

DESCRIPTION

TITLE OF INVENTION

CONTROL APPARATUS FOR GATEWAY IN MOBILE COMMUNICATION
SYSTEM

5 TECHNICAL FIELD

[0001] The present invention relates to a control
apparatus for a gateway in a mobile communication
system.

BACKGROUND ART

10 [0002] In recent years, mobile communication services
provided by an MVNO (Mobile Virtual Network Operator)
have been widespread. The MVNO provides access to an
external network to a UE (User Equipment) using network
components owned by an MNO (Mobile Network Operator),
15 and a gateway apparatus of MVNO. In addition, cloud
services such as an AWS (Amazon Web Service) have also
been spread. The AWS can change the number of servers
to be used and their performance on demand and can
provision a private network (closed network) in the
20 cloud.

SUMMARY OF INVENTION

[0003] It is assumed that a gateway apparatus which
provides an MVNO is created on a cloud using cloud
characteristics, and the processing capability of the
25 gateway apparatus is changed in accordance with a
communication demand. If an existing gateway apparatus
is simply provided on the cloud, it is difficult to

provide a service based on a demand from users. The present invention provides an object of solving various problems when providing a gateway apparatus in a mobile communication system on a cloud.

5 [0004] According to some embodiments of the present invention, there is provided a control apparatus for a gateway in a mobile communication system, characterized in that the gateway includes a first server configured to receive data from a user apparatus and one or more
10 second servers configured to transfer the data received by the first server to an external network, the one or more second servers are virtual machines on a cloud, and the control apparatus performs an operation of changing performance of the one or more second servers,
15 an operation of monitoring a processing status of the one or more second servers, and an operation of selecting, from the one or more second servers in accordance with the processing status of the one or more second servers, a second server as a transfer
20 destination of the data received from the user apparatus by the first server.

[0005] Other features and advantages of the present invention will be apparent from the following descriptions taken in conjunction with the accompanying
25 drawings, in which like reference characters designate the same or similar parts throughout the figures thereof.

BRIEF DESCRIPTION OF DRAWINGS

[0006] The accompanying drawings, which are incorporated in and constitute a part of the specification, illustrate embodiments of the invention
5 and, together with the description, serve to explain the principles of the invention.

Fig. 1 is a block diagram for explaining an example of the arrangement of a mobile communication system according to some embodiments;

10 Fig. 2 is a timing chart for explaining an example of the operation of the mobile communication system according to some embodiments;

Fig. 3 explains an example of a management table of the mobile communication system according to some
15 embodiments;

Fig. 4 is a block diagram for explaining an example of the arrangement of a mobile communication system according to some embodiments;

Fig. 5A explains an example of a management table
20 in the mobile communication system according to some embodiments;

Fig. 5B explains an example of a management table in the mobile communication system according to some embodiments;

25 Fig. 6 is a flowchart for explaining an example of the operation of a control server according to some embodiments; and

Fig. 7 is a block diagram for explaining an example of the protocol configuration of the mobile communication system according some embodiments.

DESCRIPTION OF EMBODIMENTS

5 [0007] The arrangement of a mobile communication system according to some embodiments of the present invention will be described with reference to Fig. 1. Although a mobile communication system complying with LTE will be described with reference to Fig.1, the
10 present invention is also applicable to a mobile communication system complying with 3G or any other standard. The mobile communication system mainly includes eNBs 102, an S-GW 103, a P-GW 104, and a control server 107. The eNBs 102 and the S-GW 103 are
15 provided by an MNO (Mobile Network Operator), and the P-GW 104 and the control server 107 are provided by an MVNO (Mobile Virtual Network Operator). The P-GW 104 and the control server 107 may be created by the MVNO itself, or may be created and operated by an MVNE
20 (Mobile Virtual Network Enabler) according to an MVNO request. The S-GW 103 may be provided by the MVNO in place of the MNO.

[0008] UEs 101 are devices to which IMSIs (International Mobile Subscriber Identities) are
25 assigned and can be, for example, mobile phones, personal computers, sensors, and actuators. The IMSI is identification information uniquely assigned to each

device and is incorporated in the device in a state in which the IMSI is stored in, for example, a SIM.

