g., instance 562A-R) if implemented, and that part of the hardware 540 that executes them (be it hardware dedicated to that execution and/or time slices of hardware temporally shared), forms a separate virtual network element(s) 560A-R.

[0064] The virtual network element(s) 560A-R perform similar functionality to the virtual network element(s) 530A-R - e.g., similar to the control communication and configuration module(s) 532A and forwarding table(s) 534A (this virtualization of the hardware 540 is sometimes referred to as network function virtualization (NFV)). Thus, NFV may be used to consolidate many network equipment types onto industry standard high volume server hardware, physical switches, and physical storage, which could be located in Data centers, NDs, and customer premise equipment (CPE). While embodiments of the invention are illustrated with each instance 562A-R corresponding to one VNE 560A-R, alternative embodiments may implement this correspondence at a finer level granularity (e.g., line card virtual machines virtualize line cards, control card virtual machine virtualize control cards, etc.); it should be understood that the techniques described herein with reference to a correspondence of instances 562A-R to VNEs also apply to embodiments where such a finer level of granularity and/or unikernels are used.

[0065] In certain embodiments, the virtualization layer 554 includes a virtual switch that provides similar forwarding services as a physical Ethernet switch. Specifically, this virtual switch forwards traffic between instances 562A-R and the NIC(s) 544, as well as optionally between the instances 562A-R; in addition, this virtual switch may enforce network isolation between the VNEs 560A-R that by policy are not permitted to communicate with each other (e.g., by honoring virtual local area networks (VLANs)).

[0066] The third exemplary ND implementation in Figure 3A is a hybrid network device 506, which includes both custom ASICs/special-purpose OS and COTS processors/standard OS in a single ND or a single card within an ND. In certain embodiments of such a hybrid network device, a platform VM (i.e., a VM that that implements the functionality of the special-purpose network device 502) could provide for para-virtualization to the networking hardware present in the hybrid network device 506.

[0067] Regardless of the above exemplary implementations of an ND, when a single one of multiple VNEs implemented by an ND is being considered (e.g., only one of the VNEs is part of a given virtual network) or where only a single VNE is currently being implemented by an ND, the shortened term network element (NE) is sometimes used to refer to that VNE. Also in all of

the above exemplary implementations, each of the VNEs (e.g., VNE(s) 530A-R, VNEs 560A-R, and those in the hybrid network device 506) receives data on the physical NIs (e.g., 516, 546) and forwards that data out the appropriate ones of the physical NIs (e.g., 516, 546). For example, a VNE implementing IP router functionality forwards IP packets on the basis of some of the IP header information in the IP packet; where IP header information includes source IP address, destination IP address, source port, destination port (where "source port" and "destination port" refer herein to protocol ports, as opposed to physical ports of a ND), transport protocol (e.g., user datagram protocol (UDP), Transmission Control Protocol (TCP), and differentiated services code point (DSCP) values.

[0068] Figure 3C illustrates various exemplary ways in which VNEs may be coupled according to some embodiments of the invention. Figure 3C shows VNEs 570A.1-570A.P (and optionally VNEs 570A.Q-570A.R) implemented in ND 500A and VNE 570H.1 in ND 500H. In Figure 3C, VNEs 570A.1-P are separate from each other in the sense that they can receive packets from outside ND 500A and forward packets outside of ND 500A; VNE 570A.1 is coupled with VNE 570H.1, and thus they communicate packets between their respective NDs; VNE 570A.2-570A.3 may optionally forward packets between themselves without forwarding them outside of the ND 500A; and VNE 570A.P may optionally be the first in a chain of VNEs that includes VNE 570A.Q followed by VNE 570A.R (this is sometimes referred to as dynamic service chaining, where each of the VNEs in the series of VNEs provides a different service – e.g., one or more layer 4-7 network services). While Figure 3C illustrates various exemplary relationships between the VNEs, alternative embodiments may support other relationships (e.g., more/fewer VNEs, more/fewer dynamic service chains, multiple different dynamic service chains with some common VNEs and some different VNEs).

[0069] The NDs of Figure 3A, for example, may form part of the Internet or a private network; and other electronic devices (not shown; such as end user devices including workstations, laptops, netbooks, tablets, palm tops, mobile phones, smartphones, phablets, multimedia phones, Voice Over Internet Protocol (VOIP) phones, terminals, portable media players, GPS units, wearable devices, gaming systems, set-top boxes, Internet enabled household appliances) may be coupled to the network (directly or through other networks such as access networks) to communicate over the network (e.g., the Internet or virtual private networks (VPNs) overlaid on (e.g., tunneled through) the Internet) with each other (directly or through servers) and/or access content and/or services. Such content and/or services are typically provided by one or more servers (not shown) belonging to a service/content provider or one or more end user devices (not shown) participating in a peer-to-peer (P2P) service, and may include, for example, public webpages (e.g., free content, store fronts, search services), private webpages (e.g.,

username/password accessed webpages providing email services), and/or corporate networks over VPNs. For instance, end user devices may be coupled (e.g., through customer premise equipment coupled to an access network (wired or wirelessly)) to edge NDs, which are coupled (e.g., through one or more core NDs) to other edge NDs, which are coupled to electronic devices acting as servers. However, through compute and storage virtualization, one or more of the electronic devices operating as the NDs in Figure 3A may also host one or more such servers (e.g., in the case of the general purpose network device 504, one or more of the software instances 562A-R may operate as servers; the same would be true for the hybrid network device 506; in the case of the special-purpose network device 502, one or more such servers could also be run on a virtualization layer executed by the compute resource(s) 512); in which case the servers are said to be co-located with the VNEs of that ND.

[0070] A virtual network is a logical abstraction of a physical network (such as that in Figure 3A) that provides network services (e.g., L2 and/or L3 services). A virtual network can be implemented as an overlay network (sometimes referred to as a network virtualization overlay) that provides network services (e.g., layer 2 (L2, data link layer) and/or layer 3 (L3, network layer) services) over an underlay network (e.g., an L3 network, such as an Internet Protocol (IP) network that uses tunnels (e.g., generic routing encapsulation (GRE), layer 2 tunneling protocol (L2TP), IPSec) to create the overlay network).

[0071] A network virtualization edge (NVE) sits at the edge of the underlay network and participates in implementing the network virtualization; the network-facing side of the NVE uses the underlay network to tunnel frames to and from other NVEs; the outward-facing side of the NVE sends and receives data to and from systems outside the network. A virtual network instance (VNI) is a specific instance of a virtual network on a NVE (e.g., a NE/VNE on an ND, a part of a NE/VNE on a ND where that NE/VNE is divided into multiple VNEs through emulation); one or more VNIs can be instantiated on an NVE (e.g., as different VNEs on an ND). A virtual access point (VAP) is a logical connection point on the NVE for connecting external systems to a virtual network; a VAP can be physical or virtual ports identified through logical interface identifiers (e.g., a VLAN ID).

[0072] Examples of network services include: 1) an Ethernet LAN emulation service (an Ethernet-based multipoint service similar to an Internet Engineering Task Force (IETF) Multiprotocol Label Switching (MPLS) or Ethernet VPN (EVPN) service) in which external systems are interconnected across the network by a LAN environment over the underlay network (e.g., an NVE provides separate L2 VNIs (virtual switching instances) for different such virtual networks, and L3 (e.g., IP/MPLS) tunneling encapsulation across the underlay network); and 2) a virtualized IP forwarding service (similar to IETF IP VPN (e.g., Border

Gateway Protocol (BGP)/MPLS IPVPN) from a service definition perspective) in which external systems are interconnected across the network by an L3 environment over the underlay network (e.g., an NVE provides separate L3 VNIs (forwarding and routing instances) for different such virtual networks, and L3 (e.g., IP/MPLS) tunneling encapsulation across the underlay network)). Network services may also include quality of service capabilities (e.g., traffic classification marking, traffic conditioning and scheduling), security capabilities (e.g., filters to protect customer premises from network – originated attacks, to avoid malformed route announcements), and management capabilities (e.g., full detection and processing).

[0073] Fig. 3D illustrates a network with a single network element on each of the NDs of Figure 3A, and within this straight forward approach contrasts a traditional distributed approach (commonly used by traditional routers) with a centralized approach for maintaining reachability and forwarding information (also called network control), according to some embodiments of the invention. Specifically, Figure 3D illustrates network elements (NEs) 570A-H with the same connectivity as the NDs 500A-H of Figure 3A.

[0074] Figure 3D illustrates that the distributed approach 572 distributes responsibility for generating the reachability and forwarding information across the NEs 570A-H; in other words, the process of neighbor discovery and topology discovery is distributed.

[0075] For example, where the special-purpose network device 502 is used, the control communication and configuration module(s) 532A-R of the ND control plane 524 typically include a reachability and forwarding information module to implement one or more routing protocols (e.g., an exterior gateway protocol such as Border Gateway Protocol (BGP), Interior Gateway Protocol(s) (IGP) (e.g., Open Shortest Path First (OSPF), Intermediate System to Intermediate System (IS-IS), Routing Information Protocol (RIP), Label Distribution Protocol (LDP), Resource Reservation Protocol (RSVP) (including RSVP-Traffic Engineering (TE): Extensions to RSVP for LSP Tunnels and Generalized Multi-Protocol Label Switching (GMPLS) Signaling RSVP-TE)) that communicate with other NEs to exchange routes, and then selects those routes based on one or more routing metrics. Thus, the NEs 570A-H (e.g., the compute resource(s) 512 executing the control communication and configuration module(s) 532A-R) perform their responsibility for participating in controlling how data (e.g., packets) is to be routed (e.g., the next hop for the data and the outgoing physical NI for that data) by distributively determining the reachability within the network and calculating their respective forwarding information. Routes and adjacencies are stored in one or more routing structures (e.g., Routing Information Base (RIB), Label Information Base (LIB), one or more adjacency structures) on the ND control plane 524. The ND control plane 524 programs the ND forwarding plane 526 with information (e.g., adjacency and route information) based on the

routing structure(s). For example, the ND control plane 524 programs the adjacency and route information into one or more forwarding table(s) 534A-R (e.g., Forwarding Information Base (FIB), Label Forwarding Information Base (LFIB), and one or more adjacency structures) on the ND forwarding plane 526. For layer 2 forwarding, the ND can store one or more bridging tables that are used to forward data based on the layer 2 information in that data. While the above example uses the special-purpose network device 502, the same distributed approach 572 can be implemented on the general purpose network device 504 and the hybrid network device 506.

