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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
02 July 2020 (02.07.2020)

WIPO | PCT

(10) International Publication Number
WO 2020/135221 A1

(51) International Patent Classification:
H04L 12/46 (2006.01)

(21) International Application Number:
PCT/CN2019/126622

(22) International Filing Date:
19 December 2019 (19.12.2019)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
16/236,253 28 December 2018 (28.12.2018) US

(71) Applicant: **ALIBABA GROUP HOLDING LIMITED** [—/CN]; Fourth Floor, One Capital Place, P.O. Box 847, George Town, Grand Cayman (KY).

(72) Inventor: **CHENG, Gang**; Alibaba Group Legal Department 5/F, Building 3, No. 969 West Wen Yi Road, Yu Hang District, Hangzhou, Zhejiang 311121 (CN).

(74) Agent: **BEIJING SANYOU INTELLECTUAL PROPERTY AGENCY LTD.**; 16th Fl., Block A, Corporate Square, No. 35 Jinrong Street, Beijing 100033 (CN).

(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, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, 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: METHOD, APPARATUS, AND COMPUTER-READABLE STORAGE MEDIUM FOR NETWORK OPTIMIZATION FOR ACCESSING CLOUD SERVICE FROM ON-PREMISES NETWORK

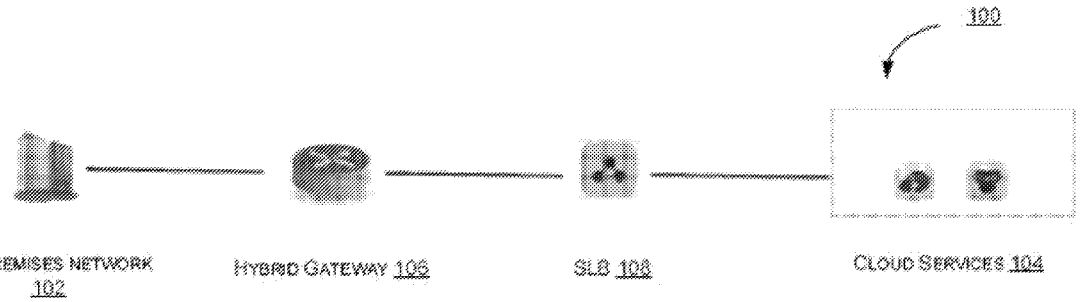

FIG. 1

(57) Abstract: Apparatus and methods include receiving a first packet, the first packet including a first source Internet protocol (IP) address and a first destination IP address; adding a first outer header to the first packet to generate an encapsulated packet, the first outer header including a second source IP address and a second destination IP address; forwarding the encapsulated packet to a Server Load Balancer (SLB) associated with the second destination IP address; receiving a response packet from a front end server having a front end server private IP address; and saving the front end server private IP address to bypass the SLB.

WO 2020/135221 A1

METHOD, APPARATUS, AND COMPUTER-READABLE STORAGE MEDIUM FOR NETWORK OPTIMIZATION FOR ACCESSING CLOUD SERVICE FROM ON-PREMISES NETWORK

CROSS REFERENCE TO RELATED APPLICATION

[1] This disclosure claims the benefits of priority to United States application number 16/236,253, filed December 28, 2018, which is incorporated herein by reference in its entirety.

BACKGROUND

[2] Virtual Extensible Local Area Network (VXLAN) may be referred to as an overlay technology because it allows stretching Layer 2 connections over an intervening Layer 3 network by encapsulating (tunneling) data into a VXLAN packet. Devices that support VXLANs are called virtual tunnel endpoints (VTEPs), which encapsulate and de-encapsulate VXLAN traffic. The migration of virtual machines is enabled between servers that exist in separate Layer 2 domains by tunneling the traffic over Layer 3 networks. This functionality allows the user to dynamically allocate resources within or between data centers without being constrained by Layer 2 boundaries or being forced to create large or geographically stretched Layer 2 domains. Using routing protocols to connect Layer 2 domains also allows load-balance of the traffic to ensure that the user gets the best use of available bandwidth.

[3] Enterprise customers can connect their on-premises network to the public cloud using a leased line-based hybrid network solution for good network performance and security. For example, an on-premises network may be connected to the cloud service through a hybrid gateway and a server load balancer (SLB). When accessing the public cloud service, all the

incoming and outgoing traffic of the customer needs to go through the SLB. In other words, the cloud service is behind the SLB from the point of view of the customer.

[4] Because the total cost increases along with the total capacity of SLBs, in practice, the total SLB capacity is preferably reduced to an affordable amount. Therefore, the SLB is usually implemented on a server which provides far less network throughput capability than regular network switches or routers. Thus, the SLB becomes the bottleneck which limits the accessibility of the cloud service from the customer. In case of large network traffic burst, the customer might experience packet loss or network congestion.

BRIEF DESCRIPTION OF THE DRAWINGS

[5] The detailed description is set forth with reference to the accompanying figures. In the figures, the left-most digit(s) of a reference number identifies the figure in which the reference number first appears. The use of the same reference numbers in different figures indicates similar or identical items or features.

[6] FIG. 1illustrates an example diagram showing a network where the on-premises network is connected to cloud service with lease lines.

[7] FIG. 2illustrates an example diagram of the format of a VXLAN packet.

[8] FIG. 3illustrates an example diagram showing a system for optimizing accessing to cloud service from the on-premises user.

[9] FIG. 4A, illustrate example flowcharts of a method for optimizing network accessing to cloud service.

[10] FIG. 4B, illustrate example flowcharts of a method for optimizing network accessing to cloud service.

[11] FIG. 4C illustrate example flowcharts of a method for optimizing network accessing to cloud service.

[12] FIG. 5A illustrates an example diagram showing the format of the first packet in FIG. 3.

[13] FIG. 5B illustrates an example diagram showing the format of the encapsulated

packet in FIG. 3.

[14] FIG. 5C illustrates an example diagram showing the format of the third packet in FIG. 3.

[15] FIG. 5D illustrates an example diagram showing the format of the fourth packet in FIG. 3.

[16] FIG. 5E illustrates an example diagram showing the format of the fifth packet in FIG. 3.

[17] FIG. 5F illustrates an example diagram showing the format of the sixth packet in FIG. 3.

[18] FIG. 5G illustrates an example diagram showing the format of the seventh packet in FIG. 3.

[19] FIG. 6illustrates an example block diagram of an apparatus for optimizing network accessing to cloud service.

DETAILED DESCRIPTION

[20] Apparatuses and methods discussed herein are directed to improving cloud service, and more specifically to network optimization for accessing cloud service from on-premises network.

[21] Apparatuses and methods discussed herein may be usable to provide a system-level solution to minimize the total workload of SLBs needed for the traffic between the on-premises network and the cloud service. The speed of data transmission and thus the speed of processing associated with a user application that depends on the cloud service may be improved by reducing the network latency. Also, the network robustness of accessing the cloud service from the on-premises network may be improved for a hybrid user. The hybrid user is a user of hybrid cloud which is a cloud computing environment that uses a mix of on-premises, private cloud and third-party, public cloud services.

[22] A hybrid gateway may receive a first packet. The first packet may include a first source Internet protocol (IP) address and a first destination IP address. The hybrid gateway

may add a first outer header to the first packet to generate an encapsulated packet. The first outer header including a second source IP address, a second destination IP address. The hybrid gateway may forward the encapsulated packet to an SLB associated with the second destination IP address. The hybrid gateway may receive a response packet from a front end server having a front end server private IP address. In a system, some servers are designated for the purpose of receiving requests and sending them to other servers. The frontend server may accept the requests connection from a user, then proxy the connection to a backend server on which the user's request may be served. The hybrid gateway may save the front end server private IP address.

[23] The hybrid gateway may receive a second packet. The second packet including the first source IP address and the first destination IP address. The hybrid gateway may add a second outer header to the second packet to generate a bypass packet. The second outer header may include a third source IP address and a third destination IP address. The third destination IP address may be the front end server private IP address. The hybrid gateway may forward the bypass packet to the front end server associated with the front end server private IP address without first sending any information related to the second packet to the SLB prior to sending the bypass packet to the front end server, i.e., bypassing the SLB.

[24] The first outer header to the first packet may be performed based on virtual extensible local area network (VXLAN) encapsulation.

[25] The first outer header may further include a VXLAN ID. the second source IP address may be a hybrid gateway private IP address. The second destination IP address may be a SLB private IP address.

[26] The first packet may be sent from an on-premises network. The first source IP address may be a private IP address of the on-premises network. The first destination IP address may be an IP address of cloud service.

[27] The front end server private IP address may be saved in a session table.

[28] The SLB may receive the encapsulated packet. The SLB may select the front end server associated with the front end server private IP address.

[29] The SLB may replace the second destination IP address of the encapsulated packet with the front end server private IP address to generate a modified packet. The SLB

may forward the modified packet to the front end server.