[0009] Each eNB 102 connects each UE 101 to the S-GW 103 and transfers a packet between the UE 101 and the S-GW 103. Each eNB 102 further provides a wireless resource management function, a mobility management function, and a scheduling function to each UE 101. The S-GW 103 provides a function of routing a packet from or to each UE 101. The S-GW 103 corresponds to the SGSN (packet access control node) of a 3G network.

[0010] The P-GW 104 has a function of providing access to an external network 108 to each UE 101. The P-GW 104 is a gateway apparatus functioning as the endpoint of a core network included in the mobile communication system. The external network 108 may be a public network such as the Internet or a private network provided by each individual enterprise. The P-GW 104 corresponds to the GGSN (packet gateway node) of the 3G network. A packet from each eNB 102 to the P-GW 104 is transferred in an encapsulated form via a tunnel (GTP tunnel) established in accordance with a GTP (GPRS Tunneling Protocol). Any other L2 tunnel may be used in place of the GTP tunnel.

[0011] The P-GW 104 is formed from one or more first servers 105 and one or more second servers 106. Each first server 105 is connected to the S-GW 103 and transfers a packet received from the S-GW 103 to any

one of the one or more second servers 106. Each first server 105 transfers, to the S-GW 103, a packet received by one of the one or more second servers 106.

[0012] Each second server 106 provides various
5 services to the communications by the UE 101. For example, the second server 106 functions as NAT which performs processing in a network layer or a proxy which performs processing in an application layer. In addition, the second server 106 may perform processing
10 for a still image and a moving image or credential assignment processing in place of the UE 101.

[0013] The upper limit of the number of servers simultaneously connectable to the S-GW 103 in the P-GW 104 is determined by the MNO. To solve this problem,
15 the P-GW 104 is arranged as follows. The servers are classified into servers (first servers 105) which perform packet (data) exchange with the S-GW 103 and servers (second servers 106) which perform access to the external network 108 and provide various additional
20 services, thereby increasing the number of second servers 106 so as to exceed the upper limit of connectable servers by the MNO.

[0014] Each of the one or more first servers 105 and each of the one or more second servers 106 can be
25 virtual machines, respectively, on a cloud. This cloud may be a public cloud such as an AWS or a private cloud provided for one enterprise. If the cloud is a public

cloud, the one or more first servers 105 and the one or more second servers 106 may belong to a virtual private network on the cloud. For example, if the cloud is an AWS, a virtual private network is provided by a VPC

5 (Virtual Private Cloud) function.

[0015] The provision of the P-GW 104 on the cloud can change the performance of the P-GW 104 at an appropriate timing in accordance with the processing status of the P-GW 104. The change in performance of
10 the P-GW 104 can be implemented by replacing (so-called scale up/scale down) one server with another server (a server having processing capability higher or lower than the previous server) or implemented by changing (so-called scale out/scale in) the number of servers.

15 [0016] The control server 107 is a server which controls the operation of the P-GW 104. The control server 107 may be a virtual machine on a cloud or an on-premise server. The control server 107 may be formed from a single server or a plurality of servers.

20 In this embodiment, the control server 107 is exemplified as a server different from the first server 105 and the second server 106. However, one of the first servers 105 or one of the second servers 106 may have a function of the control server 107.

25 [0017] The control server 107 monitors the processing status of each of the one or more first servers 105 and each of the one or more second servers 106. Examples

of the processing status of a monitoring target can be a CPU use rate, a memory use rate, a network use rate, an average throughput, the number of GTP sessions, and the like. A criterion of the processing status used to
5 change the performance of the first server 105 may be different from a criterion of the processing status used to change the performance of the second server 106. Since the first server 105 performs the function of transferring a packet to the second server 106, the
10 network performance tends to become a bottleneck. For this reason, the control server 107 may change the performance of the first server 105 using the network performance (for example, the network use rate, the average throughput, or the like) as a criterion. Since
15 the second server 106 performs various processing operations for a packet, and its calculation performance tends to become a bottleneck. For this reason, the control server 107 may change the performance of the first server 105 using the
20 calculation function (for example, the CPU use rate, the memory use rate, or the like) as a criterion.

