Office de la Propriété Intellectuelle du Canada
Un organisme d'Industrie Canada

Canadian Intellectual Property Office
An agency of Industry Canada

CA 2655984 C 2013/05/28

(11)(21) 2 655 984

(12) BREVET CANADIEN
CANADIAN PATENT

(13) C

(86) Date de dépôt PCT/PCT Filing Date: 2007/11/29
(87) Date publication PCT/PCT Publication Date: 2008/06/05
(45) Date de délivrance/Issue Date: 2013/05/28
(85) Entrée phase nationale/National Entry: 2008/12/17
(86) N° demande PCT/PCT Application No.: US 2007/085977
(87) N° publication PCT/PCT Publication No.: 2008/067493
(30) Priorités/Priorities: 2006/11/30 (US11/565,563); 2007/03/11 (US60/894,223); 2007/10/31 (US11/933,020)

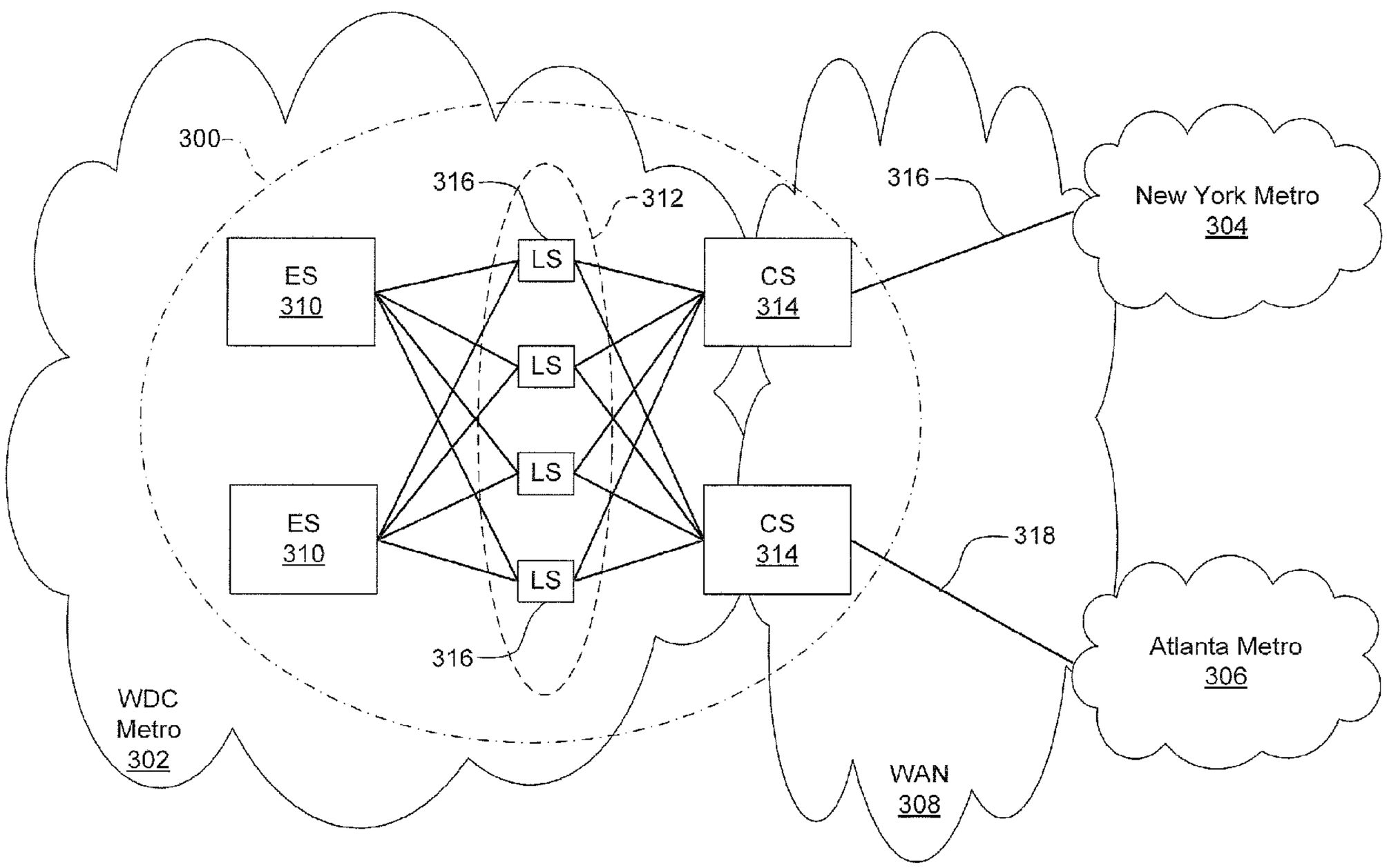
(51) Cl.Int./Int.Cl. *H04L 12/46* (2006.01)

(72) Inventeurs/Inventors:
LAWRENCE, JOSEPH, US;
GIBBINGS, CHRISTOPHER J., US;
COMSTEDT, NICLAS, US;
EL-AAWAR, NASSAR, US

(73) Propriétaire/Owner:
LEVEL 3 COMMUNICATIONS, LLC, US

(74) Agent: MARKS & CLERK

(54) Titre : SYSTEME ET PROCEDE PERMETTANT DE COMMUTER LE TRAFIC A TRAVERS UN RESEAU
(54) Title: SYSTEM AND METHOD FOR SWITCHING TRAFFIC THROUGH A NETWORK



(57) Abrégé/Abstract:
Embodiments of the present invention generally relate to network communications. More specifically, embodiments relate to a system and method for switching data through a network. An embodiment of a switching system communicatively couples an external network to a wide area network. The system includes a plurality of edge switches communicatively coupled to the external network, a plurality of core switches communicatively coupled to the wide area network, and an interconnected matπx of switches communicatively coupled to the core switches and the edge switches and configured to forward communication traffic between the edge switches and the core switches.

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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
5 June 2008 (05.06.2008)

PCT

(10) International Publication Number
WO 2008/067493 A3

(51) International Patent Classification:
H04L 12/50 (2006.01)

(21) International Application Number:
PCT/US2007/085977

(22) International Filing Date:
29 November 2007 (29.11.2007)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

11/565,563	30 November 2006 (30.11.2006)	US
60/894,223	11 March 2007 (11.03.2007)	US
11/933,020	31 October 2007 (31.10.2007)	US

(71) Applicant *(for all designated States except US)*: **LEVEL 3 COMMUNICATIONS, LLC** [US/US]; 1025 Eldorado Boulevard, Broomfield, CO 80021 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **LAWRENCE, Joseph** [US/US]; 2989 Tin Cup, Boulder, CO 80305 (US). **GIBBINGS, Christopher** [US/US]; 3074 West 113th Court, Unit D, Westminster, CO 80031 (US). **COMST-EDT, Niclas** [SE/US]; 14000 Winding River Court, Unit R2, Broomfield, CO 80020 (US). **EL-AAWAR, Nassar** [US/US]; 466 South Downing, Denver, CO 80209 (US).

(74) Agent: **RIETH, Damon A.**; Hensley Kim & Holzer, LLC, 1660 Lincoln Street, Suite 3000, Denver, CO 80264 (US).

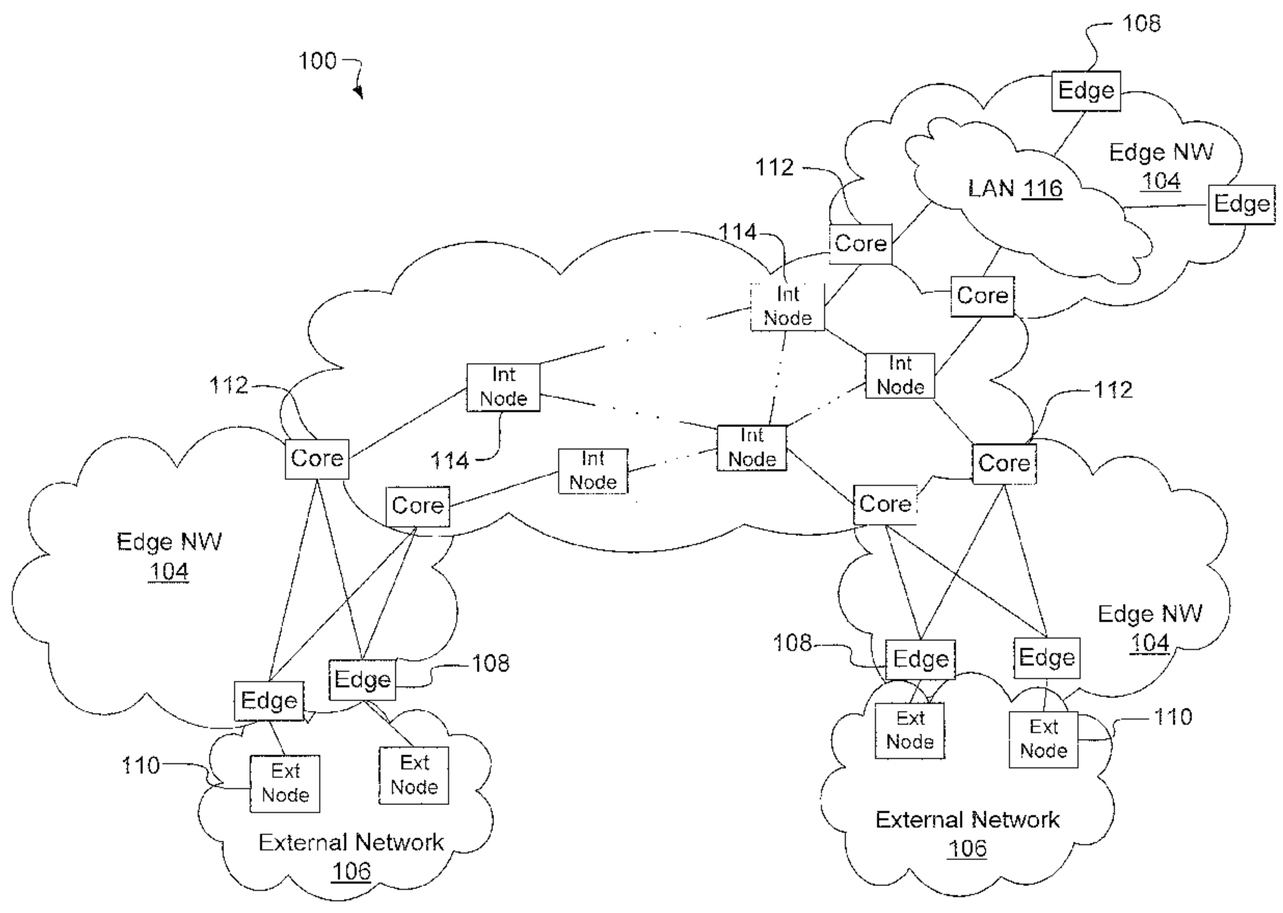
(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN,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

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

(88) Date of publication of the international search report:
17 July 2008

(54) Title: SYSTEM AND METHOD FOR SWITCHING TRAFFIC THROUGH A NETWORK



(57) Abstract: Embodiments of the present invention generally relate to network communications. More specifically, embodiments relate to a system and method for switching data through a network. An embodiment of a switching system communicatively couples an external network to a wide area network. The system includes a plurality of edge switches communicatively coupled to the external network, a plurality of core switches communicatively coupled to the wide area network, and an interconnected matπx of switches communicatively coupled to the core switches and the edge switches and configured to forward communication traffic between the edge switches and the core switches.