[30] FIG. 1illustrates an example diagram showing a network 100.The on-premises network 102 may be connected to cloud service 104 with lease or dedicated lines through a hybrid gateway 106 and a server load balancer (SLB) 108. When a hybrid user of the on-premises network 102 needs to access the cloud service 104, the hybrid user may send a data packet to the hybrid gateway 106. The hybrid gateway 106 may receive the packet and perform encapsulation, for example, the VXLAN encapsulation on the packet and send the encapsulated packet to the SLB 108. The SLB 108 may receive the encapsulated packet from the hybrid gateway 106 and replace the destination IP address of the encapsulated packet with the IP address of a selected front end server to generate a modified packet. The SLB 108 may send the modified packet to the selected front end server. The selected front end server may receive the modified packet and generate a response packet to serve the request from the hybrid user. The response packet may carry the private IP address of the selected front end server. The front end server may send the response packet to the SLB 108 which in turn forward the response packet to the hybrid gateway 106. The hybrid gateway 106 may forward the response packet to the hybrid user of the on-premises network 102. For the follow-up traffic from the hybrid user, such procedure may repeat.

[31] FIG. 2 illustrates an example diagram of the format of a VXLAN packet 200. Referring to FIG. 2, the VXLAN packet 200 may include the following components.

[32] An outer source IP address 202 and an outer destination IP address 204 may represent the two endpoints address of the tunnel. A user datagram protocol (UDP) header 206 may carry a source port number and a destination port number of two endpoints of the tunnel. In general, the destination port of a UDP header may be a fixed value, for example, 8472 by default. AVXLAN header 208 may include a VXLAN ID. Each user, i.e., each tenant, may be assigned a unique 24-bits VXLAN ID in one data center (DC). A packet may use the VXLAN ID of a user as the VXLAN header to be forwarded to the user with this VXLAN ID .An inner packet 210 may include a MAC address 212, an inner source IP address 214, and an inner destination IP address 216, which may be carried by an original Layer-2 data packet.

[33] FIG. 3illustrates an example block diagram showing a system300 for optimizing accessing of cloud service from the on-premises user.

[34] When a hybrid user of the on-premises network 302 needs to access the cloud service, the hybrid user may send out a first packet 304 to the hybrid gateway 306.

[35] The hybrid gateway 306 may receive the first packet 304 from the on-premises network 302. The hybrid gateway 306 may perform encapsulation, for example, VXLAN encapsulation on the first packet 304 by adding a first outer header to generate an encapsulated packet308. Also, the encapsulation of the first packet 304 may be performed based on other protocols or standards. The hybrid gateway 306 may send the encapsulated packet308 to the SLB 310.

[36] The SLB 310 may receive the encapsulated packet308 from the hybrid gateway 306. The SLB 310 may select a front end server 314 from a plurality of front end servers. The selection of the front end server may be performed randomly or based on any suitable criteria. The selected front end server 314 may have a private IP address, for example, 10.4.1.1. Each front end server FE1, FE2, ..., FEn may have its own private IP address, for example, 10.4.1.1, 10.4.1.2, ..., 10.4.1.n. The SLB 310 may perform a destination network address translation (DNAT) on the encapsulated packet308 by replacing the second destination IP address with a third destination IP address to generate a modified packet 312. The third destination IP address may be the private IP address of the selected front end server 314. The SLB 310may send the modified packet 312 to the selected front end server 314.

[37] The selected front end server 314 may receive the modified packet 312 from the SLB 310. The selected front end server 314 may serve the request of the hybrid user and generate a response packet 316 based on the modified packet 312. The response packet 316 may include a second outer header. The selected front end server 314 may send the response packet 316 to the hybrid gateway 306.

[38] The hybrid gateway 306 may receive the response packet 316 from the selected front end server 314. The hybrid gateway 306 may remove the second outer header of the response packet 316 to generate a modified response packet 318. The hybrid gateway 306 may save the private IP address of the selected front end server 314 in a session table. Also, the private IP address of the selected front end server 314 may be saved in a memory, a storage device, or any other suitable location. The hybrid gateway 306 may forward the modified response packet 318 to the hybrid user of on-premises network 302.

[39] Additionally or alternatively, the session table may save other information for the hybrid gateway to determine whether other data packets from a user will be sent to this address of the front end server. For example, information about relationship between the user and the frontend server may be set up. A first user may correspond to a first front end server, a second user may correspond to a second server, an n^{th} user may correspond to an n^{th} server, etc., where n may be a positive integer. The session table may be specific to a particular user or may be used for different users. Moreover, the session table may include an entry indicating that a bypass may be made. This entry may be expired after a predetermined period of time or after a predetermined number of packets from the same user are sent, etc. The predetermined period of time and the predetermined number may beset and adjusted as necessary. Under certain conditions, for example, after the predetermined period of time or after the predetermined number of data packets are sent, the load balancing may be performed to make sure the selected frontend server is not overloaded or another server that has no job task is available for the user, etc.

[40] The hybrid user of on-premises network 302 may send a second packet 320 to the hybrid gateway 306.

[41] The hybrid gateway 306 may receive the second packet 320 from the on-premises network 302. The hybrid gateway 306 may obtain the private IP address of the selected front end server 314 saved in the session table or any other proper locations. The hybrid gateway 306 may encapsulate the second packet 320 to generate a bypass packet 322 by adding a third outer header. The third outer header may include the third destination IP address which is the private IP address of the selected front end server 314. The hybrid gateway 306 may send the bypass packet 322 to the selected front end server 314 directly, bypassing the SLB 310.

[42] FIGs.4A-4Cillustrate example flowcharts of a process 400 for optimizing network accessing to cloud service. The data flow of the process 400 may be described as below.

[43] At block 402, the hybrid gateway 306 may receive a first packet 304 sent by a hybrid user of the on-premises network 302.

[44] At block 404, the hybrid gateway 306 may perform encapsulation, for example,

VXLAN encapsulation on the first packet 304 by adding a first outer header to generate an encapsulated packet308.Also, the encapsulation of the first packet 304 may be performed based on other protocols or standards.

[45] At block 406, the hybrid gateway 306 may send the encapsulated packet308 to the SLB 310.

[46] At block 408, the SLB 310 may receive the encapsulated packet308from the hybrid gateway 306.

[47] At block 410, The SLB 310 may select a front end server 314 from a plurality of front end servers. The selection of the front end server may be performed randomly or based on any suitable criteria. The selected front end server 314 may have a private IP address, for example, 10.4.1.1. Each front end server FE1, FE2, ..., FEn may have its own private IP address, for example, 10.4.1.1, 10.4.1.2, ..., 10.4.1.n.

[48] At block 412, the SLB 310 may perform a destination network address translation (DNAT) on the encapsulated packet308 by replacing the second destination IP address with a third destination IP address to generate a modified packet 312. The third destination IP address may be the private IP address of the selected front end server 314.

[49] At block 414, the SLB310 may send the modified packet 312 to the selected front end server 314.

[50] At block 416, the selected front end server 314 may receive the modified packet 312.

[51] At block 418, the selected front end server 314 may generate a response packet 316 based on the modified packet 312. The selected front end server 314 may serve the request of the hybrid user. Also, the response packet 316 may be an empty packet. The response packet 316 may include a second outer header.

[52] At block 420, the selected front end server 314 may send the response packet 316 to the hybrid gateway 306.

[53] At block 422, the hybrid gateway 306 may receive the response packet 316 from the selected front end server 314.

[54] At block 424, the hybrid gateway 306 may remove the second outer header of the response packet 316 to generate a modified response packet 318.

[55] At block 426, the hybrid gateway 306 may save the private IP address of the selected front end server 314 in a session table. Also, the private IP address of the selected front end server 314 may be saved in a memory, a storage device, or any other suitable location.

[56] Additionally or alternatively, the session table may save other information for the hybrid gateway to determine whether other data packets from a user will be sent to this address of the front end server. For example, information about relationship between the user and the frontend server may be set up. A first user may correspond to a first front end server, a second user may correspond to a second server, an n^{th} user may correspond to an n^{th} server, etc., where n may be a positive integer. The session table may be specific to a particular user or may be used for different users. Moreover, the session table may include an entry indicating that a bypass may be made. This entry may be expired after a predetermined period of time or after a predetermined number of packets from the same user are sent, etc. The predetermined period of time and the predetermined number may beset and adjusted as necessary. Under certain conditions, for example, after the predetermined period of time or after the predetermined number of data packets are sent, the load balancing may be performed to make sure the selected frontend server is not overloaded or another server that has no job task is available for the user, etc.

[57] At block 428, the hybrid gateway 306 may forward the modified response packet 318 to the on-premises network 302.

[58] Subsequently, the consecutive packets sent from the hybrid user of the on-premises network 302 may be forwarded from the gateway 306 to the front end server 314 directly without going through the SLB 310, i.e., bypassing the SLB 310. Thus, the process 400 may further include the following.

[59] Referring to FIG. 4A- FIG. 4C, at block 430, the hybrid gateway 306 may receive a second packet 320 from the on-premises network 302.

[60] At block 432, the hybrid gateway 306 may obtain the private IP address of the selected front end server 314 saved in the session table or any other proper locations.

[61] At block 434, the hybrid gateway 306 may encapsulate the second packet 320 to generate a bypass packet 322 by adding a third outer header. The third outer header may

include the third destination IP address which is the private IP address of the selected front end server 314.