[0018] The control server 107 may change the performance of the P-GW 104 regardless of the processing status of the P-GW 104. For example, the
25 control server 107 may change the performance of the P-GW 104 in accordance with an instruction from the administrator or user of the mobile communication

system. In place of this, the control server 107 may automatically change the performance of the P-GW 104 in accordance with the prescheduled settings.

[0019] Subsequently, an example of a protocol configuration of a U plane (user plane) of the mobile communication system in Fig. 1 will be described with reference to Fig. 7. An end-to-end GTP tunnel 701 is established between the UE 101 and the second server 106. A packet transmitted from the UE 101 is transferred to the second server 106 via the GTP tunnel 701. An IP address (to be referred to as a user IP address hereinafter) used in an IP connection 702 with the UE 101 and an IP address (to be referred to as a transport IP address hereinafter) used in an IP connection 703 with the first server 105 are assigned to the second server 106. An IP packet is transferred from the UE 101 via the IP connection 702, and a GTP packet is transferred from the first server 105 via the IP connection 703.

[0020] Since the virtual machines in the virtual private network communicate with each other depending on a cloud service, an IP address in a subnetwork assigned to the virtual private network must be set in each virtual machine. The upper limit of the number of IP addresses in the subnetwork is determined by a cloud service. If this IP address in the subnetwork is assigned to the UE 101, the number of UEs 101

connectable to the mobile communication system is limited by the number of IP addresses in the virtual private network. For this reason, an address irrelevant to the address in the virtual environment

5 subnet is assigned to the UE 101, NAT is performed at the virtual environment address assigned to a node when IP communication is performed with the outside of the second server 106 which performs GTP tunnel termination. In addition, in the second server 106, an IP address in

10 the subnetwork assigned to the virtual private network is set as a transport IP address for causing the first server 105 to access the second server 106. An IP address not limited by the virtual private network is set as a user IP address for causing the UE 101 to

15 access the second server 106. If the user IP address is additionally assigned to the second server 106 which performs GTP tunnel termination, the communication from the UE 101 to the user IP address of the second server 106 can reach the second server 106 without any NAT.

20 [0021] Subsequently, an example of the operation of the mobile communication system in Fig. 1 will be described with reference to Fig. 2. This operation can be performed by causing a processor such as a CPU of each UE or each server to execute a program. In place

25 of this, each UE and each server may include circuits (for example, ASICs) for executing the following operation.

[0022] In S1, the UE 101 is attached to the eNB 102 and requests to generate a session with the P-GW 104 to the core network. The UE 101 includes its own IMSI (to be referred to as an IMSI_A hereinafter) in this
5 request. The S-GW 103 establishes a C-plane GTP tunnel (GTP-C) with one of the one or more first servers 105 and transfers the request from the UE 101 to the first server 105 via this GTP tunnel. For example, the S-GW 103 selects, by a round robin method, one first server
10 105 from the one or more first servers 105 connected to the S-GW 103.

[0023] Upon reception of the request from the UE 101, the first server 105 determines whether the first server 105 refers to customer information and a session
15 is established for the IMSI included in the request. This determination may be performed by causing the first server 105 to refer to the customer information or performed by the control server 107 which has received the request from the first server 105. If the
20 session can be established, the first server 105 assigns an IP address (to be referred to as 172.24.240.100) to the UE 101 and returns this IP address to the UE 101 in S2.

[0024] In S2, the first server 105 sends, to the
25 control server 107, a request for assignment of the second server 106. In response to this request, the control server 107 selects one second server 106 of the

one or more second servers 106 and returns the transport IP address of this second server 106 to the first server 105. The control server 107 may select one second server 106 (for example, a server having a minimum load) in accordance with the processing status of the one or more second servers 106. The first server 105 establishes a U-plane GTP tunnel (GTP-U) between the UE 101 and the selected second server 106. In addition, the first server 105 notifies each second server 106 of information concerning a pair of the IMSI of the UE 101 and the IP address assigned to the UE 101. This notification can be directly performed from the first server 105 to the second server 106 or may be performed via the control server 107. For example, the first server 105 may register information concerning the pair of the IMSI and the IP address in the control server 107, and the second server 105 may read out this information from the control server 107.