[0076] Figure 3D illustrates that a centralized approach 574 (also known as software defined networking (SDN)) that decouples the system that makes decisions about where traffic is sent from the underlying systems that forwards traffic to the selected destination. The illustrated centralized approach 574 has the responsibility for the generation of reachability and forwarding information in a centralized control plane 576 (sometimes referred to as a SDN control module, controller, network controller, OpenFlow controller, SDN controller, control plane node, network virtualization authority, or management control entity), and thus the process of neighbor discovery and topology discovery is centralized. The centralized control plane 576 has a south bound interface 582 with a data plane 580 (sometime referred to the infrastructure layer, network forwarding plane, or forwarding plane (which should not be confused with a ND forwarding plane)) that includes the NEs 570A-H (sometimes referred to as switches, forwarding elements, data plane elements, or nodes). The centralized control plane 576 includes a network controller 578, which includes a centralized reachability and forwarding information module 579 that determines the reachability within the network and distributes the forwarding information to the NEs 570A-H of the data plane 580 over the south bound interface 582 (which may use the OpenFlow protocol). Thus, the network intelligence is centralized in the centralized control plane 576 executing on electronic devices that are typically separate from the NDs.

[0077] For example, where the special-purpose network device 502 is used in the data plane 580, each of the control communication and configuration module(s) 532A-R of the ND control plane 524 typically include a control agent that provides the VNE side of the south bound interface 582. In this case, the ND control plane 524 (the compute resource(s) 512 executing the control communication and configuration module(s) 532A-R) performs its responsibility for participating in controlling how data (e.g., packets) is to be routed (e.g., the next hop for the data and the outgoing physical NI for that data) through the control agent communicating with the centralized control plane 576 to receive the forwarding information (and in some cases, the reachability information) from the centralized reachability and forwarding information module 579 (it should be understood that in some embodiments of the invention, the control communication and configuration module(s) 532A-R, in addition to communicating with the

centralized control plane 576, may also play some role in determining reachability and/or calculating forwarding information – albeit less so than in the case of a distributed approach; such embodiments are generally considered to fall under the centralized approach 574, but may also be considered a hybrid approach).

[0078] While the above example uses the special-purpose network device 502, the same centralized approach 574 can be implemented with the general purpose network device 504 (e.g., each of the VNE 560A-R performs its responsibility for controlling how data (e.g., packets) is to be routed (e.g., the next hop for the data and the outgoing physical NI for that data) by communicating with the centralized control plane 576 to receive the forwarding information (and in some cases, the reachability information) from the centralized reachability and forwarding information module 579; it should be understood that in some embodiments of the invention, the VNEs 560A-R, in addition to communicating with the centralized control plane 576, may also play some role in determining reachability and/or calculating forwarding information – albeit less so than in the case of a distributed approach) and the hybrid network device 506. In fact, the use of SDN techniques can enhance the NFV techniques typically used in the general purpose network device 504 or hybrid network device 506 implementations as NFV is able to support SDN by providing an infrastructure upon which the SDN software can be run, and NFV and SDN both aim to make use of commodity server hardware and physical switches.

[0079] Figure 3D also shows that the centralized control plane 576 has a north bound interface 584 to an application layer 586, in which resides application(s) 588. The centralized control plane 576 has the ability to form virtual networks 592 (sometimes referred to as a logical forwarding plane, network services, or overlay networks (with the NEs 570A-H of the data plane 580 being the underlay network)) for the application(s) 588. Thus, the centralized control plane 576 maintains a global view of all NDs and configured NEs/VNEs, and it maps the virtual networks to the underlying NDs efficiently (including maintaining these mappings as the physical network changes either through hardware (ND, link, or ND component) failure, addition, or removal).

[0080] While Figure 3D shows the distributed approach 572 separate from the centralized approach 574, the effort of network control may be distributed differently or the two combined in certain embodiments of the invention. For example: 1) embodiments may generally use the centralized approach (SDN) 574, but have certain functions delegated to the NEs (e.g., the distributed approach may be used to implement one or more of fault monitoring, performance monitoring, protection switching, and primitives for neighbor and/or topology discovery); or 2) embodiments of the invention may perform neighbor discovery and topology discovery via both

the centralized control plane and the distributed protocols, and the results compared to raise exceptions where they do not agree. Such embodiments are generally considered to fall under the centralized approach 574, but may also be considered a hybrid approach.

[0081] While Figure 3D illustrates the simple case where each of the NDs 500A-H implements a single NE 570A-H, it should be understood that the network control approaches described with reference to Figure 3D also work for networks where one or more of the NDs 500A-H implement multiple VNEs (e.g., VNEs 530A-R, VNEs 560A-R, those in the hybrid network device 506). Alternatively or in addition, the network controller 578 may also emulate the implementation of multiple VNEs in a single ND. Specifically, instead of (or in addition to) implementing multiple VNEs in a single ND, the network controller 578 may present the implementation of a VNE/NE in a single ND as multiple VNEs in the virtual networks 592 (all in the same one of the virtual network(s) 592, each in different ones of the virtual network(s) 592, or some combination). For example, the network controller 578 may cause an ND to implement a single VNE (a NE) in the underlay network, and then logically divide up the resources of that NE within the centralized control plane 576 to present different VNEs in the virtual network(s) 592 (where these different VNEs in the overlay networks are sharing the resources of the single VNE/NE implementation on the ND in the underlay network).

[0082] On the other hand, Figures 3E and 3F respectively illustrate exemplary abstractions of NEs and VNEs that the network controller 578 may present as part of different ones of the virtual networks 592. Figure 3E illustrates the simple case of where each of the NDs 500A-H implements a single NE 570A-H (see Figure 3D), but the centralized control plane 576 has abstracted multiple of the NEs in different NDs (the NEs 570A-C and G-H) into (to represent) a single NE 570I in one of the virtual network(s) 592 of Figure 3D, according to some embodiments of the invention. Figure 3E shows that in this virtual network, the NE 570I is coupled to NE 570D and 570F, which are both still coupled to NE 570E.

[0083] Figure 3F illustrates a case where multiple VNEs (VNE 570A.1 and VNE 570H.1) are implemented on different NDs (ND 500A and ND 500H) and are coupled to each other, and where the centralized control plane 576 has abstracted these multiple VNEs such that they appear as a single VNE 570T within one of the virtual networks 592 of Figure 3D, according to some embodiments of the invention. Thus, the abstraction of a NE or VNE can span multiple NDs.

[0084] While some embodiments of the invention implement the centralized control plane 576 as a single entity (e.g., a single instance of software running on a single electronic device), alternative embodiments may spread the functionality across multiple entities for redundancy

and/or scalability purposes (e.g., multiple instances of software running on different electronic devices).

[0085] Similar to the network device implementations, the electronic device(s) running the centralized control plane 576, and thus the network controller 578 including the centralized reachability and forwarding information module 579, may be implemented a variety of ways (e.g., a special purpose device, a general-purpose (e.g., COTS) device, or hybrid device). These electronic device(s) would similarly include compute resource(s), a set or one or more physical NICs, and a non-transitory machine-readable storage medium having stored thereon the centralized control plane software. For instance, Figure 4 illustrates, a general purpose control plane device 604 including hardware 640 comprising a set of one or more processor(s) 642 (which are often COTS processors) and network interface controller(s) 644 (NICs; also known as network interface cards) (which include physical NIs 646), as well as non-transitory machine readable storage media 648 having stored therein centralized control plane (CCP) software 650.

[0086] In embodiments that use compute virtualization, the processor(s) 642 typically execute software to instantiate a virtualization layer 654 (e.g., in one embodiment the virtualization layer 654 represents the kernel of an operating system (or a shim executing on a base operating system) that allows for the creation of multiple instances 662A-R called software containers (representing separate user spaces and also called virtualization engines, virtual private servers, or jails) that may each be used to execute a set of one or more applications; in another embodiment the virtualization layer 654 represents a hypervisor (sometimes referred to as a virtual machine monitor (VMM)) or a hypervisor executing on top of a host operating system, and an application is run on top of a guest operating system within an instance 662A-R called a virtual machine (which in some cases may be considered a tightly isolated form of software container) that is run by the hypervisor ; in another embodiment, an application is implemented as a unikernel, which can be generated by compiling directly with an application only a limited set of libraries (e.g., from a library operating system (LibOS) including drivers/libraries of OS services) that provide the particular OS services needed by the application, and the unikernel can run directly on hardware 640, directly on a hypervisor represented by virtualization layer 654 (in which case the unikernel is sometimes described as running within a LibOS virtual machine), or in a software container represented by one of instances 662A-R). Again, in embodiments where compute virtualization is used, during operation an instance of the CCP software 650 (illustrated as CCP instance 676A) is executed (e.g., within the instance 662A) on the virtualization layer 654. In embodiments where compute virtualization is not used, the CCP instance 676A is executed, as a unikernel or on top of a host operating system, on the "bare metal" general purpose control plane device 604. The instantiation of the CCP instance 676A, as well as the

virtualization layer 654 and instances 662A-R if implemented, are collectively referred to as software instance(s) 652.

[0087] In some embodiments, the CCP instance 676A includes a network controller instance 678. The network controller instance 678 includes a centralized reachability and forwarding information module instance 679 (which is a middleware layer providing the context of the network controller 578 to the operating system and communicating with the various NEs), and an CCP application layer 680 (sometimes referred to as an application layer) over the middleware layer (providing the intelligence required for various network operations such as protocols, network situational awareness, and user – interfaces). At a more abstract level, this CCP application layer 680 within the centralized control plane 576 works with virtual network view(s) (logical view(s) of the network) and the middleware layer provides the conversion from the virtual networks to the physical view.

[0088] The centralized control plane 576 transmits relevant messages to the data plane 580 based on CCP application layer 680 calculations and middleware layer mapping for each flow. A flow may be defined as a set of packets whose headers match a given pattern of bits; in this sense, traditional IP forwarding is also flow–based forwarding where the flows are defined by the destination IP address for example; however, in other implementations, the given pattern of bits used for a flow definition may include more fields (e.g., 10 or more) in the packet headers. Different NDs/NEs/VNEs of the data plane 580 may receive different messages, and thus different forwarding information. The data plane 580 processes these messages and programs the appropriate flow information and corresponding actions in the forwarding tables (sometime referred to as flow tables) of the appropriate NE/VNEs, and then the NEs/VNEs map incoming packets to flows represented in the forwarding tables and forward packets based on the matches in the forwarding tables.

[0089] Standards such as OpenFlow define the protocols used for the messages, as well as a model for processing the packets. The model for processing packets includes header parsing, packet classification, and making forwarding decisions. Header parsing describes how to interpret a packet based upon a well-known set of protocols. Some protocol fields are used to build a match structure (or key) that will be used in packet classification (e.g., a first key field could be a source media access control (MAC) address, and a second key field could be a destination MAC address).

[0090] Packet classification involves executing a lookup in memory to classify the packet by determining which entry (also referred to as a forwarding table entry or flow entry) in the forwarding tables best matches the packet based upon the match structure, or key, of the forwarding table entries. It is possible that many flows represented in the forwarding table

entries can correspond/match to a packet; in this case the system is typically configured to determine one forwarding table entry from the many according to a defined scheme (e.g., selecting a first forwarding table entry that is matched). Forwarding table entries include both a specific set of match criteria (a set of values or wildcards, or an indication of what portions of a packet should be compared to a particular value/values/wildcards, as defined by the matching capabilities – for specific fields in the packet header, or for some other packet content), and a set of one or more actions for the data plane to take on receiving a matching packet. For example, an action may be to push a header onto the packet, for the packet using a particular port, flood the packet, or simply drop the packet. Thus, a forwarding table entry for IPv4/IPv6 packets with a particular transmission control protocol (TCP) destination port could contain an action specifying that these packets should be dropped.