System and Method for Switching Traffic Through A Network

CROSS REFERENCE TO RELATED APPLICATIONS

[001] This application claims the benefit of priority to United States Patent Application Serial No. 11/933,020, filed October 31, 2007, entitled "System and Method for Switching Traffic Through a Network," published as US 2008-0151863 and United States Patent Application Serial No. 11/565,563, filed November 30, 2006, entitled "Systems and Methods for Network Routing in a Multiple Backbone Network Architecture" and published as US 2007-0086429. This application further claims the benefit of priority from United States Provisional Patent Application Serial No. 60/894,223, filed March 11, 2007, entitled "Scalable Network Node Configurations".

COPYRIGHT NOTICE

[002] Contained herein is material that is subject to copyright protection. The copyright owner has no objection to the facsimile reproduction of the patent disclosure by any person as it appears in the Patent and Trademark Office patent files or records, but otherwise reserves all rights to the copyright whatsoever. Copyright (c) 2007 Level 3 Communications, LLC.

TECHNICAL FIELD

[003] Embodiments of the present invention generally relate to network communications. More specifically, embodiments relate to a system and method for switching data through a network.

BACKGROUND

[004] Communication traffic volume over the Internet continues to grow rapidly. For example, growth in demand can be attributed in part to users' increasing demand for rich content over the Internet from video sharing web sites, pod casters and other content distributors and providers. In addition, sites devoted to social networking continue to grow in popularity, thereby contributing to increasing demand in online use. Furthermore, users are accessing the Internet through more and varied types of end-user devices, such as portable handheld

computing devices, and the communication and computing options available to users continue to increase. As a result, users are accessing a broader range of Internet resources from ever- increasing points of access.

[005] Network service providers (NSPs), such as Internet service providers (ISPs), wholesale NSPs, and backbone network operators, must adapt their networks in response to increases in communication traffic volume. More specifically, providers should be able to accommodate growth in communication traffic while maintaining the same relative service levels at roughly the same cost or lower cost. Given the relatively high capital costs associated with network implementation, integration and maintenance, and uncertainty in the trend in communication traffic it can be difficult to know what configuration of network components will provide both scalability (ability to grow with increases in traffic demand) and low cost.

[006] It is with respect to these and other problems that embodiments of the present invention have been created.

SUMMARY

[007] Embodiments of the present invention generally relate to network communications. More specifically, embodiments relate to a system and method for switching data through a network.

[008] An embodiment of a switching system communicatively couples an external network to a wide area network. The system includes a plurality of edge switches communicatively coupled to the external network, a plurality of core switches communicatively coupled to the wide area network, and an interconnected matrix of switches communicatively coupled to the core switches and the edge switches and configured to forward communication traffic between the edge switches and the core switches.

[009] In an embodiment of a system each of the edge switches, core switches and interconnected matrix of switches is an independent switching element. Each of the edge switches, core switches and interconnected matrix of switches can build or maintain a forwarding table by exchanging route information with the other switches. The interconnected matrix of switches can include one or more of a full mesh, a partial mesh, a star, a ring, a n-cube, a toroid, a CLOS-like matrix,

or an arbitrary mesh configuration. The switches in the interconnected matrix may forward communication traffic using one or more of an Ethernet protocol, an MPLS protocol, on a nonproprietary protocol. The interconnected matrix of switches may be part of or form a Local Area Network (LAN). The external network may include one of a metropolitan area network, a campus area network, an enterprise network or a personal area network.

[010] In an embodiment of a system the edge switches and the core switches forward communication traffic using one or more of an Ethernet protocol, an MPLS protocol, an Internet Protocol, or a nonproprietary protocol. Each of the edge switches and the core switches may include its own forwarding table. The interconnected matrix of switches may include multiple stages of switches. In some embodiments, the interconnected matrix of switches includes ports facing the wide area network and ports facing the external network, and wherein each of the ports facing the wide area network are connected to a port on a core switch and each of the ports facing the external network are connected to a port on an edge switch.

[011] In an embodiment of a system the edge switches, the interconnected matrix of switches, and core switches are housed in separate chassis. The edge switches may form an edge switching stage, the interconnected matrix of switches may form one or more middle switching stages, and the plurality of core switches may form a core switching stage. The independent switching elements may be scaled according to the number of switches at each stage in the interconnected matrix of switches.

[012] An embodiment of a network architecture includes a wide area network, a plurality of local networks configured to communicate over the wide area network, and a plurality of switching stages communicatively coupling the local networks to each other over the wide area network, wherein local groups of switching stages communicatively couple associated local networks to the wide area network. Each of the local networks may be in a different metropolitan area. Each group of stages may include an interconnected matrix of switches. In some embodiments at least one interconnected matrix of switches is directly connected to at least one other interconnected matrix of switches over the wide area network.

[013] In an embodiment of a network architecture, connections between interconnected matrices of switches include one or more of a braid topology or a ladder topology. Each switch may be configured to build its own forwarding table. Each interconnected matrix of switches may form a topology including one or more of a ring, a star, a full mesh, a partial mesh, an arbitrary mesh, a CLOS-like matrix, an n-cube, or a hybrid. In some embodiments at least one group of stages includes an edge stage, a middle stage and a core stage. In these embodiments, switches of the middle stage may have a plurality of ports facing the wide area network and a plurality of ports facing the local network, and each port facing the wide area network may be connected to a port on a switch in the core stage and each port facing the local network may be connected to a port on a switch in the edge stage. In an embodiment of a network architecture at least one of multiple switching stages includes a plurality of Ethernet switches.

According to an aspect of the present invention there is provided a switching system coupling an external network to a wide area network, the system comprising:

a plurality of edge switches which form an edge stage of switches and which are configured to be communicatively coupled to the external network;

a plurality of core switches which form a core stage of switches and which are configured to be communicatively coupled to the wide area network; and

an interconnected matrix of switches which form a middle stage of switches and which are configured to be communicatively coupled to the core switches and the edge switches and configured to forward communication traffic between the edge switches and the core switches

wherein each stage of switches is loosely coupled to the other stages of switches whereby the edge switches do not need to be capable of selecting a port on a core switch through which to forward data, and vice versa.

According to another aspect of the present invention there is provided a network architecture comprising:

a wide area network;

a plurality of local networks configured to communicate over the wide area network; and

a plurality of switching stages which are configured to communicatively couple the local networks to each other over the wide area network,

wherein local groups of switching stages are configured to communicatively couple associated local networks to the wide area network;

wherein at least one group of switching stages comprises an edge stage comprising a plurality of edge switches, a middle stage comprising an interconnected matrix of switches and a core stage comprising a plurality of core switches and

wherein each switching stage is loosely coupled to the other switching stages whereby the edge switches do not need to be capable of selecting a port on a core switch through which to forward data, and vice versa.

BRIEF DESCRIPTION OF THE DRAWINGS

[014] Fig. 1 illustrates an operating environment suitable for implementing embodiments of the present invention.

[015] Fig. 2 illustrates an exemplary multi-stage switching system in accordance with various embodiments.

[016] Figs. 3 - 4 illustrate other exemplary multi-stage switching systems applied to metropolitan areas in accordance with various embodiments.

[017] Fig. 5A - 5D illustrate exemplary local interconnected matrix (LIM) switch topologies.

[018] Fig. 6 illustrates an exemplary multi-stage, multi-chassis switching system employing scalable packet forwarding in accordance with various embodiments.

[019] Fig. 7 is a flowchart illustrating a process of switching a packet through a multi-stage switching system including a LIM.

[020] Fig. 8 illustrates a general purpose computing device upon which one or more aspects of embodiments of the present invention may be implemented.

[021] While the invention is amenable to various modifications and alternative forms, specific embodiments have been shown by way of example in the drawings and are described in detail below. The intention, however, is not to limit the invention to the particular embodiments described.

DETAILED DESCRIPTION

[022] Embodiments of the present invention relate to switching (e.g., forwarding) data through a network. More specifically, embodiments relate to systems and methods for forwarding traffic through multi-stage switching systems. In various embodiments, the multi-stage switching systems can include one or more inner stages. The inner stages may include a cluster of parallel switches. The cluster of parallel switches may comprise an interconnected matrix of switches forming one or more topologies. One or more of the switch stages may be housed in its own chassis. Such a system is referred to as a multi-chassis switch system. Each switch stage can be loosely coupled to the other stages and a nonproprietary protocol can be used. Packets can be forwarded through the stages using Internet Protocol, MPLS, or other protocols.

[023] Using multiple chassis with standard interfaces (e.g., not proprietary) provides for multi-generation, multi-vendor configurations. In such embodiments each stage can include a different vendor or generation of line cards or switch cards. Inner stages can be composed of alternate mesh topologies, such as, but not limited to toroidal mesh, ring, star, N stage Clos matrix, or equal path Clos. Forwarding logic in an initial stage of line cards may be simplified to a “next hop” analysis, wherein the line cards determine which of N switch ports to forward each packet. The forwarding logic is scalable as the parallel cluster of switches scales.

[024] In some embodiment, commodity switches can be used. Such embodiments can reduce cost, while allowing for scaling up in response to traffic growth and achieving efficiency in forwarding traffic between large metropolitan areas. For example, Ethernet switches are typically less expensive than most commercial IP routers. For less cost, substantially the same amount of capacity can be obtained using commodity switches, as compared to many commercial routers.

In addition, the parallel clustered configuration allows for N:1 protection of the edge switches, instead of 1:1, with a relatively small reduction in capacity.

[025] Prior to describing one or more preferred embodiments of the present invention, definitions of some terms used throughout the description are presented.

Definitions

[026] The term “network service provider” refers to an organization that provides access to one or more networks. An NSP may operate, maintain or lease a network coupled to a plurality of other networks, whereby communication traffic can be communicated to and from the various networks.

[027] The term “facing” is used in a logical sense and not a physical sense to describe the communicative orientation of ports.

[028] A “bundle” or similar terms refer to one or more physical links.

[029] A “module” is a self-contained functional component. A module may be implemented in hardware, software, firmware, or any combination thereof.

[030] The terms “connected” or “coupled” and related terms are used in an operational sense and are not necessarily limited to a direct connection or coupling.