[0025] In S3, the UE 101 transmits, to the first server 105 via the GTP tunnel, an HTTP request toward the user IP address (to be referred to as 169.254.254.169) of the second server 106. The UE 101 sets the IP address (172.24.240.100) assigned in S2 in the transmission source IP address of this HTTP request. The user IP address of the second server 106 may be stored in the UE 101 before the start of the operation in S1 or may be notified from the first server 105 to

the UE 101 in S2. A single IP address may be shared as the user IP addresses of the plurality of second servers 106.

[0026] In S4, the first server 105, the first server
5 105 transfers, to the second server 106 via the GTP tunnel established in S2, the GTP packet transmitted from the UE 101 via the GTP tunnel.

[0027] In S5, the second server 106 terminates the GTP tunnel to extract the IP packet from the GTP packet.
10 In addition, the second server 106 resolves the IMSI of the UE 101 which has transmitted the HTTP request by using the information concerning the pair notified in S2 and the transmission source IP address of this HTTP request. In place of this, the second server 106 may
15 transmit this transmission source IP address (the IP address of the UE 101) to the control server 107, cause the control server 107 to resolve the IMSI, and return it to the second server 106.

[0028] In S6, the second server 106 sends an inquiry
20 about processing to be executed to the resolved IMSI to the control server 107. This processing includes processing to data transmitted from the UE 101 to the external network 108 and processing to data returned from the external network 108 to the UE 101. The
25 control server 107 manages a table 300 shown in Fig. 3. Each entry of the table 300 indicates the processing contents to be executed for the data from the UE 101.

A column 301 indicates the IMSI of the UE 101 as the processing target. A column 302 indicates a request protocol before conversion. A column 303 indicates a request protocol after conversion. A column 304 indicates a request destination. For example, the first-row entry of the table 300 indicates that an HTTP request from the UE 101 having the IMSI_A is converted into an HTTPS request and the converted request is transferred to https://server1.example.com/. The second server 106 can perform individual processing for each UE 101. For example, the second server 106 can similarly perform conversion from mqtt to mqttS, conversion from TCP to TCPS, and conversion from WebSocket to WebSocket over SSL. In addition to simple encryption, the second server 106 can perform conversion of a protocol itself such as HTTPS transfer of data received by a Raw TCP socket or data received by a UDP packet, and communication optimization between the UE 101 and the second server 106 such as compression of response contents from the external server or deletion of unnecessary header information. The latter function is particularly advantageous in a device such as a sensor or actuator aiming at energy saving.

[0029] In addition to the protocol conversion, the second server 106 can add authentication information such as the user name and password of a site to be

accessed by the UE 101, and add information referred to in processing on the server side, such as an identification number such as the IMEI (International Mobile Equipment Identifier) of the UE 101 and a time stamp indicating data reception time. The second server 106 can also add a digital signature using secret information exchanged in advance with the reception side so as to allow the reception side to verify the authentication of the added information. Accordingly, access by spoofing of the third party can be prevented.

[0030] The control server 107 provides an API for editing the table 300. Via this API (Application Programming Interface), the administrator of the UE 101 can edit (for example, add, change, or delete) the operations to be performed for his own UE 101. Similarly, by using the API, the administrator of the UE 101 can also set information such as a user name and password for authentication and secret information for verifying authenticity of the information added by the second server 106. The control server 107 may provide a GUI (Graphical User Interface) for allowing the administrator of the UE 101 to facilitate the above processing.

[0031] In S7, the control server 107 responds, to the second server 106, the operations to be performed of an entry indicating the result obtained by searching the

table 300 using, as a key, the IMSI received from the second server 106.

[0032] In S8, the second server 106 converts the HTTP request from the UE 101 into the HTTPS request in accordance with the operations to be performed received from the control server 107 and transfers the HTTPS request to the destination (https://server1.example.com/) designated by the processing contents. In S9, the second server 106 receives a response to the HTTPS request from the destination. In S10, the second server 106 adds a GTP header to this response and transfers the resultant data to the first server 105. The first server 105 as the transfer destination is the same as the first server 105 from which the HTTP request has been transferred in S4. In S11, the first server 105 transfers the response to the UE 101 via the GTP tunnel. After that, the first server 105 may cancel the GTP tunnel between the UE 101 and the second server 106.