[62] At block 436, the hybrid gateway 306 may send the bypass packet 322 to the selected front end server 1144 directly, bypassing the SLB 310.

[63] With the systems and processes discussed herein, the first packet sent from the hybrid user of the on-premises network 302 may be routed through the gateway 306 and the SLB 310. After the SLB 310 selectes a front end server 314 from the plurality of front end servers, the private IP address of the front end server 314 may be encapsulated in the response packet 316. Therefore, from the response packet 316 sent from the front end server 314, the hybrid gateway 306 may learn about the private IP address of the selected front end server 314 and may save the private IP address of the selected front end server 314 in a session table. For follow-up packets sent from the on-premises network 302, the hybrid gateway 306 may look up the session table for the private IP address of the selected front end server 314. The hybrid gateway 306 may perform the encapsulation using the private IP address of the selected front end server 314 and forward the packet to the selected front end server 314 directly, bypassing the SLB 310. Therefore, the total capacity of SLBs needed for the traffic between the on-premises network 302 and the cloud service may be reduced. The total network latency may be reduced. The speed of data transmission and thus the speed of processing associated with a user application that depends on the cloud service may be improved. The network robustness for a hybrid user of the on-premises network to access the cloud service may be improved.

[64] Systems and processes discussed herein may also be referred to as Hybrid Fast Bypass (HFB).

[65] FIG. 5A- FIG.5G illustrate formats of various packets shown in FIG. 3.

[66] FIG. 5A illustrates an example diagram showing the format of the first packet 304 in FIG. 3. Referring to FIG. 5A, the first packet 304 may include a first source IP address 502 and a first destination IP address 504. The first source IP address 502 may be a private IP address of the on-premises network 302, for example, 10.1.1.2. The first destination IP address 504 may be the IP address of cloud service, for example, 132.1.1.1.

[67] FIG. 5B illustrates an example diagram showing the format of the encapsulated

packet308 in FIG. 3. Referring to FIG. 5B, the first outer header of the encapsulated packet308 may include a second source IP address 506, a second destination IP address 508, and a VXLAN ID 510. The second source IP address 506 may be a private IP address of the hybrid gateway 306, for example, 10.0.2.1. The second destination IP address 508 may be the private IP address of SLB 310, for example, 10.3.1.1. The VXLAN ID may be a global unique VXLAN ID representing HFB between the user on-premises network 302 and cloud service. Each such VXLAN ID may be hybrid user specific and varies between the hybrid users. The first source IP address 502 may be a private IP address of the on-premises network 302, for example, 10.1.1.2. The first destination IP address 504 may be the IP address of cloud service, for example, 132.1.1.1.

[68] FIG. 5C illustrates an example diagram showing the format of the modified packet 312 in FIG. 3. Referring to FIG. 5C, the modified packet 312 may include the second source IP address, the third destination IP address 512, VXLAN ID 510, the first source IP address, and the first destination IP address. The second source IP address 506 may be a private IP address of the hybrid gateway 306, for example, 10.0.2.1.The third destination IP address 512may be the private IP address of selected front end server 314, for example, 10.4.1.1.Each such VXLAN ID may be hybrid user specific and varies between the hybrid users. The first source IP address 502 may be a private IP address of the on-premises network 302, for example, 10.1.1.2. The first destination IP address 504 may be the IP address of cloud service, for example, 132.1.1.1.

[69] FIG. 5D illustrates an example diagram showing the format of the response packet 316 in FIG. 3. Referring to FIG. 5D, the response packet 316 may include a second outer header, the first destination IP address 504, and the first source IP address 502. The second outer header may include a third source IP address 514, a fourth destination IP address 516, and the VXLAN ID 510. The third source IP address 514 may be the private IP address of the front end server 314, for example, 10.4.1.1. The fourth destination IP address 516 may be the private IP address of the hybrid gateway 306, for example, 10.0.2.1. The first destination IP address 504 may be the cloud service IP address, for example, 132.1.1.1. The first source IP address502 may be the private IP address of the on-premises network 302, for example, 10.1.1.2. All information may be obtained from the modified packet 312 sent from

the SLB 310.

[70] FIG. 5E illustrates an example diagram showing the format of the modified response packet 318 in FIG. 3. Referring to FIG. 5E, the modified response packet 318 may include the first destination IP address 504 and the first source IP address 502. The first destination IP address 504 may be the IP address of cloud service, for example, 132.1.1.1. The first source IP address502 may be the private IP address of the on-premises network 302, for example, 10.1.1.2.

[71] FIG. 5F illustrates an example diagram showing the format of the second packet 320 in FIG. 3. Referring to FIG. 5F, the second packet 320 may include the first source IP address 502 and the first destination IP address 504. The first source IP address 502 may be the private IP address of the on-premises network 302, for example, 10.1.1.2. The first destination IP address 504 may be the IP address of cloud service, for example, 132.1.1.1.

[72] FIG. 5G illustrates an example diagram showing the format of the bypass packet 322 in FIG. 3. Referring to FIG. 5G, the bypass packet 322 may include the third outer header, the first source IP address 502, and the first destination IP address 504. The third outer header may include the second source IP address 506, the third destination IP address 512, and the VXLAN ID 510. The second source IP address 506 may be a private IP address of the hybrid gateway 306, for example, 10.0.2.1. The third destination IP address 512 may be the private IP address of selected front end server 314, for example, 10.4.1.1.Each such VXLAN ID may be hybrid user specific and varies between the hybrid users. The first source IP address 502 may be a private IP address of the on-premises network 302, for example, 10.1.1.2. The first destination IP address 504 may be the IP address of cloud service, for example, 132.1.1.1.

[73] FIG. 6illustrates an example block diagram of an apparatus 600 for optimizing network accessing to cloud service.

[74] FIG. 6 is only one example of an apparatus 600 and is not intended to suggest any limitation as to the scope of use or functionality of any computing device utilized to perform the processes and/or procedures described above. Other well-known computing devices, apparatuses, environments and/or configurations that may be suitable for use with the embodiments include, but are not limited to, driver/passenger computers, server computers, hand-held or laptop devices, multiprocessor apparatuses, microprocessor-based apparatuses,

set-top boxes, game consoles, programmable consumer electronics, network PCs, minicomputers, mainframe computers, distributed computing environments that include any of the above apparatuses or devices, implementations using field programmable gate arrays ("FPGAs") and application specific integrated circuits ("ASICs"), and/or the like.

[75] The apparatus 600 may include one or more processors 602 and memory 604 communicatively coupled to the processor(s) 602. The processor(s) 602 may execute one or more modules and/or processes to cause the processor(s) 602 to perform a variety of functions. In some embodiments, the processor(s) 602 may include a central processing unit (CPU), a graphics processing unit (GPU), both CPU and GPU, or other processing units or components known in the art. Additionally, each of the processor(s) 602 may possess its own local memory, which also may store program modules, program data, and/or one or more operating apparatuses.

[76] Depending on the exact configuration and type of the apparatus 600, the memory 604 may be volatile, such as RAM, non-volatile, such as ROM, flash memory, miniature hard drive, memory card, and the like, or some combination thereof. The memory 604 may include computer-executable instructions that are executable by the processor(s) 602, when executed by the processor(s) 602, cause the processor(s) 602 to implement systems and processes described with reference to FIGs. 1 –FIGs. 5G.

[77] The apparatus 600 may additionally include an input/output (I/O) interface 606 for receiving and outputting data. The apparatus 600 may also include a communication module 608 allowing the apparatus 600 to communicate with other devices (not shown) over a network (not shown). The network may include the Internet, wired media such as a wired network or direct-wired connections, and wireless media such as acoustic, radio frequency (RF), infrared, and other wireless media.

[78] With systems and processes discussed herein, only the first packet sent out from the on-premises network needs to go through the SLB to the selected front end server. Any consecutive packets sent from the on-premises network, for example, the seventh packet may be forwarded by the hybrid gateway to the selected front end server directly, without going through the SLB anymore. Therefore, the communication between the on-premises network and the cloud service may impose less traffic load to the SLB. Hence, the total bandwidth

needed for SLB may be significantly reduced. Meanwhile, since the follow-up traffic from the on-premises network after the first packet may be communicated directly between the on-premises network and selected front end server of the cloud service, the total network latency may be reduced. Speed of data transmission and thus the speed of processing associated with a user application that depends on the cloud service. The total capacity of SLBs needed for the traffic between the on-premises network and the cloud service may be minimized. The network robustness for a hybrid user of the on-premises network to access the cloud service may be improved.

[79] Some or all operations of the methods described above can be performed by execution of computer-readable instructions stored on a computer-readable storage medium, as defined below. The term "computer-readable instructions" as used in the description and claims, include routines, applications, application modules, program modules, programs, components, data structures, algorithms, and the like. Computer-readable instructions can be implemented on various system configurations, including single-processor or multiprocessor systems, minicomputers, mainframe computers, personal computers, hand-held computing devices, microprocessor-based, programmable consumer electronics, combinations thereof, and the like.