[0091] Making forwarding decisions and performing actions occurs, based upon the forwarding table entry identified during packet classification, by executing the set of actions identified in the matched forwarding table entry on the packet.

[0092] However, when an unknown packet (for example, a "missed packet" or a "match-miss" as used in OpenFlow parlance) arrives at the data plane 580, the packet (or a subset of the packet header and content) is typically forwarded to the centralized control plane 576. The centralized control plane 576 will then program forwarding table entries into the data plane 580 to accommodate packets belonging to the flow of the unknown packet. Once a specific forwarding table entry has been programmed into the data plane 580 by the centralized control plane 576, the next packet with matching credentials will match that forwarding table entry and take the set of actions associated with that matched entry.

[0093] A network interface (NI) may be physical or virtual; and in the context of IP, an interface address is an IP address assigned to a NI, be it a physical NI or virtual NI. A virtual NI may be associated with a physical NI, with another virtual interface, or stand on its own (e.g., a loopback interface, a point-to-point protocol interface). A NI (physical or virtual) may be numbered (a NI with an IP address) or unnumbered (a NI without an IP address). A loopback interface (and its loopback address) is a specific type of virtual NI (and IP address) of a NE/VNE (physical or virtual) often used for management purposes; where such an IP address is referred to as the nodal loopback address. The IP address(es) assigned to the NI(s) of a ND are referred to as IP addresses of that ND; at a more granular level, the IP address(es) assigned to NI(s) assigned to a NE/VNE implemented on a ND can be referred to as IP addresses of that NE/VNE.

[0094] Next hop selection by the routing system for a given destination may resolve to one path (that is, a routing protocol may generate one next hop on a shortest path); but if the routing

system determines there are multiple viable next hops (that is, the routing protocol generated forwarding solution offers more than one next hop on a shortest path – multiple equal cost next hops), some additional criteria is used - for instance, in a connectionless network, Equal Cost Multi Path (ECMP) (also known as Equal Cost Multi Pathing, multipath forwarding and IP multipath) may be used (e.g., typical implementations use as the criteria particular header fields to ensure that the packets of a particular packet flow are always forwarded on the same next hop to preserve packet flow ordering). For purposes of multipath forwarding, a packet flow is defined as a set of packets that share an ordering constraint. As an example, the set of packets in a particular TCP transfer sequence need to arrive in order, else the TCP logic will interpret the out of order delivery as congestion and slow the TCP transfer rate down.

[0095] Some NDs include functionality for authentication, authorization, and accounting (AAA) protocols (e.g., RADIUS (Remote Authentication Dial-In User Service), Diameter, and/or TACACS+ (Terminal Access Controller Access Control System Plus). AAA can be provided through a client/server model, where the AAA client is implemented on a ND and the AAA server can be implemented either locally on the ND or on a remote electronic device coupled with the ND. Authentication is the process of identifying and verifying a subscriber. For instance, a subscriber might be identified by a combination of a username and a password or through a unique key. Authorization determines what a subscriber can do after being authenticated, such as gaining access to certain electronic device information resources (e.g., through the use of access control policies). Accounting is recording user activity. By way of a summary example, end user devices may be coupled (e.g., through an access network) through an edge ND (supporting AAA processing) coupled to core NDs coupled to electronic devices implementing servers of service/content providers. AAA processing is performed to identify for a subscriber the subscriber record stored in the AAA server for that subscriber. A subscriber record includes a set of attributes (e.g., subscriber name, password, authentication information, access control information, rate-limiting information, policing information) used during processing of that subscriber's traffic.

[0096] Certain NDs (e.g., certain edge NDs) internally represent end user devices (or sometimes customer premise equipment (CPE) such as a residential gateway (e.g., a router, modem)) using subscriber circuits. A subscriber circuit uniquely identifies within the ND a subscriber session and typically exists for the lifetime of the session. Thus, a ND typically allocates a subscriber circuit when the subscriber connects to that ND, and correspondingly de-allocates that subscriber circuit when that subscriber disconnects. Each subscriber session represents a distinguishable flow of packets communicated between the ND and an end user device (or sometimes CPE such as a residential gateway or modem) using a protocol, such as the

point-to-point protocol over another protocol (PPPoX) (e.g., where X is Ethernet or Asynchronous Transfer Mode (ATM)), Ethernet, 802.1Q Virtual LAN (VLAN), Internet Protocol, or ATM). A subscriber session can be initiated using a variety of mechanisms (e.g., manual provisioning a dynamic host configuration protocol (DHCP), DHCP/client-less internet protocol service (CLIPS) or Media Access Control (MAC) address tracking). For example, the point-to-point protocol (PPP) is commonly used for digital subscriber line (DSL) services and requires installation of a PPP client that enables the subscriber to enter a username and a password, which in turn may be used to select a subscriber record. When DHCP is used (e.g., for cable modem services), a username typically is not provided; but in such situations other information (e.g., information that includes the MAC address of the hardware in the end user device (or CPE)) is provided. The use of DHCP and CLIPS on the ND captures the MAC addresses and uses these addresses to distinguish subscribers and access their subscriber records.

[0097] A virtual circuit (VC), synonymous with virtual connection and virtual channel, is a connection oriented communication service that is delivered by means of packet mode communication. Virtual circuit communication resembles circuit switching, since both are connection oriented, meaning that in both cases data is delivered in correct order, and signaling overhead is required during a connection establishment phase. Virtual circuits may exist at different layers. For example, at layer 4, a connection oriented transport layer datalink protocol such as Transmission Control Protocol (TCP) may rely on a connectionless packet switching network layer protocol such as IP, where different packets may be routed over different paths, and thus be delivered out of order. Where a reliable virtual circuit is established with TCP on top of the underlying unreliable and connectionless IP protocol, the virtual circuit is identified by the source and destination network socket address pair, i.e. the sender and receiver IP address and port number. However, a virtual circuit is possible since TCP includes segment numbering and reordering on the receiver side to prevent out-of-order delivery. Virtual circuits are also possible at Layer 3 (network layer) and Layer 2 (datalink layer); such virtual circuit protocols are based on connection oriented packet switching, meaning that data is always delivered along the same network path, i.e. through the same NEs/VNEs. In such protocols, the packets are not routed individually and complete addressing information is not provided in the header of each data packet; only a small virtual channel identifier (VCI) is required in each packet; and routing information is transferred to the NEs/VNEs during the connection establishment phase; switching only involves looking up the virtual channel identifier in a table rather than analyzing a complete address. Examples of network layer and datalink layer virtual circuit protocols, where data always is delivered over the same path: X.25, where the VC is identified by a virtual channel identifier (VCI); Frame relay, where the VC is identified by a VCI; Asynchronous

Transfer Mode (ATM), where the circuit is identified by a virtual path identifier (VPI) and virtual channel identifier (VCI) pair; General Packet Radio Service (GPRS); and Multiprotocol label switching (MPLS), which can be used for IP over virtual circuits (Each circuit is identified by a label).

[0098] Certain NDs (e.g., certain edge NDs) use a hierarchy of circuits. The leaf nodes of the hierarchy of circuits are subscriber circuits. The subscriber circuits have parent circuits in the hierarchy that typically represent aggregations of multiple subscriber circuits, and thus the network segments and elements used to provide access network connectivity of those end user devices to the ND. These parent circuits may represent physical or logical aggregations of subscriber circuits (e.g., a virtual local area network (VLAN), a permanent virtual circuit (PVC) (e.g., for Asynchronous Transfer Mode (ATM)), a circuit-group, a channel, a pseudo-wire, a physical NI of the ND, and a link aggregation group). A circuit-group is a virtual construct that allows various sets of circuits to be grouped together for configuration purposes, for example aggregate rate control. A pseudo-wire is an emulation of a layer 2 point-to-point connection-oriented service. A link aggregation group is a virtual construct that merges multiple physical NIs for purposes of bandwidth aggregation and redundancy. Thus, the parent circuits physically or logically encapsulate the subscriber circuits.

[0099] Each VNE (e.g., a virtual router, a virtual bridge (which may act as a virtual switch instance in a Virtual Private LAN Service (VPLS) is typically independently administrable. For example, in the case of multiple virtual routers, each of the virtual routers may share system resources but is separate from the other virtual routers regarding its management domain, AAA (authentication, authorization, and accounting) name space, IP address, and routing database(s). Multiple VNEs may be employed in an edge ND to provide direct network access and/or different classes of services for subscribers of service and/or content providers.

[00100] Within certain NDs, "interfaces" that are independent of physical NIs may be configured as part of the VNEs to provide higher-layer protocol and service information (e.g., Layer 3 addressing). The subscriber records in the AAA server identify, in addition to the other subscriber configuration requirements, to which context (e.g., which of the VNEs/NEs) the corresponding subscribers should be bound within the ND. As used herein, a binding forms an association between a physical entity (e.g., physical NI, channel) or a logical entity (e.g., circuit such as a subscriber circuit or logical circuit (a set of one or more subscriber circuits)) and a context's interface over which network protocols (e.g., routing protocols, bridging protocols) are configured for that context. Subscriber data flows on the physical entity when some higher-layer protocol interface is configured and associated with that physical entity.

[00101] Some NDs provide support for implementing VPNs (Virtual Private Networks) (e.g., Layer 2 VPNs and/or Layer 3 VPNs). For example, the ND where a provider's network and a customer's network are coupled are respectively referred to as PEs (Provider Edge) and CEs (Customer Edge). In a Layer 2 VPN, forwarding typically is performed on the CE(s) on either end of the VPN and traffic is sent across the network (e.g., through one or more PEs coupled by other NDs). Layer 2 circuits are configured between the CEs and PEs (e.g., an Ethernet port, an ATM permanent virtual circuit (PVC), a Frame Relay PVC). In a Layer 3 VPN, routing typically is performed by the PEs. By way of example, an edge ND that supports multiple VNEs may be deployed as a PE; and a VNE may be configured with a VPN protocol, and thus that VNE is referred as a VPN VNE.

[00102] Some NDs provide support for VPLS (Virtual Private LAN Service). For example, in a VPLS network, end user devices access content/services provided through the VPLS network by coupling to CEs, which are coupled through PEs coupled by other NDs. VPLS networks can be used for implementing triple play network applications (e.g., data applications (e.g., high-speed Internet access), video applications (e.g., television service such as IPTV (Internet Protocol Television), VoD (Video-on-Demand) service), and voice applications (e.g., VoIP (Voice over Internet Protocol) service)), VPN services, etc. VPLS is a type of layer 2 VPN that can be used for multi-point connectivity. VPLS networks also allow end use devices that are coupled with CEs at separate geographical locations to communicate with each other across a Wide Area Network (WAN) as if they were directly attached to each other in a Local Area Network (LAN) (referred to as an emulated LAN).