[0033] According to the above processing, the common user IP address is assigned to the plurality of second servers 106. Since the communication between the first server 105 and the second server 106 is identified by the transport IP address assigned to each second server 106, identifiable GTP tunnels are established between the plurality of UEs 101 and the plurality of second servers 106. For this reason, even if the scale

in/scale out operation or the scale down/scale up operation of the second server 106 is performed, the endpoint used as the destination of the UE 101, that is, the user IP address of the second server 106 need not
5 be changed.

[0034] According to the above processing, since the user IP address of the second server 106 can be set in advance in the UE 101, the overhead concerning the DNS processing for obtaining this IP address can be
10 suppressed. In addition, the user IP address of the second server 106 can be set as a hardcode in the UE 101. This is particularly advantageous when the UE 101 is a weak embedded device such as a sensor. Even if the user IP address of the second server 106 set in the
15 UE 101 cannot be changed, the communication partner of the UE 101 in the external network 108 can be changed by updating the table 300 of the control server 107.

[0035] Subsequently, communication between the first server 105 and the second server 106 in the processing
20 of Fig. 2 will be described in detail. Although only one S-GW 103 is illustrated in Fig. 1, a general mobile communication system includes a plurality of S-GWs 103. A GTP tunnel between the S-GW 103 and the P-GW 104 (the first server 105 in practice) can be identified by a
25 TEID (Tunnel Endpoint Identifier). Each first server 105 uniquely sets the TEID of the GTP tunnel with the S-GW 103 to prevent the interference between the GTP

tunnels. However, if the respective first servers 105 set the same TEID for the respective GTP tunnels and packets are transferred to the same second server from the respective GTP tunnels, the interference occurs
5 between the GTP tunnels.

[0036] To solve this problem, the control server 107 may monitor the TEID assigned by each first server 105 and may select the second server 106 to which the packet is transferred from the first server 105 so as
10 to prevent the interference between the GTP tunnels in the second servers 106. More specifically, the control server 107 may select the second server 106 to which the packet is transferred from the first server 105 so as to prevent the TEIDs assigned by the respective
15 servers 105 from overlapping in the second servers 106.

[0037] In place of this, another tunnel different from the GTP tunnel between the S-GW 103 and the first server 105 may be set between the first server 105 and the second server 106, and a packet may be transferred
20 from the first server 105 to the second server 106 via the other tunnel. The tunnel between the first server 105 and the second server 106 may be a GTP tunnel. Accordingly, the same software stack can be used in the first server 105 and the second server 106.

25 [0038] In the above embodiment, the second server 106 terminates the GTP packet from the UE 101 to extract the IP address included in this GTP packet. In place

of this, a third server different from the second server 106 may terminate a GTP packet from the UE 101. In this case, the UE 101 sets the destination IP address of the IP packet as the user IP address of the third server. The second server 106 transfers, to the third server via the GTP tunnel or the other tunnel, the GTP packet transferred from the first server 105.

[0039] Subsequently, the arrangement of a mobile communication system according to another embodiment of the present invention will be described with reference to Fig. 4. The mobile communication system in Fig. 4 is different from the mobile communication system in Fig. 1 in the arrangement of a P-GW 104 and the operation of a control server 107. The rest may be the same as in the mobile communication system in Fig. 1. In Fig. 4, some constituent elements which are the same as those of the mobile communication system in Fig. 1 will not be illustrated. The same description as in Fig. 1 will be omitted.

[0040] A user having a UE 101 connectable to the mobile communication system may want to access his own private network using this UE 101 without going through the public network such as the Internet. For example, assume that a given enterprise purchases a plurality of SIMs and UEs assigned with these SIMs access the private network of this enterprise. Even if a second server 106 belongs to the virtual private network on a

cloud, the second server 106 can access another private network by a function provided by the cloud. The accessible other private network is another virtual private network present on the same cloud as the second
5 server 106 and different from the virtual private network to which the second server 106 belongs, or a private network of the on-premise or another cloud connected by a leased line to the virtual private network to which the second server 106 belongs.