[80] The computer-readable storage media may include volatile memory (such as random access memory (RAM)) and/or non-volatile memory (such as read-only memory (ROM), flash memory, etc.).The computer-readable storage media may also include additional removable storage and/or non-removable storage including, but not limited to, flash memory, magnetic storage, optical storage, and/or tape storage that may provide non-volatile storage of computer-readable instructions, data structures, program modules, and the like.

[81] A non-transient computer-readable storage medium is an example of computer-readable media. Computer-readable media includes at least two types of computer-readable media, namely computer-readable storage media and communications media. Computer-readable storage media includes volatile and non-volatile, removable and non-removable media implemented in any process or technology for storage of information such as computer-readable instructions, data structures, program modules, or other data. Computer-readable storage media includes, but is not limited to, phase change memory

(PRAM), static random-access memory (SRAM), dynamic random-access memory (DRAM), other types of random-access memory (RAM), read-only memory (ROM), electrically erasable programmable read-only memory (EEPROM), flash memory or other memory technology, compact disk read-only memory (CD-ROM), digital versatile disks (DVD) or other optical storage, magnetic cassettes, magnetic tape, magnetic disk storage or other magnetic storage devices, or any other non-transmission medium that can be used to store information for access by a computing device. In contrast, communication media may embody computer-readable instructions, data structures, program modules, or other data in a modulated data signal, such as a carrier wave, or other transmission mechanism. As defined herein, computer-readable storage media do not include communication media.

[82] The computer-readable instructions stored on one or more non-transitory computer-readable storage media that, when executed by one or more processors, may perform operations described above with reference to FIGs. 1- FIGs. 6. Generally, computer-readable instructions include routines, programs, objects, components, data structures, and the like that perform particular functions or implement particular abstract data types. The order in which the operations are described is not intended to be construed as a limitation, and any number of the described operations can be combined in any order and/or in parallel to implement the processes.

Example Clauses

[83] Clause 1. A method, comprising: receiving a first packet, the first packet including a first source Internet protocol (IP) address and a first destination IP address; adding a first outer header to the first packet to generate an encapsulated packet, the first outer header including a second source IP address and a second destination IP address; forwarding the encapsulated packet to a Server Load Balancer (SLB) associated with the second destination IP address; receiving a response packet from a front end server having a front end server private IP address; and saving the front end server private IP address to bypass the SLB.

[84] Clause 2. The method of clause 1, further comprising: receiving a second packet, the second packet including the first source IP address and the first destination IP address; adding a second outer header to the second packet to generate a bypass packet, the

second outer header including a third source IP address and a third destination IP address, the third destination IP address being the front end server private IP address; and forwarding the bypass packet to the front end server associated with the front end server private IP address to bypass the SLB.

[85] Clause 3. The method of clause 1, wherein adding the first outer header to the first packet is performed based on virtual extensible local area network (VXLAN) encapsulation.

[86] Clause 4The method of clause 3, wherein the first outer header further includes a VXLAN identification (ID), the second source IP address being a hybrid gateway private IP address, the second destination IP address being a SLB private IP address.

[87] Clause 5. The method of clause 1, wherein the first packet is sent from an on-premises network, the first source IP address being a private IP address of the on-premises network, the first destination IP address being an IP address of cloud service.

[88] Clause 6. The method of clause 1, wherein the front end server private IP address is saved in a session table.

[89] Clause 7. The method of clause 2, wherein after forwarding the encapsulated packet to the SLB, the method further comprises: receiving, by the SLB, the encapsulated packet; selecting, by the SLB, the front end server associated with the front end server private IP address; replacing the second destination IP address of the encapsulated packet with the front end server private IP address to generate a modified packet; and forwarding the modified packet to the front end server.

[90] Clause 8. An apparatus, comprising: one or more processors; memory coupled to the one or more processors, the memory storing computer-readable instructions executable by the one or more processors that when executed by the one or more processors, cause the one or more processors to perform acts comprising: receiving a first packet, the first packet including a first source Internet protocol (IP) address and a first destination IP address; adding a first outer header to the first packet to generate an encapsulated packet, the first outer header including a second source IP address and a second destination IP address; forwarding the encapsulated packet to a Server Load Balancer (SLB) associated with the second destination IP address; receiving a response packet from a front end server having a front end

server private IP address; and saving the front end server private IP address to bypass the SLB.

[91] Clause 9. The apparatus of clause 8, wherein the acts further comprise: receiving a second packet, the second packet including the first source IP address and the first destination IP address; adding a second outer header to the second packet to generate a bypass packet, the second outer header including a third source IP address and a third destination IP address, the third destination IP address being the front end server private IP address; and forwarding the bypass packet to the front end server associated with the front end server private IP address to bypass the SLB.

[92] Clause 10. The apparatus of clause 8, wherein adding the first outer header to the first packet is performed based on virtual extensible local area network (VXLAN) encapsulation.

[93] Clause 11. The apparatus of clause 10, wherein the first outer header further includes a VXLAN identification (ID), the second source IP address being a hybrid gateway private IP address, the second destination IP address being a SLB private IP address.

[94] Clause 12. The apparatus of clause 8, wherein the first packet is sent from an on-premises network, the first source IP address being a private IP address of the on-premises network, the first destination IP address being an IP address of cloud service.

[95] Clause 13. The apparatus of clause 8, wherein the front end server private IP address is saved in a session table.

[96] Clause 14. The apparatus of clause 9, wherein after forwarding the encapsulated packet to the SLB, the method further comprises: receiving, by the SLB, the encapsulated packet; selecting, by the SLB, the front end server associated with the front end server private IP address; replacing the second destination IP address of the encapsulated packet with the front end server private IP address to generate a modified packet; and forwarding the modified packet to the front end server.

[97] Clause 15. A computer-readable storage medium storing computer-readable instructions executable by one or more processors, that when executed by the one or more processors, cause the one or more processors to perform acts comprising: receiving a first packet, the first packet including a first source Internet protocol (IP) address and a first

destination IP address; adding a first outer header to the first packet to generate an encapsulated packet, the first outer header including a second source IP address and a second destination IP address; forwarding the encapsulated packet to a Server Load Balancer (SLB) associated with the second destination IP address; receiving a response packet from a front end server having a front end server private IP address; and saving the front end server private IP address to bypass the SLB.

[98] Clause 16. The computer-readable storage medium of clause 15, wherein the acts further comprise: receiving a second packet, the second packet including the first source IP address and the first destination IP address; adding a second outer header to the second packet to generate a bypass packet, the second outer header including a third source IP address and a third destination IP address, the third destination IP address being the front end server private IP address; and forwarding the bypass packet to the front end server associated with the front end server private IP address to bypass the SLB.

[99] Clause 17. The computer-readable storage medium of clause 15, wherein adding the first outer header to the first packet is performed based on virtual extensible local area network (VXLAN) encapsulation.

[100] Clause 18. The computer-readable storage medium of clause 17, wherein the first outer header further includes a VXLAN identification (ID), the second source IP address being a hybrid gateway private IP address, the second destination IP address being a SLB private IP address.

[101] Clause 19. The computer-readable storage medium of clause 15, the first packet is sent from an on-premises network, the first source IP address being a private IP address of the on-premises network, the first destination IP address being an IP address of cloud service.

[102] Clause 20. The computer-readable storage medium of clause 15, wherein the front end server private IP address is saved in a session table.

Conclusion

[103] Although the subject matter has been described in language specific to structural features and/or methodological acts, it is to be understood that the subject matter defined in

the appended claims is not necessarily limited to the specific features or acts described. Rather, the specific features and acts are disclosed as exemplary forms of implementing the claims.

CLAIMS

1. A method, comprising:

receiving a first packet, the first packet including a first source Internet protocol (IP) address and a first destination IP address;

adding a first outer header to the first packet to generate an encapsulated packet, the first outer header including a second source IP address and a second destination IP address;

forwarding the encapsulated packet to a Server Load Balancer (SLB) associated with the second destination IP address;

receiving a response packet from a front end server having a front end server private IP address; and

saving the front end server private IP address to bypass the SLB.

2. The method of claim 1, further comprising:

receiving a second packet, the second packet including the first source IP address and the first destination IP address;

adding a second outer header to the second packet to generate a bypass packet, the second outer header including a third source IP address and a third destination IP address, the third destination IP address being the front end server private IP address; and

forwarding the bypass packet to the front end server associated with the front end server private IP address to bypass the SLB.

3. The method of claim 1, wherein adding the first outer header to the first packet is performed based on virtual extensible local area network (VXLAN) encapsulation.

4. The method of claim 3, wherein the first outer header further includes a VXLAN identification (ID), the second source IP address being a hybrid gateway private IP address, the second destination IP address being a SLB private IP address.

5. The method of claim 1, wherein the first packet is sent from an on-premises

network, the first source IP address being a private IP address of the on-premises network, the first destination IP address being an IP address of cloud service.