[00103] In VPLS networks, each CE typically attaches, possibly through an access network (wired and/or wireless), to a bridge module of a PE via an attachment circuit (e.g., a virtual link or connection between the CE and the PE). The bridge module of the PE attaches to an emulated LAN through an emulated LAN interface. Each bridge module acts as a "Virtual Switch Instance" (VSI) by maintaining a forwarding table that maps MAC addresses to pseudowires and attachment circuits. PEs forward frames (received from CEs) to destinations (e.g., other CEs, other PEs) based on the MAC destination address field included in those frames.

[00104] While the flow diagrams in the figures show a particular order of operations performed by certain embodiments of the invention, it should be understood that such order is exemplary (e.g., alternative embodiments may perform the operations in a different order, combine certain operations, overlap certain operations, etc.).

[00105] While the invention has been described in terms of several embodiments, those skilled in the art will recognize that the invention is not limited to the embodiments described, can be

practiced with modification and alteration within the spirit and scope of the appended claims. The description is thus to be regarded as illustrative instead of limiting.

CLAIMS

What is claimed is:

1. A method in a first network device (101B) of an inter-chassis redundancy (ICR) system (101) coupled with a second network device (101A) of the ICR system (101), of enabling traffic reroute in the ICR system (101), wherein each one of the first and the second network devices is coupled with a third network device (103A) through a Multi-Chassis Link Aggregation Group (MC-LAG)(114), the method comprising:

monitoring (302) an ICR link (112) coupling the first network device (101B) to the second network device (101A) of the ICR system (101);

detecting (304) a failure of the ICR link; and

in response to detecting the failure of the ICR link (112), and to determining (306) that the first network device (101B) is in a standby mode,

transmitting (308) to the third network device (103A) an indication that one or more links (114B) associated with the first network device (101B) are no longer operative to join the MC-LAG (114).

2. The method of claim 1, wherein the third network device (103A) is operative to load balance downstream traffic to the first network device (101B) and the second network device (101A), and wherein transmitting to the third network device (103A) the indication that one or more links associated with the first network device (101B) are no longer operative to join the MC-LAG (114) causes the third network device (103A) to redirect downstream traffic only towards the second network device (101A).

3. The method of claim 1, wherein the indication that one or more links associated with the first network device (101B) are no longer operative to join the MC-LAG (114) is included in an out-of-sync field of a Link Aggregation Control Protocol (LACP) protocol data unit packet.

4. The method of claim 1, wherein transmitting (308) to the third network device (103A) the indication includes:

identifying (314) the one or more links associated with the first network device, wherein the one or more links are active links of the MC-LAG (114) coupling the first network device (101B) to the third network device (103A).

5. The method of claim 4, wherein transmitting to the third network device (103A) the indication further includes:

setting (316) the one or more links to be in a temporary halted state; and

sending (318) Link Aggregation Control Protocol (LACP) protocol data units with out-of-sync bit set to respective MC-LAG partners for each one of the one or more links.

6. The method of claim 5, the method further comprising:
detecting (404) that the ICR link has resumed normal operations; and
transmitting (408) to the third network device (103A) a second indication that the one or more links associated with the first network device (101B) are operative to join the MC-LAG (114).

7. The method of claim 6, wherein transmitting to the third network device (103A) a second indication that the one or more links associated with the first network device (101B) are operative to join the MC-LAG (114) causes the third network device (103A) to load balance downstream traffic over the first network device (101B) and the second network device (101A).

8. A first network device (101B) of an inter-chassis redundancy (ICR) system to be coupled with a second network device (101A) of the ICR system (101), for enabling traffic reroute in the ICR system (101), wherein each one of the first and the second network devices is to be coupled with a third network device (103A) through a Multi-Chassis Link Aggregation Group (MC-LAG), (114) the first network device (101B) comprising:
one or more processors and a non-transitory computer readable storage medium, said non-transitory computer readable storage medium containing instructions, which when executed by the one or more processors, causes the first network device (101B) to:
monitor (302) an ICR link (112) coupling the first network device (101B) to the second network device (101A) of the ICR system (101),
detect (304) a failure of the ICR link, and
in response to detecting the failure of the ICR link, and determining (306) that the first network device (101B) is in a standby mode,
transmit (308) to the third network device (103A) an indication that one or more links associated with the first network device (101B) are no longer operative to join the MC-LAG (114).

9. The first network device of claim 8, wherein the third network device (103A) is operative to load balance downstream traffic to the first network device (101B) and the second network device, and wherein to transmit to the third network device (103A) the indication that one or more links (114B) associated with the first network device (101B) are no longer operative

to join the MC-LAG (114) causes the third network device (103A) to redirect downstream traffic only towards the second network device(101A).

10. The first network device of claim 8, wherein the indication that one or more links (114B) associated with the first network device (101B) are no longer operative to join the MC-LAG (114) is included in an out-of-sync field of a Link Aggregation Control Protocol (LACP) protocol data unit.

11. The first network device of claim 8, wherein to transmit to the third network device (103A) the indication includes:

to identify (314) the one or more links associated with the first network device, wherein the one or more links are active links of the MC-LAG (114) coupling the first network device (101B) to the third network device.

12. The first network device of claim 11, wherein to transmit to the third network device (103A) the indication further includes:

to set (316) the one or more links to be in a temporary halted state; and

to send (318) Link Aggregation Control Protocol (LACP) protocol data units with out-of-sync bit set to respective MC-LAG partners for each one of the one or more links.

13. The first network device of claim 12, wherein the first network device (101B) is further to:

detect (404) that the ICR link (112) has resumed normal operations; and

transmit (408) to the third network device (103A) a second indication that the one or more links associated with the first network device (101B) are operative to join the MC-LAG (114).

14. The first network device of claim 13, wherein to transmit to the third network device (103A) a second indication that the one or more links (114B) associated with the first network device (101B) are operative to join the MC-LAG (114) causes the third network device (103A) to load balance downstream traffic over the first network device (101B) and the second network device(101A).

15. A non-transitory computer readable storage medium that provide instructions, which when executed by a processor of a first network device (101B) of an inter-chassis redundancy (ICR) system (101) to be coupled with a second network device (101A) of the ICR system (101), for

enabling traffic reroute in the ICR system (101), wherein each one of the first and the second network devices is to be coupled with a third network device (103A) through a Multi-Chassis Link Aggregation Group (MC-LAG)(114), cause said processor to perform operations comprising:

monitoring (302) an ICR link (112) coupling the first network device (101B) to the second network device (101A) of the ICR system (101);

detecting (304) a failure of the ICR link; and

in response to detecting the failure of the ICR link, and to determining (306) that the first network device (101B) is in a standby mode,

transmitting (308) to the third network device (103A) an indication that one or more links associated with the first network device (101B) are no longer operative to join the MC-LAG (114).

16. The non-transitory computer readable storage medium of claim 15, wherein the third network device (103A) is operative to load balance downstream traffic to the first network device (101B) and the second network device(101A), and wherein transmitting to the third network device (103A) the indication that one or more links (114B) associated with the first network device (101B) are no longer operative to join the MC-LAG (114) causes the third network device (103A) to redirect downstream traffic only towards the second network device (101A).

17. The non-transitory computer readable storage medium of claim 15, wherein the indication that one or more links (114B) associated with the first network device (101B) are no longer operative to join the MC-LAG (114) is included in an out-of-sync field of a Link Aggregation Control Protocol (LACP) protocol data unit.

18. The non-transitory computer readable storage medium of claim 15, wherein transmitting to the third network device (103A) the indication includes:

identifying (314) the one or more links associated with the first network device, wherein the one or more links are active links of the MC-LAG (114) coupling the first network device (101B) to the third network device (103A).

19. The non-transitory computer readable storage medium of claim 18, wherein transmitting to the third network device (103A) the indication further includes:

setting (316) the one or more links to be in a temporary halted state; and

sending (318) Link Aggregation Control Protocol (LACP) protocol data units with out-of-sync bit set to respective MC-LAG partners for each one of the one or more links (114B).

20. The non-transitory computer readable storage medium of claim 19, wherein the operations further comprise:

detecting (404) that the ICR link (112) has resumed normal operations; and

transmitting (408) to the third network device (103A) a second indication that the one or more links (114B) associated with the first network device (101B) are operative to join the MC-LAG (114).

21. The non-transitory computer readable storage medium of claim 20, wherein transmitting to the third network device (103A) a second indication that the one or more links (114B) associated with the first network device (101B) are operative to join the MC-LAG (114) causes the third network device (103A) to load balance downstream traffic over the first network device (101B) and the second network device (101A).