[031] The phrases “in one embodiment,” “according to one embodiment,” and the like generally mean the particular feature, structure, or characteristic following the phrase is included in at least one embodiment of the present invention, and may be included in more than one embodiment of the present invention. Importantly, such phases do not necessarily refer to the same embodiment.

[032] If the specification states a component or feature “may”, “can”, “could”, or “might” be included or have a characteristic, that particular component or feature is not required to be included or have the characteristic.

[033] The terms “responsive” and “in response to” includes completely or partially responsive.

[034] The term “computer-readable media” is media that is accessible by a computer, and can include, without limitation, computer storage media and communications media. Computer storage media generally refers to any type of computer-readable memory, such as, but not limited to, volatile, non-volatile, removable, or non-removable memory. Communication media refers to a modulated signal carrying computer-readable data, such as, without limitation, program modules, instructions, or data structures.

Exemplary System

[035] Fig. 1 illustrates an exemplary operating environment 100 in which embodiments of the present invention may be implemented. The operating environment 100 shown in Fig. 1 is presented for illustrative purposes only and is not intended to limit the invention for use in any particular environment. Embodiments may be used in many other environments, such as, but not limited to, data centers, corporate networks, and interexchange points.

[036] The illustrated operating environment 100 includes a wide area network composed of a wide area network (WAN) 102 (e.g., a backbone network) and multiple edge networks 104. The edge networks 104 are communicatively located between external networks 106 and the WAN 102, whereby communication traffic traveling to and from the WAN 102 to the external networks 106 (and vice versa) traverse the respective edge networks 104. The term “external” is used to indicate a node is external to the WAN formed by the WAN 102 and the edge networks 104.

[037] In some embodiments one or more of the external networks 106 is an Internet service provider (ISP) network (e.g., tier 1, tier 2 or tier 3 ISP). In other embodiments, one or more of the external networks 106 include, by way of example, but not limitation, one or more of a metropolitan area networks (MAN), campus area network (CAN), personal area networks (PAN), enterprise network, home or office network.

[038] In accordance with various embodiments, the WAN 102 may span a relatively large geographic area that may include multiple distinct areas, such as metropolitan areas, campuses or enterprise locations. The edge networks 104 provide links between the WAN 102 and external networks 106 in the geographic regions. An example of an edge network 104 is a Point-of-Presence (POP) facility. As discussed further below, an edge network 104 can include one or more switching nodes that include multiple switching stages. These stages generally include one or more edge nodes 108 and one or more core nodes 112.

[039] In the illustrated embodiment, the edge nodes 108 are coupled to one or more external nodes 110 of external networks 106. External nodes 110 may be routers, switches or other network devices. Core nodes 112 communicatively couple the edge network 104 to the WAN 102. In some embodiments core nodes 112 are

communicatively coupled to internal nodes 114 of the WAN 102. As discussed further below, core nodes 112 in one area (e.g., metropolitan area) may be connected directly to core nodes 112 of another area via link bundles (e.g., link aggregation group (LAG)) in accordance with a multi-stage switching embodiment.

[040] Internal nodes 114 are typically, by way of example, routers or switches that preferably forward traffic at high speed through the WAN 102. The embodiment shown in Fig. 1 includes only a small number of internal nodes 114, core nodes 112, edge nodes 108 and external nodes 110, but in actual operation, there will be many more of these types of nodes.

[041] In accordance with some embodiments, edge nodes 108 may be communicatively coupled to core nodes 112 via a local area network (LAN) 116. In such embodiments, the LAN 116 includes switches that forward packets between edge nodes 108 and core nodes 112 using one or more communication protocols. In various embodiments, the switches in the LANs 116 may perform Internet protocol (IP) forwarding, or some other buffered forwarding protocol, such as Multi-protocol label switching (MPLS), or some combination of those protocols.

[042] In various embodiments, edge nodes 108 and core nodes 112 of the edge network 104 form a switch cluster. In accordance with these embodiments, multiple switches can be installed at each of one or more switching stages in the edge network 104. Packets traversing the edge network 104 can be switched in a "next-hop" fashion whereby packets are forwarded to the next stage using a non-proprietary protocol. Each stage may be implemented in a separate chassis that is accessible for adding and removing switch cards (e.g., Ethernet cards), to allow for easy scaling and multi-generational and multi-vendor implementations.

[043] In some embodiments, switch stages can be logically extended across the WAN 102. Switch stages in one local area may be directly connected to switch stages in another local area across the WAN 102. In such embodiments, switch stages in the two areas are part of a single mesh of switches.

[044] Fig. 2 illustrates an exemplary network configuration 200 including a wide area network 202 communicatively coupled to a first network site 204 and a second network site 206. A set of switch stages communicatively couples the network sites to the WAN 202. For example, stage 1A 208, stage 1B 210 and stage 1C 212 facilitate communications between the WAN 202 and the first network site

204. Stage 2A 214, stage 2B 216 and stage 2C 218 facilitate communications between WAN 202 and the second network site 206.

[045] The embodiment of Fig. 2 can be applied to more than two network sites. Switch stages may be added or removed. The network sites may be different enterprise sites, metropolitan area networks, campus area networks or others. Many different network topologies may be implemented within each of the stages, as well as between the network sites. Exemplary embodiments are discussed below.

[046] Fig. 3 illustrates an exemplary switching system 300 in accordance with the embodiment of Fig. 2. This embodiment generally depicts three metropolitan area networks, Washington DC 302, New York City 304 and Atlanta 306, communicatively coupled over a WAN 308. The switching system 300 resides generally between the WDC network 302 and the WAN 308.

[047] The switching system 300 has a first stage of provider edge switches 310 (e.g., edge routers), a middle stage including a local interconnect matrix (LIM) 312, and a third stage of provider core switches 314 (e.g., core routers). The LIM 312 includes a cluster of LIM switches (LS) 316. In the exemplary embodiment, the switching system 300 links the Washington DC metro network 302, the New York metro network 304 and the Atlanta metro network 306 via link bundle 316 and link bundle 318, respectively (link bundles are shown with thicker lines). The links between the LIM switches 316 are also link bundles, and, in some embodiments, may comprise 4 parallel 10 Gig-E links.

[048] The core switches 314 form a core stage of switches and the edge switches 310 form an edge stage of switches. In some embodiments, the LIM switches 316 can comprise multiple stages of switches. Only one stage is illustrated in Fig. 3 for ease of description.

[049] In one embodiment, the edge switches 310 form a first stage in a Clos-like matrix, the LIM 312 forms a second stage in the matrix, and the core switches 314 form a third stage in the matrix. In other embodiments, more stages may be included. For example, the local interconnect matrix 312 may include multiple switch stages in any topology. The stage or stages of the LIM 312 may include an aggregation layer that provides intelligent packet buffering or other functionality. For example, more complex functionality that is often attributed to commercial routers, such as Quality of Service (QoS), may be incorporated into the LIM 312. The LIM

312 may also perform Equal-cost Multi-path (ECMP) routing to balance the traffic load across edge switches 310 and core switches 314.

[050] The switches 316 of the LIM 312 may form one or more meshes, which may be of one or more different topologies. Such alternate mesh topologies may or may not include a Clos-like matrix. By way of example, but without limitation, a mesh of switches 316 may form an equal Clos path or an NxM Clos matrix. For example, the Clos matrix may be made up of a number of stages of switches 316, in which each stage is made up of the same or a different number of parallel switches. Alternatively, the switches 316 of the LIM 312 may be in ring, star, n-cube, toroidal, full mesh, partial mesh, arbitrary mesh, or hybrid topologies.

[051] In one embodiment the LIM 312 is Clos-like, with each edge switch 310 connecting to the center stage with different capacity according to the edge switch's 310 bandwidth requirements. In this embodiment it is not strictly a Clos matrix, as each element may have differing bandwidth that can vary over time, and the LIM 312 can be blocking.

[052] The switches 316 may be pure layer 2 or hybrid layer 2/layer 3. The switches 316 of the switch fabric may each be configured with its own broadcast LAN facing the edge switches 310. The edge switches 310 can set up adjacencies for protocols not necessarily carried by the switches 316, such as MPLS.

[053] In one embodiment of the switching system 300, the edge switches 310, local interconnect matrix 312 and core switches 314 are each comprised of independent switching elements in separate chassis. In this embodiment, the output ports of each switch 316 are coupled to input ports of core switches 314, and vice versa. Similarly, output ports of each edge switch 310 are coupled to input ports of the switches 306, and vice versa. In this regard, each switch 316 in the middle stage has multiple ports logically facing the WAN, which are each connected to a port of a core switch 314; and each switch 316 in the middle stage has multiple port logically facing the metropolitan (or local) network, which are each connected to a port on an edge switch 310.

[054] With regard to data (e.g., packet) forwarding through the switching system 300, the configuration shown in Fig. 3 can be loosely coupled, whereby edge switches 310 do not need to be capable of selecting a port on a core switch 314 through which to forward data, and vice versa. The edge switches 310 and core switches 314 can perform simple next-hop forwarding through the LIM 312. In

addition, a nonproprietary protocol can be used for packet forwarding at each stage, and chassis slots and ports are available for connection and reconnection of components. For example, the core switches 314 may use a layer 2 (e.g., Ethernet) or layer 2½ switching protocol (e.g., MPLS), rather than a layer 3 routing protocol (e.g., Internet protocol). In some embodiments, the LIM switches 316 are relatively simple commodity switches. The open protocol and decoupling of core switches 314 from edge switches 310 can enable flexible scaling in the LIM 312, as well as at the edge and the core.

[055] Such a configuration allows for greater flexibility with regard to component selection and interconnection. For example, individual components at each stage can be selectively interchanged. In addition, the components may be from different vendors or different generations of at different stages of the system 300. For example, edge switches 310 could be simple commodity switches, or edge switches 310 could be more sophisticated routers, such as Juniper™ routers. Similarly, core routers 314 could be simple commodity switches, or more sophisticated routers, such as Cisco™ routers.

[056] The configuration shown in Fig. 3 also allows for greater degrees of data protection at substantially the same throughput and potentially lower cost than conventional configurations. At a minimum, the multi-stage, multi-chassis configuration provides a network capacity of N times CSW, where N is the number of switches 306 in the local interconnect matrix 304, or a network capacity of N-1 times CSW with data protection. In this context, CSW refers to uplink LAG capacity to the core switch or uplink bundle capacity to the core switch. For example, with 4 switches in the inner stage, network capacity is 4 times CSW. Further efficiency is gained by local switching within the ES 310 and CS 314 stages.