6. The method of claim 1, wherein the front end server private IP address is saved in a session table.

7. The method of claim 2, wherein after forwarding the encapsulated packet to the SLB, the method further comprises:

receiving, by the SLB, the encapsulated packet;

selecting, by the SLB, the front end server associated with the front end server private IP address;

replacing the second destination IP address of the encapsulated packet with the front end server private IP address to generate a modified packet; and

forwarding the modified packet to the front end server.

8. An apparatus, comprising:

one or more processors;

memory coupled to the one or more processors, the memory storing computer-readable instructions executable by the one or more processors that when executed by the one or more processors, cause the one or more processors to perform acts comprising:

receiving a first packet, the first packet including a first source Internet protocol (IP) address and a first destination IP address;

adding a first outer header to the first packet to generate an encapsulated packet, the first outer header including a second source IP address and a second destination IP address;

forwarding the encapsulated packet to a Server Load Balancer (SLB) associated with the second destination IP address;

receiving a response packet from a front end server having a front end server private IP address; and

saving the front end server private IP address to bypass the SLB.

9. The apparatus of claim 8, wherein the acts further comprise:

receiving a second packet, the second packet including the first source IP address and the first destination IP address;

adding a second outer header to the second packet to generate a bypass packet, the second outer header including a third source IP address and a third destination IP address, the third destination IP address being the front end server private IP address; and

forwarding the bypass packet to the front end server associated with the front end server private IP address to bypass the SLB.

10. The apparatus of claim 8, wherein adding the first outer header to the first packet is performed based on virtual extensible local area network (VXLAN) encapsulation.

11. The apparatus of claim 10, wherein the first outer header further includes a VXLAN identification (ID), the second source IP address being a hybrid gateway private IP address, the second destination IP address being a SLB private IP address.

12. The apparatus of claim 8, wherein the first packet is sent from an on-premises network, the first source IP address being a private IP address of the on-premises network, the first destination IP address being an IP address of cloud service.

13. The apparatus of claim 8, wherein the front end server private IP address is saved in a session table.

14. The apparatus of claim 9, wherein after forwarding the encapsulated packet to the SLB, the method further comprises:

receiving, by the SLB, the encapsulated packet;

selecting, by the SLB, the front end server associated with the front end server private IP address;

replacing the second destination IP address of the encapsulated packet with the front

end server private IP address to generate a modified packet; and

forwarding the modified packet to the front end server.

15. A computer-readable storage medium storing computer-readable instructions executable by one or more processors, that when executed by the one or more processors, cause the one or more processors to perform acts comprising:

receiving a first packet, the first packet including a first source Internet protocol (IP) address and a first destination IP address;

adding a first outer header to the first packet to generate an encapsulated packet, the first outer header including a second source IP address and a second destination IP address;

forwarding the encapsulated packet to a Server Load Balancer (SLB) associated with the second destination IP address;

receiving a response packet from a front end server having a front end server private IP address; and

saving the front end server private IP address to bypass the SLB.

16. The computer-readable storage medium of claim 15, wherein the acts further comprise:

receiving a second packet, the second packet including the first source IP address and the first destination IP address;

adding a second outer header to the second packet to generate a bypass packet, the second outer header including a third source IP address and a third destination IP address, the third destination IP address being the front end server private IP address; and

forwarding the bypass packet to the front end server associated with the front end server private IP address to bypass the SLB.

17. The computer-readable storage medium of claim 15, wherein adding the first outer header to the first packet is performed based on virtual extensible local area network

(VXLAN) encapsulation.

18. The computer-readable storage medium of claim 17,wherein the first outer header further includes a VXLAN identification (ID), the second source IP address being a hybrid gateway private IP address, the second destination IP address being a SLB private IP address.

19. The computer-readable storage medium of claim 15, the first packet is sent from an on-premises network, the first source IP address being a private IP address of the on-premises network, the first destination IP address being an IP address of cloud service.

20. The computer-readable storage medium of claim 15, wherein the front end server private IP address is saved in a session table.

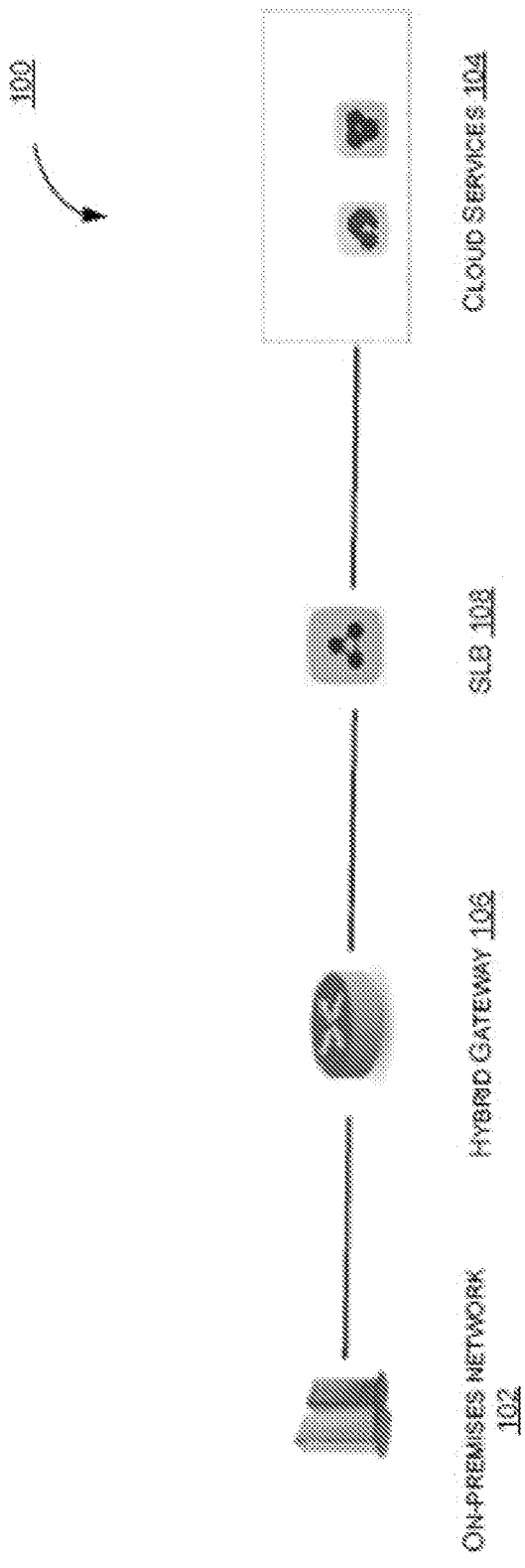
100
CLOUD SERVICES 104
SLB 108
HYBRID GATEWAY 106
ON-PREMISES NETWORK 102

FIG. 1

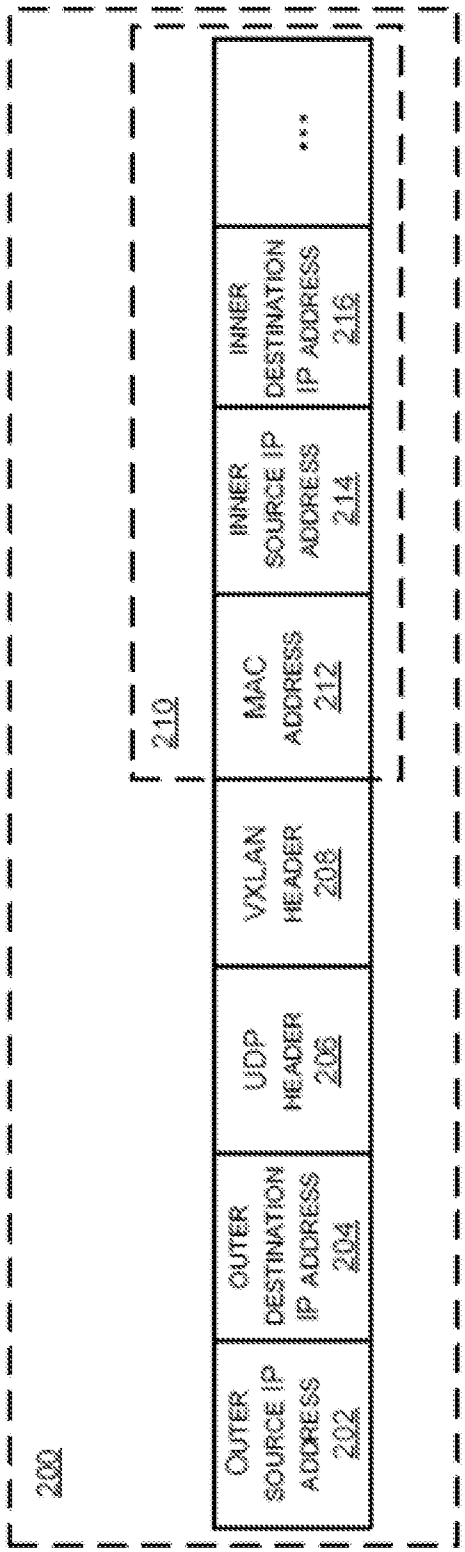
200
OUTER SOURCE IP ADDRESS 202
OUTER DESTINATION IP ADDRESS 204
UDP HEADER 206
VXLAN HEADER 208
210
MAC ADDRESS 212
INNER SOURCE IP ADDRESS 214
INNER DESTINATION IP ADDRESS 216
...