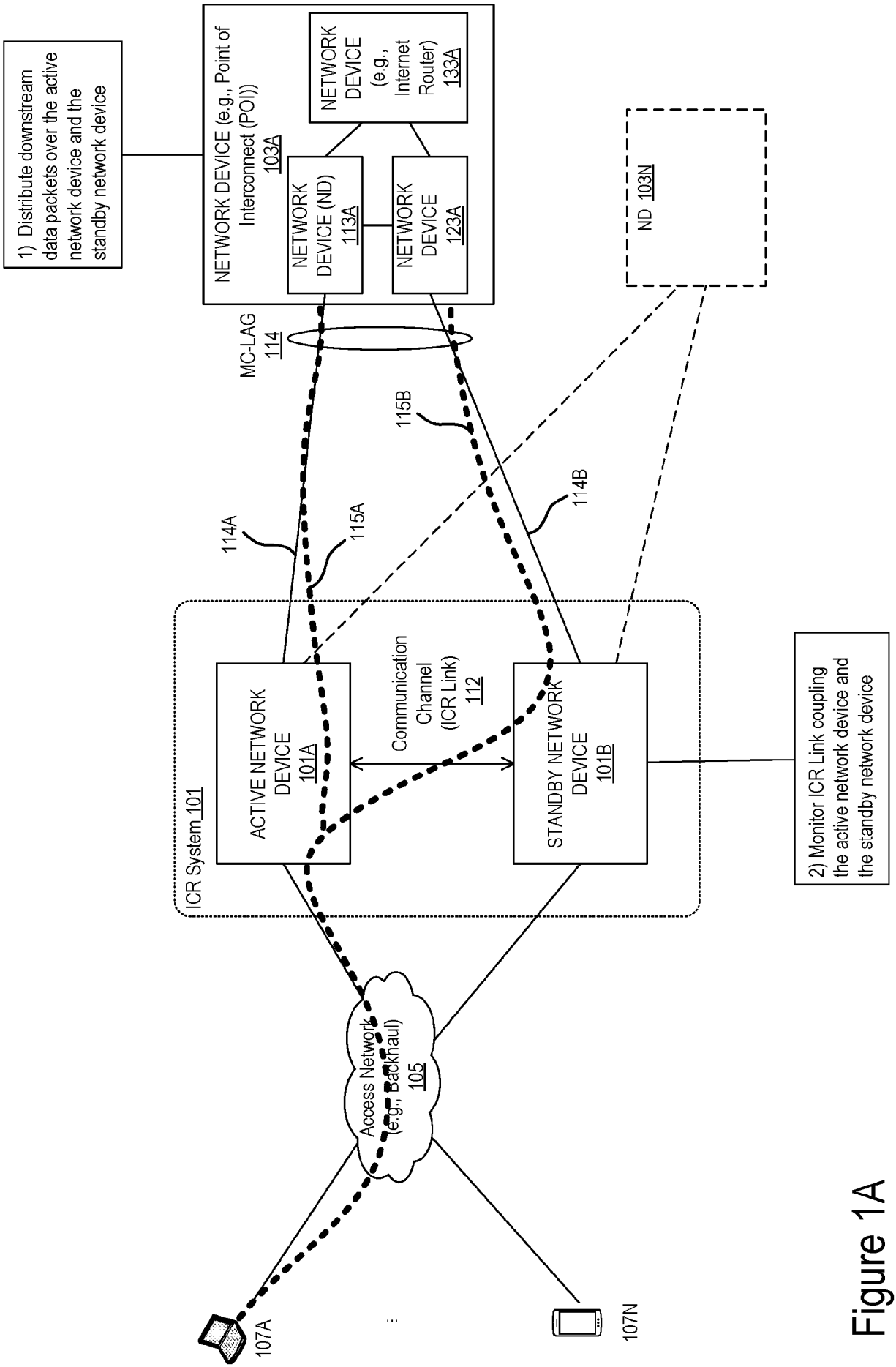

1) Distribute downstream data packets over the active network device and the standby network device
NETWORK DEVICE (e.g., Point of Interconnect (POI)) 103A
NETWORK DEVICE (ND) 113A
NETWORK DEVICE 123A
NETWORK DEVICE (e.g., Internet Router) 133A
ND 103N
MC-LAG 114
114A
115A
115B
114B
ICR System 101
ACTIVE NETWORK DEVICE 101A
Communication Channel (ICR Link) 112
STANDBY NETWORK DEVICE 101B
2) Monitor ICR Link coupling the active network device and the standby network device
Access Network (e.g., Backhaul) 105
107A
107N

Figure 1A

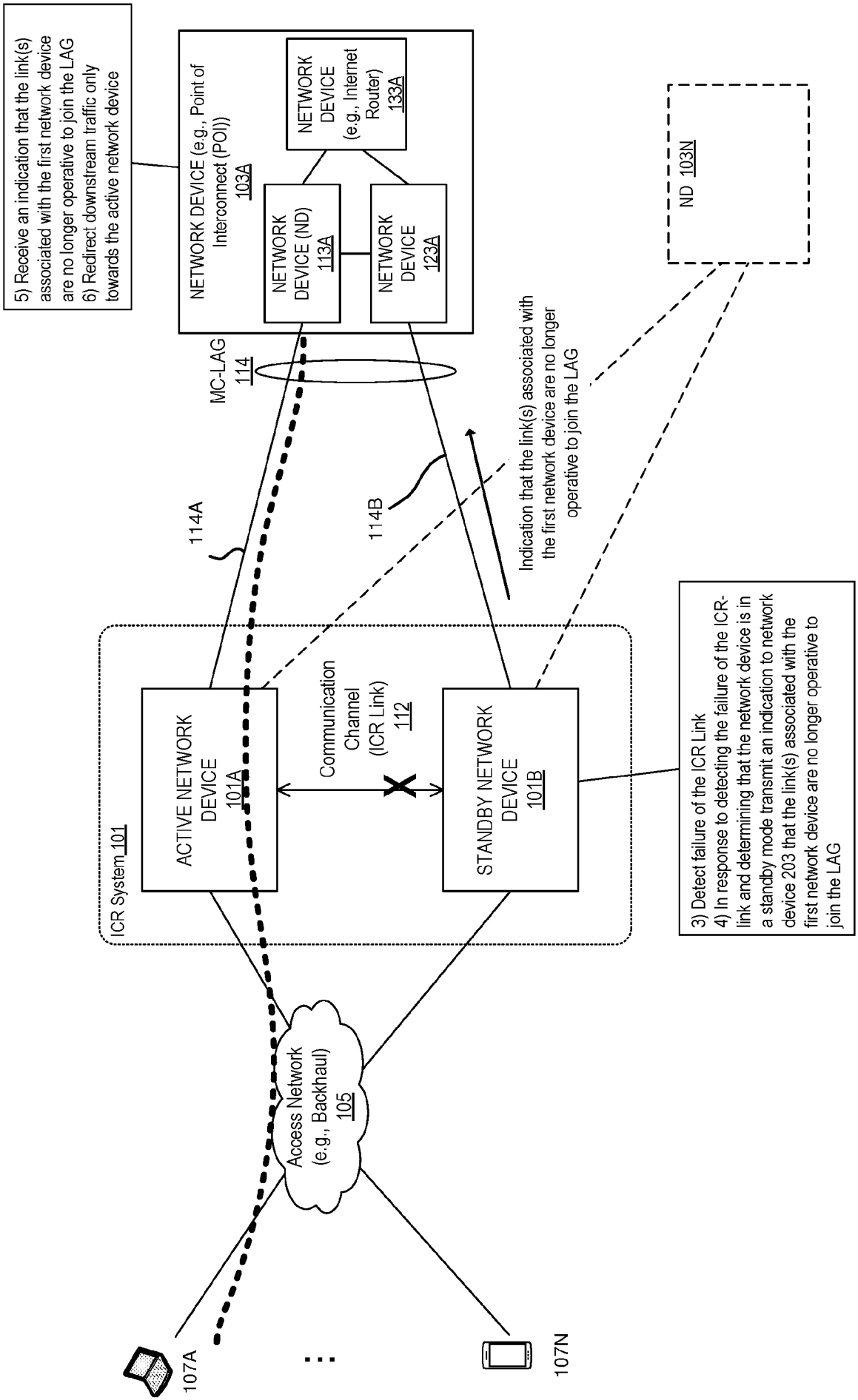

5) Receive an indication that the link(s) associated with the first network device are no longer operative to join the LAG
6) Redirect downstream traffic only towards the active network device
NETWORK DEVICE (e.g., Point of Interconnect (POI)) 103A
NETWORK DEVICE (e.g., Internet Router) 133A
NETWORK DEVICE (ND) 113A
NETWORK DEVICE 123A
ND 103N
MC-LAG 114
114A
114B
Indication that the link(s) associated with the first network device are no longer operative to join the LAG
ICR System 101
ACTIVE NETWORK DEVICE 101A
Communication Channel (ICR Link) 112
STANDBY NETWORK DEVICE 101B
3) Detect failure of the ICR Link
4) In response to detecting the failure of the ICR-link and determining that the network device is in a standby mode transmit an indication to network device 203 that the link(s) associated with the first network device are no longer operative to join the LAG
Access Network (e.g., Backhaul) 105
107A
107N

Figure 1B

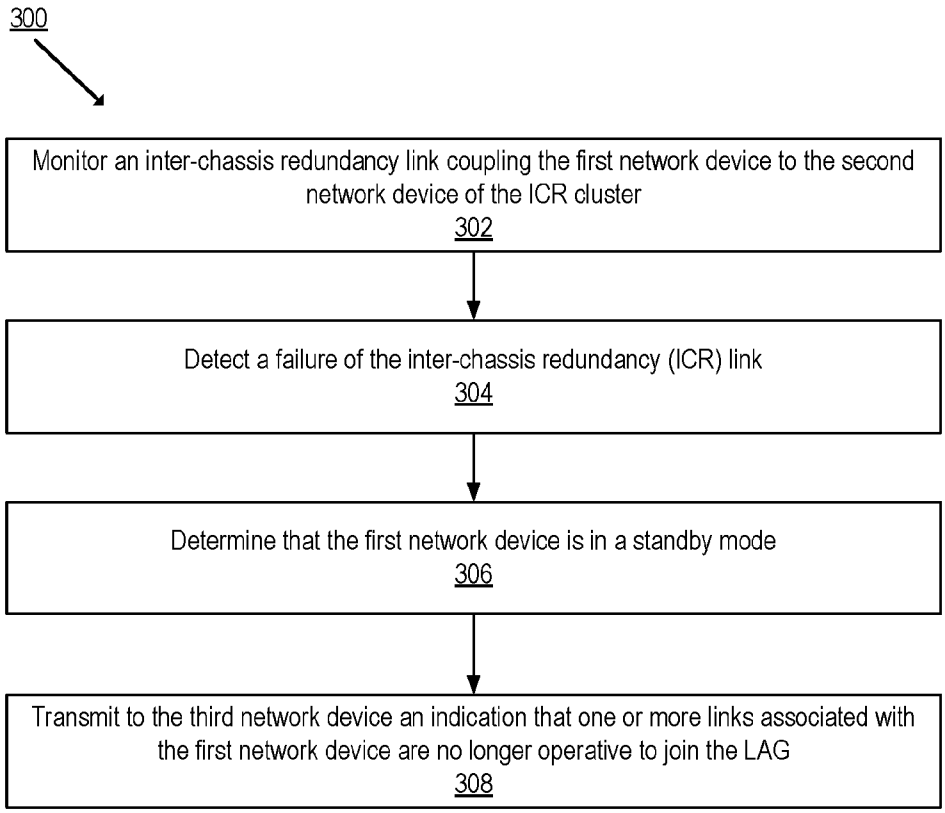

300
Monitor an inter-chassis redundancy link coupling the first network device to the second network device of the ICR cluster
302
Detect a failure of the inter-chassis redundancy (ICR) link
304
Determine that the first network device is in a standby mode
306
Transmit to the third network device an indication that one or more links associated with the first network device are no longer operative to join the LAG
308