10 [0041] The UE 101 of another user having no access right to a given private network of a given user should not access this private network. Due to the limitation of the cloud service, only single transfer information may be set for all servers belonging to the same
15 subnetwork in the virtual private network. The transfer information here is information for defining an external network which the UE 101 can access. The transfer information includes, for example, a routing policy defined by a routing table, and an NACL (Network
20 Access Control List) for access limitation. In the embodiment shown in Fig. 1, if all the second servers 106 are included in the same subnetwork, the single transfer information is applied to all the UEs 101 which can use this mobile communication system. As a
25 result, the UEs 101 which can access the external private network cannot be limited.

[0042] In this embodiment, the plurality of second

servers 106 are divided into a plurality of groups, the respective groups are included in different subnetworks, and the transfer information is defined for each group. In the embodiment shown in Fig. 4, the plurality of
5 second servers 106 are classified into groups 401A, 401B, 401C, and 401X. Each group includes at least one second server 106. As in the embodiment shown in Fig. 1, the control server 107 can change the performance of each second server 106 included in each group, as
10 needed. The second servers 106 belonging to the group 401A can access only a private network 108A. For example, the private network 108A is another virtual private network of the same cloud as the second servers 106. If this cloud is the AWS, these virtual private
15 networks can be implemented by the VPC function and are connected by the VPC peering function. The second servers 106 included in the group 401B can access only a private network 108B. For example, the private network 108B is an on-premise private network. If the
20 second servers 106 are included in the AWS, the second servers 106 and the private network 108B are connected by dedicated lines by the AWS Direct Connect function. The second servers 106 included in the group 401C can access only a private network 108C. For example, the
25 private network 108A is another virtual private network of the same cloud as the second servers 106. If this cloud is the AWS, these virtual private networks are

implemented by the VPC function and are connected by the VPN connection function. The private networks 108A to 108C are networks managed by different users. The second servers 106 included in the group 401X can
5 access only the Internet 108X. The one or more first servers 105 may belong to a subnetwork different from those of the second servers 106 or may belong to a virtual network different from those of the second servers 106.

10 [0043] Subsequently, the above transfer information will be described with reference to Figs. 5A and 5B. The transfer information is formed from two tables 500 and 510 and managed by the control server 107. The table 500 defines that a packet from each UE 101 should
15 be processed by the second server 106 of a specific group. Each entry (row) of the table 500 is generated for each IMSI. A column 501 indicates an IMSI. A column 502 indicates the group of the second server 106 which processes a packet from the UE 101 having each
20 IMSI.

[0044] The table 510 indicates information concerning second servers 106. Each entry (row) of the table 510 is generated for each second server 106. A column 511 indicates an identifier of the second server 106. The
25 identifier of the second server 106 is unique in the mobile communication system. A column 512 indicates the transport IP address of the second server 106. The

transport IP address is set for each second server 106.
A column 513 indicates the user IP address of the
second server 106. The common user IP address may be
set for the plurality of second servers 106 or user IP
5 addresses may be set for the plurality of second
servers, respectively, as in the above embodiment. If
the user IP address of the second server 106 is set in
the UE 101 in advance, the table 510 need not include
the column 513. A column 514 indicates the identifier
10 of a group to which the second server 106 belongs. The
group identifier is unique in the mobile communication
system. A column 515 indicates a subnetwork to which
the second server 106 belongs. A column 516 indicates
the external network of packet transfer destination of
15 the second server 106. The common transfer destination
is set for the second servers 106 belonging to the same
group. The administrator (for example, MVNO) of the
mobile communication system creates and updates the
tables 500 and 510 in accordance with a request from
20 the user of the mobile communication system.

[0045] Subsequently, an example of the operation of
the control server 107 will be described with reference
to Fig. 6. This operation can be done by causing a
processor such as the CPU of the control server 107 to
25 execute a program. In place of this, the control
server 107 may include a circuit (for example, ASIC)
which execute the following operation. The operation

in Fig. 6 is performed in S2 of Fig. 2. In the following description, a case in which the first server 105 sends, to the control server 107, an inquiry about the transport IP address of the second server 106 and
5 the user IP address of the second server 106 will be described below. As described above, however, the user IP address of the second server 106 may be set in the UE 101 in advance. In this case, processing about the user IP address of the second server 106 in the
10 following description may be omitted.