FIG. 2

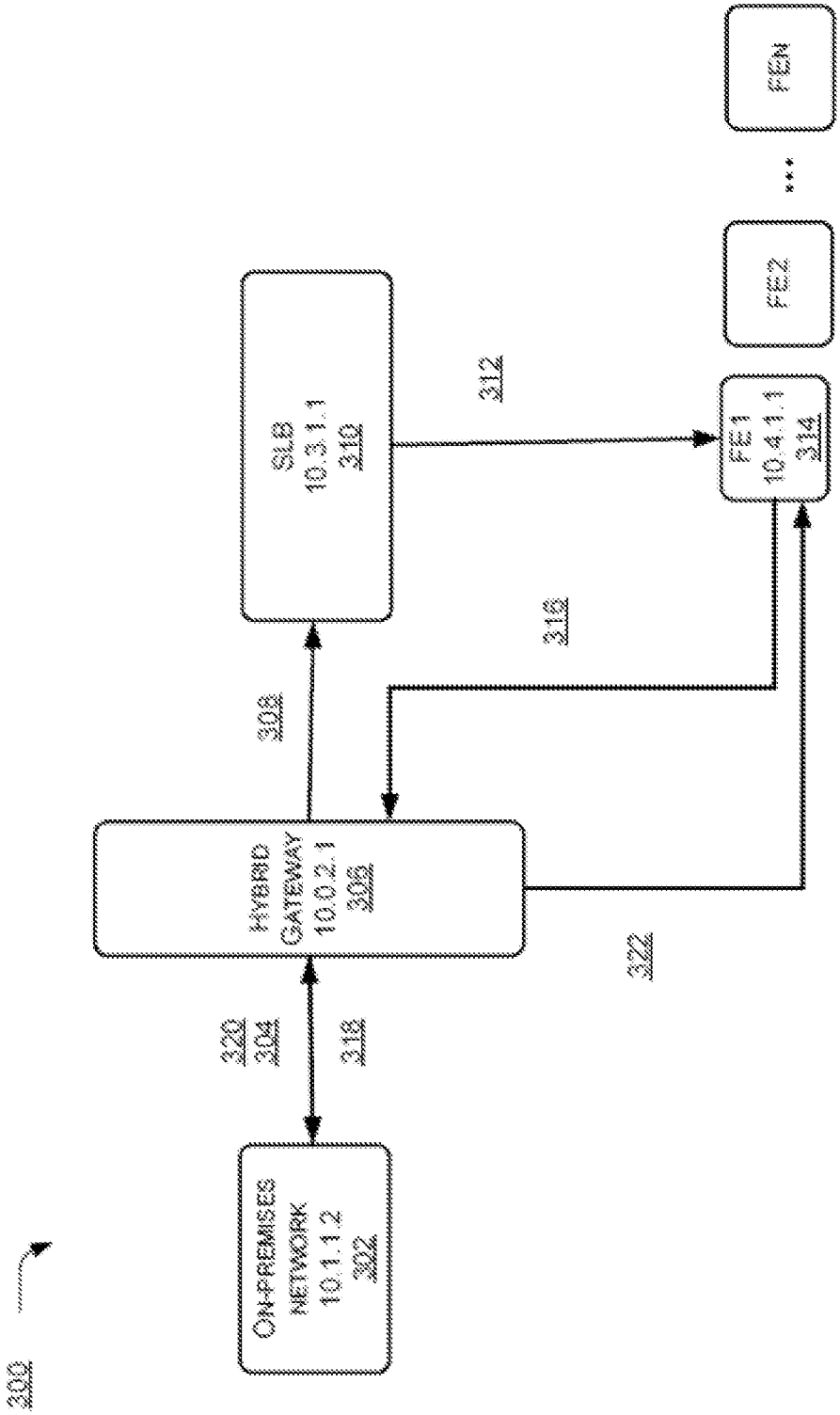
300
ON-PREMISES NETWORK 10.1.1.2 302
320 304
318
HYBRID GATEWAY 10.0.2.1 306
308
SLB 10.3.1.1 310
312
316
322
FE1 10.4.1.1 314
FE2
FEN

FIG. 3

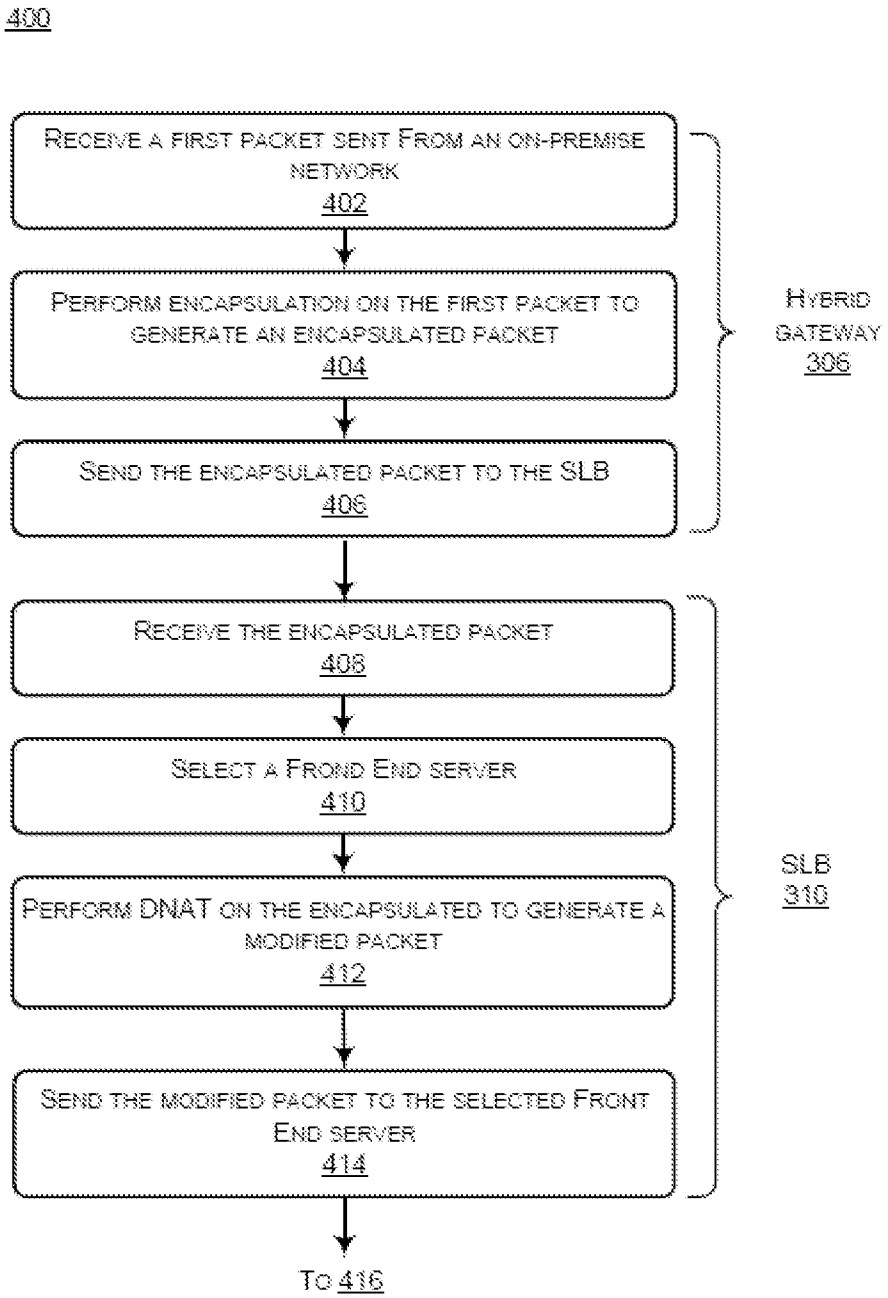
400
Receive a first packet sent from an on-premise network
402
Perform encapsulation on the first packet to generate an encapsulated packet
404
Send the encapsulated packet to the SLB
406
Hybrid gateway
306
Receive the encapsulated packet
408
Select a Frond End server
410
Perform DNAT on the encapsulated to generate a modified packet
412
Send the modified packet to the selected Front End server
414
SLB
310
To 416