Figure 1C

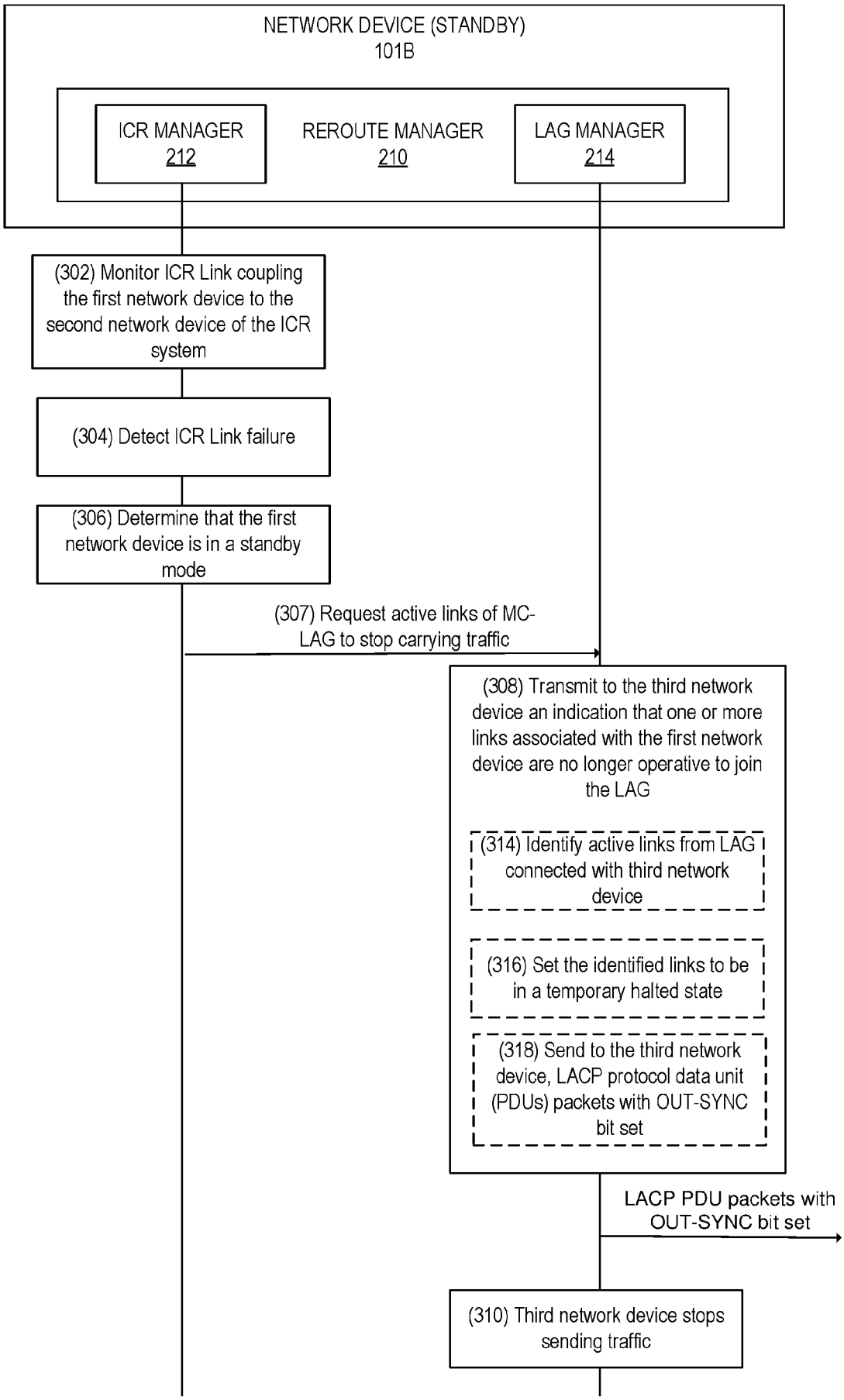

NETWORK DEVICE (STANDBY)
101B
ICR MANAGER
212
REROUTE MANAGER
210
LAG MANAGER
214
(302) Monitor ICR Link coupling the first network device to the second network device of the ICR system
(304) Detect ICR Link failure
(306) Determine that the first network device is in a standby mode
(307) Request active links of MC-LAG to stop carrying traffic
(308) Transmit to the third network device an indication that one or more links associated with the first network device are no longer operative to join the LAG
(314) Identify active links from LAG connected with third network device
(316) Set the identified links to be in a temporary halted state
(318) Send to the third network device, LACP protocol data unit (PDUs) packets with OUT-SYNC bit set
LACP PDU packets with OUT-SYNC bit set
(310) Third network device stops sending traffic

Figure 1D

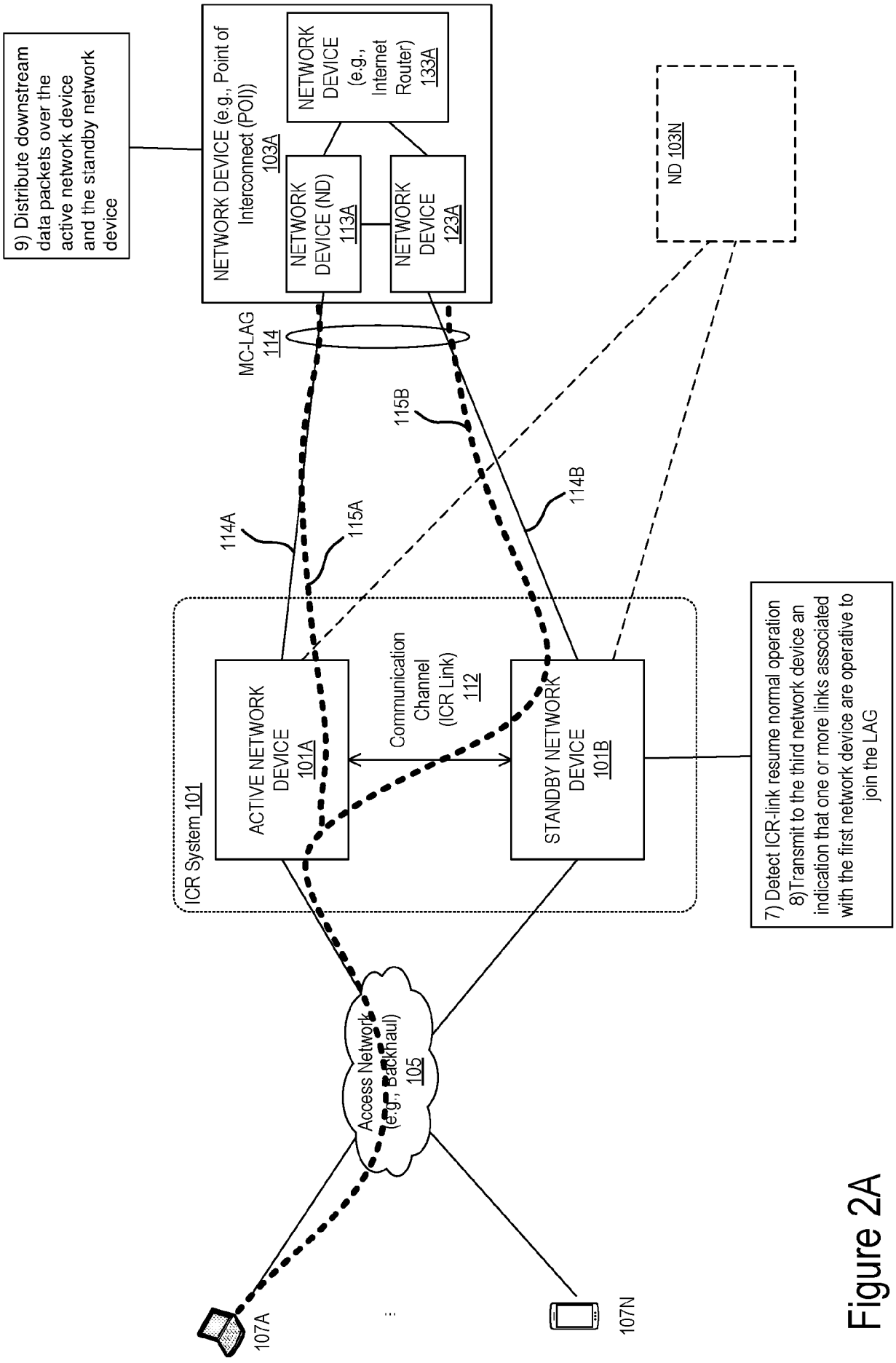

9) Distribute downstream data packets over the active network device and the standby network device
NETWORK DEVICE (e.g., Point of Interconnect (POI)) 103A
NETWORK DEVICE (ND) 113A
NETWORK DEVICE 123A
NETWORK DEVICE (e.g., Internet Router) 133A
ND 103N
MC-LAG 114
114A
115A
115B
114B
ICR System 101
ACTIVE NETWORK DEVICE 101A
Communication Channel (ICR Link) 112
STANDBY NETWORK DEVICE 101B
7) Detect ICR-link resume normal operation
8) Transmit to the third network device an indication that one or more links associated with the first network device are operative to join the LAG
Access Network (e.g., Backhaul) 105
107A
107N

Figure 2A

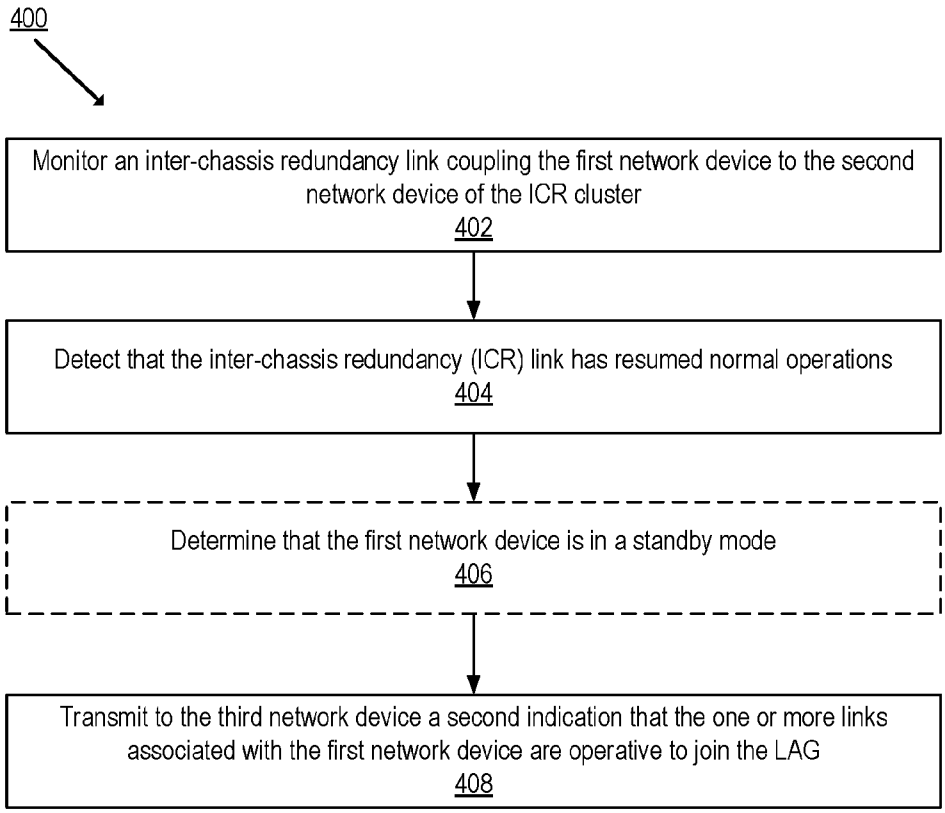

400
Monitor an inter-chassis redundancy link coupling the first network device to the second network device of the ICR cluster
402
Detect that the inter-chassis redundancy (ICR) link has resumed normal operations
404
Determine that the first network device is in a standby mode
406
Transmit to the third network device a second indication that the one or more links associated with the first network device are operative to join the LAG
408