[057] In contrast to a conventional implementation with a tightly coupled control plane, the described embodiments can allow for higher system reliability, in part because the control and forwarding protocol functions are loosely coupled and can operate independently. This enables the use of N:1 protection of the ES 310 and CS 314 uplinks in contrast to a less efficient 1:1 protection of uplinks to a traditional node design with dual core routers. For example with 4 uplinks, if there is 1:1 protection 2 uplink are available for working traffic while 2 uplinks are required for protection. If there is 3:1 protection, one link bundle is required for redundancy, and

the network capacity is 3 uplinks. This uplink efficiency reduces costs on the ES as well as increasing the capacity of the overall system.

[058] Fig. 4 illustrates another exemplary multi-stage switching system 400 in accordance with various embodiments. The particular embodiment illustrates variations on the basic LIM configuration. In addition, the mesh of a LIM at an edge network (e.g., a POP) may be extended across a wide area network (WAN). Such variations are illustrated in Fig. 4.

[059] Fig. 4 again illustrates metropolitan networks in Washington DC 402, New York City 404 and Atlanta 406 communicatively coupled via a WAN 408. Switch stages are shown at each of the metro areas. One of the stages at each metro area is a LIM. for example, there is a WDC LIM 410, a NYC LIM 412, and an Atlanta LIM 414. Each of the metro areas has core switches as well.

[060] LIM switches may be arranged in any of a number of topologies. For example, in the illustrated embodiment, the WDC LIM switches 416 are arranged as two n-cube configurations 418. It is to be understood that any type of configuration can be used in a LIM, including, but not limited to, full mesh, partially connected mesh, star, ring, toroid, arbitrary mesh, or any combination thereof (e.g., hybrid configurations). Some exemplary topologies are shown in Fig. 5.

[061] Referring to Fig. 5, there are shown just a few possible topologies for LIMs. Fig. 5A illustrates a partially connected mesh of switches. Fig. 5B illustrates a toroid topology. Fig. 5C illustrates a star ring topology. Fig. 5D illustrates and arbitrary mesh topology. Of course, the invention is not limited to any of these, and hybrid topologies can be implemented. The choice of the LIM switch configuration may depend on numerous factors, such as traffic volume, cost of infrastructure, location or others.

[062] In the illustrated embodiment, the WDC core switches 420 are coupled to switches 422 in the New York City metro network 404 and switches 424 in the Atlanta metro network 406 via link bundles 426 and 428, respectively. This linking configuration, in which the LIM mesh is extended out across the WAN is referred to as “braided”. The braided configuration can enable efficient communication switching in the WDC network 402 for traffic between the New York City network 404 and the Atlanta network 406. This configuration enables the WAN through traffic to be locally switched in the CS 420. This reduces the CS 420 uplink capacity requirements to the LIM 410 and so allows for increasing the number of core

switches 420 (e.g., scaling up) in response to traffic increases in either metro market. The braiding concept can be extended across multiple WAN hops and throughout the entire network. In one embodiment, referred to as a “ladder” configuration, the braiding is implemented as parallel links between different metro locations and the parallel braiding is continued across multiple hops throughout the network. In this embodiment the CS 420 switches may be reside in different locations within the metro and the parallel links may be routed over diverse fiber paths to enable efficient protection within the WAN

[063] In an alternative embodiment, the core switches may be directly linked to the center stage; e.g., the LIM stage of other metro networks. For example, the core switches 420 can be directly connected to the switches in the NYC LIM 412. Further still, in another embodiment, middle stages may be directly connected to other middle stages. For example, the WDC core the LIM switches 416 can be directly linked to the switches of the NYC LIM 412 and the Atlanta LIM 414.

[064] Fig. 6 illustrates an exemplary multi-chassis switching system 600 in accordance with one embodiment. The multi-chassis switching system 600 includes a first forwarding card chassis 602, a switch fabric chassis 604 and a second forwarding card chassis 606. Forwarding card chassis 602 includes a plurality of independent forwarding elements (IFE) 608 and forwarding card chassis 606 includes another plurality of IFEs 610. The IFEs 608 and IFEs 610 are independent of the other IFEs. Each IFE has its own forwarding table, which may be built using a link advertising protocol. Switch fabric 604 includes a cluster of multiple parallel switching devices 612. In some embodiments, the cluster of switching devices form a local interconnect matrix (LIM).

[065] The switching devices 612 perform parallel switching of packets (or datagrams or frames) between IFEs 608 and IFEs 610. In one embodiment, the switching fabric 612 includes a designated number (e.g., 4, 8) of Ethernet switches connected in such a way to form one or more stages of a Clos-like matrix. The system in this embodiment may be blocking or non-blocking. The components of the system 600 (i.e., forwarding card chassis 602, switch chassis 604 and forwarding card chassis 606) are divided into separate chassis that are individually accessible and functionally independent. In addition, packets can be forwarded through the switching system 600 using IP protocol or another protocol that is not proprietary. As

such, the embodiment shown in Fig. 6 allows for easy upgrading and scalability to the system 600.

[066] In addition, functionality can be included in the switch devices 612 for performing higher level functions in addition to switching. For example, switch devices 612 can perform buffering. As another example, switch devices 612 can make switching decisions based on demand and capacity. That is, the switch devices 612 can determine demand placed on other stages of the system and make traffic forwarding decisions based on that demand and available and/or necessary capacity.

[067] Stages in the switching fabric 604 can receive demand or capacity information from other stages via traffic engineering tunnels (e.g., MPLS tunnels) or via layer two VLANs. Alternatively, changes to IGP can be leveraged to communicate bandwidth information to the switching stage 604. For example, a "downstream" stage switch can communicate to upstream stage switches (e.g., via IGP or other protocols) that it is connected to a New York-based site with 30 Gb/s of traffic. The upstream stage switches can use this protocol information, as well as information about other switches, to perform load balancing across stages of the switch fabric 604.

[068] With regard to the IFEs 608, packet forwarding is simplified and scalable. In the illustrated embodiment, the line cards serve as Ethernet switch interconnects. Packet forwarding through the IFEs 608 is a "next-hop" decision, meaning that the IFEs 608 only need to determine which port on the switch devices 612 to forward the packet to. For example, if the switch fabric 604 includes a 4 x 4 parallel cluster of Ethernet switches, the line cards only need to determine which of the 4 ports in the first stage of the switch fabric 604 to forward the packets to. As such, the forwarding table in the IFE 608 only needs an entry for each port connected to the next stage. However, the forwarding table may still be a complete forwarding table. The IFE 608 can perform load balancing across the switch devices 612, for example, by performing round robin distribution or some other load balancing policy.

Exemplary Operations

[069] Fig. 7 is a flowchart illustrating an algorithm for switching a packet through a multi-stage and/or multi-chassis switching system such as those shown in

Figs. 1 – 6. For ease of illustration, it is assumed that a middle stage cluster of parallel switches is communicatively coupled between an initial stage and a final stage of the switching system. It will be understood that numerous other switch topologies may be used.

[070] Initially in a receiving operation 702 a packet is received at a first stage of the switching system. The first stage may include one or more routers, one or more forwarding elements, or switches. In a determining operation 704 it is determined which port in the next stage the packet should be forwarded to. The determining operation 704 may involve a table look up using a next hop IP address to determine which of N ports to forward the packet. The determining operation 704 may also perform flow load balancing if multiple links exist between the switch and the next stage. A forwarding operation 706 forwards the packet to the determined port in a parallel cluster of switches.

[071] A switching operation 708 switches the packet through 1 or more stages of the parallel switch cluster. The switching may involve the MPLS protocol or other buffered protocol. At each switch, ternary content addressable memory (TCAM) is typically accessed to determine the next port to forward the packet to. In another forwarding operation 710, the packet is forwarded to a core stage in the network. From the core stage, the packet can be forwarded to its destination through a backbone or other wide area network.

[072] In an alternative embodiment, in which middle LIM stages of different metropolitan areas are connected, the switching operation 708 may forward the packet directly to a parallel switch cluster (or other middle stage LIM) of another metropolitan area. In this embodiment, the packet bypasses core routers or switches enroute to the destination metropolitan area.

Exemplary Computing Device

[073] Fig. 8 is a schematic diagram of a computing device 800 upon which embodiments of the present invention may be implemented and carried out. For example, the computing device 800 may be located at an initial stage of a network and perform packet forwarding. As discussed herein, embodiments of the present invention include various steps or operations. A variety of these steps may be performed by hardware components or may be embodied in machine-executable instructions, which may be used to cause a general-purpose or special-purpose

processor programmed with the instructions to perform the operations. Alternatively, the steps may be performed by a combination of hardware, software, and/or firmware.

[074] According to the present example, the computing device 800 includes a bus 801, at least one processor 802, at least one communication port 803, a main memory 804, a removable storage media 805, a read only memory 806, and a mass storage 807. Processor(s) 802 can be any known processor, such as, but not limited to, an Intel® Itanium® or Itanium 2® processor(s), AMD® Opteron® or Athlon MP® processor(s), or Motorola ® lines of processors. Communication port(s) 803 can be any of an RS-232 port for use with a modem based dialup connection, a 10/100 Ethernet port, a Gigabit port using copper or fiber, or a USB port. Communication port(s) 803 may be chosen depending on a network such a Local Area Network (LAN), Wide Area Network (WAN), or any network to which the computing device 800 connects. The computing device 800 may be in communication with peripheral devices (not shown) such as, but not limited to, printers, speakers, cameras, microphones, or scanners.

[075] Main memory 804 can be Random Access Memory (RAM), or any other dynamic storage device(s) commonly known in the art. Read only memory 806 can be any static storage device(s) such as Programmable Read Only Memory (PROM) chips for storing static information such as instructions for processor 802. Mass storage 807 can be used to store information and instructions. For example, hard disks such as the Adaptec® family of SCSI drives, an optical disc, an array of disks such as RAID, such as the Adaptec family of RAID drives, or any other mass storage devices may be used.

[076] Bus 801 communicatively couples processor(s) 802 with the other memory, storage and communication blocks. Bus 801 can be a PCI /PCI-X, SCSI, or USB based system bus (or other) depending on the storage devices used. Removable storage media 805 can be any kind of external hard-drives, floppy drives, IOMEGA® Zip Drives, Compact Disc – Read Only Memory (CD-ROM), Compact Disc – Re-Writable (CD-RW), Digital Video Disk – Read Only Memory (DVD-ROM).

[077] Embodiments of the present invention include various steps, which are described in this specification. The steps may be performed by hardware components or may be embodied in machine-executable instructions, which may be

used to cause a general-purpose or special-purpose processor programmed with the instructions to perform the steps. Alternatively, the steps may be performed by a combination of hardware, software and/or firmware.

[078] Embodiments of the present invention may be provided as a computer program product, which may include a machine-readable medium having stored thereon instructions, which may be used to program a computer (or other electronic devices) to perform a process. The machine-readable medium may include, but is not limited to, floppy diskettes, optical disks, compact disc read-only memories (CD-ROMs), and magneto-optical disks, ROMs, random access memories (RAMs), erasable programmable read-only memories (EPROMs), electrically erasable programmable read-only memories (EEPROMs), magnetic or optical cards, flash memory, or other type of media/machine-readable medium suitable for storing electronic instructions. Moreover, embodiments of the present invention may also be downloaded as a computer program product, wherein the program may be transferred from a remote computer to a requesting computer by way of data signals embodied in a carrier wave or other propagation medium via a communication link (e.g., a modem or network connection).