[0046] In S601, the control server 107 receives an inquiry to obtain the transport IP address and user IP address of the second server 106 from the first server 105. This inquiry includes the IMSI of the UE 101
15 received from the UE 101 in S1 of Fig. 2 to the first server 105.

[0047] In S602, the control server 107 specifies the group of the second servers 106 which are to process a packet from the UE 101 having the IMSI received in S601
20 with reference to the table 500. For example, if the IMSI received by the control server 107 is "IMSI_1A", the control server 107 specifies that the second servers 106 included in "Group A" should process the packet from the UE 101 having this IMSI.

25 [0048] In S603, the control server 107 specifies the second servers included in the specified group and selects one of the second servers with reference to the

table 510. For example, if the group specified by the control server 107 is "Group A", the control server 107 specifies "Node A1" and "Node A2" as the second servers 106 included in this group. The control server 107 may
5 select one second server 106 based on the processing status of each second server 106 included in the specified group. For example, the control server 107 selects the second server 106 having a lowest load.

[0049] In S604, the control server 107 reads out the
10 transport IP address and user IP address of the selected second server 106 from the table 510 and returns this transport IP address to the first server 105 as the inquiry source.

[0050] In S3 of Fig. 2, the first server 105 then
15 notifies the UE 101 of the user IP address of the selected second server 106. In S4 of Fig. 2, the UE 101 transmits, to the first server 105, the HTTP request toward the user IP address of the selected second server 106. In S5 of Fig. 2, the first server
20 105 transfers the packet to the selected second server 106 using the transport IP address of the second server 106 obtained in S604. After that, the second server 106 transfers the packet to the external network 108 in accordance with the transfer information set in the
25 subnetwork to which this second server 106 belongs.

[0051] The packet from a UE 101A whose IMSI is "IMSI_1A" is transferred to the second servers 106

included in the group 401A. Since the second servers 106 included in the group 401A can access only the private network 108A, the packet from the UE 101A is transferred to only the private network 108A, as indicated by a dotted line 402, but is not transmitted to other private networks, as indicated by a dotted line 403. Similarly, a UE 101B is transferred to only the private network 108B via the second servers 106 included in the group 401B. A UE 101C is transferred to only the private network 108C via the second servers 106 included in the group 401C. A UE 101X in which access to a private network is not set is transferred to only the Internet 108X via the second servers 106 included in the default group 401X.

[0052] In the embodiment shown in Fig. 4, the second server 106 terminates the GTP tunnel, extracts the IP packet, and then transfers the IP packet to the external network via the L3 connection. However, depending on a cloud service, the number of peer-connectable virtual networks and the number of virtual interfaces for private connections are limited. As a result, the number of private networks connectable to the P-GW 104 is limited. To solve this problem, the second server 106 may set the L2 tunnel with the server in the external private network and transfer the packet via the L2 tunnel. If the L2 tunnel is used, processing using the IP address of the UE 101 can be

performed on the private network side. Whether the L2 tunnel or L3 connection is used can be set for each private network.

[0053] In the above description, a case in which the
5 MVNO manages the P-GW 104 has been described. However, the present invention is also applicable when the MNO manages the P-GW 104. A case in which P-GW 104 is L2-connected to the UE 101 has been described above.
However, the present invention is also applicable when
10 the P-GW 104 is L3-connected to the UE 101.

[0054] In each embodiment described above, the first servers 105 may be classified into a C-plane server for performing communication with the S-GW 103 on the C plane and a P-plane server for performing communication
15 with the S-GW 103 on the P plane. In addition, the C-plane servers can also be classified into a server for receiving data from the S-GW 103 via the GTP tunnel and a server for decapsulating this data and extracting the IP packet.

20 [0055] The present invention is not limited to the above-described embodiments, and various changes and modifications can be made within the spirit and scope of the present invention. Therefore, to apprise the public of the scope of the present invention, the
25 following claims are made.