FIG. 4A

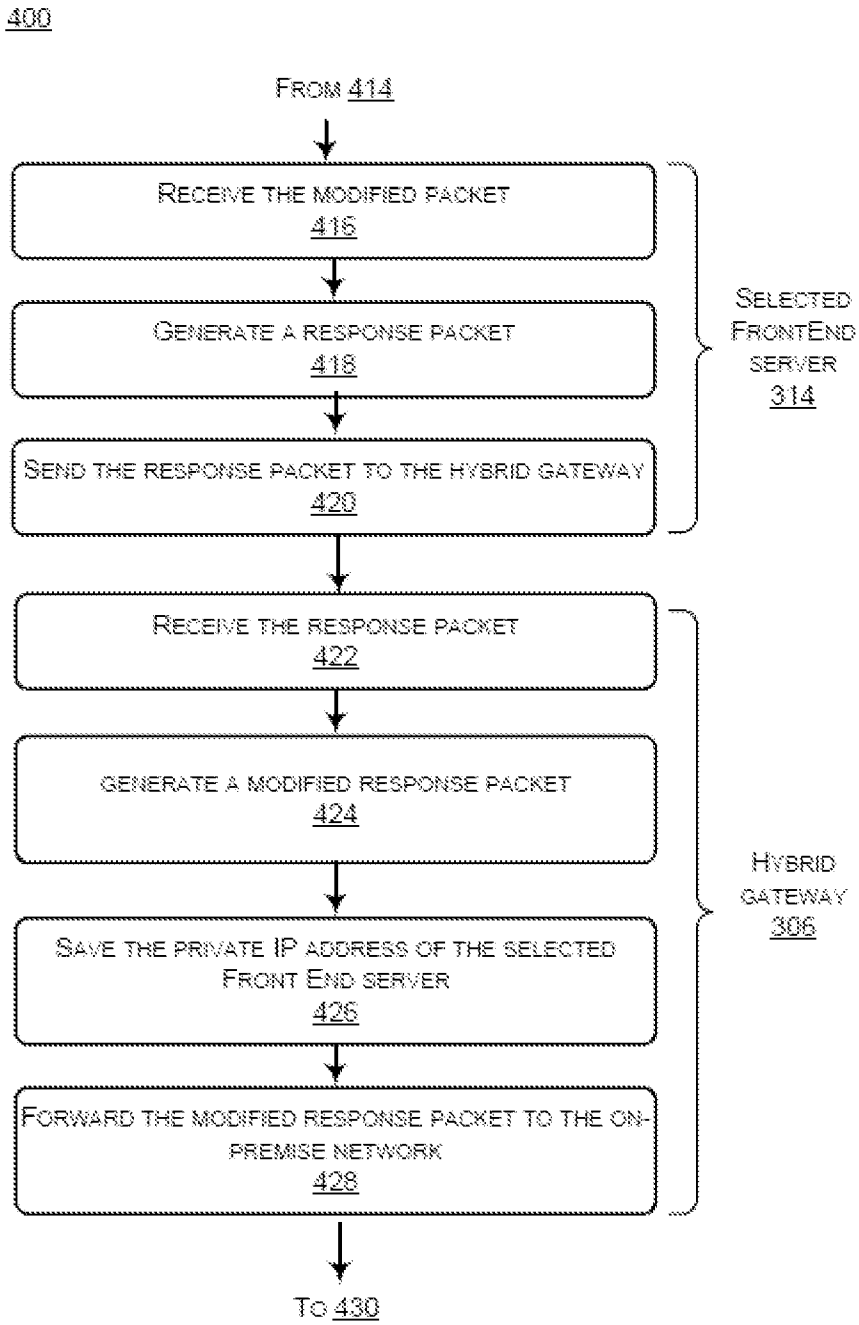
400
FROM 414
RECEIVE THE MODIFIED PACKET
416
GENERATE A RESPONSE PACKET
418
SEND THE RESPONSE PACKET TO THE HYBRID GATEWAY
420
SELECTED FRONTEND SERVER
314
RECEIVE THE RESPONSE PACKET
422
GENERATE A MODIFIED RESPONSE PACKET
424
SAVE THE PRIVATE IP ADDRESS OF THE SELECTED FRONT END SERVER
426
FORWARD THE MODIFIED RESPONSE PACKET TO THE ON-PREMISE NETWORK
428
HYBRID GATEWAY
306
TO 430

FIG. 4B

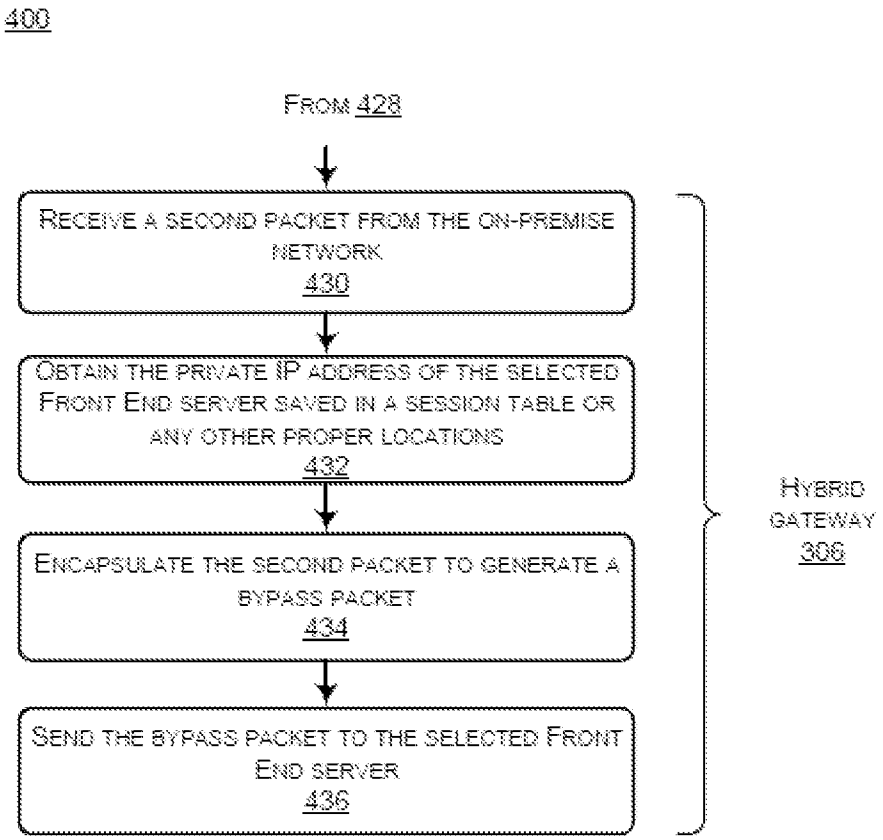
400
FROM 428
RECEIVE A SECOND PACKET FROM THE ON-PREMISE NETWORK
430
OBTAIN THE PRIVATE IP ADDRESS OF THE SELECTED FRONT END SERVER SAVED IN A SESSION TABLE OR ANY OTHER PROPER LOCATIONS
432
ENCAPSULATE THE SECOND PACKET TO GENERATE A BYPASS PACKET
434
SEND THE BYPASS PACKET TO THE SELECTED FRONT END SERVER
436
HYBRID GATEWAY
306

FIG. 4C

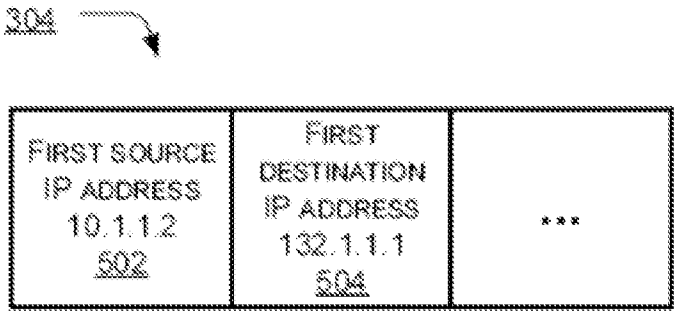
304
FIRST SOURCE IP ADDRESS
10.1.1.2
502
FIRST DESTINATION IP ADDRESS
132.1.1.1
504
***