[079] Various modifications and additions can be made to the exemplary embodiments discussed without departing from the scope of the present invention. For example, while the embodiments described above refer to particular features, the scope of this invention also includes embodiments having different combinations of features and embodiments that do not include all of the described features. Accordingly, the scope of the present invention is intended to embrace all such alternatives, modifications, and variations together with all equivalents thereof.

The embodiments of the invention in which an exclusive property or privilege is claimed are defined as follows:

1. A switching system coupling an external network to a wide area network, the system comprising:

a plurality of edge switches which form an edge stage of switches and which are configured to be communicatively coupled to the external network;

a plurality of core switches which form a core stage of switches and which are configured to be communicatively coupled to the wide area network; and

an interconnected matrix of switches which form a middle stage of switches and which are configured to be communicatively coupled to the core switches and the edge switches and configured to forward communication traffic between the edge switches and the core switches

wherein each stage of switches is loosely coupled to the other stages of switches whereby the edge switches do not need to be capable of selecting a port on a core switch through which to forward data, and vice versa.

2. The system of claim 1, wherein each of the edge switches, core switches and interconnected matrix of switches is an independent switching element.

3. The system of claim 2, wherein each of the edge switches, core switches and interconnected matrix of switches develops a forwarding table by exchanging route information with the other switches.

4. The system of claim 1, wherein the interconnected matrix of switches comprises a full mesh, a partial mesh, a star, a ring, a n-cube, a toroid, a CLOS-like matrix, or an arbitrary mesh configuration, or any combination thereof.

5. The system of claim 1, wherein the switches in the interconnected matrix

forward communication traffic using an Ethernet protocol, an MPLS protocol, or a nonproprietary protocol, or any combination thereof.

6. The system of claim 1, wherein the interconnected matrix of switches comprises a Local Area Network (LAN).

7. The system of claim 1, wherein the external network comprises a metropolitan area network, a campus area network, an enterprise network or a personal area network.

8. The system of claim 1, wherein the edge switches and the core switches forward communication traffic using an Ethernet protocol, an MPLS protocol, an Internet Protocol, or a nonproprietary protocol, or any combination thereof.

9. The system of claim 1, wherein each of the edge switches and the core switches includes its own forwarding table.

10. The system of claim 1, wherein the interconnected matrix of switches comprises multiple stages of switches.

11. The system of claim 1, wherein the interconnected matrix of switches comprises ports facing the wide area network and ports facing the external network, and wherein each of the ports facing the wide area network are connected to a port on a core switch and each of the ports facing the external network are connected to a port on an edge switch.

12. The system of claim 1, wherein the edge switches, the interconnected matrix of switches, and core switches are housed in separate chassis.

13. The system of claim 2, wherein the independent switching elements are scaled according to the number of switches at each stage in the interconnected matrix of switches.

14. A network architecture comprising:

a wide area network;

a plurality of local networks configured to communicate over the wide area network; and

a plurality of switching stages which are configured to communicatively couple the local networks to each other over the wide area network,

wherein local groups of switching stages are configured to communicatively couple associated local networks to the wide area network;

wherein at least one group of switching stages comprises an edge stage comprising a plurality of edge switches, a middle stage comprising an interconnected matrix of switches and a core stage comprising a plurality of core switches and

wherein each switching stage is loosely coupled to the other switching stages whereby the edge switches do not need to be capable of selecting a port on a core switch through which to forward data, and vice versa.

15. The network architecture of claim 14, wherein each of the local networks is in a different metropolitan area.

16. The network architecture of claim 14, wherein each group of stages comprises an interconnected matrix of switches.

17. The network architecture of claim 16, wherein at least one interconnected matrix of switches is directly connected to at least one other interconnected matrix of switches over the wide area network.

18. The network architecture of claim 17, wherein connections between interconnected matrices of switches comprise a braid topology or a ladder topology, or a combination thereof.

19. The network architecture of claim 16, wherein each interconnected matrix

of switches forms a topology comprising a ring, a star, a full mesh, a partial mesh, an arbitrary mesh, a CLOS-like matrix, an n-cube, or a hybrid, or any combination thereof.

20. The network architecture of claim 14, wherein switches of the middle stage have a plurality of ports facing the wide area network and a plurality of ports facing the local network, and wherein each port facing the wide area network is connected to a port on a switch in the core stage and each port facing the associated local network is connected to a port on a switch in the edge stage.

21. The network architecture of claim 14, wherein at least one of the plurality of switching stages comprises a plurality of Ethernet switches.

22. The network architecture of claim 14, wherein each of the edge switches and the core switches is configured to build its own forwarding table.