Figure 2B

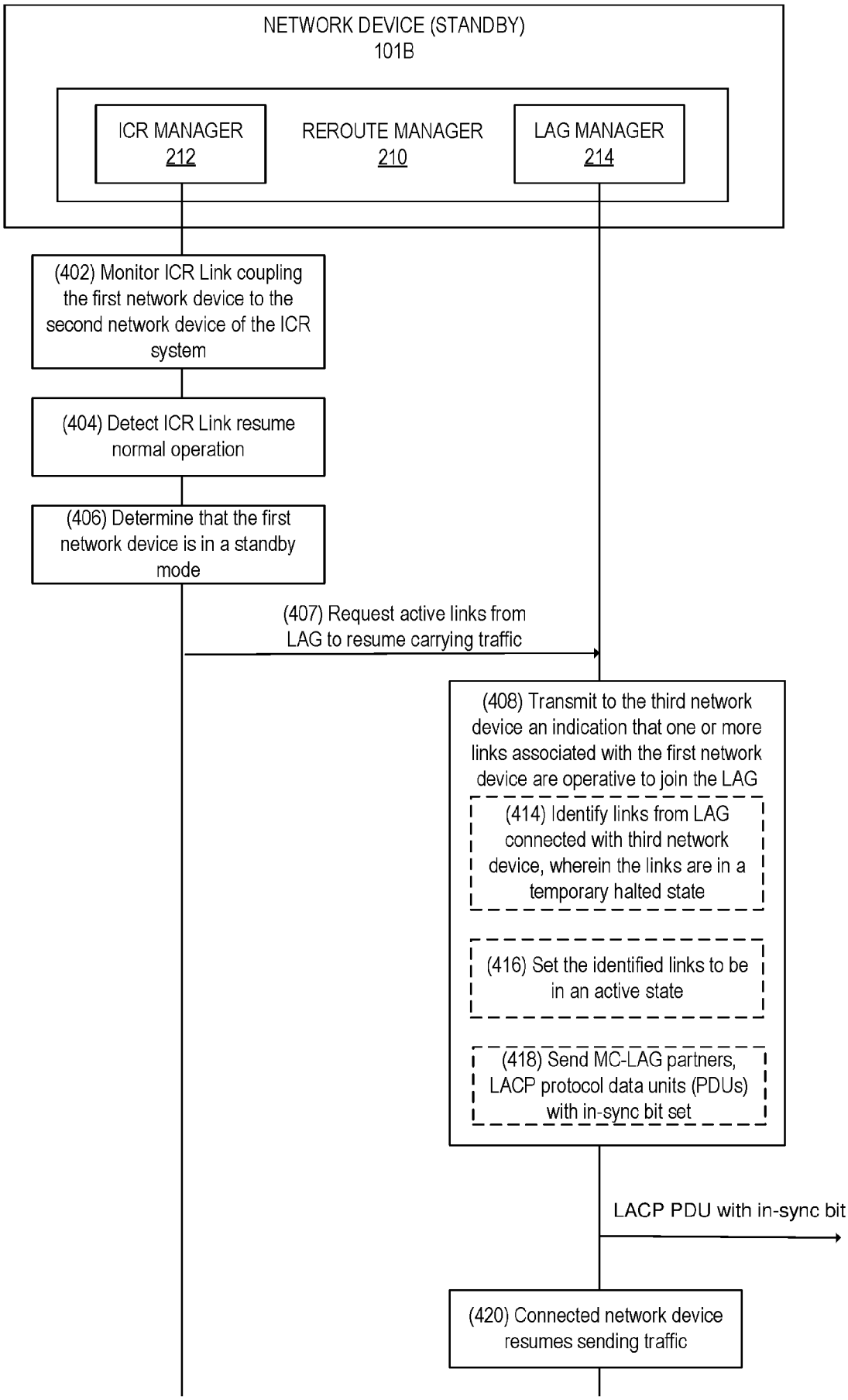

NETWORK DEVICE (STANDBY)
101B
ICR MANAGER
212
REROUTE MANAGER
210
LAG MANAGER
214
(402) Monitor ICR Link coupling the first network device to the second network device of the ICR system
(404) Detect ICR Link resume normal operation
(406) Determine that the first network device is in a standby mode
(407) Request active links from LAG to resume carrying traffic
(408) Transmit to the third network device an indication that one or more links associated with the first network device are operative to join the LAG
(414) Identify links from LAG connected with third network device, wherein the links are in a temporary halted state
(416) Set the identified links to be in an active state
(418) Send MC-LAG partners, LACP protocol data units (PDUs) with in-sync bit set
LACP PDU with in-sync bit
(420) Connected network device resumes sending traffic

Figure 2C

Fig. 3A

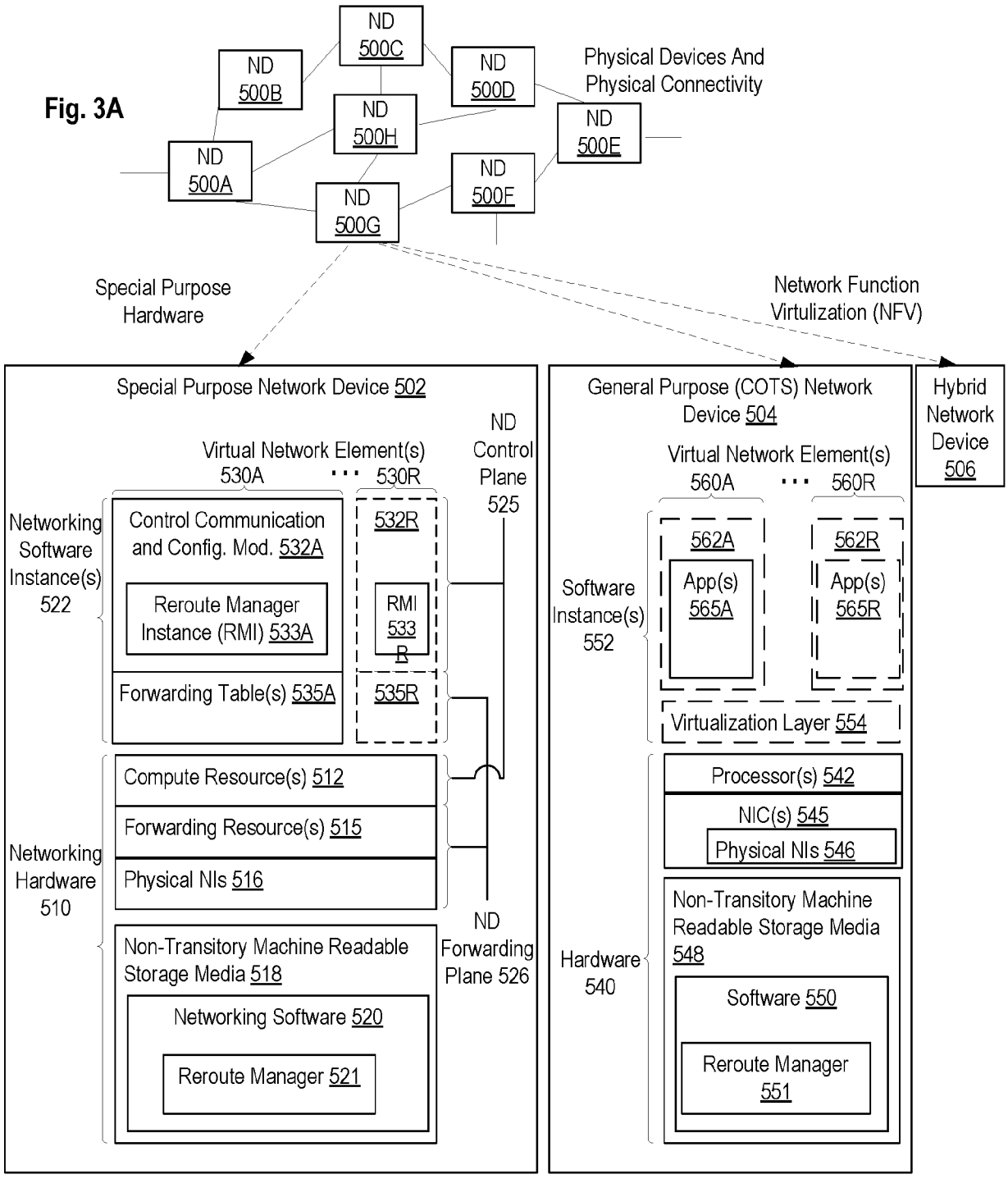

ND 500A
ND 500B
ND 500C
ND 500D
ND 500E
ND 500F
ND 500G
ND 500H
Physical Devices And Physical Connectivity
Special Purpose Hardware
Network Function Virtulization (NFV)
Special Purpose Network Device 502
Virtual Network Element(s) 530A ... 530R
ND Control Plane 525
Networking Software Instance(s) 522
Control Communication and Config. Mod. 532A
Reroute Manager Instance (RMI) 533A
Forwarding Table(s) 535A
532R
RMI 533R
535R
Networking Hardware 510
Compute Resource(s) 512
Forwarding Resource(s) 515
Physical NIs 516
Non-Transitory Machine Readable Storage Media 518
Networking Software 520
Reroute Manager 521
ND Forwarding Plane 526
General Purpose (COTS) Network Device 504
Virtual Network Element(s) 560A ... 560R
Software Instance(s) 552
562A
App(s) 565A
562R
App(s) 565R
Virtualization Layer 554
Hardware 540
Processor(s) 542
NIC(s) 545
Physical NIs 546
Non-Transitory Machine Readable Storage Media 548
Software 550
Reroute Manager 551
Hybrid Network Device 506