CLAIMS

1. A control apparatus for a gateway in a mobile communication system, characterized in that
the gateway includes a first server configured to
5 receive data from a user apparatus and one or more second servers configured to transfer the data received by the first server to an external network,
the one or more second servers are virtual machines on a cloud, and
10 the control apparatus performs
an operation of changing performance of the one or more second servers,
an operation of monitoring a processing status of the one or more second servers, and
15 an operation of selecting, from the one or more second servers in accordance with the processing status of the one or more second servers, a second server as a transfer destination of the data received from the user apparatus by the first server.
- 20 2. The control apparatus according to claim 1, characterized in that
the control apparatus
receives an identifier of the user apparatus from the gateway, and
25 returns, to the gateway, operations to be performed for data from the user apparatus having the identifier.

3. The control apparatus according to claim 2,
characterized in that the operations to be performed
include at least one of protocol conversion, data
compression, header deletion, addition of user
5 authentication information, addition of an
identification number of the user apparatus, or
addition of a digital signature.
4. The control apparatus according to claim 2 or 3,
characterized in that the control apparatus provides an
10 interface configured to edit the operations to be
performed.
5. The control apparatus according to any one of
claims 1 to 4, characterized in that
a common address is set in the one or more second
15 servers, and
the first server receives data having the common
address as a destination from the user apparatus.
6. The control apparatus according to claim 5,
characterized in that the common address is different
20 from an address set in the one or more second servers
in order to cause the first server to communicate with
the one or more second servers.
7. The control apparatus according to claim 5 or 6,
characterized in that
25 the first server transfers the received data to
the second server selected by the control apparatus,
and

the second server selected by the control apparatus performs processing in one of a network layer and an application layer for data transferred from the first server.

- 5 8. The control apparatus according to claim 5 or 6, characterized in that

the first server transfers the received data to the second server selected by the control apparatus, and

- 10 the second server selected by the control apparatus transfers, to another server configured to perform processing in one of a network layer and an application layer for data, the data transferred from the first server.

- 15 9. The control apparatus according to any one of claims 1 to 8, characterized in that

the first server receives data from the user apparatus via an L2 tunnel, and

- 20 the first server transfers the data, via an L2 tunnel different from the L2 tunnel, to the second server selected by the control apparatus.

10. The control apparatus according to any one of claims 1 to 8, characterized in that

the gateway includes a plurality of first servers,

- 25 the first server receives data from the user apparatus via an L2 tunnel, and

the control apparatus selects the second server

serving as a transfer destination of data received from the user apparatus by the first server so as not to overlap identifiers of the L2 tunnels in the second servers.

- 5 11. The control apparatus according to any one of claims 1 to 10, characterized in that
- the gateway includes one or more first servers,
 - the one or more first servers are virtual machines on a cloud,
 - 10 the control apparatus further performs an operation of changing performance of the one or more second servers, and
 - an operation of monitoring a processing status of the one or more second servers, and
 - 15 the control apparatus performs a change in performance of the one or more first servers and the change in performance of the one or more second servers based on different criteria.
- 20 12. The control apparatus according to claim 11, characterized in that the control apparatus performs the change in the performance of the one or more first servers based on network performance, and the change in the performance of the one or more second servers based on calculation performance.
- 25 13. The control apparatus according to any one of claims 1 to 12, characterized in that
- the one or more second servers are divided into a

plurality of groups, and

the control apparatus selects, from a group corresponding to identification information of the user apparatus, the second server serving as a transfer
5 destination of the data received from the user apparatus by the first server.

14. The control apparatus according to claim 13, characterized in that different pieces of transfer information are set in the groups, respectively.

10 15. The control apparatus according to claim 14, characterized in that the transfer information defines an external network which the user apparatus can access.

16. The control apparatus according to any one of claims 1 to 15, characterized in that the second server
15 and the external network communicate with each other via the L2 tunnel.

17. The control apparatus according to any one of claims 1 to 15, characterized in that the second server and the external network communicate with each other
20 via an L3 connection.

18. A program for causing a computer to operate as a control apparatus of a gateway in a mobile communication system, characterized in that
the gateway includes a first server configured to
25 receive data from a user apparatus and one or more second servers configured to transfer the data received by the first server to an external network,

the one or more second servers are virtual machines on a cloud, and

the program causes the computer to perform an operation of changing performance of the one

5 or more second servers,

an operation of monitoring a processing status of the one or more second servers, and

an operation of selecting, from the one or more second servers in accordance with the processing status
10 of the one or more second servers, a second server as a transfer destination of the data received from the user apparatus by the first server.