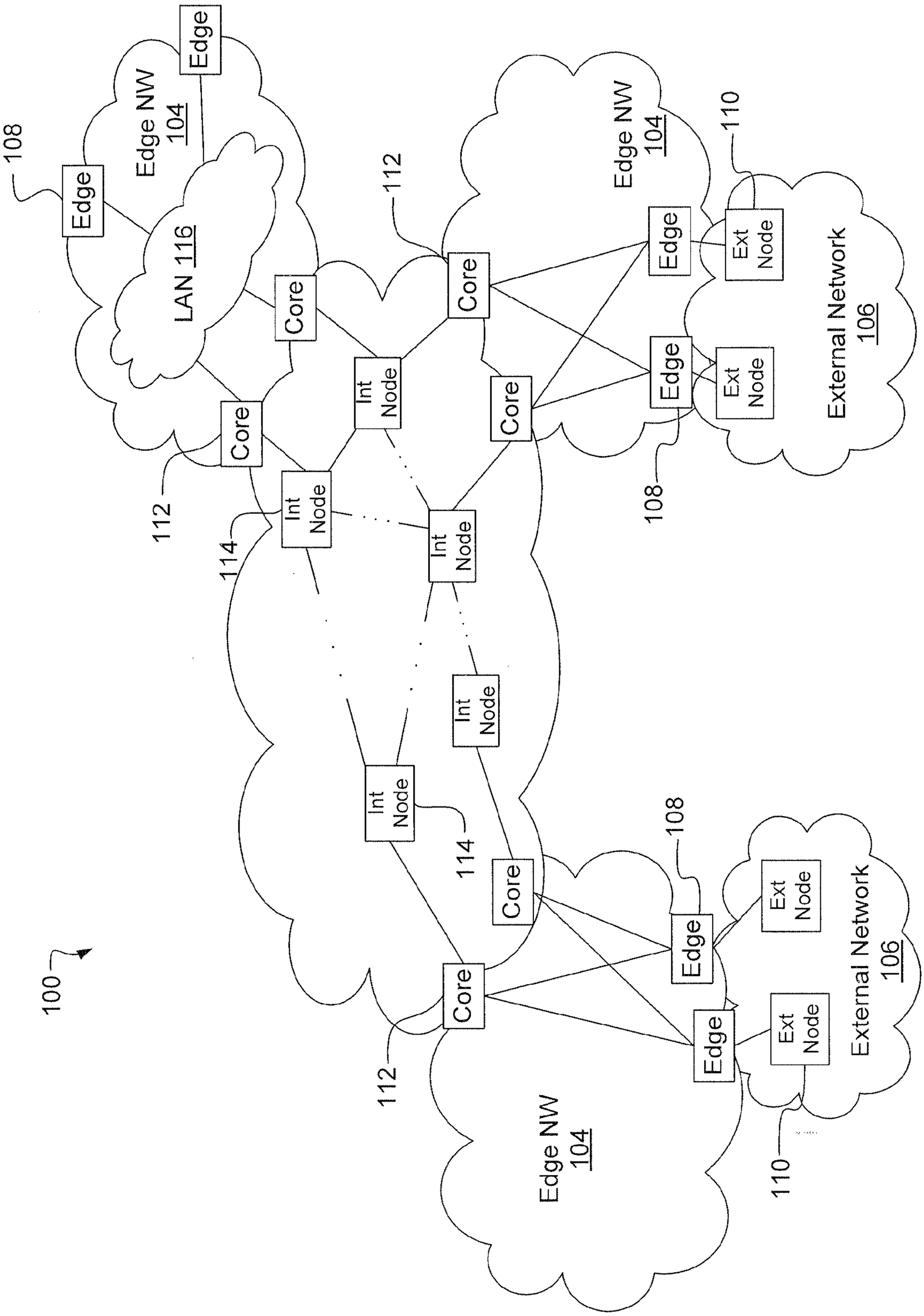

100
Edge NW 104
LAN 116
Edge
Core
Int Node
Ext Node
External Network 106
108
110
112
114

FIG. 1

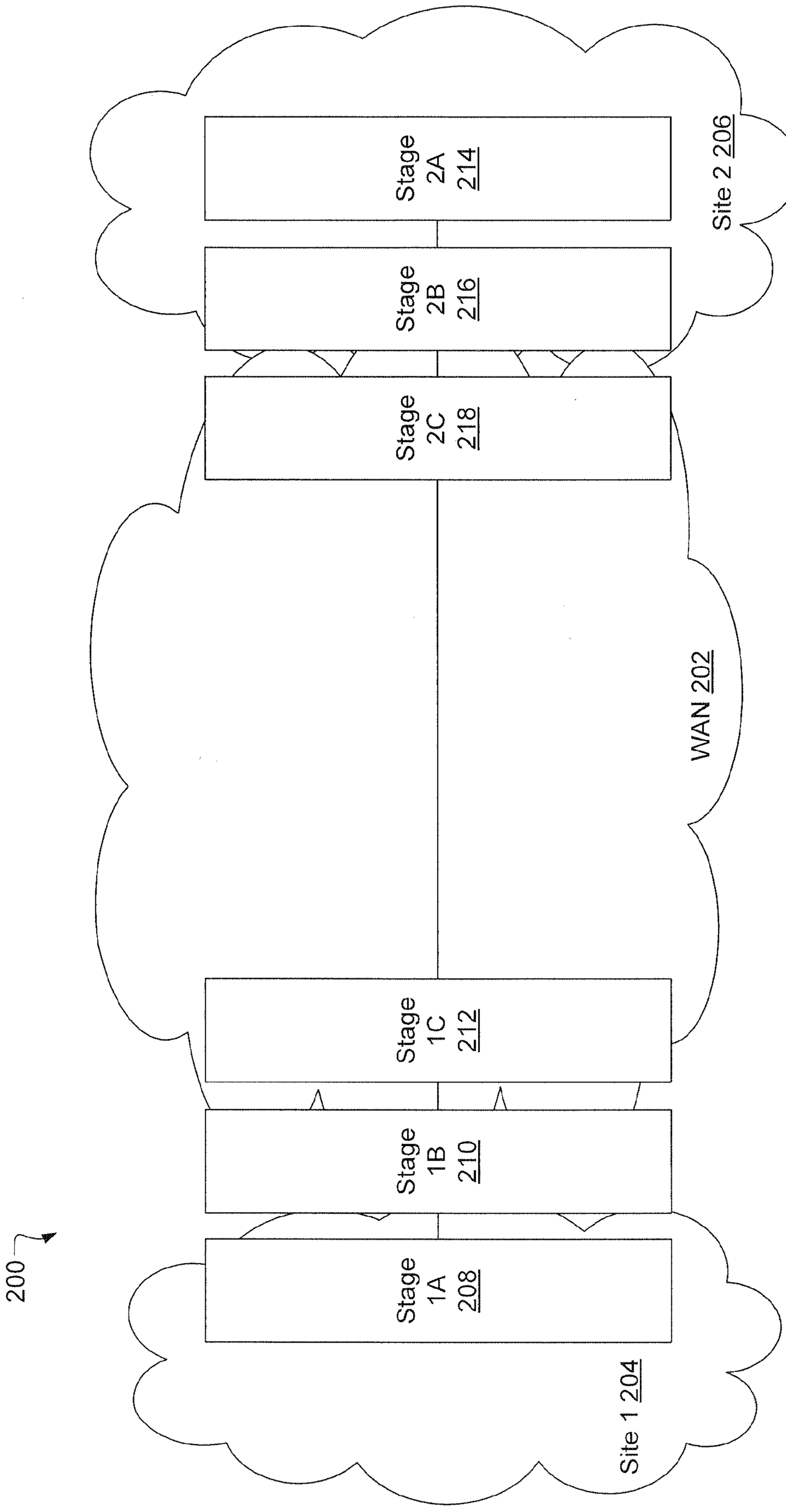

200
Site 1 204
Stage 1A 208
Stage 1B 210
Stage 1C 212
WAN 202
Site 2 206
Stage 2A 214
Stage 2B 216
Stage 2C 218

FIG. 2

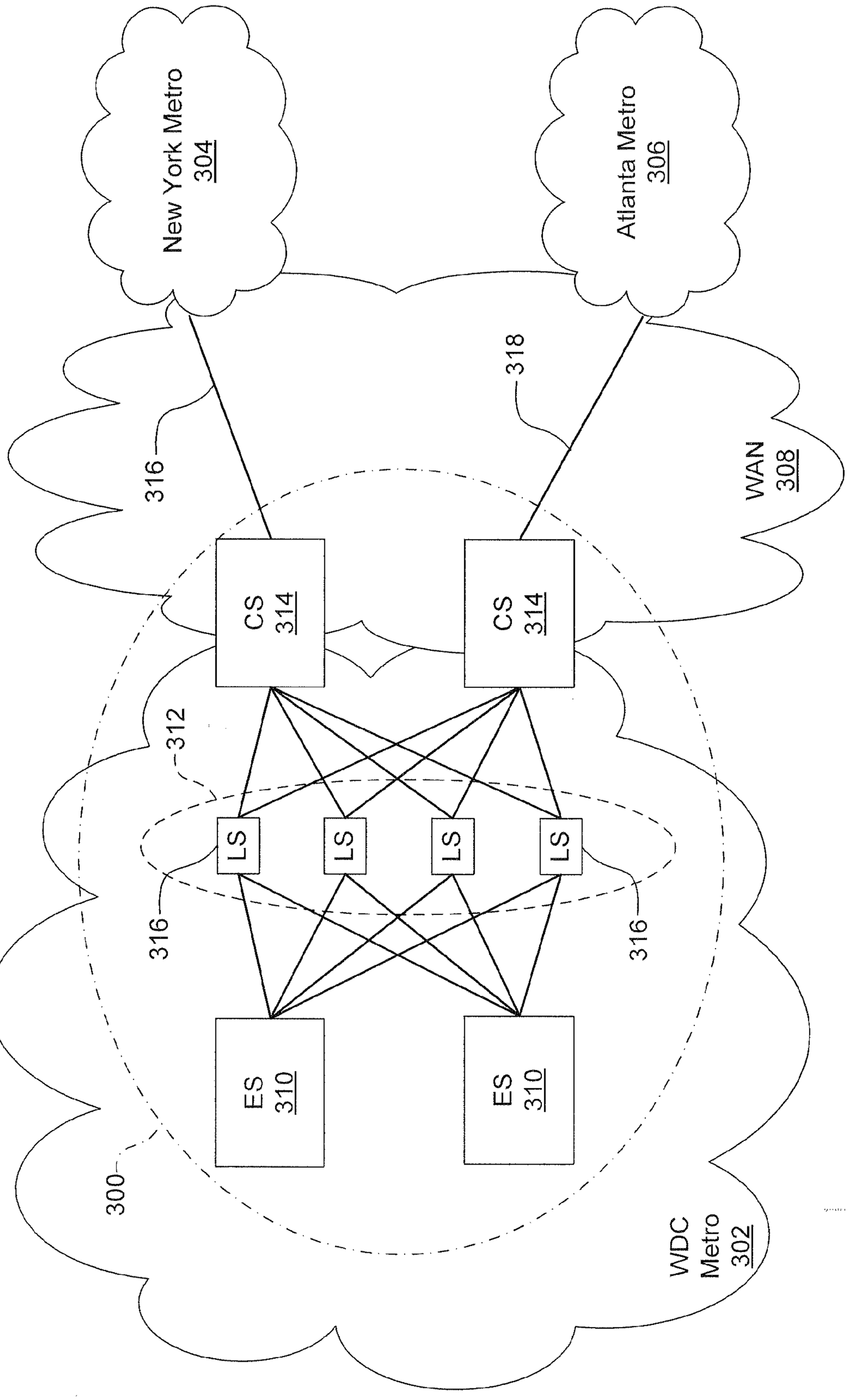

New York Metro
304
Atlanta Metro
306
316
318
WAN
308
CS
314
CS
314
312
LS
LS
LS
LS
316
316
ES
310
ES
310
300
WDC
Metro
302

FIG. 3

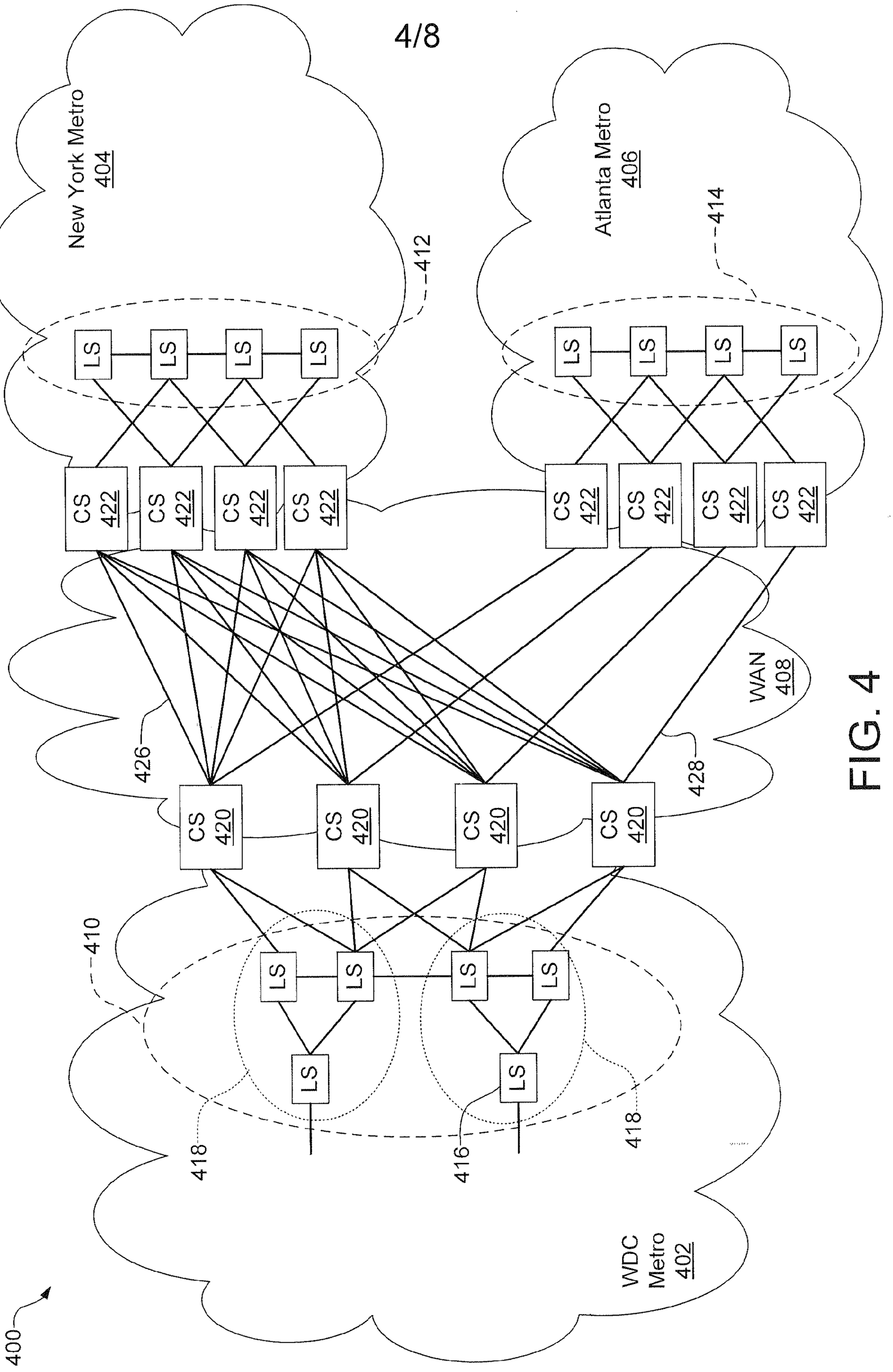

New York Metro
404
412
Atlanta Metro
406
414
LS
CS
422
426
428
WAN
408
410
418
416
WDC
Metro
402
400