Fig. 3B

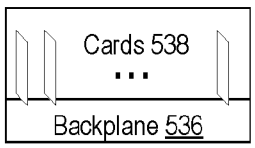

Cards 538
...
Backplane 536

Fig. 3C

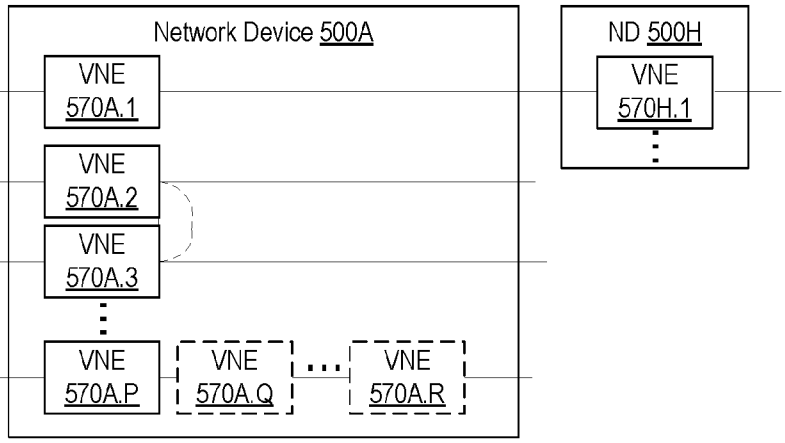


Fig. 3D

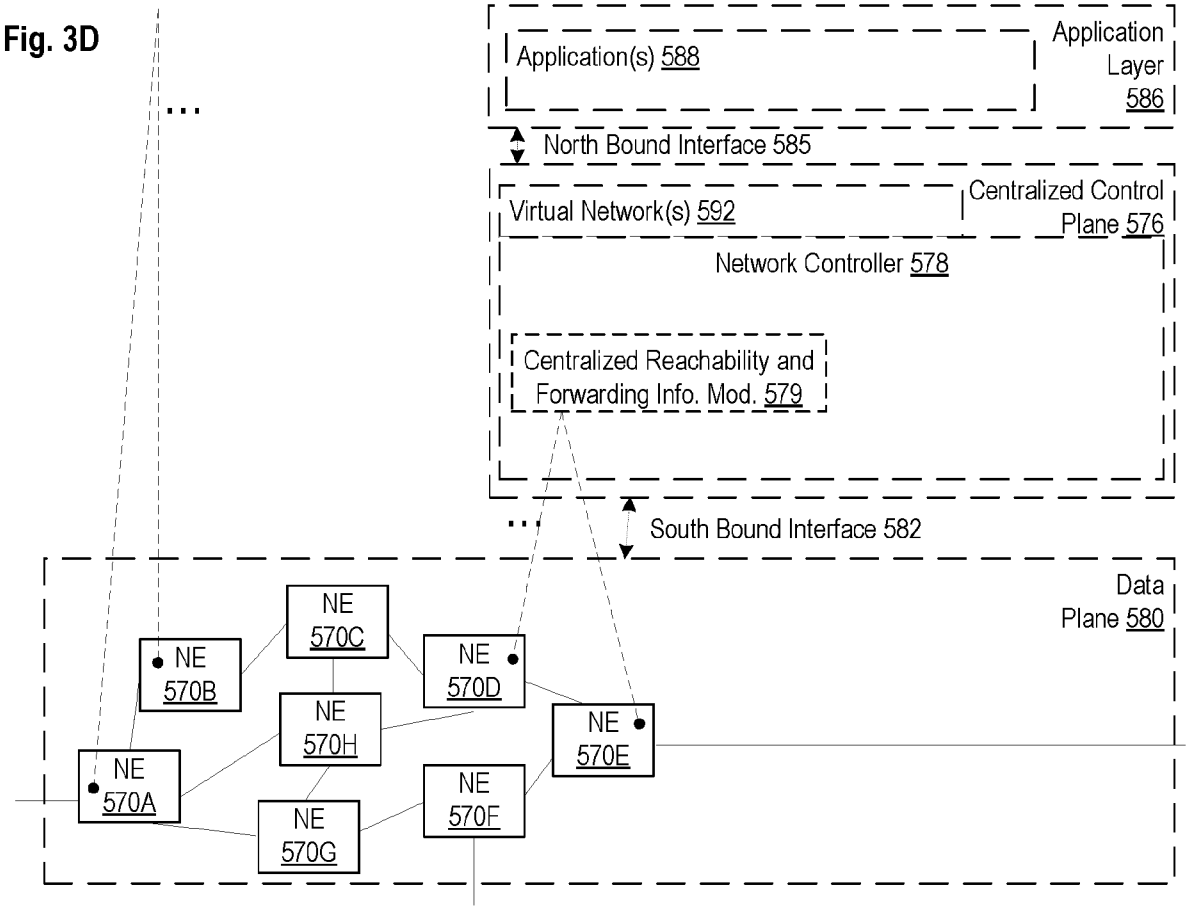


Fig. 3E

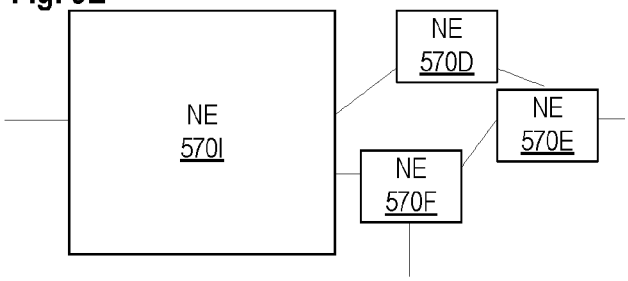


Fig. 3F

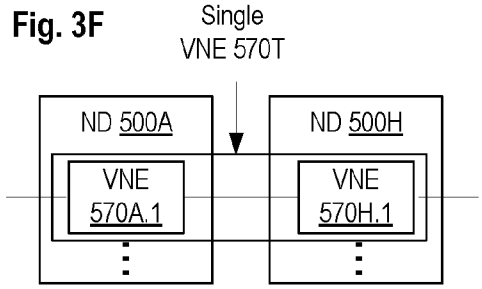
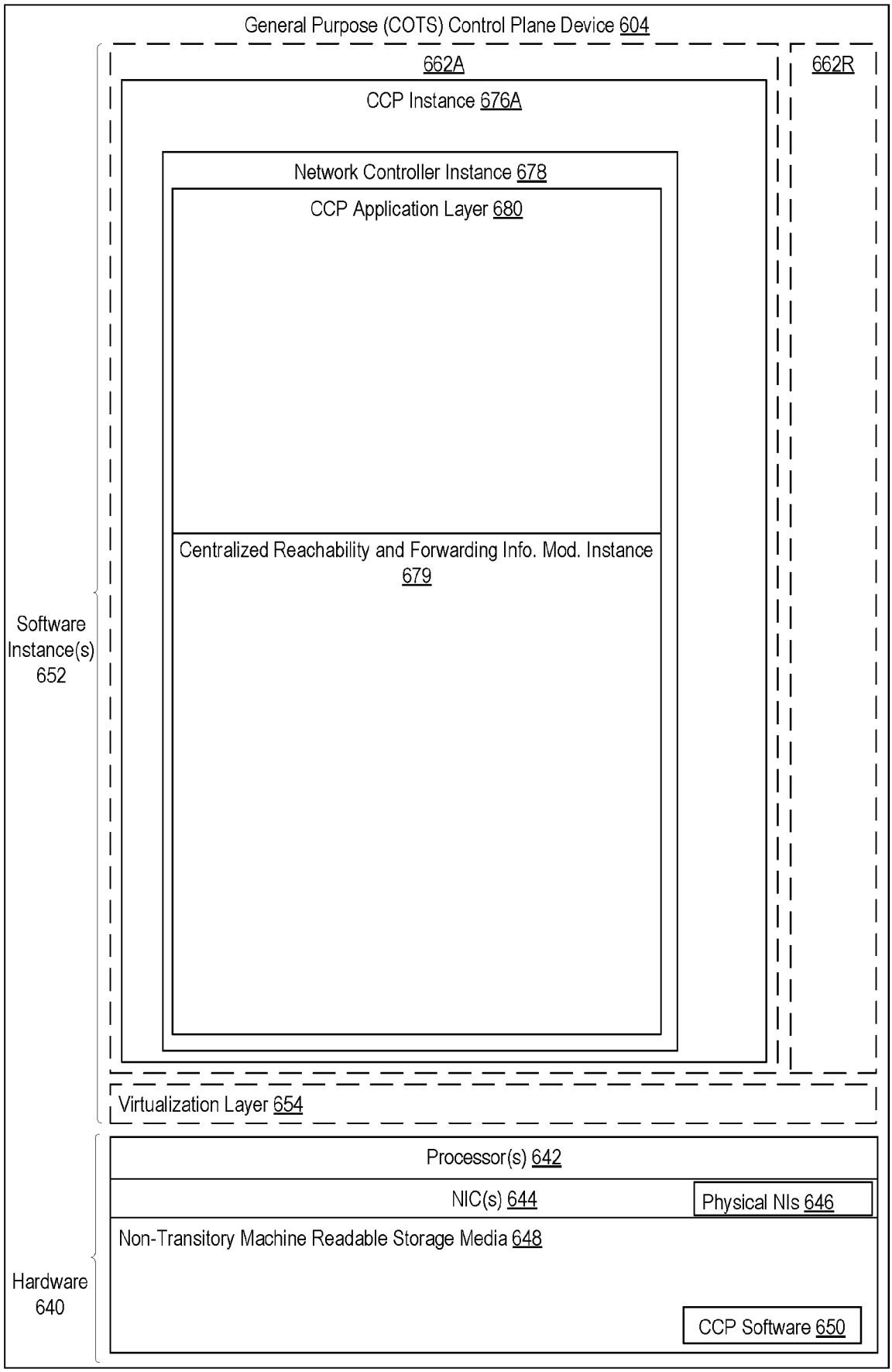


Fig. 4


General Purpose (COTS) Control Plane Device 604
662A
662R
CCP Instance 676A
Network Controller Instance 678
CCP Application Layer 680
Centralized Reachability and Forwarding Info. Mod. Instance 679
Software Instance(s) 652
Virtualization Layer 654
Processor(s) 642
NIC(s) 644
Physical NIs 646
Non-Transitory Machine Readable Storage Media 648
CCP Software 650
Hardware 640

INTERNATIONAL SEARCH REPORT

International application No
PCT/IB2016/052308

A. CLASSIFICATION OF SUBJECT MATTER
INV. H04L12/709 H04L12/939
ADD.

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)
H04L

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

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2013/148303 A1 (ALCATEL LUCENT [FR]) 3 October 2013 (2013-10-03) figure 1 page 7, line 29 - page 8, line 8 page 23, line 27 - page 24, line 4 page 25, line 3 - line 7 page 25, line 28 - page 26, line 6 figure 9 page 5, line 27 - line 30 page 26, line 4 - line 6 figure 11 page 29, line 1 - line 19 ----- -/--	1-21

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

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
15 November 2016	21/11/2016

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Bos, Jürgen

INTERNATIONAL SEARCH REPORT

International application No
PCT/IB2016/052308

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	"IEEE Std 802.1AX (TM) -2008 IEEE Standard for Local and metropolitan area networks-", , 3 November 2008 (2008-11-03), pages 1-163, XP055216396, 3 Park Avenue, New York, NY 10016-5997, USA Retrieved from the Internet: URL:http://standards.ieee.org/getieee802/download/802.1AX-2008.pdf [retrieved on 2015-09-28] chapter 5.4.15; page 54	1-21

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/IB2016/052308

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2013148303 A1	03-10-2013	CN 104221336 A	17-12-2014
		EP 2832059 A1	04-02-2015
		JP 5873597 B2	01-03-2016
		JP 2015515809 A	28-05-2015
		KR 20140127904 A	04-11-2014
		WO 2013148303 A1	03-10-2013