FIG. 5A

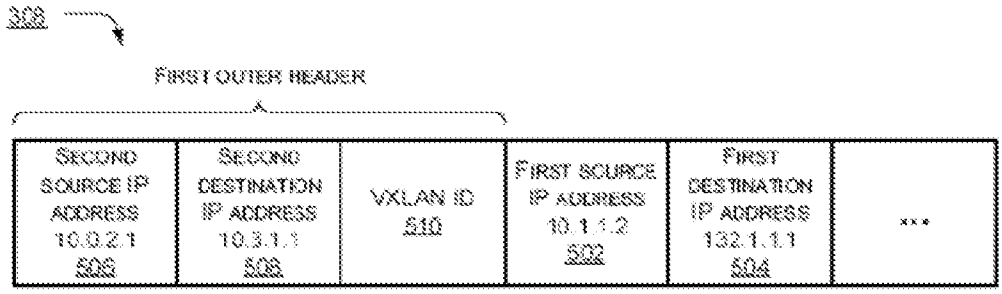

FIG. 5B

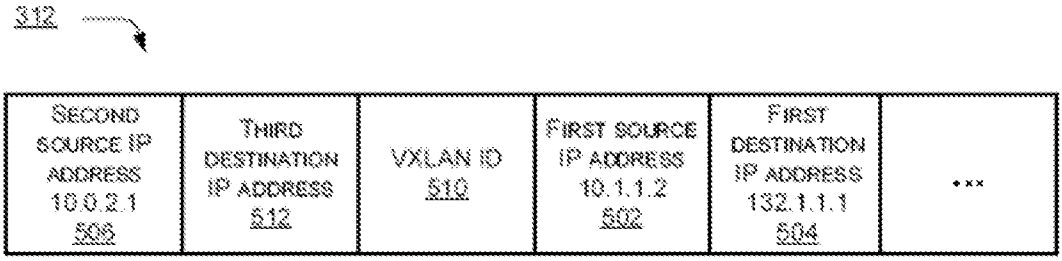

FIG. 5C

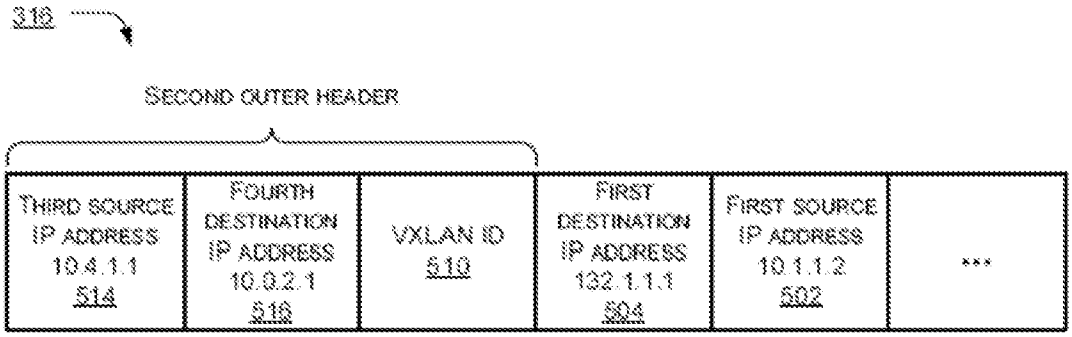

FIG. 5D

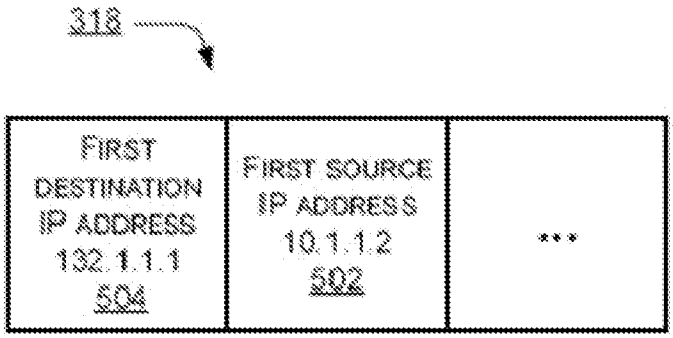
318
FIRST DESTINATION IP ADDRESS 132.1.1.1 504
FIRST SOURCE IP ADDRESS 10.1.1.2 502
...

FIG. 5E

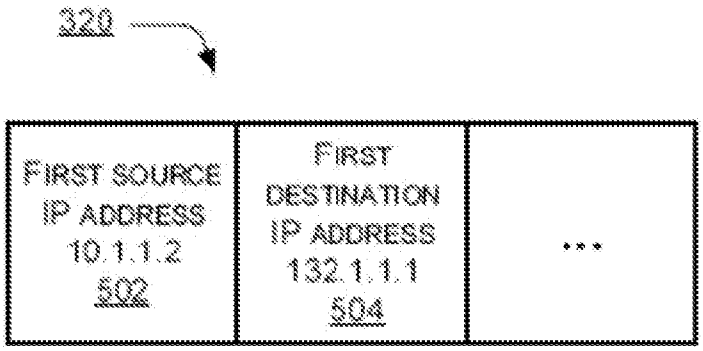
320
FIRST SOURCE IP ADDRESS 10.1.1.2 502
FIRST DESTINATION IP ADDRESS 132.1.1.1 504
...

FIG. 5F

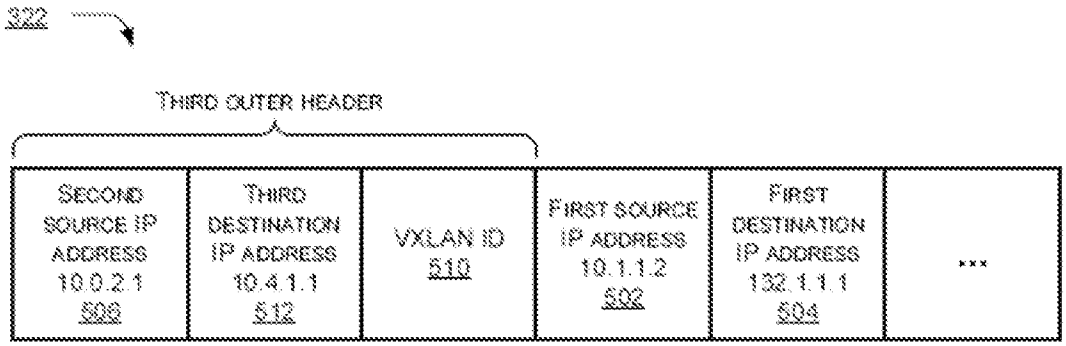
322
THIRD OUTER HEADER
SECOND SOURCE IP ADDRESS 10.0.2.1 508
THIRD DESTINATION IP ADDRESS 10.4.1.1 512
VXLAN ID 510
FIRST SOURCE IP ADDRESS 10.1.1.2 502
FIRST DESTINATION IP ADDRESS 132.1.1.1 504
...

FIG. 5G

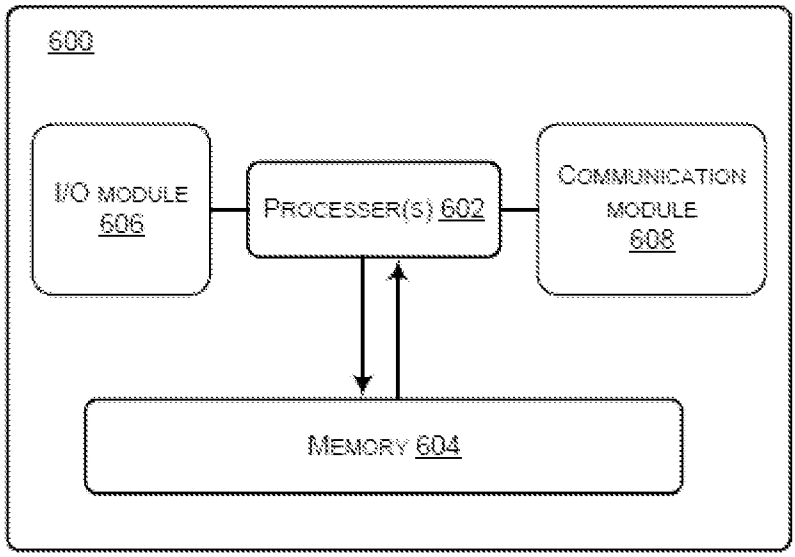
600
I/O MODULE 606
PROCESSER(S) 602
COMMUNICATION MODULE 608
MEMORY 604

FIG. 6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2019/126622

A. CLASSIFICATION OF SUBJECT MATTER

H04L 12/46(2006.01)i

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)

WPI, EPODOC, CNPAT, CNKI, IEEE: VXLAN, SLB, server, load balancer, IP, encapsulated, packet, address, bypass

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2018176307 A1 (NICIRA, INC.) 21 June 2018 (2018-06-21) description, paragraphs [0043]-[0052], [0088]-[0106], [0166], figures 1, 3	1-20
A	US 2018139272 A1 (MICROSOFT TECHNOLOGY LICENSING, LLC) 17 May 2018 (2018-05-17) the whole document	1-20
A	CN 107465590 A (ALIBABA GROUP HOLDING LIMITED) 12 December 2017 (2017-12-12) the whole document	1-20
A	CN 104521195 A (HUAWEI TECHNOLOGIES CO., LTD.) 15 April 2015 (2015-04-15) the whole document	1-20
A	US 2017099188 A1 (CISCO TECHNOLOGY, INC.) 06 April 2017 (2017-04-06) the whole document	1-20

☐ Further documents are listed in the continuation of Box C. ☑ 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
09 March 2020	**26 March 2020**
Name and mailing address of the ISA/CN **National Intellectual Property Administration, PRC** **6, Xitucheng Rd., Jimen Bridge, Haidian District, Beijing** **100088** **China**	Authorized officer **WU,Xue**
Facsimile No. **(86-10)62019451**	Telephone No. **86-(10)-53961406**

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2019/126622

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
US	2018176307	A1	21 June 2018	None			
US	2018139272	A1	17 May 2018	US	10320895	B2	11 June 2019
				CN	109937401	A	25 June 2019
				EP	3542268	A1	25 September 2019
				WO	2018093616	A1	24 May 2018
CN	107465590	A	12 December 2017	US	2017353351	A1	07 December 2017
				US	10348556	B2	09 July 2019
CN	104521195	A	15 April 2015	US	2015156035	A1	04 June 2015
				EP	2875615	A1	27 May 2015
				US	9705702	B2	11 July 2017
				EP	2875615	B1	21 March 2018
				US	8989192	B2	24 March 2015
				US	2014050223	A1	20 February 2014
				CN	104521195	B	09 March 2018
				WO	2014028612	A1	20 February 2014
US	2017099188	A1	06 April 2017	None			