FIG. 4

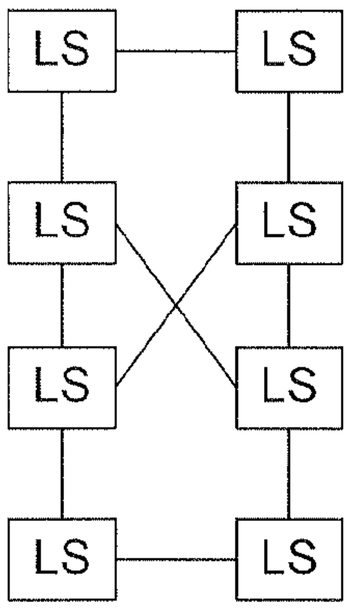


FIG. 5A

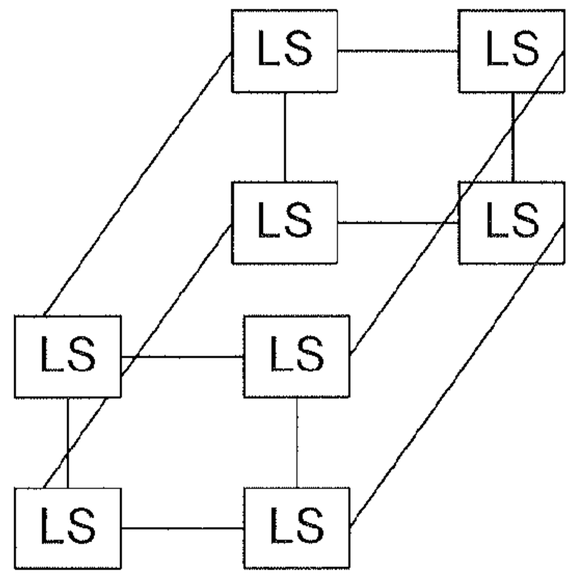


FIG. 5B

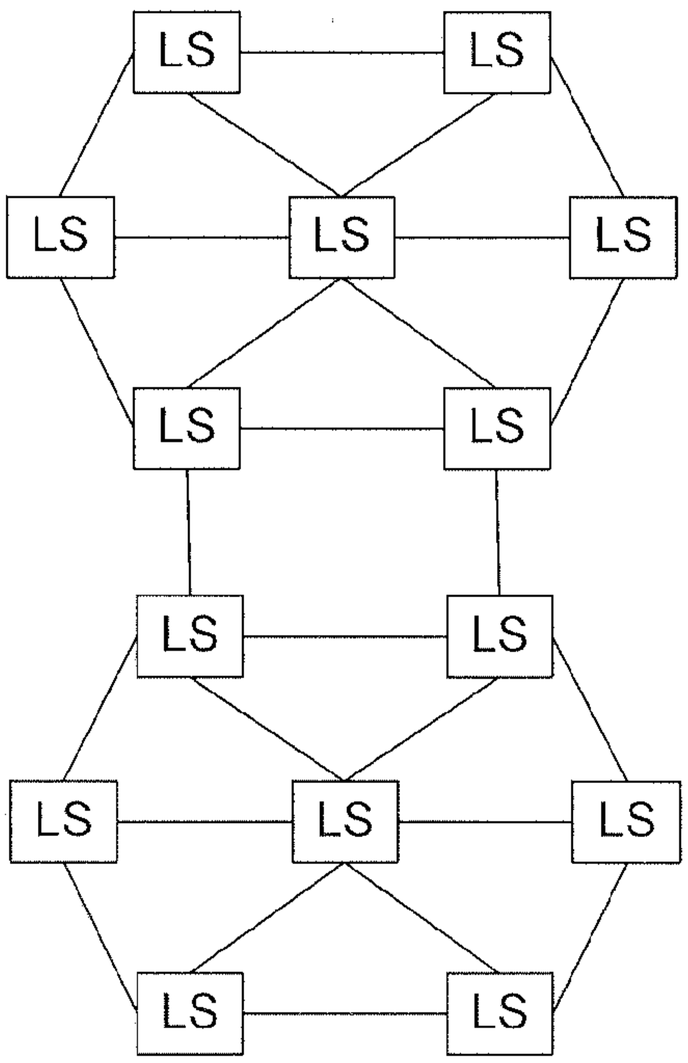


FIG. 5C

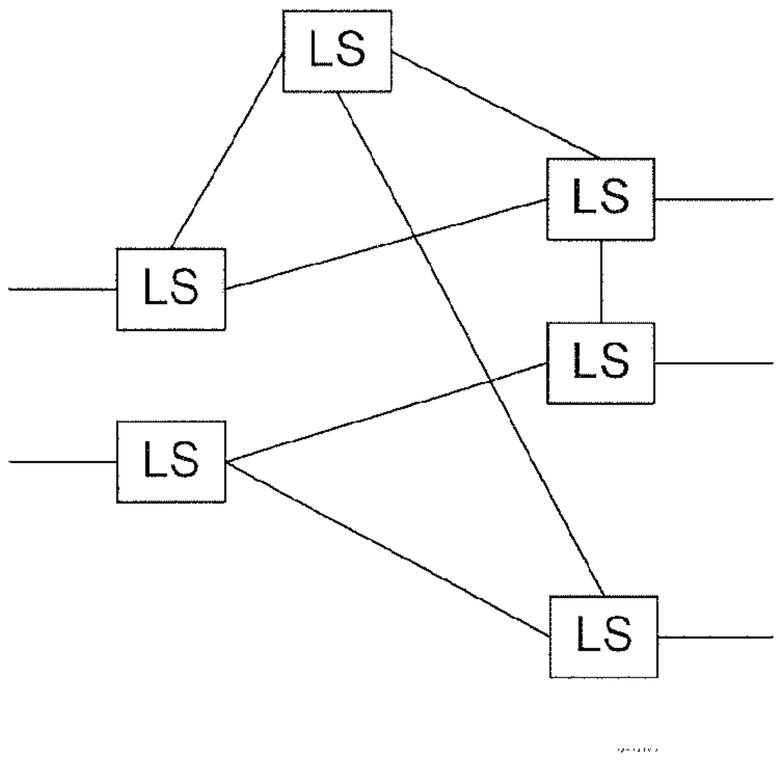


FIG. 5D

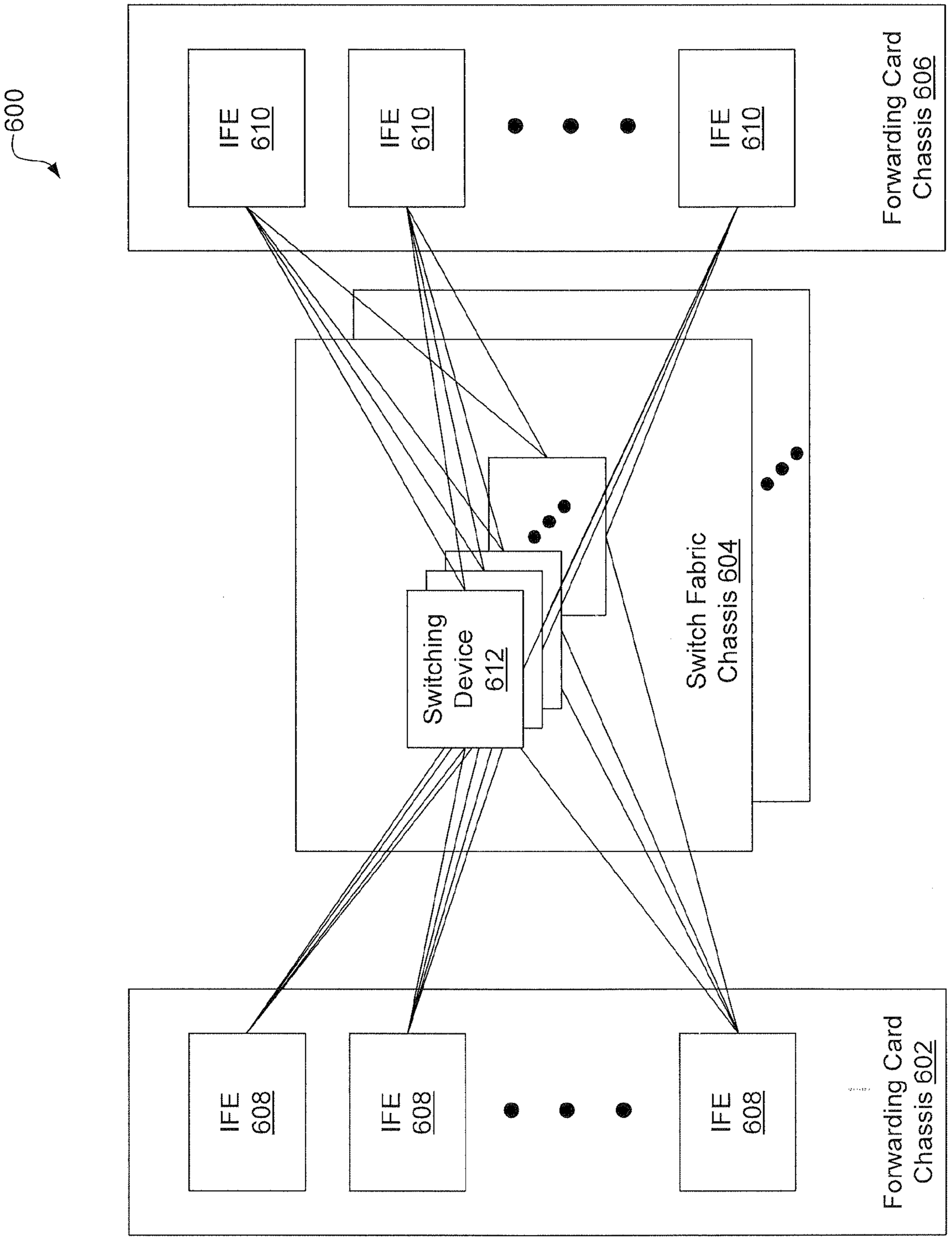

600
IFE 610
IFE 610
IFE 610
Forwarding Card Chassis 606
Switching Device 612
Switch Fabric Chassis 604
IFE 608
IFE 608
IFE 608
Forwarding Card Chassis 602

FIG. 6

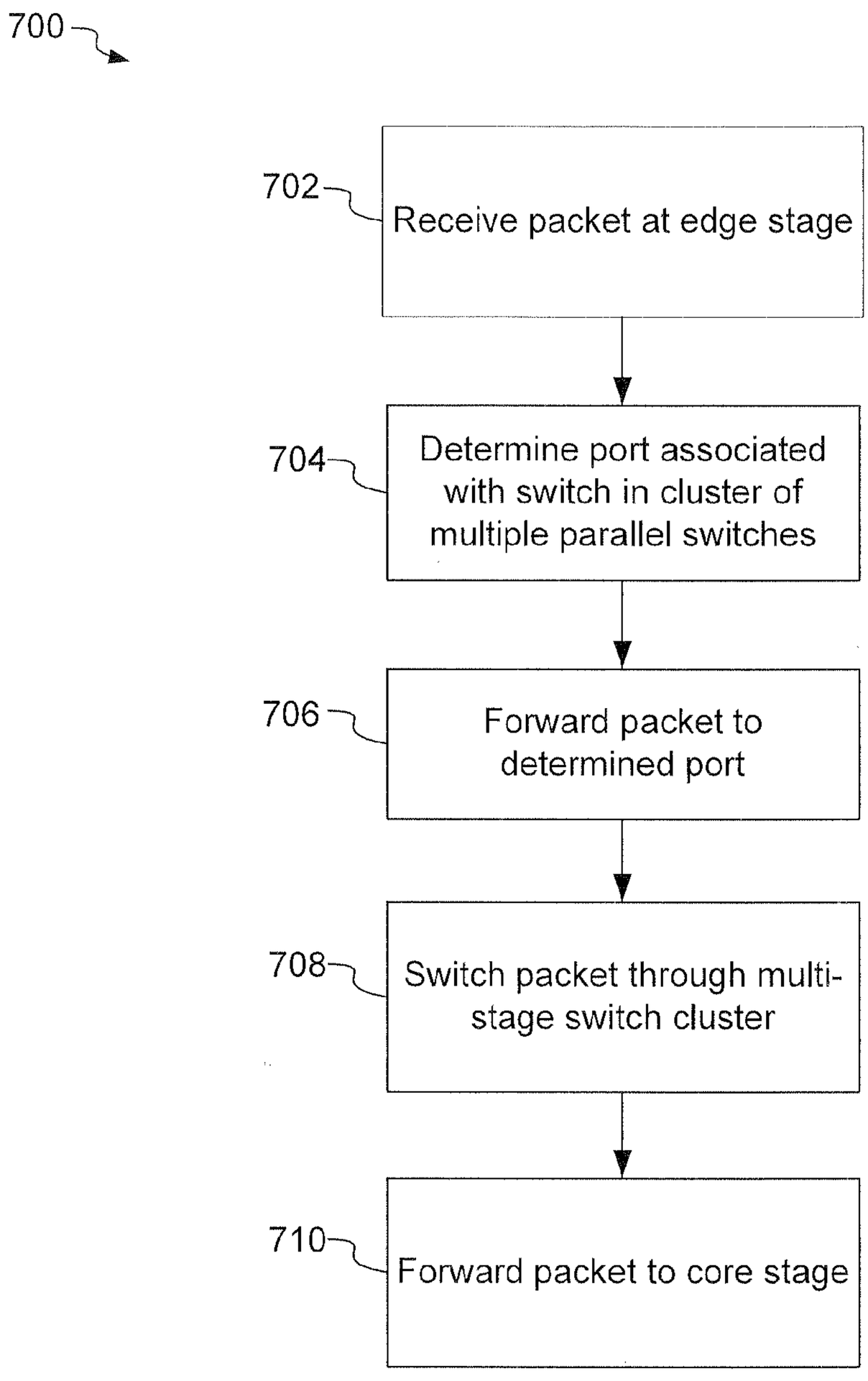

700
702
Receive packet at edge stage
704
Determine port associated with switch in cluster of multiple parallel switches
706
Forward packet to determined port
708
Switch packet through multi-stage switch cluster
710
Forward packet to core stage

FIG. 7

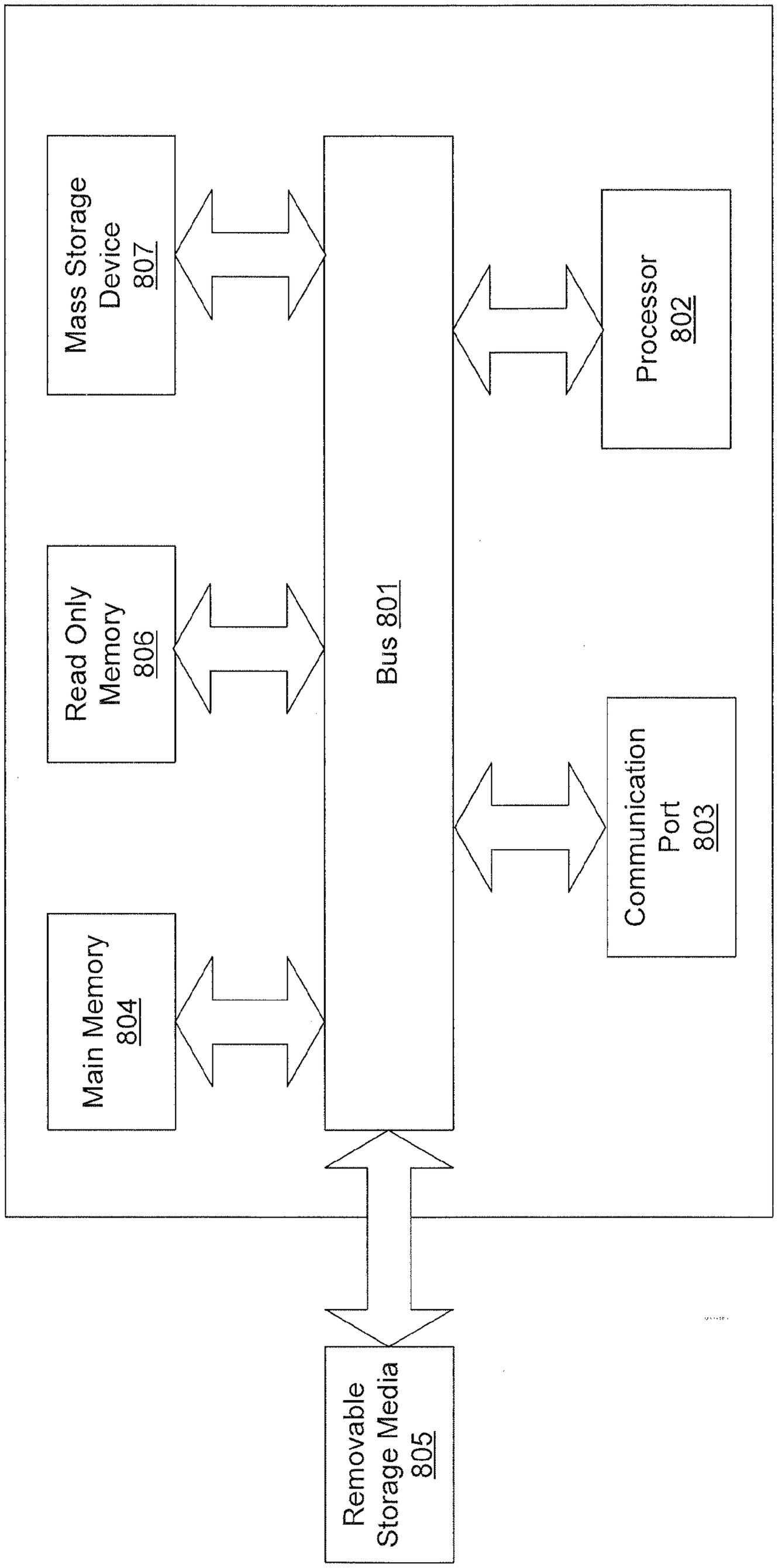
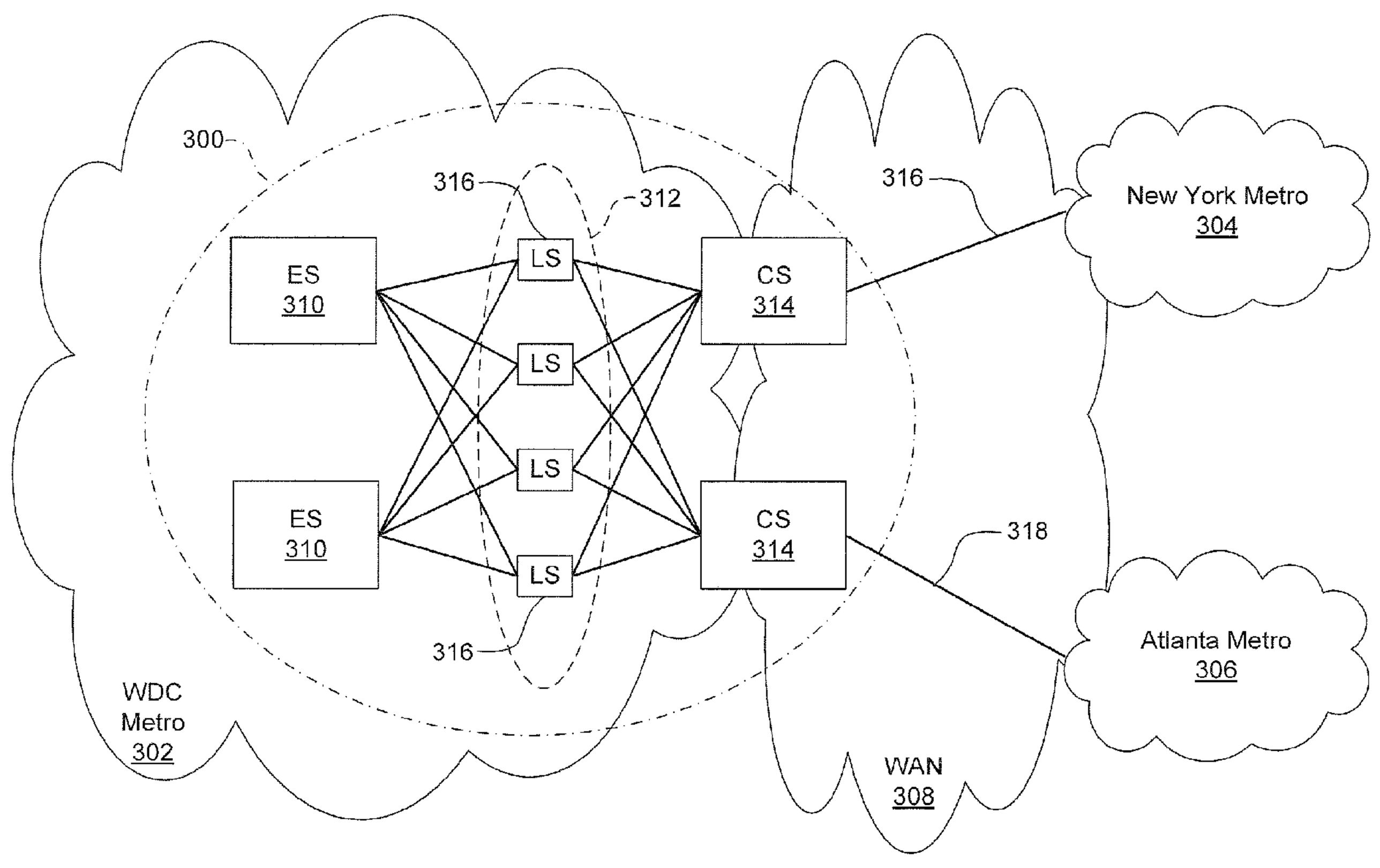

800
Removable Storage Media 805
Main Memory 804
Read Only Memory 806
Mass Storage Device 807
Bus 801
Communication Port 803
Processor 802

FIG. 8


300
316
312
316
New York Metro
304
ES
310
LS
CS
314
LS
LS
ES
310
CS
314
318
LS
316
Atlanta Metro
306
WDC
Metro
302
WAN
308